UNIVERSIDAD CARLOS III DE MADRID

TESIS DOCTORAL

Marginal Productivity Index Policies for Dynamic Priority Allocation in Restless Bandit Models

Autor:
Peter Jacko

Director:
José Niño Mora

DEPARTAMENTO DE ESTADÍSTICA

Getafe, Mayo 2009

TESIS DOCTORAL

Marginal Productivity Index Policies for Dynamic Priority Allocation in Restless Bandit Models

Autor: Peter Jacko

Director: José Niño Mora

Firma del Tribunal Calificador:

Firma

Presidente:

Vocal:

Vocal:

Vocal:

Secretario:

Calificación:

Getafe, de de

To my parents
for their endless support

Acknowledgements

The direction towards completion of this dissertation has been greatly influenced by my Ph.D. advisor prof. José Niño. I am grateful for his effort to help me get closer to clear thinking and clear presentation of ideas. I am happy to have been introduced by him to this exciting interdisciplinary topic motivated by real-life problems in diverse areas and putting together my favorite subjects: math, computer science and economics.

I would like to thank prof. Brunilde Sansò for warmly hosting me at Gerad in 2007 and for introducing me to telecommunications engineering in an extremely efficient way. Her open-mindedness was encouraging and essential for me to step into and to maintain the interest in the field of telecommunications.

During the four years I have been working on this dissertation I had motivating discussions with several members of the Department of Statistics. Talks with Sofía Villar and Bernardo D'Auria led to contributions in this dissertation, and Dae-Jin Lee was a great office-mate to share ideas on a lot of topics and to help improve my Spanish. I am glad for having shared the lunch time in those years with such friendly and interesting people like Kenedy Alva, Ignacio Cascos, María Durbán, Sergio García, Emilio Letón, Elisa Molanes, and Ismael Sánchez.

Two thousands kilometers away from my home country is quite enough to miss it, and even more if one arrives without speaking a Spanish word. It was simply great to have here my home-city friends Peter and Zuzana during all the time. Several others (especially Vlado, Katka, Lucia, Lenka and Barbora) brought me a piece of Slovakia over these years and reminded me of my deteriorating mother language, though Marcel has convinced me that it can get even worse. I am very glad that my old-time friends (B0, Martin, Marek, Gabča, Katka, Tomáš, Martina, Yuyka, Rado) find time for meeting me whenever we happen to be at the same place, even if the long-distance communication with me is so sparse.

The first who took me out of the office apart from my master class-mates was Giovanna and as a unique person from outside the university shared with me all these years here, in and around Madrid. Isa is the very best Spanish friend and it is not only because she speaks Slovak better than I do. In Montreal I felt like at home thanks to

Afshin, Valerio, Alice and Tolga, though only Afshin had his home there.

Meeting Vincenzo led to a breakthrough in my life in Madrid. Our "A la montaña" project has been so successful that we would have never imagined. Dozens of nice events resulted in making good friends and enjoying sport and the nature with them. Especially Erika, André, Patricia, Goki, Ana and Sara are those I would mind not meeting otherwise.

This period of personal development and focus on investigation would be impossible without the predoctoral research training grant BES-2005-7026 linked to the research project MTM2004-02334 from the Spanish Ministry of Education and Science, whose support is acknowledged. The funding support of the research project MTM2007-63140 from the Spanish Ministry of Education and Science and of the European Commission Networks of Excellence Euro-NGI and Euro-FGI is further acknowledged. The working conditions provided by the Department of Statistics and the Department of Business Administration were excellent and have also contributed to successful progress in this dissertation.

Gracias.

Abstract

This dissertation addresses three complex stochastic and dynamic resource allocation problems: (i) Admission Control and Routing with Delayed Information, (ii) Dynamic Product Promotion and Knapsack Problem for Perishable Items, and (iii) Congestion Control in Routers with Future-Path Information. Since these problems are intractable for finding an optimal solution at middle and large scale, we instead focus on designing tractable and well-performing heuristic priority rules.

We model the above problems as the multi-armed restless bandit problems in the framework of Markov decision processes with special structure. We employ and enrich existing results in the literature, which identified a unifying principle to design dynamic priority index policies based on the Lagrangian relaxation and decomposition of such problems. This decomposition allows one to consider parametric-optimization subproblems and, in certain "indexable" cases, to solve them optimally via the marginal productivity (MP) index. The MP index is then used as a dynamic priority measure to define heuristic priority rules for the original intractable problems.

For each of the problems considered we perform such a decomposition, identify indexability conditions, and obtain formulae for the MP indices or tractable algorithms for their computation. The MP indices admit the following priority interpretations in the three respective problems: (i) undesirability for routing a job to a particular queue, (ii) promotion necessity of a particular perishable product, and (iii) usefulness of a particular flow transmission.

Apart from the practical contribution of deriving the heuristic priority rules for the three intractable problems considered, our main theoretical contributions are the following: (i) a linear-time algorithm for computing MP indices in the admission control problem with delayed information, matching thus the complexity of the best existing algorithm under no delays, (ii) a new type of priority index policy based on solving a (deterministic) knapsack problem, and (iii) a new extension of the existing multi-armed restless bandit model by incorporating random arrivals of restless bandits.

Resumen

Esta tesis estudia tres complejos problemas dinámicos y estocásticos de asignación de recursos: (i) Enrutamiento y control de admisión con información retrasada, (ii) Promoción dinámica de productos y el Problema de la mochila para artículos perecederos, y (iii) Control de congestión en "routers" con información del recorrido futuro. Debido a que la solución óptima de estos problemas no es asequible computacionalmente a gran y mediana escala, nos concentramos en cambio en diseñar políticas heurísticas de prioridad que sean computacionalmente tratables y cuyo rendimiento sea cuasi-óptimo.

Modelizamos los problemas arriba mencionados como problemas de "multi-armed restless bandit" en el marco de procesos de decisión Markovianos con estructura especial. Empleamos y enriquecemos resultados existentes en la literatura, que constituyen un principio unificador para el diseño de políticas de índices de prioridad basadas en la relajación Lagrangiana y la descomposición de dichos problemas. Esta descomposición permite considerar subproblemas de optimización paramétrica, y en ciertos casos "indexables", resolverlos de manera óptima mediante el índice de productividad marginal (MP). El índice MP es usado como medida de prioridad dinámica para definir reglas heurísticas de prioridad para los problemas originales intratables.

Para cada uno de los problemas bajo consideración realizamos tal descomposición, identificamos las condiciones de indexabilidad, y obtenemos fórmulas para los índices MP o algoritmos computacionalmente tratables para su cálculo. Los índices MP correspondientes a cada uno de estos tres problemas pueden ser interpretados en términos de prioridades como el nivel de: (i) la penalización de dirigir un trabajo a una cola particular, (ii) la necesidad de promocionar un cierto artículo perecedero, y (iii) la utilidad de una transmisión de flujo particular.

Además de la contribución práctica de la obtención de reglas heurísticas de prioridad para los tres problemas analizados, las principales contribuciones teóricas son las siguientes: (i) un algoritmo lineal en el tiempo para el cómputo de los índices MP en el problema de control de admisión con información retrasada, igualando, por lo tanto, la complejidad del mejor algoritmo existente para el caso sin retrasos, (ii) un nuevo tipo de política de índice de prioridad basada en la resolución de un problema (determinista)

de la mochila, y (iii) una nueva extensión del modelo existente de “multi-armed restless bandit” a través de la incorporación de las llegadas aleatorias de los “restless bandits”.

Contents

List of Figures

Be simple.
Be happy.

Chapter 1

Introduction

Economic decision making under uncertainty is one of the most important challenges of everyday life. People have developed, based on their beliefs or intuition, ways of ordering alternatives by assigning them *priorities* in order to deal with complex decisions. Therefore, it is of great practical interest to have a general methodology for designing tractable priority rules for relevant intractable problems. Moreover, we would like to be able to identify assumptions under which such priority rules may lead to optimal decisions, and to provide suboptimality bounds for these heuristics in general.

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

In this work we aim to answer the above-mentioned questions by *dynamic priority* rules, that is, reconsidering the priority-order of alternatives regularly in time. The need for *dynamic (priority) allocation* arises whenever the activities one invests in have any of the following features: (1) the decision-maker does not have perfect information about the reward that the activity yields, (2) the reward is known, but subject to a random factor, (3) the reward is known, but changes over time. Thus, we will deal with those cases, in which the decision-maker faces a trade-off between exploitation (taking a reward today) and exploration (obtaining a possibly higher reward tomorrow). This is captured by the *restless bandit models* we will use in this work.

Analysis of such problems is of both theoretical and practical value. From a practical point of view, *stochastic and dynamic resource allocation problems* arise in areas as diverse as product (R&D) management, marketing, financial economics, optimal con-

sumption planning, telecommunications, engineering systems, medicine, etc., where a well-reasoned advice is more than needed. Three applications in stochastic scheduling, marketing, and telecommunications addressed in this work are outlined in Section 1.1.

Stochastic and dynamic resource allocation problems are naturally modeled in the framework of Markov decision processes (MDPs), which briefly surveys Section 1.2. We will focus on a particular family of MDPs with special structure, called the *multi-armed restless bandit problems*. Section 1.3 describes the key concepts of the restless bandit models and their priority solutions based on the *marginal productivity index*. We emphasize that index policies we propose to use are usually *simple though suboptimal* priority rules for the complex and intractable problems considered. Finally, Section 1.4 outlines the results for our three problems obtained by employing and enriching these methods.

1.1 Description of Three Problems

We address the following three problems in this dissertation.

1.1.1 Admission Control and Routing with Delayed Information

Chapter 4 addresses the problem of dynamic job admission control and/or routing in a model of parallel loss queues with delayed state observation and/or delayed action implementation. Two versions of the model are considered, depending on whether the admission control capability is enabled or not. The queue servers may be endowed with finite or infinite buffer space. Such problems are relevant in a variety of application domains, most notably in the operation of packet-switched communication networks and distributed computer systems.

In addition, in such systems there are nonnegligible propagation delays, which force the controller to take decisions based on stale system state information and cannot take effect before a time lag. This is the main distinctive feature of the problem addressed with respect to existing literature, where little has been done on problems with delays. Recent applications in which the delays are of special importance include satellite communications, long-distance-controlled robots, and situations in which an advanced processing of observations is necessary.

Note that in problems with delayed state observation or delayed action implementation, the decision-maker does not have perfect information about the reward that the activity yields. Such problems are naturally formulated as *partially observed Markov decision processes* (POMDPs), which in turn are readily reformulated as conventional MDPs.

1.1.2 Dynamic Product Promotion and Knapsack Problem for Perishable Items

Chapter 5 introduces the Knapsack Problem for Perishable Items addressing the stochastic combinatorial problem of choosing a collection of perishable products to be allocated at a promotion location with limited space (called knapsack). Such a dynamic problem of expected revenue maximization subject to a physical space constraint arises in a variety of industries, where products have an associated lifetime and cannot be sold afterwards.

We design a finite-horizon model in which demand is altered not by price changes, but rather by moving a number of product units to a scarce promotion space, where they are likely to attract extra customers. Examples of the promotion space include shelves close to the cash register, promotion kiosks, or a depot used for selling via the Internet.

As a special case of the knapsack problem, we also deal with the problem of dynamic promotion of a single perishable product. This may be of interest if there is no knapsack (or budget) restriction, and perishability is an important factor in altering the customer demand.

1.1.3 Congestion Control in Routers with Future-Path Information

Chapter 6 considers a bottleneck router with a scarce resource that is given by the bandwidth available for which several flows compete. Each flow generates certain *goodput* reward for its receiver, if it is delivered, which can be achieved fully only if the flow is transmitted by the router. The difficulty is that these flows are stochastically and dynamically changing transmission rates, so the rewards may increase or decrease over time.

This work is concerned with congestion control which tries to exploit flows' future-path information about network congestion at routers. In the periods of network congestion, a use of future-path information may be highly valuable for the network performance, since dropping a packet too late on its route implies that all the scarce resources (bandwidth and buffer space) it has consumed so far are wasted.

Thus, the question is whether to exploit the present rewards by transmitting at the arrival rate, or to take a locally-suboptimal action of packet dropping or marking which may yield higher rewards further downstream or to the following packets arriving at the router.

We consider the problem under three basic router variants with the following network congestion control functions:

(i) *TD router*: congestion control based on tail dropping (buffer overflow),

(ii) *ICN router*: congestion avoidance with implicit congestion notification (packet dropping), and

(iii) *ECN router*: congestion avoidance with explicit congestion notification (packet marking).

1.2 Markov Decision Process Framework

In stochastic and dynamic resource allocation, the controller may influence by her *actions* the future evolution of an underlying system at various points in time. In such a *sequential decision process*, there are rewards (or costs) incurred over time that depend on the actions taken and the way in which the system evolves. The goal of the controller may be to maximize the expected total[1] reward or to minimize the expected total cost over a certain time horizon. If the horizon is infinite, then one may need to use discounting or long-run averaging in order to have a finite-valued objective (Stidham, 2002). Nevertheless, such alternatives of objective function may also be relevant in some finite horizon problems.

When the information needed to predict the future evolution of a system is contained in the current *state* of the system and depends on the current action, we call such a sequential decision process a *Markov decision process* (MDP).[2] MDPs have a great modeling power, which can provide results on the existence and structure of good policies and on methods for the computation of optimal policies. Therefore, it has naturally been used in a variety of applications in areas including engineering systems, operations research, management science, economics and applied probability.

MDP theory has been developed in two separate streams, for discrete-time and continuous-time models, respectively. In further discussion we will focus on *discrete-time* MDPs, which is an important setting from at least two points of view: (1) there is a large number of interesting problems being naturally modeled in the discrete time setting and (2) important classes of continuous-time problems can be exactly reformulated into discrete-time setting using the *uniformization* technique.

In stochastic dynamic systems it is typically not possible to have information about future states at a decision moment, and therefore decisions should not be based on them. Hence, a useful solution concept for an MDP is a *non-anticipative policy* (or, *history-*

[1] Throughout this work, the term *total* is reserved to mean the sum over all the time epochs.

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

dependent policy), which is defined as a set of rules specifying the action to be taken for each decision point in time and for each possible state of the system, using only current and past information.

A policy thus answers the following question: What action should be taken at a given time if the system is in a given state? As we will see later, a class of *stationary policies* is often of high interest. A policy is stationary if the answer to the question just stated does not depend on the point in time (i.e., it is time-homogeneous). Such a policy is appropriate, because MDPs are of Markovian nature, i.e., the future evolution of the system depends on history only through the current state.

The breakthrough in MDPs was made by an approach, now called *dynamic programming*, developed by Richard Bellman in the 1950's. The idea of dynamic programming is based on the *Principle of optimality*: at any point in time, an optimal policy must prescribe an action that optimizes the sum of immediate reward and (expected) total reward obtained if an optimal policy is applied from the subsequent point in time on. The mathematical concept associated to the Principle of optimality is the optimality equations of dynamic programming, called the *Bellman equations*. For infinite-horizon problems, Bellman equations simplify so that they are not time-dependent; indeed, the optimal objective value is a unique fixed point solution.

The value of dynamic programming is due to its both theoretical and practical power. Dynamic programming provides a coherent theoretical framework for studying Markov decision processes. As such, it leads to several general theoretical results including a necessary and sufficient condition for optimality of a stationary policy in some broad cases. For instance, it implies that for finite-state and finite-action MDPs there is an optimal policy that is deterministic, stationary, and independent of the initial state. From practical point of view it is remarkable that the dynamic programming approach reduces optimization over the sequence of decisions in various points in time to a sequence of parameter optimizations for every time point, thus, it may significantly decrease the problem complexity.

Still, for many problems this may be not enough to make the solution of the problem tractable. A typical knot arising in their use is that the dynamic programming recursions may be too many (or infinitely many) to allow actual computation. The size of dynamic programming formulation is typically exponentially growing with the size of the model, which is known as the *curse of dimensionality*. Here comes out a necessity for other approaches. One of the solution approach alternatives is *linear programming* (LP) reformulation of Bellman equations. Since each Bellman equation includes an optimization term, it can be relaxed to a set of linear inequalities, one for each action. Once this has been done with all Bellman equations, one adds an objective function that

forces at least one inequality to be satisfied sharply for each state. From the solution to this associated LP problem, one can readily get the optimal policy for the original MDP. As Stidham (2002) points out, the LP approach is especially well suited to constrained MDPs, in which the optimal policy must satisfy side constraints, what allows to reduce the set of policies.

However, the LP reformulation as such does not help to deal with the curse of dimensionality. Section 3.1 discusses the Lagrangian relaxation, an approach we will adopt in this work helping decompose complex problems with special structure in order to obtain well-performing suboptimal solutions.

1.3 A Unifying Principle to Design Dynamic Priority Index Policies

Due to the curse of dimensionality, complex resource allocation problems, such as those described in Section 1.1, are typically addressed, analyzed, and solved by ad-hoc techniques. Moreover, even if the focus is on finding good (i.e., not necessarily optimal) policies based on priorities, many people may have proposed *ad hoc* priority rules. This section briefly outlines a unifying principle to design priority rules described in more detail in Chapter 3, which has been developed in the literature on the (restless) bandit problems.

In this work we develop MDP models with a special structure, falling into the framework of the *multi-armed restless bandit problem*. This is a fundamental model of allocating a scarce resource to stochastic and dynamic alternatives. We assume that the alternatives evolve independent of each other, except for a sample path constraint on the resource capacity.

We will formally define the multi-armed restless bandit problem after the review of its historical development in Chapter 2. At this place we only anticipate that it is PSPACE-hard, and therefore intractable on a medium and large scale. We will thus focus attention on the more realistic and practical goals of designing and computing well-grounded heuristic policies that are readily implementable.

For the multi-armed restless bandit problem there is a tractable relaxation, which help decompose it into separate *parametric optimization* subproblems for each alternative. This relaxation has two steps: first, we relax the sample path constraint requiring it only *on average*; in the second step we apply the standard Lagrangian relaxation. Such a relaxed problem then decomposes into subproblems due to the independence of alternatives, where the Lagrangian multiplier appears as a parameter. This parameter has an interpretation of *price* or *wage* paid for allocating the scarce resource.

The parametric optimization subproblems corresponding to all the alternatives can often be solved optimally under certain *convexity* assumptions that are natural for the problem in hand. Moreover, as the pricing parameter changes, the optimal solutions may change in a *monotonic* way, so that it is optimal to allocate the scarce resource less when the price is higher. In such a case we will be interested in the break-even values of the price parameter at which the optimal solution changes. Such values measure the marginal productivity of allocating the scarce resource.

We will call these values the *marginal productivity (MP) indices*, and let us say that a parametric subproblem is *indexable* if they exist. Then, we can use the MP indices in dynamic priority rules to measure the priority of allocating the scarce resource to a given alternative. Given the economic interpretation of MP indices, giving priority correspondingly to them results in allocating the scarce resource to alternatives with currently highest productivity of using the resource.

The MP index is thus established as a unifying design principle of dynamic priority policies in intractable stochastic and dynamic resource allocation problems. This approach has been proved recently to be well-grounded and tractable in a variety of problems of increasing complexity, and is enriched by our work. Experimental studies suggest that MP indices are typically close to optimal, and performing better or at least not worse than other rules that may exist for certain applications. Moreover, several existing well-performing rules obtained by ad-hoc methods have been shown to be special cases of MP.

Let us emphasize that MP indices are given by optimal solutions in a parametric subproblem case, whereas they are used to define heuristic priority rules for the intractable resource allocation problems. Thus, several questions need to be addressed for a given problem:

(i) [Mathematical question.] For a given optimization criterion, under what conditions are the alternatives indexable?

(ii) [Algorithmic question.] How to calculate the MP indices quickly?

(iii) [Experimental question.] How close to optimal are the resulting MP index priority policies? And how do they compare to alternative policies?

1.4 Results Outline for Three Problems

We obtain the following results for the three problems considered in this dissertation.

1.4.1 Admission Control and Routing with Delayed Information

A priority policy in terms of MP indices is derived for this problem under the following three performance objectives: (i) minimization of the expected total discounted sum of holding costs and rejection costs, (ii) minimization of the expected time-average sum of holding costs and rejection costs, and (iii) maximization of the expected time-average number of job completions. Our employment of existing theoretical and algorithmic results on restless bandit indexation together with some new results yields a fast algorithm that computes the MP index for an admission control of a queue in linear time with respect to the buffer size.

Such MP index values can be used both to immediately obtain the optimal thresholds for a single-queue admission control problem, and to design an index priority policy for the routing problem (with possible admission control) in the multi-queue system. The MP index can be thought of as a measure of *undesirability of routing* a job to a particular queue, given as a function of the queue's augmented state, which refers to the observed action-state pair at the previous period.

Our approach seems to be tractable also for the analogous problems with larger delays and, more generally, for arbitrary restless bandits with delays.

This work was presented at the Third International Conference on Performance Evaluation Methodologies and Tools (ValueTools) in 2008, and an extended abstract was published as Jacko and Niño-Mora (2008).

1.4.2 Dynamic Product Promotion and Knapsack Problem for Perishable Items

For the dynamic product promotion problem we derive an optimal MP index policy with closed-form indices. The MP indices, that can be interpreted in this setting as *promotion priority indices*, capture the marginal rate of promotion as a function of its price, salvage value, lifetime, expected demand, and expected promotion power. For a single product we obtain structural results analogous to those reported in dynamic pricing literature: the promotion priority increases with shorter product lifetime.

For the Knapsack Problem for Perishable Items we propose a new MP-index-based heuristic that includes solving a deterministic knapsack problem and whose nearly-optimal performance and superiority to other heuristics is demonstrated in a computational study.

This work was presented at the Sixth Czech-Slovak International Symposium on Combinatorics, Graph Theory, Algorithms and Applications in 2006 and at the VIIIth Annual Conference of INFORMS Revenue Management and Pricing Section in 2008. An extended abstract was published as Jacko and Niño-Mora (2007).

1.4.3 Congestion Control in Routers with Future-Path Information

We model the congestion control problem of multiple flows in the framework of multi-armed restless bandit problem with an additional feature of random arrivals of flows with random finite length. We set out to maximize the expected time-average network goodput so that the expected time-average router's throughput is below its bandwidth and so that the router's buffer space is not overflowed. We discuss how a novel concept of *network-capability fairness* arises by implementing at routers a congestion control mechanism that maximizes the expected time-average network goodput.

Relaxation and decomposition of this problem allows us to introduce a *transmission index* which evaluates the usefulness of each flow's transmission at the moment. The transmission index is the MP index arising in the context of this problem. Moreover, argue that the transmission index identifies locally optimal router actions for the decentralized network problem. Since the index captures the value of network services to users, it can be interpreted as a network *congestion price* (when multiplied by the packet size).

We apply these general results to derive closed-form expressions of the transmission index for TD, ICN, and ECN routers, and we present two situations in which this index defines an optimal transmission priority policy for a multiple-flow problem at a bottleneck router. Finally, we discuss proposals for practical implementation of the transmission index in existing congestion avoidance mechanisms. Such changes are arguably expected to lead both to a lower delay and higher network throughput, though the implementation may be costly at the moment due to the necessity of information gathering by network nodes.

This work was presented at the EuroFGI Workshop on IP QoS and Traffic Control in 2007, and an extended abstract was published as Jacko and Sansò (2007).

Learn by play.
< J. A. Comenius >

Chapter 2

Multi-Armed Restless Bandit Problem and Marginal Productivity Index Policies

In this work we develop MDP models with a special structure, falling into the framework of the *multi-armed restless bandit problem*—a fundamental stochastic and dynamic resource allocation model. In this chapter we outline important contributions in its historical development, review its peculiarities in order to highlight its broad applicability and present the framework it offers for dynamic resource allocation problems, based on the survey Niño-Mora (2007b).

2.1 Multi-Armed Bandit Problem

The *multi-armed bandit problem*, originally described by Robbins (1952), is a model of a controller optimizing her decisions while acquiring knowledge at the same time. Although it is a simply-stated problem of stochastic dynamic optimization, its solution had been a challenging open problem for a considerably long time, until the celebrated result of Gittins and Jones (1974) (reviewed below) appeared. This problem models the fundamental trade-off between *exploitation* (getting the highest immediate rewards) and *exploration* (learning about the system and receiving possibly even higher rewards later).

The multi-armed bandit problem is named after a *one-armed bandit slot machine* one can find in casinos (see Figure 2.1). In the multi-armed case, the gambler has to decide which arm to pull (exactly one at a time) in order to maximize her total reward in a series of trials. So, we can rephrase the multi-armed bandit problem as the problem

Figure 2.1: The multi-armed bandit problem is named after a one-armed bandit slot machine one can find in casinos.

concerned with the question of how to dynamically allocate a *single* scarce resource amongst several stochastic alternative projects (Weber, 1992).

Each bandit is modeled as a random reward yielding process whenever played, whereas it remains *frozen* (no evolution, no rewards) whenever not played. Such a bandit is called *classic* as opposed to its extension called the *restless bandit* introduced in Whittle (1988), which admits evolution and rewards/costs even if not played.

2.2 Multi-Armed Restless Bandit Problem

The *multi-armed restless bandit problem* is a natural generalization of the multi-armed bandit problem, which is capable to cover considerably broader set of practical situations. To the classical model we add just two simply-stated features: (1) bandits are allowed to evolve and yield rewards when not played (no freezing anymore), and (2) we are to allocate the scarce resource parallely to a fixed number of bandits (instead of playing only one bandit). Nevertheless, the increased modeling power comes at the expense of tractability: the multi-armed restless bandit problem is *P-SPACE hard*, even in the deterministic case (Papadimitriou and Tsitsiklis, 1999). The research focus must thus shift to the design of well-grounded, tractable heuristic policies.

In this work we consider the *work-reward restless bandits* generalized by Niño-Mora (2002), which significantly expand the modeling scope of the *restless bandits* introduced

in Whittle (1988), which in turn are a generalization of the *classic bandits* of Robbins (1952). In the following, we will use the term *bandit* (without adjectives) as a generic term for any of the above three classes.

2.3 Index Policies

An appealing feature of the multi-armed classic bandit problem is optimality of the *index priority policy* (also called *index rule*), obtained by Gittins and presented in a series of papers in the early 1970s (as documented in Whittle (1980)). Since bandits are competing for a scarce resource, we assign to each of them a *dynamic allocation index*, and then apply the index priority policy, defined as follows: "Assign the scarce resource to a bandit of highest current index value." The index he proposed became known as the *Gittins index*, and the solution to the multi-armed classic bandit problem as the *Gittins index (priority) policy*. See Gittins (1979) for a reconciliation of the ideas. The significance of the Gittins index is that it can be computed for each bandit in isolation, i.e., it is independent of the other bandits. An excellent presentation of the multi-armed classic bandit problem introducing an intuitive (almost verbal) proof of optimality of the Gittins index policy was given in Weber (1992).

A classical example of optimality of an index priority policy is the *$c\mu$-rule* for the job sequencing problem (Smith, 1956). In that problem, jobs labeled by k with linear holding costs c_k and mean service time μ_k^{-1} must be scheduled for service at a single server so that the expected total holding cost is minimized. The $c\mu$-rule prescribes to schedule the jobs as follows: "Assign the server to an uncompleted job of highest index $c_k\mu_k$." Note, however, that such an allocation index is *static*, as opposed to the Gittins index, which is dynamic (depending on the actual bandit's state).

We will use the approach introduced by Whittle (1988), who proposed to solve the multi-armed restless bandit problem by solving its relaxation by Lagrangian methods. The Whittle relaxation was to replace a family of sample-path constraints (of playing a fixed number of bandits at every period) by a unique one (of playing the required number of bandits *on average*). Then, using a Lagrangian multiplier, such a constraint can be dualized and included in the objective. This allows to decompose the multi-armed problem and significantly simplify the solution procedure by considering bandits in isolation. Whittle (1988) further proposed an index, which in the case of classic bandits recovers the Gittins index, to be used in an index priority policy: "Assign the scarce resource to bandits of highest current index values."

Identification of tractable indices that make an index priority policy well-performing is a central issue in the literature concerning extensions of the multi-armed bandit prob-

lem. Unless we are lucky to find a particular class of bandits with special structure as in the case of classic bandits, in general we can only expect a form of asymptotic optimality of an index policy for the multi-armed restless bandit problem, as was shown in Weber and Weiss (1990).

The above-mentioned authors realized that well-performing indices often have an economic interpretation. The $c\mu$-rule takes into account expected savings of serving a given queue with respect to not serving it. Gittins (1979) characterized his index as the maximal reward rate, because it is calculated as the maximal rate of expected rewards per unit of expected time. Whittle (1988) interpreted the index he proposed as a fair charge for assigning the scarce resource to the bandit. Niño-Mora (2002, 2006b) coined the term *marginal productivity (MP)* index for his index that generalizes all the above indices. It is "marginal" because it captures the effect of employing the scarce resource at a given moment with respect to not doing so, and "productivity" reflects that it is computed as the maximum rate of marginal rewards per unit of marginal work. Put in an economics jargon, the MP index is nothing but the marginal rate of transformation of employing the scarce resource at a given state of a bandit.

Whittle (1988) further realized that not all restless bandits are *indexable*; for non-indexable bandits such indices simply do not exist. Niño-Mora (2001, 2002, 2006b) introduced methods of analysis to determine a priori whether a given restless bandit model is indexable, and gave an adaptive-greedy algorithm based on Klimov (1974) to calculate the MP indices. That approach, which we follow in this work, has been shown in a growing variety of applications to yield an index priority policy which is well-grounded, intuitive, easy-to-implement, and nearly-optimal, i.e., well-suited for practical purposes. We dedicate a separate chapter (Chapter 3) to the methodology of establishing existence and computation of MP indices as applied in this work.

2.4 Mathematical Programming Formulation and Conservation Laws

The mathematical programming approach we review in this section is closely connected to graphical interpretation of problems and is thus very well suited for providing insights of the solution methods and for helping to exploit the problem structure. With each policy and each initial state, one can associate a *performance vector*, for instance, the expected average reward under the policy if starting from the initial state. Then, a set of admissible policies (which depends on a given problem) defines a *performance region* (or *achievable region*), i.e., the space of all possible system performance vectors which are achievable under admissible policies.

Structural properties of such a performance region lead to structural properties in the given problem. We may therefore be interested in describing the performance region so that the optimization problem can be efficiently solved by classical mathematical programming methods. When an analysis via this methodology is available, one can typically make clear and strong statements about optimal policies.

The earliest intentions to use such an approach were made in queueing theory, originated in Klimov (1974) and Coffman and Mitrani (1980), later followed by Federgruen and Groenevelt (1988) in a more general framework of a certain family of queueing models. In the latter contribution it was shown that the performance region in those models is a polytope of special type. An important concept of (strong) *conservation laws* was introduced in Shanthikumar and Yao (1992), where the previous results were extended by proving a powerful result about the achievable region approach. When the performance vectors satisfy strong conservation laws, the achievable region is a particular polytope (called the *base of a polymatroid*, previously known in combinatorial optimization), completely characterized by those laws, and the set of vertices of the achievable region is equivalent to the set of performance vectors obtained by all the *static index policies*. Then, optimization of a linear objective can be accomplished by a greedy algorithm, which indeed finds an optimum in a vertex, hence ensuring that there is an optimal index policy (Stidham, 2002).

Bertsimas and Niño-Mora (1996) drawing on the work of Tsoucas (1991), extended those results to a more complex class of stochastic dynamic problems. They defined *generalized conservation laws*, whose satisfaction by performance vectors implies that the performance region is a polytope of special structure. Moreover, optimization of a linear objective over such a polytope is solved by an *adaptive-greedy algorithm* based on Klimov (1974), which leads to an optimal *dynamic index policy*. More general results in a similar fashion introducing *partial conservation laws* were obtained in Niño-Mora (2001, 2002, 2006b), in which the analysis is closely tied to *restless bandits*.

Conservation laws and polytopes treated in the listed papers were exploited mainly in the context of queueing systems and networks. An early survey of the achievable region approach was given in Dacre et al. (1999). An updated exposition can be found in Niño-Mora (2009), which stresses that conservation laws help exploit the problem structure in order to design and compute optimal or nearly optimal policies.

2.5 Single Restless Bandit Model

In what follows, *project* is used as a short name for the most general work-reward restless bandit, to emphasize its broad applicability in problems of effort allocation to alter-

native projects. Consider the time slotted into time epochs $t \in \mathcal{T} := \{0, 1, 2, \dots\}$.[1] The time epoch t corresponds to the beginning of a time period t.

We can either *work* or *not work* on a given project. We denote by $\mathcal{A} := \{0, 1\}$ the *action space*, i.e., the set of allowable actions, where 1 corresponds to working, and 0 to not working (resting). This action space is the same for every project. A project labeled by $k \in \mathcal{K}$ can be modeled independently of other projects as the tuple

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

where

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

The dynamics of project k is thus captured by the *state process* $X_k(t) \in \mathcal{N}_k$ and the *action process* $a_k(t) \in \mathcal{A}$ at all time epochs $t \in \mathcal{T}$. As a result of choosing action $a_k(t)$ in state $X_k(t)$ at time epoch t, the project expends the work, earns the reward, and evolves its state for the time epoch $t + 1$. It is natural to require that the expected one-period work is nonnegative, therefore $0 \le W_{k,n}^a$. The classic bandits and the restless bandits in Whittle (1988) satisfy $W_{k,n}^a = a$ for all k, n, a, while this requirement was dropped in the work-reward restless bandits introduced in Niño-Mora (2002). In this work we will also assume that $R_{k,n}^a$ is bounded.

Note that we have the same action space $\mathcal{A}$ available at every state. Though at first glance this may appear as a limitation on the applicability of the model, the opposite is true. Notice that to effectively restrict the number of allowable actions at certain states, we can define actions 1 and 0 as duplicates by having the same one-period consequences. If actions 1 and 0 at state n of project k satisfy $W_{k,n}^1 = W_{k,n}^0$, $R_{k,n}^1 = R_{k,n}^0$, and $p_{k,n,m}^1 = p_{k,n,m}^0$ for all m, then we say that state $n \in \mathcal{N}_k$ of project k is *uncontrollable*.

[1] Throughout this work we stick to the following notational conventions to ease the reading: every set is typeset in calligraphic font (e.g., $\mathcal{T}, \mathcal{N}$), the corresponding uppercase letter denotes the number of its elements (T, N), and their generic element is written in lowercase (t, n). Vectors are in lowercase boldface in row/column form as necessary ($\boldsymbol{n}, \boldsymbol{z}$) , and matrices are in uppercase boldface ($\boldsymbol{P}$). In problem parameters and variables, superscripts are reserved to actions or policies, and subscripts to states.

In some applications it may be more natural to consider $C^a_{k,n} := -R^a_{k,n}$, the expected one-period *cost* paid by project k at state n if action a is decided at the beginning of a period. We, however, prefer using rewards during this exposition since they provide a more intuitive interpretation of certain concepts obtained in the solution.

Example 2.1 (Job Sequencing Problem: MDP Formulation). Consider K jobs waiting at the beginning (i.e., at time epoch $t = 0$) for service at a server that can serve one job at every time period. Let $0 < \mu_k < 1$ be the probability that the service of job k is completed within one period and let $0 < c_k$ be the holding cost per period incurred for job k waiting. These jobs can be viewed as competing projects and while there are at least two jobs waiting, one must decide to which job the server should be allocated. Note that otherwise the decision is trivial.

Suppose that the server is preemptive (i.e., the service of a job can be interrupted at any time epoch even if not completed). We have the action space $\mathcal{A} := \{0, 1\}$, where action 0 means "not serving" a job, and action 1 means "serving" a job.

Thus, we define job k with

- state space $\mathcal{N}_k := \{\text{'completed'}, \text{'waiting'}\}$;

- expected one-period works (job completions)

$$\begin{aligned} W^1_{k,\text{'completed'}} &:= 0, & W^1_{k,\text{'waiting'}} &:= \mu_k, \\ W^0_{k,\text{'completed'}} &:= 0, & W^0_{k,\text{'waiting'}} &:= 0; \end{aligned}$$

- expected one-period rewards (the negative of holding costs)

$$\begin{aligned} R^1_{k,\text{'completed'}} &:= 0, & R^1_{k,\text{'waiting'}} &:= -c_k \cdot (1 - \mu_k) - 0 \cdot \mu_k, \\ R^0_{k,\text{'completed'}} &:= 0, & R^0_{k,\text{'waiting'}} &:= -c_k; \end{aligned}$$

- one-period state-transition probability matrix if serving a job,

$$\boldsymbol{P}^1_k := \begin{array}{r} \\ \text{'completed'} \\ \text{'waiting'} \end{array} \begin{array}{c} \begin{array}{cc} \text{'completed'} & \text{'waiting'} \end{array} \\ \begin{pmatrix} 1 & 0 \\ \mu_k & 1 - \mu_k \end{pmatrix} \end{array},$$

and if not serving,

$$\boldsymbol{P}_k^0 := \begin{array}{r} \\ \text{'completed'} \\ \text{'waiting'} \end{array} \begin{array}{c} \begin{array}{cc} \text{'completed'} & \text{'waiting'} \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}.$$

Notice that when the server is allocated to job k, the expected one-period work is only $W^1_{k,\text{'waiting'}} = \mu_k < 1$ if the job is still waiting, and $W^1_{k,\text{'completed'}} = 0$ if it is already completed. In this example, the state 'completed' is uncontrollable, because actions 1 and 0 have the same one-period consequences in that state.

• • •

2.6 MDP Formulation of Multi-Armed Restless Bandit Problem

Let $\mathbb{E}^\pi_{\boldsymbol{n}}$ denote the expectation conditioned on a policy π and on a joint initial state $\boldsymbol{n} := (n_k)_{k\in\mathcal{K}}$, where $X_k(0) = n_k$. For any initial joint state $\boldsymbol{n} := (n_k)_{k\in\mathcal{K}}$ and for any discount factor $0 < \beta < 1$, the problem under the *β-discounted criterion* is to find an admissible policy $\pi \in \Pi$ maximizing the objective given by the expected aggregate total β-discounted reward, i.e.,

$$\max_\pi \mathbb{E}^\pi_{\boldsymbol{n}} \left[\sum_{k\in\mathcal{K}} \sum_{t=0}^{\infty} \beta^t R^{a_k(t)}_{k,X_k(t)} \right], \tag{2.1}$$

subject to the *sample path capacity* constraint,

$$\sum_{k\in\mathcal{K}} W^{a_k(t)}_{k,X_k(t)} \le W, \text{ for all } t = 0, 1, 2, \ldots \tag{2.2}$$

where W is the available capacity to be used every period. Note that the sample path capacity constraint is conditioned on the policy π and the joint initial state $\boldsymbol{n}$.

Analogously we can formulate the problem under the *time-average criterion*. For any initial joint state $\boldsymbol{n} := (n_k)_{k\in\mathcal{K}}$, the problem is to find a policy π maximizing the objective given by the expected aggregate time-average reward, i.e.,

$$\max_\pi \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}^\pi_{\boldsymbol{n}} \left[\sum_{k\in\mathcal{K}} \sum_{t=0}^{T-1} R^{a_k(t)}_{k,X_k(t)} \right], \tag{2.3}$$

subject to the same sample path capacity constraint (2.2).

Note that the sample path capacity constraint may be required as *equality* instead of inequality. Indeed, such an equality constraint was considered in the multi-armed classic bandit problem in Gittins (1979) and in the multi-armed restless bandit problem in Whittle (1988).

Further, in some applications one may require a sample path constraint on *actions* instead of on expected one-period works, i.e.,

$$\sum_{k \in \mathcal{K}} a_k(t) \leq W, \text{ for all } t = 0, 1, 2, \ldots$$

where W is interpreted as the maximum number of projects upon which action 1 can be applied during a period.

Connect the dots.
< S. Jobs >

Chapter 3

Design of Index Policies: Lagrangian Relaxation, Decomposition, and Restless Bandit Indexation

3.1 Whittle Relaxation, Lagrangian Relaxation, and Decomposition

Whittle (1988) proposed to relax the sample path capacity constraint (2.2) so that under the β-discounted criterion we require this constraint to hold only in "expected total β-discounted" terms,

$$\mathbb{E}_{\boldsymbol{n}}^{\pi}\left[\sum_{k\in\mathcal{K}}\sum_{t=0}^{\infty}\beta^{t}W_{k,X_k(t)}^{a_k(t)}\right]\leq\mathbb{E}_{\boldsymbol{n}}^{\pi}\left[\sum_{t=0}^{\infty}\beta^{t}W\right]=\frac{W}{1-\beta} \tag{3.1}$$

and under the time-average criterion only in "expected time-average" terms,

$$\lim_{T\to\infty}\frac{1}{T}\,\mathbb{E}_{\boldsymbol{n}}^{\pi}\left[\sum_{k\in\mathcal{K}}\sum_{t=0}^{T-1}W_{k,X_k(t)}^{a_k(t)}\right]\leq\lim_{T\to\infty}\frac{1}{T}\,\mathbb{E}_{\boldsymbol{n}}^{\pi}\left[\sum_{t=0}^{T-1}W\right]=W. \tag{3.2}$$

The *Whittle relaxation* (2.1)-(3.1) (or (2.3)-(3.2)) can be solved by the Lagrangian relaxation (see, e.g., Guignard, 2003; Visweswaran, 2009), introducing a nonnegative Lagrangian multiplier, say ν, to dualize the constraint. Thus, under the β-discounted criterion we obtain

$$\max_{\pi}\mathbb{E}_{\boldsymbol{n}}^{\pi}\left[\sum_{k\in\mathcal{K}}\sum_{t=0}^{\infty}\beta^{t}\left(R_{k,X_k(t)}^{a_k(t)}-\nu W_{k,X_k(t)}^{a_k(t)}\right)\right]+\nu\frac{W}{1-\beta}.$$

and under the time-average criterion we obtain

$$\max_{\pi} \lim_{T\to\infty} \frac{1}{T}\, \mathbb{E}^{\pi}_{\boldsymbol{n}} \left[\sum_{k\in\mathcal{K}} \sum_{t=0}^{T-1} \left(R^{a_k(t)}_{k,X_k(t)} - \nu W^{a_k(t)}_{k,X_k(t)} \right) \right] + \nu W.$$

The Lagrangian relaxation for every $\nu \geq 0$ yields an *upper bound* for the (maximization) objective value of the multi-armed restless bandit problem. Indeed, it finds a family of solutions (one for each ν) that are better or as good as the optimal solution of the original problem. These solutions, however, typically do not satisfy the sample path capacity constraint, i.e. they are typically not feasible for the multi-armed restless bandit problem. (Note that if the dualized constraint is an equality, then ν is allowed to be negative.)

Strong LP duality yields that there exists $\nu^* \geq 0$ (which under the β-discounted criterion depends on the joint initial state), for which the Lagrangian relaxation achieves the same objective value as the Whittle relaxation. Further, if $\nu^* \neq 0$, then LP complementary slackness assures that any optimal solution to the Lagrangian relaxation satisfies the constraint in the Whittle relaxation (see Niño-Mora, 2001). In other words, any optimal solution to the Lagrangian relaxation is an optimal solution to the Whittle relaxation if $\nu^* \neq 0$.

Finally, for any fixed ν, the Lagrangian relaxation decomposes into single-project subproblems due to their mutual independence. If $\pi_k \in \Pi_k$ is an admissible policy for project k, then the single-project subproblem under the β-discounted criterion is

$$\max_{\pi_k} \sum_{t=0}^{\infty} \beta^t\, \mathbb{E}^{\pi_k}_{n_k} \left[R^{a_k(t)}_{k,X_k(t)} - \nu W^{a_k(t)}_{k,X_k(t)} \right] \tag{3.3}$$

and under the time-average criterion is

$$\max_{\pi_k} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}^{\pi_k}_{n_k} \left[R^{a_k(t)}_{k,X_k(t)} - \nu W^{a_k(t)}_{k,X_k(t)} \right]. \tag{3.4}$$

This parametric optimization problem is addressed in the following sections.

3.2 Restless Bandit Indexation

In the rest of this chapter we outline the *restless bandit indexation* methodology which is crucial to design and efficiently compute marginal productivity (MP) indices for their use in dynamic priority policies. As we have seen in Section 3.1, the multi-armed restless bandit problem can be relaxed and decomposed into subproblems, which can be

viewed as optimization problems of binary-action MDPs with ν-parametric one-period rewards, where ν is interpreted as *wage* per unit of work. In the following we focus on such a single-bandit problem, and we drop the project label k.

The MDP formulation is the following:

- *State space* is $\mathcal{N}$, with N possible states;
- *Actions* $\mathcal{A} := \{0, 1\}$ are available in each state, which are called the passive action (0; resting) and the active action (1; working);
- *Active dynamics*: If the project is in state n and the active action is employed at a given period, then during that period it generates reward R_n^1, a wage for work νW_n^1 must be paid, and the project moves to another state according to the active transition probability matrix $\boldsymbol{P}^1$ for the next period;
- *Passive dynamics*: If the project is in state n and the passive action is employed at a given period, then during that period it generates reward R_n^0, a wage for work νW_n^0 must be paid, and the project moves to another state according to the passive transition probability matrix $\boldsymbol{P}^0$ for the next period.

Note that $R_n^a - \nu W_n^a$ is a parametric one-period (net) reward if action a is applied in state n.

Problems (3.3) and (3.4) are problems of finding an optimal policy for the MDP described above over an infinite time horizon, under the *β-discounted criterion* with discount factor $0 < \beta < 1$ and under the *time-average criterion*, respectively. Our analytical focus will be on the discounted criterion, whose optimal policy given in terms of MP indices can be directly extended to the latter by taking limit $\beta \to 1$. See, e.g., Puterman (2005) for a more thorough discussion on MDP optimization criteria.

Since for finite-state finite-action MDPs there exists an optimal policy that is deterministic, stationary, and independent of the initial state, we narrow our focus only to those policies and represent them via *active sets* $\mathcal{S} \subseteq \mathcal{N}$. In other words, a policy $\mathcal{S}$ prescribes to be active in states in $\mathcal{S}$ and passive in states in $\mathcal{S}^{\mathrm{C}} := \mathcal{N} \setminus \mathcal{S}$. This view is crucial in this approach, as it admits a combinatorial optimization formulation of the ν-wage problem, which we develop next.

Let $\mathbb{E}_i^{\mathcal{S}}$ denote the expectation over the state process $X(t)$ evolving according to $\boldsymbol{P}^0$ and $\boldsymbol{P}^1$ and over the action process $a(t)$ evolving according to $\mathcal{S}$, conditioned on policy $\mathcal{S}$ and on initial state $X(0) = i$. Let us denote by $f_i^{\mathcal{S}}$ the total β-discounted expected reward earned under policy $\mathcal{S}$ starting from initial state i, defined by

$$f_i^{\mathcal{S}} := \mathbb{E}_i^{\mathcal{S}} \left[\sum_{t=0}^{\infty} \beta^t R_{X(t)}^{a(t)} \right]. \tag{3.5}$$

Similarly, $g_i^{\mathcal{S}}$ is the total β-discounted expected work expended under policy $\mathcal{S}$ starting from initial state i, defined by

$$g_i^{\mathcal{S}} := \mathbb{E}_i^{\mathcal{S}}\left[\sum_{t=0}^{\infty} \beta^t W_{X(t)}^{a(t)}\right]. \tag{3.6}$$

Then, formulated for initial state i, the problem (3.3) is solved by solving the following *ν-wage problem*:

$$\max_{\mathcal{S}\subseteq\mathcal{N}} f_i^{\mathcal{S}} - \nu g_i^{\mathcal{S}}. \tag{3.7}$$

Section A.1 shows that (under a regularity condition) this problem can be *normalized*, i.e., it is possible to formulate an equivalent problem by redefining the active-action rewards and works, so that the passive-action rewards and works are zero. This often leads to a simpler analysis than in the non-normalized case.

3.3 Marginal Productivity Index and Indexability

Niño-Mora (2006b) coined the term marginal productivity index based on the methodology outlined next. Note that, for simplicity of exposition, we assume that there is no uncontrollable state (i.e., there is no state n satisfying $R_n^0 = R_n^1$, $W_n^0 = W_n^1$, and $\boldsymbol{P}_n^0 = \boldsymbol{P}_n^1$ at the same time).

The crucial observation for solving problem (3.7) was made by Whittle (1988) and is the following. Under some (problem-specific) natural regularity conditions, the (minimal) optimal active sets $\mathcal{S}(\nu)$ for increasing values of the wage parameter ν will be nested: either monotonically expanding or diminishing. Then, we can attach to each state n a (minimum) value of the wage parameter ν, called the MP index ν_n, below which n enters $\mathcal{S}(\nu)$. Note that if the wage $\nu = \nu_n$, then both the actions are optimal in state n.

Definition 3.1. We say that the project is *indexable* if there exists an index ν_n for $n \in \mathcal{N}$ such that

$$\mathcal{S}(\nu) = \{n \in \mathcal{N} : \nu_n > \nu\}, \text{ for all } \nu.$$

In such a case we say that ν_n is the project's marginal productivity (MP) index.

That is, the set of MP indices ν_n for all $n \in \mathcal{N}$ (if they exist) defines an optimal *MP index policy*: "Work if and only if the MP index of the current state is greater than the wage parameter ν."

Let $\langle a, \mathcal{S}\rangle$ be the policy that implements action a in the initial period and policy $\mathcal{S}$ proceeds. We consider the $(n, \mathcal{S})$-*marginal reward* defined as $r_n^{\mathcal{S}} := f_n^{\langle 1,\mathcal{S}\rangle} - f_n^{\langle 0,\mathcal{S}\rangle}$, and the $(n, \mathcal{S})$-*marginal work* defined as $w_n^{\mathcal{S}} := g_n^{\langle 1,\mathcal{S}\rangle} - g_n^{\langle 0,\mathcal{S}\rangle}$. Finally, let $\nu_n^{\mathcal{S}} := r_n^{\mathcal{S}}/w_n^{\mathcal{S}}$ be the $(n, \mathcal{S})$-*marginal productivity rate* provided that the denominator does not vanish.

If the problem is indexable, then the index ν_n must be equal to the $(n, \mathcal{S})$-marginal productivity rate for some active set $\mathcal{S}$. Indeed: suppose that the wage parameter $\nu = \nu_n$, then by the definition of ν_n there is a policy $\mathcal{S}$ such that the objective is not altered by adding or removing state n from $\mathcal{S}$, i.e.,

$$f_n^{\mathcal{S}\cup\{n\}} - \nu_n g_n^{\mathcal{S}\cup\{n\}} = f_n^{\mathcal{S}\setminus\{n\}} - \nu_n g_n^{\mathcal{S}\setminus\{n\}}. \tag{3.8}$$

Moreover, notice that under $\nu = \nu_n$ any action can be applied anytime the system is in state n without altering the value of the objective function. In particular, we can apply policies $\langle 1, \mathcal{S}\setminus\{n\}\rangle$, $\langle 0, \mathcal{S}\setminus\{n\}\rangle$, $\langle 1, \mathcal{S}\cup\{n\}\rangle$, $\langle 0, \mathcal{S}\cup\{n\}\rangle$ interchangeably. Hence we have

$$f_n^{\langle 1,\mathcal{S}\setminus\{n\}\rangle} - \nu_n g_n^{\langle 1,\mathcal{S}\setminus\{n\}\rangle} = f_n^{\langle 0,\mathcal{S}\setminus\{n\}\rangle} - \nu_n g_n^{\langle 0,\mathcal{S}\setminus\{n\}\rangle} \tag{3.9}$$

and

$$f_n^{\langle 1,\mathcal{S}\cup\{n\}\rangle} - \nu_n g_n^{\langle 1,\mathcal{S}\cup\{n\}\rangle} = f_n^{\langle 0,\mathcal{S}\cup\{n\}\rangle} - \nu_n g_n^{\langle 0,\mathcal{S}\cup\{n\}\rangle}. \tag{3.10}$$

So, using the definition of the marginal productivity rate, we can write

$$\nu_n = \nu_n^{\mathcal{S}\setminus\{n\}} = \nu_n^{\mathcal{S}\cup\{n\}} \text{ for some } \mathcal{S}, \tag{3.11}$$

and therefore the term *marginal productivity* index. Notice from the identities above that index ν_n is the value of the wage parameter ν for which marginal reward equals marginal work cost.

3.4 Sufficient Indexability Condition and Adaptive-Greedy Algorithm

Establishing indexability in general is a cumbersome task, however, we implement an analytically tractable sufficient condition called *PCL($\mathcal{F}$)-indexability*. This sufficient condition draws on polyhedral arguments of having a problem satisfied the *partial conservation laws* (PCL) for a postulated family of active sets $\mathcal{F} \subseteq 2^{\mathcal{N}}$ (see Niño-Mora, 2001, 2002, 2006b). An adaptive-greedy algorithm $AG_{\mathcal{F}}$ exhibited in Figure 3.1 calculates candidates $\widehat{\nu}_{n_k}$ for MP indices and candidates $\widehat{\mathcal{S}}_k$ for optimal active sets.

The postulated family of active sets $\mathcal{F}$ must satisfy certain connectivity assumptions, so that all the steps in the adaptive-greedy algorithm are well-defined. As a simplest

```
set  $\widehat{\mathcal{S}}_0 := \emptyset$;
for  $k := 0$  to  $N-1$  do
   choose  $n_k \in \operatorname{argmax}\{\nu_n^{\widehat{\mathcal{S}}_k} : n \in \widehat{\mathcal{S}}_k^{\mathrm{C}}$  and  $\widehat{\mathcal{S}}_k \cup \{n\} \in \mathcal{F}\}$;
   set  $\widehat{\mathcal{S}}_{k+1} := \widehat{\mathcal{S}}_k \cup \{n_k\}$;
   set  $\widehat{\nu}_{n_k} := \nu_{n_k}^{\widehat{\mathcal{S}}_k}$;
end {for};
```

Figure 3.1: Adaptive-greedy algorithm $AG_{\mathcal{F}}$.

example, in a single-queue admission control problem we may postulate the family $\mathcal{F}$ to be the set of all the threshold policies that prescribe to admit customers while the queue is shorter than a particular threshold.

Definition 3.2. Let $\widehat{\nu}_{n_k}$ under a sequence of choices of n_k for $k = 0, 1, \ldots, N-1$ be the output of the adaptive-greedy algorithm $AG_{\mathcal{F}}$ for the ν-wage problem (3.7). The problem (3.7) is called *PCL($\mathcal{F}$)-indexable* if the following two conditions hold:

(i) the $(n, \mathcal{S})$-marginal work $w_n^{\mathcal{S}}$ is positive for all $n \in \mathcal{N}$ and $\mathcal{S} \in \mathcal{F}$;

(ii) the quantities $\widehat{\nu}_{n_k}$ are nonincreasing in k.

The following is the main methodological result (see Niño-Mora, 2001, 2002).

Proposition 3.1. *If a problem is PCL($\mathcal{F}$)-indexable, then it is indexable with marginal productivity indices $\nu_n := \widehat{\nu}_n$, where $\widehat{\nu}_n$ is the output of the adaptive-greedy algorithm $AG_{\mathcal{F}}$.*

3.5 Geometric Interpretation

Under indexability, each of the active sets $\widehat{\mathcal{S}}_0, \widehat{\mathcal{S}}_1, \ldots, \widehat{\mathcal{S}}_N$ calculated by the algorithm $AG_{\mathcal{F}}$ is optimal for some value of the wage parameter ν. In particular, the active set $\widehat{\mathcal{S}}_0$ is optimal for $\nu \geq \nu_{n_0}$, the active set $\widehat{\mathcal{S}}_1$ is optimal for $\nu_{n_0} \geq \nu \geq \nu_{n_1}$, etc., and the active set $\widehat{\mathcal{S}}_N$ is optimal for $\nu_{n_{N-1}} \geq \nu$. In geometric terms, these sets determine the *upper boundary* of the achievable work-reward region of the ν-wage parametric problem (3.7), and the MP indices are the *slopes* of the lines connecting the performance vectors of subsequent optimal policies.

These concepts, well-known in the bi-criteria (parametric) linear programming literature, are illustrated in Figure 3.2. In fact, Niño-Mora (2007a) elucidated that the adaptive-greedy algorithm is analogous to the parametric Simplex method introduced in Saaty and Gass (1954).

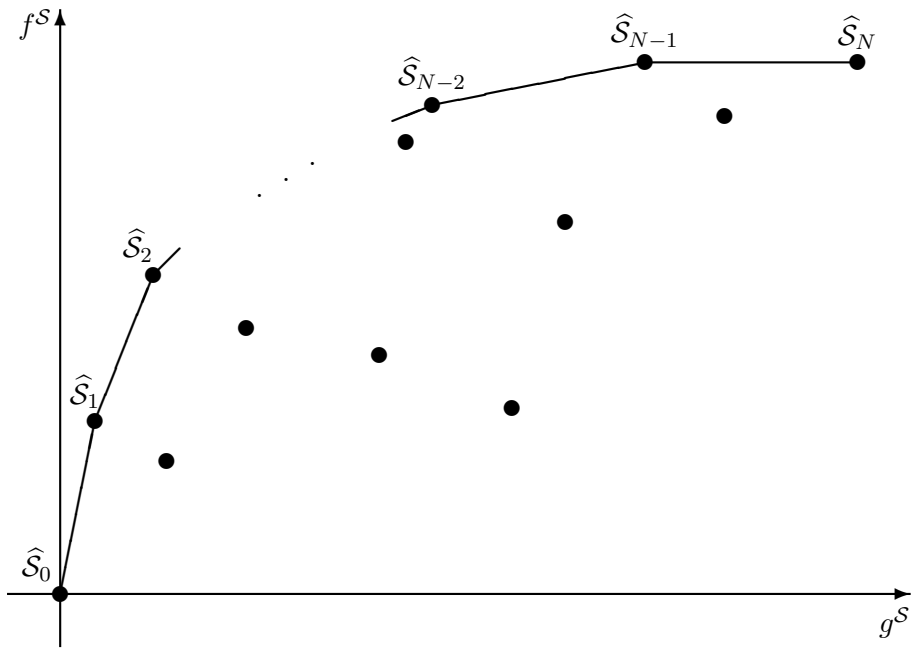


Figure 3.2: An illustration of the achievable work-reward region leading to optimal active sets $\widehat{\mathcal{S}}_0, \widehat{\mathcal{S}}_1, \ldots, \widehat{\mathcal{S}}_N$.

Road is created walking.
< A. Machado >

Chapter 4

Admission Control and Routing with Delayed Information

4.1 Introduction

This chapter addresses the problem of designing and computing a tractable heuristic policy for dynamic job admission control and/or routing in a discrete time Markovian model of parallel loss queues with one-period delayed state observation and/or action implementation, which comes close to optimizing an infinite-horizon problem under several objectives. Two versions of the model are considered, depending on whether the admission control capability is enabled or not. The queue servers may be endowed with finite or infinite buffer space.

We consider the following three performance objectives: (i) minimization of the expected total discounted sum of holding costs and rejection costs, (ii) minimization of the expected time-average sum of holding costs and rejection costs, and (iii) maximization of the expected time-average number of job completions. Holding costs are assumed to be convex and nondecreasing in the number of jobs queued in the buffer space.

Such problems are relevant in a variety of application domains, most notably in the operation of packet-switched communication networks and distributed computer systems. In such systems there are nonnegligible propagation delays, which force the controller to make decisions based on stale system state information and take effect only after a time lag. Additional recent applications include long-distance-controlled robots, and situations in which an advanced processing of observations is necessary.

As for our considering joint admission control and routing problems, instead of restricting attention to the conventional pure-routing case, the motivation is that it allows the system designer to take into account the tradeoff between rejection and holding

costs. The key insight is that, when the system is heavily congested, denying access to further arrivals until the congestion is sufficiently reduced can substantially decrease holding costs at a relatively small expense in terms of increased rejection costs.

The above problems are naturally formulated as *partially observed Markov decision processes* (POMDPs), which in turn are readily reformulated as conventional *Markov decision processes* (MDPs) by redefining the state of each queue as the *augmented state* build up of the last observed queue length and the actions applied since then (Brooks and Leondes, 1972). Computation of optimal policies for the resultant multidimensional MDPs by solving the associated *dynamic programming* (DP) equations is, however, hindered by the curse of dimensionality in large-scale models. We will thus focus attention on the more realistic and practical goals of designing and computing well-grounded heuristic policies that are readily implementable. Since in such problems the controller must dynamically assess the relative values of alternative rejection and routing actions, it is intuitively appealing to do so based on an *index policy*, defined after the model description below.

4.1.1 Model Description

Time is slotted into discrete-time *epochs* $t = 0, 1, 2, \ldots$ The system consists of N independent parallel queues with servers and a gate. Queue $n \in \mathcal{N} := \{1, \ldots, N\}$ is endowed with a (possibly infinite) buffer with room for holding $I_n \geq 1$ jobs waiting or in service, and has a single geometric server, which serves jobs in FCFS order and completes the service of a job at the end of a period with probability $0 < \mu_n < 1$.

Jobs arrive to the system as a Bernoulli stream with probability $0 < \lambda \leq 1$ of arrival at the beginning of each period. Upon a job's arrival to the gate, a central controller (gatekeeper) must decide: (i) in the case that admission control is enabled, whether to admit the job or to reject it (admission function); and, if admitted, (ii) to which of N queues in parallel to route the job for service (router function). We assume that a customer that is admitted and routed to an empty queue starts to be served immediately, and therefore may leave the system at the end of the same period if the service is completed.

Denote by $X_n(t)$ the *state* of queue n at the beginning of period t, given by the number of jobs it holds waiting or in service, and by $a_n(t) \in \{0, 1\}$ the *action* indicator that takes the value 1 when a job arriving at time t *is not to be routed to queue* n. We assume that at the start of period t the controller does not know the current state, but has information on previous states and actions, knowing in particular $X_n(t-1)$ and $a_n(t-1)$ for each queue n. Thus, we deal with the problem with a *one-period delay*.

Action choice is based on adoption of an admission and routing policy (if admission

function is enabled), or just a routing policy (if it is not), denoted by π. This is to be chosen from the corresponding class Π of (possibly randomized) policies that use previous state and action information.

4.1.2 Performance Objectives

For two of our performance objectives defined below we will assume that the system incurs per-queue *holding costs* at rate $c_n(i_n)$ per period during which i_n jobs are held in queue n, such that $c_n(i_n)$ is convex and nondecreasing in i_n for each queue n. The system further incurs *loss costs* at rate ν per rejected job, due either to active rejection (not admitting) or to forced rejection (when an admitted job finds the buffer to which it is routed full).

We will find it convenient to formulate the *overal cost* incurred in a period in which the joint system state is $\boldsymbol{i} = (i_n)$ and action $\boldsymbol{a} = (a_n)$ prevails as a constant plus a term that is separably additive across queues, using the identity

$$\sum_{n\in\mathcal{N}} c_n(i_n) + \nu\lambda\left[1 - \sum_{n\in\mathcal{N}}(1-a_n)\right] = -(N-1)\lambda\nu + \sum_{n\in\mathcal{N}}\left[c_n(i_n) + \nu\lambda a_n\right].$$

Note that the term $1 - \sum_{n\in\mathcal{N}}(1-a_n)$ in the above equality takes the value 1 if an arrival is to be rejected ($a_n = 1$ for every queue n), and takes the value 0 otherwise ($a_n = 0$ for exactly one queue n).

For the third performance objective we will denote by $c'_n(a_n, i_n)$ the expected rate of *job completions* per period during which i_n jobs are held in queue n and action a_n prevails, i.e.,

$$c'_n(a_n, i_n) := \begin{cases} 0, & \text{if } i_n = 0 \text{ and } a_n = 1, \\ \lambda\mu_n, & \text{if } i_n = 0 \text{ and } a_n = 0, \\ \mu_n, & \text{if } i_n \geq 1. \end{cases}$$

Let $\mathbb{E}^{\pi}_{(\boldsymbol{a},\boldsymbol{i})}[\cdot]$ denote expectation under policy π conditioned on the initial previous joint action and state vectors being equal to $\boldsymbol{a}(-1) := \boldsymbol{a} = (a_n)$ and $\boldsymbol{X}(-1) := \boldsymbol{i} = (i_n)$. The operation of such a system raises the following performance optimization problems:

(i) find a policy minimizing the expected total discounted sum of holding costs and

rejection costs,

$$\min_{\pi\in\Pi} \mathbb{E}^{\pi}_{(\boldsymbol{a},\boldsymbol{i})} \left[\sum_{t=0}^{\infty} \sum_{n\in\mathcal{N}} \{c_n(X_n(t)) + \nu\lambda a_n(t)\}\, \beta^t \right], \tag{4.1}$$

where $0 < \beta < 1$ is the discount factor;

(ii) find a policy minimizing the expected time-average sum of holding costs and rejection costs,

$$\min_{\pi\in\Pi} \limsup_{T\to\infty} \frac{1}{T}\, \mathbb{E}^{\pi}_{(\boldsymbol{a},\boldsymbol{i})} \left[\sum_{t=0}^{T} \sum_{n\in\mathcal{N}} \{c_n(X_n(t)) + \nu\lambda a_n(t)\} \right]; \tag{4.2}$$

(iii) find a policy maximizing the expected time-average number of job completions,

$$\max_{\pi\in\Pi} \liminf_{T\to\infty} \frac{1}{T}\, \mathbb{E}^{\pi}_{(\boldsymbol{a},\boldsymbol{i})} \left[\sum_{t=0}^{T} \sum_{n\in\mathcal{N}} c'_n(a_n(t), X_n(t)) \right]. \tag{4.3}$$

4.1.3 Index Policies

As mentioned earlier, a way to formulate present model as an MDP is to redefine the state of each queue n as the augmented state $\widetilde{X}_n(t) := (a_n(t-1), X_n(t-1))$, and use the joint state and action process $\widetilde{\boldsymbol{X}}(t) := (\widetilde{X}_n(t))$ and $\boldsymbol{a}(t) := (a_n(t))$.

In the present model, index policies are based on attaching to each queue n a numeric index $\nu_n(a_n, i_n)$, which can be thought of as a *measure of undesirability of routing a job to queue* n, given as a function of the queue's augmented state, which we denote by (a_n, i_n) and emphasize that it refers to the observed action-state pair at the previous period. We note that we allow the index to be undefined for certain (uncontrollable) states. Further, under the time-average criteria (4.2) and (4.3), if the (first-order) index $\nu_n(a_n, i_n)$ is defined and constant, then we define a second-order index $\gamma_n(a_n, i_n)$. (When the first-order index is undefined, then the second-order index is undefined as well.)

The resultant index policy prescribes the following actions, when at time t the augmented state of each queue n is known to be $\widetilde{X}_n(t) = (a_n, i_n)$:

- under objective (4.1),
  - in the problem version with admission control capability, the policy prescribes to admit an arriving job if $\nu > \nu_n(a_n, i_n)$ for at least one queue n such that $\nu_n(a_n, i_n)$ is defined, i.e., if the cost of rejecting the job exceeds the undesirability of routing it to some queue; otherwise, the job is rejected;

  - if admitted, the job is routed to a queue
    - of *lowest index* $\nu_n(a_n, i_n)$, breaking ties arbitrarily, among those queues n for which $\nu_n(a_n, i_n)$ is defined and $\nu > \nu_n(a_n, i_n)$, if at least one such queue exists;
    - of *undefined index* $\nu_n(a_n, i_n)$, breaking ties arbitrarily, if there is no queue n for which $\nu_n(a_n, i_n)$ is defined and $\nu > \nu_n(a_n, i_n)$ and if at least one queue with undefined index exists;
    - of *lowest index* $\nu_n(a_n, i_n)$, breaking ties arbitrarily, in all the remaining cases;

- under objectives (4.2) and (4.3),
  - in the problem version with admission control capability, the policy prescribes to admit an arriving job if $\nu > \nu_n(a_n, i_n)$ for at least one queue n such that $\nu_n(a_n, i_n)$ is defined, i.e., if the cost of rejecting the job exceeds the undesirability of routing it to some queue; otherwise, the job is rejected;
  - if admitted, the job is routed to a queue
    - of *lexicographically lowest index pair* $(\nu_n(a_n, i_n), \gamma_n(a_n, i_n))$, breaking ties arbitrarily, among those queues n for which $\nu_n(a_n, i_n)$ and $\gamma_n(a_n, i_n)$ are defined and $\nu > \nu_n(a_n, i_n)$, if at least one such queue exists;
    - of *undefined index* $\nu_n(a_n, i_n)$, breaking ties arbitrarily, if there is no queue n for which $\nu_n(a_n, i_n)$ and $\gamma_n(a_n, i_n)$ are defined and $\nu > \nu_n(a_n, i_n)$, and if at least one queue with undefined index exists;
    - of *lexicographically lowest index pair* $(\nu_n(a_n, i_n), \gamma_n(a_n, i_n))$, breaking ties arbitrarily, in all the remaining cases.

Note that such policies may well prescribe to admit and route a job to a queue that is actually full, unbeknownst to the controller, in which case the job will be blocked and hence rejected.

For the special case of the pure-routing problem under objectives (4.1) and (4.2) in which there are two symmetric infinite-buffer queues and linear holding cost ($N = 2$, $\mu_n \equiv \mu$, and $c_n(i) \equiv i$), it was shown in Kuri and Kumar (1995) that an index policy is optimal: the *Join the Shortest Expected Queue* (JSEQ) rule, where the JSEQ index of a

queue n is defined as

$$\nu_n^{\text{JSEQ}}(a_n, i_n) := \begin{cases} i_n - \mu, & \text{if } a_n = 1, i_n \geq 1, \\ 0, & \text{if } a_n = 1, i_n = 0, \\ i_n + \lambda - \mu, & \text{if } a_n = 0, i_n \geq 1, \\ \lambda(1-\mu), & \text{if } a_n = 0, i_n = 0, \end{cases}$$

where the index represents the expected value of $X_n(t)$ conditioned on $(a_n(t-1), X_n(t-1)) = (a_n, i_n)$. Such a result partially extends to queues with delays classical results in (Winston, 1977; Hordijk and Koole, 1990) for symmetric queues without delays on optimality of the *Join the Shortest (Nonfull) Queue* (JSQ) rule.

For the case of routing to two nonsymmetric queues with infinite buffers, in which index policies need no longer be optimal, Artiges (1995) showed (in a variation on the above model) that the optimal routing policy is characterized by a monotone switching curve, extending a classical result in Hajek (1984) for a model without delayed information. Still, one can easily devise a variety of heuristic routing index rules by defining indices based on ad hoc arguments, analogously to the *Shortest Expected Delay* routing rule in Houck (1987). Yet, a drawback of such conventional indices, which typically measure a queue's expected weighted load, is that they only give a routing rule, being of no use to obtain a reasonable combined admission control and routing rule as outlined above, since consideration of rejection costs does not play a role in their definition.

Hence, we are led to address the issue of defining appropriate indices $\nu_n(a_n, i_n)$ for the above admission control and routing problems. Instead of proposing some ad hoc index via heuristic arguments, we will deploy a unifying fundamental design principle for priority allocation policies in *multiarmed restless bandit problems* (MARBPs), of which (4.1), (4.2), and (4.3) are special cases, based on the economically intuitive concept of *marginal productivity* (MP) index.

Such an approach was introduced in Whittle (1988), and has been developed and applied in a variety of models by the second author in work including Niño-Mora (2001, 2002, 2006b,a), which was reviewed in Niño-Mora (2007b). In particular, Niño-Mora (2002, 2007c) introduced such an approach to the design of index policies for admission control and routing to parallel exponential queues without delayed information. As for use of MP index policies for problems with delayed state information, they were introduced in Niño-Mora (2007d) in the setting of a dynamic scheduling model.

In the present setting, and focusing for concreteness on discounted problem (4.1) under combined admission control and routing, such a restless bandit indexation approach is based on decoupling the problem into individual single-queue admission con-

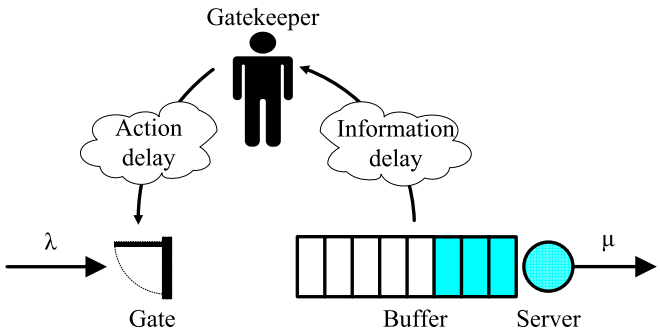


Figure 4.1: A design of the single-queue admission control problem with delay. The gatekeeper's work consists of shutting and opening the entry gate, thus rejecting some of the arriving customers.

trol subproblems, one for each queue $n \in \mathcal{N}$:

$$\min_{\pi_n \in \Pi_n} \mathbb{E}^{\pi_n}_{(a_n,i_n)} \left[\sum_{t=0}^{\infty} \{c_n(X_n(t)) + \nu\lambda a_n(t)\} \beta^t \right], \tag{4.4}$$

where Π_n denotes the class of admission control policies based on one-period delayed state observation for operating queue n *in isolation*, and $\mathbb{E}^{\pi_n}_{(a_n,i_n)}[\cdot]$ denotes expectation conditioned on the initial observed state and action pair being equal to $\widetilde{X}_n(0) := (a_n(-1), X_n(-1)) = (a_n, i_n)$. Note that, in such a setting, taking action $a_n(t) = 1$ at period t means denying access to potential arrivals, which can be conveniently visualized as the action of *shutting the queue's entry gate* which is taken by the *gatekeeper* (see Figure 4.1).

Problem (4.4) is a single *restless bandit problem* (RBP), i.e., a binary-action ($a_n(t) = 1$: active; $a_n(t) = 0$: passive) MDP, on which we can deploy the powerful theoretical and algorithmic results available for *restless bandit indexation* (cf. Niño-Mora, 2007b). Let us say that problem (4.4) is *indexable* if there exists an index $\nu_n^{\text{MPI}}(a_n, i_n)$ that characterizes its optimal policies for every real value of the rejection cost parameter ν, as follows: it is optimal to take the active action (shut the entry gate) in augmented state $\widetilde{X}_n(t) = (a_n, i_n)$ if $\nu_n^{\text{MPI}}(a_n, i_n) \geq \nu$ and it is optimal to take the passive action (open the entry gate) in augmented state $\widetilde{X}_n(t) = (a_n, i_n)$ if $\nu_n^{\text{MPI}}(a_n, i_n) \leq \nu$.

In such a case, we term $\nu_n^{\text{MPI}}(a_n, i_n)$ the queue's MP index, due to its economic interpretation as a measure of the rate of marginal reduction in expected holding cost relative to the marginal increase in expected rejections that results from shutting the gate in state (i_n, a_n) instead of opening it, which characterizes the expected holding cost versus rejections tradeoff curve. Such is the index we propose to use as the basis for designing an index rule for admission control and/or routing for the multi-queue

problems of concern.

4.1.4 Goals and Contributions

Two issues need thus be addressed: (i) show that problem (4.4) is indeed indexable; and (ii) design an efficient index-computing algorithm. As for the first issue, we will deploy the sufficient indexability conditions based on *partial conservation laws* (PCLs) introduced in Niño-Mora (2001, 2002). Such conditions require one to identify a family of stationary deterministic policies among which an optimal policy for problem (4.4) exists for every value of the parameter ν.

For such a purpose, we draw on results in Altman and Nain (1992) and Kuri and Kumar (1995) that characterize the structure of optimal policies for such a single-queue admission control problem (in an infinite-buffer model) with one-period delayed state information. Such work shows that it suffices to consider policies that are characterized by two thresholds $k_1 \geq k_0 \geq 0$, as follows: if the previous observed number of jobs in the system was i and the previous action was to open, i.e., $a = 0$ (resp. shut, i.e., $a = 1$) the queue's entry gate, the (k_0, k_1)-policy prescribes to shut the gate iff $i > k_0$ (resp. iff $i > k_1$).

The intuition behind such a result is that, if it is optimal to shut the entry gate given that it was previously shut, then, other things being equal, it should also be optimal to shut it when it was previously open, as in the latter case the actual number of jobs in the system cannot be smaller than in the former. It is further shown in Altman and Nain (1992) that one need only consider threshold pairs that differ in at most one unit: $0 \leq k_1 - k_0 \leq 1$. Note that, in order to be consistent with such *bi-threshold policies*, the MP index $\nu_n^{\text{MPI}}(a_n, i_n)$ must be monotone nondecreasing in i_n for both $a_n \in \{0, 1\}$, and must satisfy $\nu_n^{\text{MPI}}(0, i_n) \geq \nu_n^{\text{MPI}}(1, i_n)$.

As for the second issue, that of index computation, provided PCL-indexability is established relative to such a family of policies, one can use the adaptive-greedy algorithm introduced in Niño-Mora (2001, 2002) to compute the MP indices. Using the general fast-pivoting implementation given in Niño-Mora (2007a) such an algorithm has a cubic arithmetic operation complexity in the number of restless bandit states, which in the present setting corresponds to an $\mathcal{O}(I^3)$ operation count. While tractable, such a complexity can be overly burdensome for online computation in high-speed communication switches.

Relative to the above two issues, this chapter presents the following contributions: (i) it shows that problem (4.4) is PCL-indexable relative to bi-threshold policies, which ensures both existence of the MP index and the validity of the adaptive-greedy algorithm for its computation; (ii) by exploiting special structure, a substantially faster in-

dex algorithm is presented that computes the MP indices in $\mathcal{O}(I)$ operations; (iii) we present an algorithm to calculate the MP indices for any particular state by performing at most $\mathcal{O}(\log_2 I)$ arithmetic operations; (iv) validity of the same algorithm for the MP indices under the time-average criterion to be used for (4.2) is established; (v) the MP index to be used for (4.3) is obtained by the same algorithm and shown to be constant, and the second-order MP index is derived.

An extensive computational study testing the performance of the proposed index policies will be included in the final (paper) version of this chapter.

4.2 MDP Formulations

In order to draw an analogy, in this section we formulate as a Markov decision process (MDP) both the admission control problem without delay and the admission control problem with a one-period delay. Since the problem considered in the following is a single-queue case, we drop the subscript n from the notation. For concreteness, in this section we focus on the objective (4.1); for the problem of maximum expected number of job completions one would simply replace the holding cost c_i for the completion reward $-c'(a,i)$.

4.2.1 Admission Control Problem

First we formulate as an MDP the no-delay admission control problem. Let $X(t)$ be the state process, denoting the *queue length* (including customers in service, if any) at time epoch t. If $a(t)$ denotes the action process, then the task at time epoch t is to choose between closing the gate ($a(t) = 1$) and letting the gate open ($a(t) = 0$). The MDP elements are as follows:

- The *action space* is denoted by $\mathcal{A} := \{0, 1\}$.
- The *state space* is $\mathcal{I} := \{0, 1, \ldots, I\}$, where state $i \in \mathcal{I}$ represents the number of customers in the buffer or in service.
- Denoting by $\zeta := \lambda(1-\mu)$, $\eta := \mu(1-\lambda)$, and $\varepsilon := 1-\zeta-\eta$, the *one-period transition probabilities* $p_{ij}^a := \mathbb{P}\left[X(t) = j | X(t-1) = i, a(t-1) = a\right]$ from state $1 \leq i \leq I-1$ to state j under action a are

$$p_{ij}^0 = \begin{cases} \eta & \text{if } j = i-1 \\ \varepsilon & \text{if } j = i \\ \zeta & \text{if } j = i+1 \end{cases} \qquad p_{ij}^1 = \begin{cases} \mu & \text{if } j = i-1 \\ 1-\mu & \text{if } j = i \end{cases} \tag{4.5}$$

and for the boundary cases, $p^1_{00} = 1$, and

$$p^0_{0j} = \begin{cases} 1-\zeta & \text{if } j = 0 \\ \zeta & \text{if } j = 1 \end{cases} \qquad p^a_{Ij} = \begin{cases} \mu & \text{if } j = I-1 \\ 1-\mu & \text{if } j = I \end{cases} \tag{4.6}$$

The remaining transition probabilities are zero.

- If the queue length is $i \in \mathcal{I}$ and action $a \in \mathcal{A}$ is chosen, then the gatekeeper's *one-period reward* is defined as the negative of the expected holding cost at the current epoch,

$$R^a_i := -c_i.$$

At the same time, the gatekeeper's *one-period work* is defined as the expected number of rejected customers during the current period,

$$W^1_i := \lambda \qquad W^0_i := \begin{cases} \lambda & \text{if } i = I \\ 0 & \text{otherwise} \end{cases}$$

Thus, for rejection cost (gatekeeper's wage) ν, the *one-period overal cost* is

$$-R^a_i + \nu W^a_i = c_i + \lambda\nu a + (1-a)\mathbf{1}\{i = I\}\lambda\nu,$$

where $\mathbf{1}\{Y\}$ is the $0/1$ indicator function of statement Y.

Given the definition above, we call state I *uncontrollable*, because in this state both the actions result in identical consequences (for having identical one-period reward, one-period work, and transition probabilities), and there is actually no decision to make. This is not the case for the remaining states, henceforth called *controllable*.

Finally, to ease later reference we summarize here our model parameters assumptions:

$$0 < \beta < 1, \quad 0 < \lambda \leq 1, \quad 0 < \mu < 1, \quad 0 \leq \eta < 1, \quad 0 < \varepsilon < 1, \quad 0 < \zeta < 1. \tag{4.7}$$

4.2.2 Admission Control Problem with Delay

In this subsection we follow the classic reformulation as MDP of problems with a discrete-time delay, which is a special case of *partially observed MDPs*, by augmenting the state space (Brooks and Leondes, 1972).

In the admission control problem with delay[1], the decision at epoch t is based on $\widetilde{X}(t) := (a(t-1), X(t-1))$, which is henceforth called an *augmented state* process. Thus, $\widetilde{X}(t)$ is the observed state at time epoch t, while $X(t)$ is the actual (hidden) queue length process. The MDP elements of the admission control problem with delay are as follows:

- The *action space* is $\mathcal{A}$ as in the no-delay problem.
- Recall that in the no-delay problem, state I is uncontrollable. Consequently, states $(0, I)$ and $(1, I)$ in the problem with delay are *duplicates*, having identical one-period reward, one-period work, and transition probabilities, so they can and should be merged into a unique state $(*, I)$. We therefore define the *augmented state space*

$$\widetilde{\mathcal{I}} := (\mathcal{A} \times \{0, 1, \ldots, I-1\}) \cup \{(*, I)\}.$$

- The *one-period transition probabilities* are

$$\begin{aligned} p^{a'}_{(a,i),(b,j)} &:= \mathbb{P}\left[\widetilde{X}(t+1) = (b,j) | \widetilde{X}(t) = (a,i), a(t) = a'\right] \\ &= \mathbb{P}\left[X(t) = j, a(t) = b | X(t-1) = i, a(t-1) = a, a(t) = a'\right] \\ &= p^a_{ij} \cdot \mathbf{1}\{a' = b\}. \end{aligned}$$

  For the merged state $(*, I)$, we have $p^{a'}_{(a,i),(*,I)} := p^{a'}_{(a,i),(0,I)} + p^{a'}_{(a,i),(1,I)} = p^a_{ij}$.

- If the current-epoch augmented state is (a, i), then the gatekeeper's *one-period reward* is defined as the negative of the expected holding cost at the current epoch,

$$\overline{R}^b_{(a,i)} := \mathbb{E}\left[R^b_{X(t)} | a(t-1) = a, X(t-1) = i\right].$$

  Similarly, the gatekeeper's *one-period work* is defined as the expected number of rejected customers during the current period,

$$\overline{W}^b_{(a,i)} := \mathbb{E}\left[W^b_{X(t)} | a(t-1) = a, X(t-1) = i\right].$$

  Thus, for rejection cost (gatekeeper's wage) ν, the *one-period overal cost* is $-\overline{R}^b_{(a,i)} + \nu\overline{W}^b_{(a,i)}$.

The above one-period reward and one-period work can be explicitly stated as fol-

[1] We use the 'tilded' notation for the delayed version when not doing so might be confusing; note that state-dependent quantities are easy to distinguish since the original state is uni-dimensional, while the augmented state of the delayed problem is bi-dimensional.

lows:

$$\overline{R}^b_{(a,i)} := \begin{cases} 0, & \text{if } (a,i) = (1,0), \\ -[(1-\zeta)c_0 + \zeta c_1], & \text{if } (a,i) = (0,0), \\ -[\mu c_{i-1} + (1-\mu)c_i], & \text{if } a = 1 \text{ and } 1 \le i \le I-1, \\ -[\eta c_{i-1} + \varepsilon c_i + \zeta c_{i+1}], & \text{if } a = 0 \text{ and } 1 \le i \le I-1, \\ -[\mu c_{I-1} + (1-\mu)c_I], & \text{if } (a,i) = (*,I). \end{cases}$$

$$\overline{W}^b_{(a,i)} := \begin{cases} \lambda, & \text{if } b = 1, \\ \zeta, & \text{if } b = 0 \text{ and } (a,i) = (*,I), \\ \zeta\lambda, & \text{if } b = 0 \text{ and } (a,i) = (0,I-1), \\ 0, & \text{otherwise.} \end{cases}$$

To evaluate a policy π under the discounted criterion, we consider the following two measures. Let $\overline{g}^\pi_{(a,i)}$ be the *expected total β-discounted work* (or, the expected total β-discounted number of rejected customers) if starting from state $(a(-1), X(-1)) := (a,i)$ under policy π,

$$\overline{g}^\pi_{(a,i)} := \mathbb{E}^\pi_{(a,i)} \left[\sum_{t=0}^{\infty} \beta^t \overline{W}^{a(t)}_{(a(t-1),X(t-1))} \right].$$

Analogously is defined $\overline{f}^\pi_{(a,i)}$, the *expected total β-discounted reward* if starting from state $(a(-1), X(-1)) := (a,i)$ under policy π,

$$\overline{f}^\pi_{(a,i)} := \mathbb{E}^\pi_{(a,i)} \left[\sum_{t=0}^{\infty} \beta^t \overline{R}^{a(t)}_{(a(t-1),X(t-1))} \right].$$

If the rejection cost ν is interpreted as the wage paid to gatekeeper for each rejected customer, then the objective is to solve the following ν-wage problem for each ν:

$$\min_{\pi \in \Pi} -\overline{f}^\pi_{(a,i)} + \nu \overline{g}^\pi_{(a,i)}, \tag{4.8}$$

where Π is the set of all non-anticipative control policies.

Next we present alternative, simpler definitions of one-period work and one-period reward, and show that they lead to an equivalent problem. These alternative definitions capture the reward and work one period earlier comparing to the original ones. If the current-epoch augmented state is (a,i), then the alternative gatekeeper's one-period

reward is defined as the negative of the expected holding cost at the previous epoch,

$$R^b_{(a,i)} := \beta(-c_i/\beta) = -c_i. \tag{4.9}$$

Similarly, the alternative gatekeeper's one-period work is defined as the expected number of rejected customers during the previous period,

$$W^b_{(1,i)} := \lambda \qquad W^b_{(0,i)} := \begin{cases} \lambda & \text{if } i = I \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

Notice that we have $R^b_{(a,i)} = R^a_i$ and $W^b_{(a,i)} = W^a_i$.

Then, the alternative expected total β-discounted work if starting from state $(a,i) := (a(-1), X(-1))$ under policy π is

$$g^\pi_{(a,i)} := \mathbb{E}^\pi_{(a,i)} \left[\sum_{t=0}^{\infty} \beta^t W^{a(t)}_{(a(t-1),X(t-1))} \right]. \tag{4.11}$$

Analogously, the alternative expected total β-discounted reward if starting from state $(a,i) := (a(-1), X(-1))$ under policy π is

$$f^\pi_{(a,i)} := \mathbb{E}^\pi_{(a,i)} \left[\sum_{t=0}^{\infty} \beta^t R^{a(t)}_{(a(t-1),X(t-1))} \right]. \tag{4.12}$$

Then, the alternative objective is

$$\min_{\pi \in \Pi} -f^\pi_{(a,i)} + \nu g^\pi_{(a,i)}, \tag{4.13}$$

where Π is the set of all non-anticipative control policies, and the following proposition demonstrates its equivalence to (4.8).

Proposition 4.1.

(i) For any state $(a,i) \in \widetilde{\mathcal{I}}$ and any policy $\pi \in \Pi$, $f^\pi_{(a,i)} = R^a_i + \beta \overline{f}^\pi_{(a,i)}$.

(ii) For any state $(a,i) \in \widetilde{\mathcal{I}}$ and any policy $\pi \in \Pi$, $g^\pi_{(a,i)} = W^a_i + \beta \overline{g}^\pi_{(a,i)}$.

(iii) Problems (4.8) and (4.13) are equivalent.

Proof. (i) Using the above definitions, we can write

$$\overline{f}^\pi_{(a,i)} = \mathbb{E}^\pi_{(a,i)} \left[\sum_{t=0}^{\infty} \beta^t \, \mathbb{E}\left[R^{a(t)}_{X(t)} | a(t-1), X(t-1) \right] \right] = \mathbb{E}^\pi_{(a,i)} \left[\sum_{t=0}^{\infty} \beta^t R^{a(t)}_{X(t)} \right],$$

where the last equality follows from Fubini's theorem and from the law of total expectation.

On the other hand, we have

$$f^{\pi}_{(a,i)} = \mathbb{E}^{\pi}_{(a,i)}\left[\sum_{t=0}^{\infty} \beta^t R^{a(t-1)}_{X(t-1)}\right],$$

hence we obtain identity $f^{\pi}_{(a,i)} = R^a_i + \beta\overline{f}^{\pi}_{(a,i)}$. □

(ii) Analogously to (i). □

(iii) Since, for all (a,i), $-R^a_i + \nu W^a_i$ is constant, $\beta > 0$, and the identities in (i) and (ii) hold, (4.13) is equivalent to (4.8). □

Notice that the alternative one-period reward $R^b_{(a,i)}$ and the one-period work $W^b_{(a,i)}$ are independent of the current-epoch action (superscript b), therefore we will omit the superscript in the remaining sections.

4.3 Restless Bandit Indexation

In the previous section we have formulated the admission control problem with delay as a binary-action Markov decision process (MDP), i.e., a *restless bandit*, where shutting the entry gate corresponds to the active action, and opening it as the passive action.

We next address such a problem by deploying a restless bandit indexation approach, following the seminal idea introduced in Whittle (1988) and developed by the second author, in work surveyed in Niño-Mora (2007b). We focus on the finite-buffer problem under the discounted criterion. The solution to the problem under the time-average criterion is treated in subsection 4.4.5.

MDP theory ensures existence of an optimal policy that is stationary, deterministic and independent of the initial state. We represent a stationary deterministic policy in terms of an *active set* $\mathcal{S} \subseteq \widetilde{\mathcal{I}}$, i.e., the set of states in which it prescribes to shut the gate; in the remaining states it prescribes to let the gate open. The problem to find an optimal admission control policy is thus reduced to finding an optimal active set,

$$\min_{\mathcal{S}\subseteq\widetilde{\mathcal{I}}} -f^{\mathcal{S}}_{(a,i)} + \nu g^{\mathcal{S}}_{(a,i)}. \tag{4.14}$$

For every rejection cost ν, the optimal policy is characterized by the unique solution

vector $(v^*_{(a,i)}(\nu))_{(a,i)\in\widetilde{\mathcal{I}}}$ to the *Bellman equations*

$$v^*_{(a,i)}(\nu) = \min_{a'\in\mathcal{A}} \left[-R_{(a,i)} + \nu W_{(a,i)} - \beta \sum_{(b,j)\in\widetilde{\mathcal{I}}} p^{a'}_{(a,i),(b,j)} v^*_{(b,j)}(\nu) \right], \quad (a,i) \in \widetilde{\mathcal{I}} \tag{4.15}$$

where $v^*_{(a,i)}(\nu)$ denotes the optimal value of (4.13) starting at (a,i) under rejection cost ν. Hence, there exists a *maximal optimal active set* (i.e., a set of states in which it is optimal to close the gate) $\mathcal{S}^*(\nu) \subseteq \widetilde{\mathcal{I}}$ for (4.13), which is characterized by

$$\mathcal{S}^*(\nu) := \left\{ (a,i) \in \widetilde{\mathcal{I}} : \sum_{(b,j)\in\widetilde{\mathcal{I}}} p^0_{(a,i),(b,j)} v^*_{(b,j)}(\nu) \leq \sum_{(b,j)\in\widetilde{\mathcal{I}}} p^1_{(a,i),(b,j)} v^*_{(b,j)}(\nu) \right\}.$$

Problem (4.14) can be viewed as a bi-criteria parametric optimization problem. Intuitively, if the rejection cost $\nu \to -\infty$, the optimal active set should be $\widetilde{\mathcal{I}}$, whereas if the rejection cost $\nu \to \infty$, the optimal active set should be the empty set. In fact, we set out to show a stronger, so-called *indexability* property: Active sets $\mathcal{S}^*(\nu)$ diminish monotonically from $\widetilde{\mathcal{I}}$ to the empty set as the rejection cost ν increases from $-\infty$ to ∞. Such a property was introduced in Whittle (1988) for the restless bandits with one-periods works equal to 1 under the active action, and equal to 0 under the passive action, and extended to restless bandits without these limitations in Niño-Mora (2002).

Such an indexability property is equivalent to existence of break-even values $\nu^{\text{MPI}}_{(a,i)}$ of the rejection cost ν attached to augmented states $(a,i) \in \widetilde{\mathcal{I}}$, which characterize the optimal policies for (4.14) as follows: it is optimal to take the active action when the system occupies augmented state (a,i) if $\nu^{\text{MPI}}_{(a,i)} \geq \nu$, and it is optimal to take the passive action when the system occupies augmented state (a,i) if $\nu^{\text{MPI}}_{(a,i)} \leq \nu$. Since we have defined $\mathcal{S}^*(\nu)$ as the *maximal* optimal active set, state $(a,i) \in \mathcal{S}^*(\nu)$ if $\nu = \nu_{(a,i)}$, though this choice is arbitrary. We will refer to index $\nu^{\text{MPI}}_{(a,i)}$ as the *marginal productivity index* (MPI), after its economic interpretation as the marginal productivity of work at state (a,i), as elucidated in Niño-Mora (2002, 2006b).

4.3.1 Exploiting Special Structure

While one could test numerically whether a given instance is indexable and calculate the indices $\nu_{(a,i)}$ for all $(a,i) \in \widetilde{\mathcal{I}}$, we aim instead to establish analytically indexability of the admission control problem with delay in general. This will further allow us to achieve our second objective of obtaining a fast way of computing the indices. In this subsection we present how to exploit special structure of the model by aligning indexability to a known family of optimal bi-threshold policies.

Suppose that we postulate a family $\mathcal{F} \subseteq 2^{\widetilde{\mathcal{I}}}$ of active sets, satisfying certain connectivity conditions (see Niño-Mora (2007b) for the details). Before presenting such a family for the admission control problem with delay, we review a test (deployed in Section 4.4) to verify whether a postulated family $\mathcal{F}$ can be used to establish indexability, via the sufficient conditions termed PCL($\mathcal{F}$)-indexability introduced in Niño-Mora (2001, 2002).

Let policy $\langle a, \mathcal{S} \rangle$ be the policy where action a is applied in the current period and policy $\mathcal{S}$ proceeds. Notice that policy $\langle a, \mathcal{S} \rangle$ implies that the next-epoch augmented state will be (a, j) for some state $j \in \mathcal{I}$. We define the *marginal work* of closing the gate instead of letting it open (or, of rejecting possible customers instead of admitting them), if starting from state (a, i) under active-set policy $\mathcal{S}$, as

$$w^{\mathcal{S}}_{(a,i)} := g^{\langle 1,\mathcal{S} \rangle}_{(a,i)} - g^{\langle 0,\mathcal{S} \rangle}_{(a,i)}, \tag{4.16}$$

i.e., as the increment in total work that results from closing the gate instead of opening it at current epoch. Analogously, we define the *marginal reward*,

$$r^{\mathcal{S}}_{(a,i)} := f^{\langle 1,\mathcal{S} \rangle}_{(a,i)} - f^{\langle 0,\mathcal{S} \rangle}_{(a,i)}, \tag{4.17}$$

as the analogous increment in total reward. Finally, we define the *marginal productivity rate*

$$\nu^{\mathcal{S}}_{(a,i)} := \frac{r^{\mathcal{S}}_{(a,i)}}{w^{\mathcal{S}}_{(a,i)}}, \tag{4.18}$$

provided that the denominator does not vanish. As we will see, the denominator is positive for the admission control problem with delay. It can be shown that if the indices exist, then $\nu_{(a,i)} = \nu^{\mathcal{S}}_{(a,i)}$ for some active set $\mathcal{S}$.

In Figure 4.2 is given a scheme of the *adaptive-greedy* algorithm $AG_{\mathcal{F}}$, which calculates the candidates for the maximal optimal active sets $\{\widehat{\mathcal{S}}_k\}_{k=0}^{2I+1}$ and the candidates for the MP indices $\{\widehat{\nu}_{i_k}\}_{k=1}^{2I+1}$. It is greedy, since in each step it picks the state with the lowest marginal productivity rate $\nu^{\widehat{\mathcal{S}}_{k-1}}_{(a_k,i_k)}$ (out of the feasible ones), and it is adaptive, because in each step it updates the marginal productivity rates for the actual active set $\widehat{\mathcal{S}}_{k-1}$.

Now we are ready to define PCL($\mathcal{F}$)-indexability, based on partial conservation laws (PCL), which determines both the computational and analytical value of the adaptive-greedy algorithm $AG_{\mathcal{F}}$.

Definition 4.1 (PCL($\mathcal{F}$)-indexability). The admission control problem with delay is called *PCL($\mathcal{F}$)-indexable*, if

$\widehat{\mathcal{S}}_0 := \widetilde{\mathcal{I}}$;
for $k = 1$ to $2I + 1$ do
  pick $(a_k, i_k) \in \arg\min \left\{ \nu_{(a,i)}^{\widehat{\mathcal{S}}_{k-1}} : (a,i) \in \widehat{\mathcal{S}}_{k-1} \text{ and } \widehat{\mathcal{S}}_{k-1} \setminus \{(a,i)\} \in \mathcal{F} \right\}$;
  $\widehat{\nu}_{(a_k,i_k)} := \nu_{(a_k,i_k)}^{\widehat{\mathcal{S}}_{k-1}}$;
  $\widehat{\mathcal{S}}_k := \widehat{\mathcal{S}}_{k-1} \setminus \{(a_k, i_k)\}$;
end {for};
{Output $\{\widehat{\mathcal{S}}_k\}_{k=0}^{2I+1}, \{\widehat{\nu}_{(a_k,i_k)}\}_{k=1}^{2I+1}$}

Figure 4.2: Algorithmic scheme of $AG_{\mathcal{F}}$.

(i) [Positive Marginal Works under $\mathcal{F}$] for each active set $\mathcal{S} \in \mathcal{F}$ and for each controllable state $(a,i) \in \widetilde{\mathcal{I}}$, the marginal work $w_{(a,i)}^{\mathcal{S}} > 0$;

and either of the following conditions holds:

(ii) for every rejection cost ν, there exists an optimal active set $\mathcal{S} \in \mathcal{F}$;

(ii') the output $\{\widehat{\nu}_{(a_k,i_k)}\}_{k=1}^{2I+1}$ of the algorithm $AG_{\mathcal{F}}$ are marginal productivity indices in nondecreasing order.

Niño-Mora (2001, 2002, 2007b) introduced variants of PCL($\mathcal{F}$)-indexability and proved that PCL($\mathcal{F}$)-indexability implies indexability, i.e., the existence of MP indices that are calculated as $\{\widehat{\nu}_{(a_k,i_k)}\}_{k=1}^{2I+1}$ by the adaptive-greedy algorithm $AG_{\mathcal{F}}$. To ease later reference, we summarize the above in the following theorem.

Theorem 4.1. *If marginal works are positive under $\mathcal{F}$ (cf. Definition 4.1(i)) for problem (4.13), then for that problem the following statements are equivalent:*

(i) for every rejection cost ν, there exists a maximal optimal active set $\mathcal{S} \in \mathcal{F}$;

(ii) the problem is indexable and all active sets $\mathcal{S}^*(\nu) \in \mathcal{F}$;

(iii) the output $\{\widehat{\nu}_{(a_k,i_k)}\}_{k=1}^{2I+1}$ of the algorithm $AG_{\mathcal{F}}$ are marginal productivity indices in nondecreasing order.

In Section 4.4 we show that for a certain family $\mathcal{F}$ (defined below), Definition 4.1(i) holds and, given the existing results, Theorem 4.1(i) is true. In this way indexability of the admission control problem with delay will be established, and the algorithm $AG_{\mathcal{F}}$ can be used to obtain the indices.

Definition 4.1(i) has an intuitive interpretation (cf. Niño-Mora, 2002, Proposition 6.2): positivity of marginal work $w^{\mathcal{S}}_{(a,i)}$ (where $\mathcal{S} \in \mathcal{F}$ and state $(a,i) \in \widetilde{\mathcal{I}}$ is controllable) is equivalent to monotonicity of total work,

$$\begin{aligned} g^{\mathcal{S}\setminus\{(a,i)\}}_{(a,i)} &< g^{\mathcal{S}}_{(a,i)}, && \text{if } (a,i) \in \mathcal{S}, \\ g^{\mathcal{S}}_{(a,i)} &< g^{\mathcal{S}\cup\{(a,i)\}}_{(a,i)}, && \text{if } (a,i) \notin \mathcal{S}. \end{aligned}$$

Informally stated, rejecting in a larger number of states corresponds to a larger expected total discounted number of rejected customers. Definition 4.1(i) is a natural assumption in many models, though, in general, it is neither a sufficient nor a necessary condition for indexability.

4.3.2 Postulated Active-Set Family

We use the results of Altman and Nain (1992), who characterized the optimal bi-threshold policies, and identify an active-set family $\mathcal{F}$ for which Theorem 4.1(i) holds. A bi-threshold active-set policy with open-gate threshold K_0 and closed-gate threshold K_1 will be denoted by

$$\widetilde{\mathcal{I}}_{K_0,K_1} := \{(0,K_0),(0,K_0+1),\ldots,(0,I)\} \cup \{(1,K_1),(1,K_1+1),\ldots,(1,I)\}, \tag{4.19}$$

which is well-defined for all $0 \leq K_0, K_1 \leq I+1$ except the active sets $\widetilde{\mathcal{I}}_{I+1,I}$ and $\widetilde{\mathcal{I}}_{I,I+1}$, because states $(1,I)$ and $(0,I)$ are duplicates, and by definition either both or none of them can belong to $\widetilde{\mathcal{I}}_{K_0,K_1}$.

In words, active set $\widetilde{\mathcal{I}}_{K_0,K_1}$ prescribes to open or close the gate depending on the previous-epoch action and previous-epoch state. If the gate was open in the previous period, then we open the gate if and only if the queue length in the previous epoch was equal to or larger than the open-gate threshold K_0. Similarly, if the gate was closed in the previous period, then we open the gate if and only if the queue length in the previous epoch was equal to or larger than the closed-gate threshold K_1.

Intuitively, if an active set $\widetilde{\mathcal{I}}_{K_0,K_1}$ is optimal for some rejection cost ν, then $K_0 \leq K_1$. Indeed, for a given previous-epoch queue length, we would be less prone to close the gate if it was closed than if it was open in the preceding period, because the queue length could not get larger under a closed gate, and therefore the rejection costs become relatively more harmful than the holding costs. On the other hand, it can be shown that $K_1 \leq K_0 + 1$ (see below). Thus, the postulated family of optimal active sets for the

admission control problem with delay is

$$\mathcal{F} := \{\widetilde{\mathcal{I}}_{K,K} : K = 0, 1, \ldots, I+1\} \cup \{\widetilde{\mathcal{I}}_{K,K+1} : K = 0, 1, \ldots, I-1\}. \tag{4.20}$$

Theorem 4.2 (Altman and Nain (1992), Theorem 3.1). *If the holding cost c_i is nondecreasing and convex on $\mathcal{I}$, then $\mathcal{F}$ as defined in (4.20) contains an optimal active set for every rejection cost ν.*

Though the above result was shown for the problem with infinite buffer, it directly applies to the finite-buffer variant. Notice that if a bi-threshold policy is optimal for the infinite-buffer problem, then it is also optimal for all problems with buffer equal to or larger than both the thresholds. If the buffer is smaller than the larger optimal threshold (K_1), then it is optimal to open the gate all the time.

For active-set family $\mathcal{F}$ given in (4.20), picking (a_k, i_k) becomes trivial, because there is only a unique feasible augmented state in each step. For instance, in step $k = 1$, only state $(1, 0)$ belongs both to $\widehat{\mathcal{S}}_0$ and $\widehat{\mathcal{S}}_0 \setminus \{(1, 0)\} = \widetilde{\mathcal{I}}_{0,1} \in \mathcal{F}$, since $\widehat{\mathcal{S}}_0 := \widetilde{\mathcal{I}} = \widetilde{\mathcal{I}}_{0,0}$. Similarly, in step $k = 2$, only state $(0, 0)$ belongs both to $\widehat{\mathcal{S}}_1$ and $\widehat{\mathcal{S}}_1 \setminus \{(0, 0)\} = \widetilde{\mathcal{I}}_{1,1} \in \mathcal{F}$. In general, $(a_k, i_k) = (0, (k/2) - 1)$ for all even $1 \le k \le 2I$, and $(a_k, i_k) = (1, (k-1)/2)$ for all odd $1 \le k \le 2I$. Finally, in step $k = 2I + 1$, the picked state is $(*, I)$.

To summarize, the sequence of candidate active sets $\{\widehat{\mathcal{S}}_k\}_{k=0}^{2I+1}$ in algorithm $AG_{\mathcal{F}}$ under active-set family $\mathcal{F}$ given in (4.20) is

$$\begin{aligned} \widehat{\mathcal{S}}_0 = \widetilde{\mathcal{I}} = \widetilde{\mathcal{I}}_{0,0}, \widehat{\mathcal{S}}_1 = \widetilde{\mathcal{I}}_{0,1}, \widehat{\mathcal{S}}_2 = \widetilde{\mathcal{I}}_{1,1}, \widehat{\mathcal{S}}_3 = \widetilde{\mathcal{I}}_{1,2}, \widehat{\mathcal{S}}_4 = \widetilde{\mathcal{I}}_{2,2}, \ldots \\ \ldots, \widehat{\mathcal{S}}_{2I-1} = \widetilde{\mathcal{I}}_{I-1,I}, \widehat{\mathcal{S}}_{2I} = \widetilde{\mathcal{I}}_{I,I}, \widehat{\mathcal{S}}_{2I+1} = \widetilde{\mathcal{I}}_{I+1,I+1} = \emptyset, \end{aligned} \tag{4.21}$$

and the sequence of picked states $\{(a_k, i_k)\}_{k=1}^{2I+1}$ is

$$\begin{aligned} (a_1, i_1) = (1, 0), (a_2, i_2) = (0, 0), (a_3, i_3) = (1, 1), (a_4, i_4) = (0, 1), \ldots \\ \ldots, (a_{2I-1}, i_{2I-1}) = (1, I-1), (a_{2I}, i_{2I}) = (0, I-1), (a_{2I+1}, i_{2I+1}) = (*, I). \end{aligned}$$

Given the above, Figure 4.3 presents the reduction of the algorithmic scheme $AG_{\mathcal{F}}$ as it applies to the postulated family $\mathcal{F}$ given in (4.20). Notice that the computational complexity remains at the same level since the main difficulty lies in the calculation of $\nu_{(a_k,i_k)}^{\widehat{\mathcal{S}}_{k-1}}$, for which no computational details are given. Therefore we also call them algorithmic schemes, not algorithms. The goal of this chapter is to establish the validity of $AG_{\mathcal{F}}$ for our problem and to develop its implementation of low computational complexity.

```
for K = 1 to I do
   ν̂_(1,K−1) := ν^{Ĩ_{K−1,K−1}}_(1,K−1);
   ν̂_(0,K−1) := ν^{Ĩ_{K−1,K}}_(0,K−1);
end {for};
ν̂_(∗,I) := ν^{Ĩ_{I,I}}_(∗,I);
{Output {ν̂_(a,i)}_{(a,i)∈Ĩ}}
```

Figure 4.3: Algorithmic scheme of $AG_{\mathcal{F}}$ under active-set family $\mathcal{F}$ given in (4.20).

4.4 Results

In this section we focus on the admission control problem with delay to a buffer (i.e., $I \geq 2$) under the discounted criterion. The results under the time-average criterion are summarized in subsection 4.4.5.

Our main results are twofold. First, we prove the positivity of marginal works (cf. Definition 4.1(i)) for $\mathcal{F}$ given in (4.20), so that the algorithm $AG_{\mathcal{F}}$ can be applied to compute the indices. Second, we simplify $AG_{\mathcal{F}}$ obtaining a procedure that performs only a linear number of arithmetic operations to compute all the indices and the optimal thresholds.

Let us introduce a more compact notation. For any augmented-state-dependent variable $x_{(a,i)}$, we will use the backward difference operator in the first dimension, i.e., the *action-difference operator*,

$$\Delta_1 x_{(1,i)} := x_{(1,i)} - x_{(0,i)} \tag{4.22}$$

and in the second dimension, i.e., the *state-difference operator*,

$$\Delta_2 x_{(a,i)} := x_{(a,i)} - x_{(a,i-1)} \tag{4.23}$$

whenever the right-hand side expressions are defined. For definiteness, we further let $\Delta_2 x_{(a,0)} := 0$ for $a \in \mathcal{A}$. Directly from these definitions we obtain the following auxiliary identity,

$$\Delta_2 x_{(1,i)} - \Delta_2 x_{(0,i)} = \Delta_1 x_{(1,i)} - \Delta_1 x_{(1,i-1)}. \tag{4.24}$$

In the following we list our main results, drawing on the technical analysis of work measures presented in the appendix (Section B.1).

Proposition 4.2.

(i) The marginal works in problem (4.14) are positive under the active-set family $\mathcal{F}$ given in (4.20), i.e., Definition 4.1(i) holds.

(ii) If the holding cost c_i is nondecreasing and convex on $\mathcal{I}$, then the admission control problem with delay in (4.14) is PCL($\mathcal{F}$)-indexable, and therefore it is indexable and algorithm $AG_{\mathcal{F}}$ calculates the marginal productivity indices for this problem.

Proof.

(i) By B.2(iii) and B.3(iv), $\Delta_1 g^{\mathcal{S}}_{(1,i)} > 0$ for all $0 \le i \le I-1$ and $\Delta_1 g^{\mathcal{S}}_{(1,I)} = 0$ under every active set $\mathcal{S} \in \mathcal{F}$. Then, B.4 establishes the positivity of marginal works for all states. □

(ii) Due to (i), Theorem 4.1(i)-(iii) are equivalent. The validity of Theorem 4.1(i) was established in Theorem 4.2, therefore Theorem 4.1(iii) holds, and implies the above claim. □

4.4.1 A Fast Algorithm for Calculation of All Marginal Productivity Indices

Suppose that the holding cost c_i is nondecreasing and convex on $\mathcal{I}$. In the following we develop an algorithm for calculation of *all* MP indices in $\mathcal{O}(I)$, which is two orders of magnitude faster that the best general implementation of algorithm $AG_{\mathcal{F}}$ performing $\mathcal{O}(I^3)$ arithmetic operations.

The algorithmic scheme $AG_{\mathcal{F}}$ in Figure 4.3 is exhibited in its *bottom-up* version, as it calculates the MP indices in nondecreasing order (cf. Definition 4.1(ii')). This is closely related to our definition of indexability in Section 4.3 as the property that "active sets $\mathcal{S}^*(\nu)$ diminish monotonically from $\widetilde{\mathcal{I}}$ to the empty set as the rejection cost ν increases from $-\infty$ to ∞," being emulated by the bottom-up version of the algorithm. Notice that we could equivalently define indexability as "active sets $\mathcal{S}^*(\nu)$ expand monotonically from the empty set to $\widetilde{\mathcal{I}}$ as the rejection cost ν decreases from ∞ to $-\infty$." This intuitively leads to consideration of algorithm $AG_{\mathcal{F}}$ in its equivalent, *top-down* version, starting with the empty set and calculating the indices in nonincreasing order.

In other words, while the bottom-up version of algorithm $AG_{\mathcal{F}}$ traverses the active-set family $\mathcal{F}$ in the order (cf. (4.21))

$$\widetilde{\mathcal{I}}_{0,0}, \widetilde{\mathcal{I}}_{0,1}, \widetilde{\mathcal{I}}_{1,1}, \widetilde{\mathcal{I}}_{1,2}, \ldots, \widetilde{\mathcal{I}}_{I-1,I}, \widetilde{\mathcal{I}}_{I,I}, \widetilde{\mathcal{I}}_{I+1,I+1},$$

the top-down version does that in the reverse order

$$\widetilde{\mathcal{I}}_{I+1,I+1}, \widetilde{\mathcal{I}}_{I,I}, \widetilde{\mathcal{I}}_{I-1,I}, \ldots, \widetilde{\mathcal{I}}_{1,2}, \widetilde{\mathcal{I}}_{1,1}, \widetilde{\mathcal{I}}_{0,1}, \widetilde{\mathcal{I}}_{0,0}.$$

$\nu_{(1,0)} := \nu_{(1,0)}^{\widetilde{\mathcal{I}}_{0,0}}$;
for $K = 1$ to $I - 1$ do
  $\nu_{(0,K-1)} := \nu_{(0,K-1)}^{\widetilde{\mathcal{I}}_{K,K}}$;
  $\nu_{(1,K)} := \nu_{(1,K)}^{\widetilde{\mathcal{I}}_{K,K}}$;
end {for};
$\nu_{(0,I-1)} := \nu_{(0,I-1)}^{\widetilde{\mathcal{I}}_{I,I}}$;
$\nu_{(*,I)} := \nu_{(*,I)}^{\widetilde{\mathcal{I}}_{I+1,I+1}}$;
{Output $\left\{\nu_{(a,i)}\right\}_{(a,i)\in\widetilde{\mathcal{I}}}$}

Figure 4.4: Algorithmic scheme for the calculation of MP indices of the admission control problem with delay in terms of active sets $\widetilde{\mathcal{I}}_{K,K}$ only.

For instance, index $\nu_{(1,0)}$ is calculated as the marginal productivity rate $\nu_{(1,0)}^{\widetilde{\mathcal{I}}_{0,0}}$ in the bottom-up version, while the same index is calculated as the marginal productivity rate $\nu_{(1,0)}^{\widetilde{\mathcal{I}}_{0,1}}$ in the top-down version. In fact, Niño-Mora (2002, Theorem 6.4(b)) implies that $\nu_{(a_k,i_k)}^{\widehat{\mathcal{S}}_{k-1}} = \nu_{(a_k,i_k)}^{\widehat{\mathcal{S}}_k}$, using the notation of Figure 4.2. Thus, since the active set of type $\widetilde{\mathcal{I}}_{K,K}$ is efficient every two steps of the algorithm (except for the last step, where $\widetilde{\mathcal{I}}_{I+1,I+1}$ follows $\widetilde{\mathcal{I}}_{I,I}$), we can formulate the indices in terms of marginal productivity rates under active sets $\widetilde{\mathcal{I}}_{K,K}$ only. Such an algorithmic scheme is presented in Figure 4.4.

Next we develop an efficient implementation of the algorithmic scheme $AG_{\mathcal{F}}$, which we present in Figure 4.5. The algorithm FA is two orders of magnitude faster than the best existing general implementation of the algorithm $AG_{\mathcal{F}}$. We characterize the MP indices calculated as indicated in Figure 4.4 in terms of closed-form expressions of pivot state-differences given in B.13 and B.7. Notice that one iteration of the algorithm solves the problem (4.14) for the entire range of the real-valued rejection cost parameter ν.

Proposition 4.3.

(i) The algorithm FA in Figure 4.5 computes the marginal productivity indices for problem (4.14) under the discounted criterion.

(ii) The algorithm FA in Figure 4.5 performs $\mathcal{O}(I)$ arithmetic operations.

Proof. (i) The algorithm FA is an implementation of expressions of marginal productivity rates $\nu_{(a,i)}^{\widetilde{\mathcal{I}}_{K,K}}$ developed below into the algorithmic scheme given in Figure 4.4. The marginal productivity rates are, by definition (4.18), computed as the ratio of marginal rewards to marginal works. These quantities are given in B.11 and

{Input I, λ, μ, β}
{Initialization}
$\zeta := \lambda(1-\mu)$; $\eta := \mu(1-\lambda)$; $\varepsilon := 1-\zeta-\eta$;
$A_0 := 0$; $A'_0 := \beta\zeta$; $B := \beta\mu/(1-\beta+\beta\mu)$; $B' := \beta\zeta B + \beta(\mu-\eta)$;
$C''_{0,0} := \Delta_1 R_{(1,0)} - \zeta B \Delta_2 R_{(1,1)}/\mu$; $D_0 := 0$; $D'_0 := \Delta_1 R_{(1,0)}$;
$\nu_{(1,0)} := C''_{0,0}/\lambda$;
{Loop}
for $K = 1$ to $I-1$ do
$A_K := \beta\zeta/[1-\beta+\beta\zeta+\beta\eta(1-A_{K-1})]$; $A'_K := \beta\zeta + \beta(\mu-\eta)A_K$;
$Z_K := A_K A'_{K-1}/A'_K$;
$C''_{K,K} := \Delta_1 R_{(1,K)} - \zeta B \Delta_2 R_{(1,K+1)}/\mu$;
$C''_{K,K+1} := \Delta_1 R_{(1,K+1)} - \zeta B^2 \Delta_2 R_{(1,K+1)}/\mu - \zeta B \Delta_2 R_{(1,K+2)}/\mu$;
$D_K := \left(\beta\eta D_{K-1} - \Delta_2 R_{(0,K)}\right) A_K/(\beta\zeta)$; $D'_K := \Delta_1 R_{(1,K)} + \beta(\mu-\eta)D_K$;

$$f^0 := -\frac{\frac{\beta\zeta}{A_K} D_K - \Delta_1 R_{(1,K)} + (1-\beta)C''_{K,K+1} - (\beta D'_{K-1} + \Delta_2 R_{(1,K)})B' + \beta\mu(\beta\zeta D_{K-1} B + C''_{K,K+1} - \Delta_1 R_{(1,K-1)})}{\frac{A'_K}{A_K} + \beta A'_{K-1} B' + \beta\zeta\beta\mu(1 - BA_{K-1})};$$

$$f^1 := -\frac{\frac{\beta\zeta}{A_K} D_K - \Delta_1 R_{(1,K)} + (1-\beta)D'_{K-1} - (\beta C''_{K,K+1} + \Delta_2 R_{(1,K)})A'_{K-1} + \beta\mu(\beta\zeta D_{K-1} + (C''_{K,K+1} - \Delta_1 R_{(1,K-1)})A_{K-1})}{\frac{A'_K}{A_K} + \beta A'_{K-1} B' + \beta\zeta\beta\mu(1 - BA_{K-1})};$$

$$g^0 := \frac{\beta\lambda\,(1+B')}{\frac{A'_K}{A_K} + \beta A'_{K-1} B' + \beta\zeta\beta\mu(1 - BA_{K-1})}; \quad g^1 := \frac{1 + A'_{K-1}}{1+B'} g^0;$$

if $K = 1$ then
$$\nu_{(0,0)} := \frac{(1-\zeta)D'_0 + \zeta C''_{1,1} - (1-\zeta)A'_0 f^0 - \zeta B' f^1}{\lambda - (1-\zeta)A'_0 g^0 - \zeta B' g^1};$$
else
$$\nu_{(0,K-1)} := \frac{\eta\left[D'_{K-2} + D_{K-1}A'_{K-2}\right] + \varepsilon D'_{K-1} + \zeta C''_{K,K} - \left[\eta Z_{K-1} + \varepsilon\right] A'_{K-1} f^0 - \zeta B' f^1}{\lambda - \left[\eta Z_{K-1} + \varepsilon\right] A'_{K-1} g^0 - \zeta B' g^1};$$
end {if};
$$\nu_{(1,K)} := \frac{\mu D'_{K-1} + (1-\mu)C''_{K,K} - \mu A'_{K-1} f^0 - (1-\mu)B' f^1}{\lambda - \mu A'_{K-1} g^0 - (1-\mu)B' g^1};$$
end {for};
{Finalization}
$A_I := \beta\zeta/[1-\beta+\beta\zeta+\beta\eta(1-A_{I-1})]$; $A'_I := \beta\zeta + \beta(\mu-\eta)A_I$; $Z_I := A_I A'_{I-1}/A'_I$;
$D_I := \left(\beta\eta D_{I-1} - \Delta_2 R_{(0,I)}\right) A_I/(\beta\zeta)$;

$$f^0 := -\frac{\frac{\beta\zeta}{A_I} D_I - \beta\mu D'_{I-1}}{\frac{A'_I}{A_I} + \beta\mu A'_{I-1}}; \quad g^0 := \frac{\lambda(1+\beta\mu)}{\frac{A'_I}{A_I} + \beta\mu A'_{I-1}};$$

$$\nu_{(0,I-1)} := \frac{\eta\left[D'_{I-2} + D_{I-1}A'_{I-2}\right] + \varepsilon D'_{I-1} - \left[\eta Z_{I-1} + \varepsilon\right] A'_{I-1} f^0}{(\eta+\varepsilon)\lambda - \left[\eta Z_{I-1} + \varepsilon\right] A'_{I-1} g^0};$$

$$\nu_{(*,I)} := \frac{D'_{I-1} + \frac{\beta\zeta}{A_I} D_I Z_I}{\lambda(1 - Z_I)};$$

{Output $\left\{\nu_{(a,i)}\right\}_{(a,i)\in\widetilde{\mathcal{I}}}$}

Figure 4.5: Fast algorithm FA for the calculation of MP indices under general rewards.

B.4, respectively, in terms of their action-differences, which can be expressed in terms of pivot state-differences due to B.4 and B.1. The pivot state-differences $\Delta_2 f^{\widetilde{\mathcal{I}}_{K,K}}_{(0,K)}, \Delta_2 f^{\widetilde{\mathcal{I}}_{K,K}}_{(1,K)}, \Delta_2 g^{\widetilde{\mathcal{I}}_{K,K}}_{(0,K)}, \Delta_2 g^{\widetilde{\mathcal{I}}_{K,K}}_{(1,K)}$ given in B.13 and B.7 are in Figure 4.5 briefly denoted by f^0, f^1, g^0, g^1, respectively. □

(ii) The number of arithmetic operations in "Initialization" and in "Finalization" is constant (with respect to I). Similarly, at each step of "Loop", a constant number of arithmetic operations is performed. Since there are $\mathcal{O}(I)$ steps of "Loop", the overal complexity of the algorithm is $\mathcal{O}(I)$. □

Once the optimal index policy is known, the optimal thresholds for a given rejection cost ν can easily be obtained. The optimal open-gate threshold is

$$K_0 := \min\{i \in \mathcal{I} : \nu_{(0,i)} \geq \nu\}.$$

Similarly, the optimal closed-gate threshold is

$$K_1 := \min\{i \in \mathcal{I} : \nu_{(1,i)} \geq \nu\}.$$

If $\nu > \nu_{(*,I)}$, then $K_0 := I + 1$ and $K_1 := I + 1$.

4.4.2 A Fast Algorithm for Calculation of One Marginal Productivity Index

In this subsection we develop an algorithm for calculation of *one* MP index, say of state (a, K) in isolation. We show that it performs at most $\mathcal{O}(\log_2 K)$ arithmetic operations, given that it may require to calculate an integer power, for which the best known algorithm (*exponentiation by squaring*) needs $\mathcal{O}(\log_2 K)$ operations.

The idea is to perform step K of the "Loop" of algorithm FA in Figure 4.5, which performs a constant number of arithmetic operations, having calculated A_K and D_K using their respective closed-form formulae. If $\eta = 0$, then A_K is constant and requires a constant number of arithmetic operations to calculate. In the following we assume $\eta > 0$ and denote by

$$s := \frac{\zeta}{\eta}, \quad t := \frac{1 - \beta + \beta\zeta + \beta\eta}{\beta\eta}, \quad u := \sqrt{t^2 - 4s}, \quad u_+ := \frac{t + u}{2}, \quad u_- := \frac{t - u}{2}. \tag{4.25}$$

Note that the above quantities are well defined given the model parameters assumptions. The following lemma[2] characterizes A_K and D_K in terms of K-th powers of u_+ and u_-.

[2] I am grateful to Sofía Villar for bringing to my attention the Möbius transformation, crucial for obtaining closed-form solutions of the recurrences in this lemma.

Lemma 4.1.

(i) For any $K \geq 1$,

$$A_K = s\frac{u_+^K - u_-^K}{u_+^{K+1} - u_-^{K+1}}.$$

(ii) If $\Delta_2 R_{(1,i)} = R$ for all $i \geq 1$, then for any $K \geq 1$,

$$D_K = -\frac{R}{1-\beta}\left[\frac{u + (u_+^K - u_-^K)u_- u_+}{u_+^{K+1} - u_-^{K+1}} - 1\right].$$

(iii) Sequences A_K and (if $\Delta_2 R_{(1,i)} = R$ for all $i \geq 1$) D_K converge as $K \to \infty$ to their respective limits

$$A = \frac{s}{u_+}, \qquad D = \frac{c}{1-\beta}(u_- - 1).$$

Proof. (i) We start by developing the formula for A_K. By definition of A_K in (B.14) and those of s and t in (4.25) we can write $A_K = \dfrac{s}{-A_{K-1} + t}$. Notice that this is a well-defined Möbius transformation $m(x) := \dfrac{0 \cdot x + s}{-1 \cdot x + t}$ represented in the matrix form as

$$\boldsymbol{M} := \begin{pmatrix} 0 & s \\ -1 & t \end{pmatrix}$$

Therefore, A_K expressed in terms of A_0 is given by the K-th functional power of $m(x)$ for $x = A_0$, i.e., $A_K = m^K(A_0) = \dfrac{a_K \cdot A_0 + b_K}{c_K \cdot A_0 + d_K}$. Since $A_0 = 0$ by definition, we obtain $A_K = b_K/d_K$. By properties of Möbius transformation, we have

$$\begin{pmatrix} a_K & b_K \\ c_K & d_K \end{pmatrix} = \boldsymbol{M}^K.$$

Since

$$\boldsymbol{M}^K = \begin{pmatrix} 0 & s \\ -1 & t \end{pmatrix}\begin{pmatrix} a_{K-1} & b_{K-1} \\ c_{K-1} & d_{K-1} \end{pmatrix} = \begin{pmatrix} s \cdot c_{K-1} & s \cdot d_{K-1} \\ -a_{K-1} + t \cdot c_{K-1} & -b_{K-1} + t \cdot d_{K-1} \end{pmatrix}$$

we have $b_K = s \cdot d_{K-1}$ and hence $A_K = sd_{K-1}/d_K$.

Next, we set out to obtain a closed-form solution for sequence d_K. Using the above identity for b_{K-1}, we have $d_K = -s \cdot d_{K-2} + t \cdot d_{K-1}$. This recurrence is

one of the two Lucas sequences (cf. Lucas, 1878; Kalman and Mena, 2003) with the initial values $d_0 = 1$ and $d_1 = t$ obtained from the definition of matrix M and the relationship $b_1 = s \cdot d_0$. Its closed-form solution is

$$d_K = \frac{u_+^{K+1} - u_-^{K+1}}{u}$$

and therefore

$$A_K = s\frac{u_+^K - u_-^K}{u_+^{K+1} - u_-^{K+1}}.$$

□

(ii) Now we develop the formula for D_K. By recursive implementation of the definition of D_K in (B.53) and using definition of s in (4.25) we have

$$D_K = -\frac{1}{\beta\eta}\left[\frac{A_K}{s}\Delta_2 R_{(1,K)} + \frac{A_K}{s}\frac{A_{K-1}}{s}\Delta_2 R_{(1,K-1)} + \dots + \frac{A_K}{s}\frac{A_{K-1}}{s}\dots\frac{A_1}{s}\Delta_2 R_{(1,1)}\right].$$

Since $A_K = sd_{K-1}/d_K$, this simplifies to

$$\begin{aligned} D_K &= -\frac{1}{\beta\eta}\left[\frac{d_{K-1}}{d_K}\Delta_2 R_{(1,K)} + \frac{d_{K-2}}{d_K}\Delta_2 R_{(1,K-1)} + \dots + \frac{d_0}{d_K}\Delta_2 R_{(1,1)}\right] \\ &= -\frac{1}{\beta\eta d_K}\sum_{k=0}^{K-1} d_k \Delta_2 R_{(1,k+1)}, \end{aligned}$$

which, under constant $\Delta_2 R_{(1,i)} = R$ for all $i \geq 1$, is

$$D_K = -\frac{R}{\beta\eta d_K}\sum_{k=0}^{K-1} d_k.$$

Plugging the expression for d_k and simplifying the constant terms gives

$$\sum_{k=0}^{K-1} d_k = \frac{1}{1-t+s}\left[1 - \frac{u_+^{K+1}(1-u_-) - u_-^{K+1}(1-u_+)}{u}\right].$$

The last two identities together with $(1-t+s)\beta\eta = 1-\beta$ imply

$$\begin{aligned} D_K &= -\frac{R}{1-\beta}\frac{u - u_+^{K+1}(1-u_-) + u_-^{K+1}(1-u_+)}{u_+^{K+1} - u_-^{K+1}} \\ &= -\frac{R}{1-\beta}\left[\frac{u + (u_+^K - u_-^K)u_-u_+}{u_+^{K+1} - u_-^{K+1}} - 1\right]. \end{aligned}$$

□

(iii) The Lucas sequence d_K defined in part (i) satisfies $d_{K+1}/d_K = u_+ > 1$ as $K \to \infty$. Therefore, we obtain the limits

$$A = \frac{s}{u_+}, \qquad\qquad D = \frac{c}{1-\beta}(u_- - 1).$$

□

Finally we note that calculation of all MP indices using this method would require $\mathcal{O}(\log_2(I!))$ arithmetic operations, which is more than the linear number performed by algorithm FA in Figure 4.5.

4.4.3 Fast Algorithm under Convex Non-Decreasing Holding Costs in Admission Control Problem with Delay

Under convex non-decreasing holding costs, the immediate reward is $R_{(a,i)} = R_i^a : -c_i$ under any $a \in \mathcal{A}, i \in \mathcal{I}$. Therefore, we have

$$\begin{aligned} \Delta_1 R_{(1,i)} &= 0, && i \geq 0, \\ \Delta_2 R_{(0,i)} = \Delta_2 R_{(1,i)} &= -c_i + c_{i-1} =: -\Delta c_i, && i \geq 1, \end{aligned}$$

and the fast algorithm simplifies to the one shown in Figure 4.6. This includes the special case of linear holding costs, when $\Delta c_i := c$ for all $i \in \mathcal{I}$.

The algorithm can also be used to derive the *myopic index*, which only looks one period ahead. Such an index is defined as $\nu_{(a,i)}^{\text{MYOPIC}} := \lim_{\beta \to 0} \frac{\nu_{(a,i)}}{\beta}$ for all $(a,i) \in \widetilde{\mathcal{I}}$. In the

{Input $I, \lambda, \mu, \beta, \{c_i\}_{i\in\mathcal{I}}$}
{Initialization}
$\zeta := \lambda(1-\mu)$; $\eta := \mu(1-\lambda)$; $\varepsilon := 1-\zeta-\eta$;
$A_0 := 0$; $A'_0 := \beta\zeta$; $B := \beta\mu/(1-\beta+\beta\mu)$; $B' := \beta\zeta B + \beta(\mu-\eta)$;
$C''_{0,0} := \zeta B\Delta c_1/\mu$; $D_0 := 0$; $D'_0 := 0$;
$\nu_{(1,0)} := C''_{0,0}/\lambda$;
{Loop}
for $K = 1$ to $I-1$ do
$A_K := \beta\zeta/[1-\beta+\beta\zeta+\beta\eta(1-A_{K-1})]$; $A'_K := \beta\zeta+\beta(\mu-\eta)A_K$;
$Z_K := A_K A'_{K-1}/A'_K$;
$C''_{K,K} := \zeta B\Delta c_{K+1}/\mu$; $C''_{K,K+1} := \zeta B^2\Delta c_{K+1}/\mu + \zeta B\Delta c_{K+2}/\mu$;
$D_K := (\beta\eta D_{K-1} + \Delta c_K)\, A_K/(\beta\zeta)$; $D'_K := \beta(\mu-\eta)D_K$;
$$f^0 := -\frac{\frac{\beta\zeta}{A_K}D_K + (1-\beta)C''_{K,K+1} - (\beta D'_{K-1} - \Delta c_K)B' + \beta\mu(\beta\zeta D_{K-1}B + C''_{K,K+1})}{\frac{A'_K}{A_K} + \beta A'_{K-1}B' + \beta\zeta\beta\mu(1-BA_{K-1})};$$
$$f^1 := -\frac{\frac{\beta\zeta}{A_K}D_K + (1-\beta)D'_{K-1} - (\beta C''_{K,K+1} - \Delta c_K)A'_{K-1} + \beta\mu(\beta\zeta D_{K-1} + C''_{K,K+1}A_{K-1})}{\frac{A'_K}{A_K} + \beta A'_{K-1}B' + \beta\zeta\beta\mu(1-BA_{K-1})};$$
$$g^0 := \frac{\beta\lambda\,(1+B')}{\frac{A'_K}{A_K} + \beta A'_{K-1}B' + \beta\zeta\beta\mu(1-BA_{K-1})}; \quad g^1 := \frac{1+A'_{K-1}}{1+B'}g^0;$$
if $K = 1$ then
$$\nu_{(0,0)} := \frac{(1-\zeta)D'_0 + \zeta C''_{1,1} - (1-\zeta)A'_0 f^0 - \zeta B' f^1}{\lambda - (1-\zeta)A'_0 g^0 - \zeta B' g^1};$$
else
$$\nu_{(0,K-1)} := \frac{\eta\left[D'_{K-2} + D_{K-1}A'_{K-2}\right] + \varepsilon D'_{K-1} + \zeta C''_{K,K} - [\eta Z_{K-1} + \varepsilon]\, A'_{K-1}f^0 - \zeta B' f^1}{\lambda - [\eta Z_{K-1} + \varepsilon]\, A'_{K-1}g^0 - \zeta B' g^1};$$
end {if};
$$\nu_{(1,K)} := \frac{\mu D'_{K-1} + (1-\mu)C''_{K,K} - \mu A'_{K-1}f^0 - (1-\mu)B' f^1}{\lambda - \mu A'_{K-1}g^0 - (1-\mu)B' g^1};$$
end {for};
{Finalization}
$A_I := \beta\zeta/[1-\beta+\beta\zeta+\beta\eta(1-A_{I-1})]$; $A'_I := \beta\zeta+\beta(\mu-\eta)A_I$; $Z_I := A_I A'_{I-1}/A'_I$;
$D_I := (\beta\eta D_{I-1} + \Delta c_I)\, A_I/(\beta\zeta)$;
$$f^0 := -\frac{\frac{\beta\zeta}{A_I}D_I - \beta\mu D'_{I-1}}{\frac{A'_I}{A_I} + \beta\mu A'_{I-1}}; \quad g^0 := \frac{\lambda(1+\beta\mu)}{\frac{A'_I}{A_I} + \beta\mu A'_{I-1}};$$
$$\nu_{(0,I-1)} := \frac{\eta\left[D'_{I-2} + D_{I-1}A'_{I-2}\right] + \varepsilon D'_{I-1} - [\eta Z_{I-1} + \varepsilon]\, A'_{I-1}f^0}{(\eta+\varepsilon)\lambda - [\eta Z_{I-1} + \varepsilon]\, A'_{I-1}g^0};$$
$$\nu_{(*,I)} := \frac{D'_{I-1} + \frac{\beta\zeta}{A_I}D_I Z_I}{\lambda(1-Z_I)};$$
{Output $\left\{\nu_{(a,i)}\right\}_{(a,i)\in\widetilde{\mathcal{I}}}$}

Figure 4.6: Fast algorithm FA for the calculation of MP indices under convex non-decreasing holding costs.

case of linear holding costs, it is straightforward to obtain the myopic index as follows.

$$\begin{aligned}
\nu^{\text{MYOPIC}}_{(1,0)} &= c(1-\mu);\\
\nu^{\text{MYOPIC}}_{(0,0)} &= c(1-\eta-\mu^2\lambda);\\
\nu^{\text{MYOPIC}}_{(1,1)} &= c(1-\mu^2);\\
\nu^{\text{MYOPIC}}_{(0,1)} &= c(1-\frac{\mu\eta}{1-\zeta}),\ \text{if } I=2;\\
\nu^{\text{MYOPIC}}_{(0,1)} &= c(1-\mu\eta),\ \text{if } I\geq 3;\\
\nu^{\text{MYOPIC}}_{(a,i)} &= c,\ \text{for all } a\in\mathcal{A} \text{ and } 2\leq i\leq I;
\end{aligned}$$

4.4.4 Admission Control Problem with Delay to Server with an Infinite Buffer

In this subsection we assume linear holding costs.

Notice that the indices calculated in the algorithm FA's "Loop" are *independent* of the buffer length I (only the indices of states $(0, I-1)$ and $(*, I)$ in "Finalization" depend on I). In other words, considering two buffers with lengths $I_1 < I_2$, the MP indices of states $(1,0),(0,0),(1,1),\ldots,(0,I_1-2),(1,I_1-1)$ are the same for both buffers; the indices of states $(0,I_1-1)$ and $(*,I_1)$ would differ, while the remaining states only exist under buffer I_2. Therefore, the algorithm FA can be used to obtain the indices for infinite-length buffer. However, in such a case, "Loop" would never stop.

We present a simple algorithmic check (Figure 4.7) that can be run before "Loop" (and after "Initialization") to verify whether $K_0 = K_1 = \infty$, i.e., whether it is optimal to let the gate open always. It is due to the fact that the indices are calculated in nondecreasing order and they converge as the buffer length $I \to \infty$.

Lemma 4.2. *If the buffer length $I=\infty$, the marginal productivity indices calculated in "Loop" of algorithm FA under the discounted criterion in Figure 4.5 converge.*

Proof. We prove that A_K and D_K converge, and that their convergence implies the convergence of the MP indices. B.5(ii) implies that A_K converges to a limit, say, $A \leq \beta$ as $K \to \infty$. This limit must satisfy

$$A = \frac{\beta\zeta}{1-\beta+\beta\zeta+\beta\eta(1-A)}.$$

This equation has two solutions for A,

$$\frac{1-\beta+\beta\zeta+\beta\eta \pm \sqrt{(1-\beta+\beta\zeta+\beta\eta)^2-4\beta\eta\beta\zeta}}{2\beta\eta}.$$

$$A := \left[1-\beta+\beta\zeta+\beta\eta-\sqrt{(1-\beta+\beta\zeta+\beta\eta)^2-4\beta\eta\beta\zeta}\right]/(2\beta\eta);$$
$$A' := \beta\zeta+\beta(\mu-\eta)A; \quad D := cA/(\beta\zeta-\beta\eta A);$$
$$f^0 := -\frac{\frac{\beta\zeta}{A}D+\beta\zeta(c+\beta\mu BD)+[c-\beta(\mu-\eta)\beta D]\,B'}{\frac{A'}{A}+\beta A'B'+\beta\zeta\beta\mu(1-BA)};$$
$$f^1 := -\frac{\frac{\beta\zeta}{A}D+c\beta\zeta BA+[\beta\mu\beta\zeta+(1-\beta)\beta(\mu-\eta)]\,D+(c-\beta\zeta\beta C)A'}{\frac{A'}{A}+\beta A'B'+\beta\zeta\beta\mu(1-BA)};$$
$$g^0 := \frac{\beta\lambda\,(1+B')}{\frac{A'}{A}+\beta A'B'+\beta\zeta\beta\mu(1-BA)}; \quad g^1 := \frac{1+A'}{1+B'}g^0;$$
$$\nu_{(1,\infty)} := \frac{[\beta(1-\mu)\beta\zeta C+\beta\mu\beta(\mu-\eta)D]-\beta\mu A'f^0-\beta(1-\mu)B'f^1}{\beta\lambda-\beta\mu A'g^0-\beta(1-\mu)B'g^1};$$
$$\texttt{if } \nu \geq \nu_{(1,\infty)} \texttt{ then } K_0 := \infty; \quad K_1 := \infty; \quad \texttt{end \{if\};}$$

Figure 4.7: Algorithmic check for the problem with infinite-length buffer.

However, it can be shown that $1 > \zeta - \beta\eta$ and $\beta < 1$ (which is true by the model parameter assumptions) implies

$$\frac{1-\beta+\beta\zeta+\beta\eta-\sqrt{(1-\beta+\beta\zeta+\beta\eta)^2-4\beta\eta\beta\zeta}}{2\beta\eta} < \beta$$
$$< \frac{1-\beta+\beta\zeta+\beta\eta+\sqrt{(1-\beta+\beta\zeta+\beta\eta)^2-4\beta\eta\beta\zeta}}{2\beta\eta}.$$

By B.5 it must be $A \leq \beta$, therefore the limit is

$$A = \frac{1-\beta+\beta\zeta+\beta\eta-\sqrt{(1-\beta+\beta\zeta+\beta\eta)^2-4\beta\eta\beta\zeta}}{2\beta\eta}.$$

Similarly, it can be shown (see the next subsection) that D_K converges and the limit is therefore

$$D = \frac{cA}{\beta\zeta-\beta\eta A}.$$

As a consequence of the above, the remaining expressions, including those of the MP indices, converge. □

If the algorithmic check does not confirm the infinite thresholds, the algorithm FA can be run, stopping the loop once an index greater than ν is found and omitting "Finalization" part.

4.4.5 Admission Control Problem with Delay under Time-Average Criterion

Our results extend directly to the admission control with delay under the time-average criterion.

Proposition 4.4. *By setting $\beta := 1$, the algorithm FA in Figure 4.5 computes the marginal productivity indices for problem (4.14) under the time-average criterion.*

Proof. The algorithm FA is valid by setting $\beta := 1$ for the time-average criterion because the MP indices under that criterion are obtained in the limit $\beta \to 1$ of the MP indices under the discounted criterion, and the limits of all the expressions exist and are finite. □

In case $I = \infty$, the algorithmic check in Figure 4.7 is only valid under $\beta < 1$, and therefore is not suitable for the time-average criterion. In fact, it is not necessary to perform such a check, because under the time-average criterion the indices diverge.

4.4.6 Further Remarks

If the system is in state $(1, 0)$, then the buffer is empty, because it was empty a period ago and the gate has been closed since then. Therefore, one could expect that the index of state $(1, 0)$ is the same as the index of state 0 in the no-delay problem, which is in fact true. Moreover, there is a simple interpretation of that expression.

If the buffer is empty, the expected total β-discounted holding cost is

$$\zeta\beta c\left[1 + \beta(1-\mu) + (\beta(1-\mu))^2 + \dots\right] = \frac{\beta\zeta c}{1-\beta+\beta\mu},$$

because ζ is the probability that the customer remains in the buffer for more than a period. The above expression is equal to $\lambda\nu_{(1,0)}$, the expected (total β-discounted) rejection cost if the rejection cost $\nu = \nu_{(1,0)}$. Thus, in state $(1, 0)$ it is optimal to close the gate if the expected rejection cost is lower than the expected discounted total holding cost of an admitted customer. Further, in state $(1, 0)$ it is optimal to let the gate open if the expected rejection cost is greater than the expected discounted total holding cost of an admitted customer. If the two expected costs are equal, then both closing and opening are optimal. It is also clear that under the former condition it is optimal to close the gate in *any* state, and therefore the indices of all states must not be smaller than $\nu_{(1,0)}$.

Figure 4.8 shows the indices for a number of instances of the admission control problem with delay. An extensive simulation study we have performed suggests a conver-

gence of the indices:

$$\begin{aligned}
&\nu_{(1,i)} \to \nu_{(0,i)} && \text{as } \lambda \to 0,\\
&\nu_{(1,i)} \to \nu_{(0,i-1)} && \text{as } \zeta \to 1,\\
&\nu_{(0,i)} \to \frac{\beta c}{1-\beta} && \text{as } i \to \infty,\\
&\nu_{(1,i)} \to \frac{\beta c}{1-\beta} && \text{as } i \to \infty.
\end{aligned}$$

The convergence of the MP indices to $\beta c/(1-\beta)$ is intuitive. If the buffer is almost full (say, the pervious-epoch queue length is $I-2$), then admitting a customer means to increase the overal holding cost by c at least in the following $I-2$ periods, because the admitted customer cannot leave the system earlier that the previous $I-2$ customers. Therefore, the expected total β-discounted holding cost is at least

$$\beta c\left[1+\beta+\beta^2+\cdots+\beta^{I-2}\right] = \frac{\beta c(1-\beta^{I-1})}{1-\beta}.$$

On the other hand, it is not greater than the expected holding cost of remaining in the buffer forever, which is

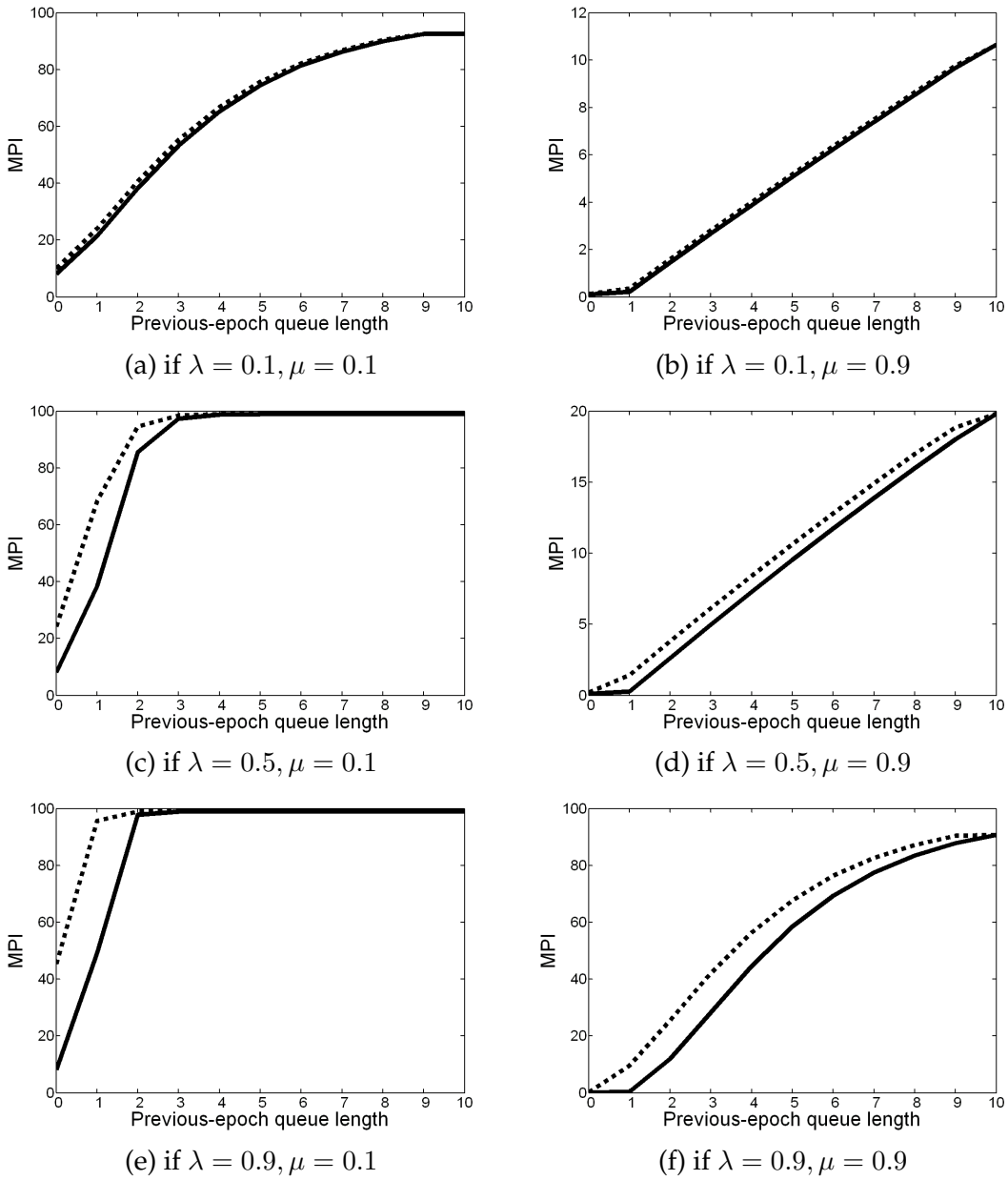
$$\beta c\left[1+\beta+\beta^2+\ldots\right] = \frac{\beta c}{1-\beta}.$$

Now it is clear that the MP indices converge to $\beta c/(1-\beta)$ as $I \to \infty$.

4.5 Fast Algorithm for the Job Completions Problem with Delay

In the job completions problem with delay, we define the one-period rewards in the following way. If the queue length is $i \in \mathcal{I}$ and action $a \in \mathcal{A}$ is chosen, then the gatekeeper's *one-period reward* is defined as the expected number of job completions during the current period,

$$R_{(a,i)} = R_i^a := \begin{cases} 0, & \text{if } i = 0 \text{ and } a = 1,\\ \lambda\mu, & \text{if } i = 0 \text{ and } a = 0,\\ \mu, & \text{if } i \geq 1. \end{cases}$$



(a) if $\lambda = 0.1, \mu = 0.1$

(b) if $\lambda = 0.1, \mu = 0.9$

(c) if $\lambda = 0.5, \mu = 0.1$

(d) if $\lambda = 0.5, \mu = 0.9$

(e) if $\lambda = 0.9, \mu = 0.1$

(f) if $\lambda = 0.9, \mu = 0.9$

Figure 4.8: Optimal MP indices for the admission control problem with delay with parameters $I = 10, c = 1, \beta = 0.99$. The solid line exhibits indices $\nu_{(1,i)}$ and the dotted line exhibits indices $\nu_{(0,i)}$.

Therefore, we have

$$\Delta_1 R_{(1,i)} = \begin{cases} -\lambda\mu, & \text{if } i = 0, \\ 0, & \text{if } i \geq 1, \end{cases}$$

and

$$\Delta_2 R_{(0,i)} = \begin{cases} \mu(1-\lambda), & \text{if } i = 1, \\ 0, & \text{if } i \geq 2, \end{cases} \qquad \Delta_2 R_{(1,i)} = \begin{cases} \mu, & \text{if } i = 1, \\ 0, & \text{if } i \geq 2. \end{cases}$$

The algorithm FA is presented in Figure 4.9, after substituting the following expressions: $C''_{0,0} := -\lambda B/\beta$; $C''_{K,K} := 0$; $C''_{K,K+1} := 0$.

However, we are only interested in the job completions problem under the long-run average criterion. Then, setting $\beta := 1$ in the fast algorithm, yields the constant index $\nu_{(a,i)} = -1$ for all $(a,i) \in \widetilde{\mathcal{I}}$. Figure 4.10 shows the simplified quantities that are obtained in this case in the fast algorithm FA. Since such an index is noninformative, we set out to obtain an alternative, second-order index in the following subsection.

4.5.1 Second-Order Marginal Productivity Index

Since the (first-order) index is noninformative, we proceed by introducing a second-order MP index $\gamma_{(a,i)}$, based on the Taylor series of $\nu_{(a,i)}$ at $\beta = 1$,

$$\nu_{(a,i)} = -1 + \gamma_{(a,i)}(1-\beta) + \mathcal{O}((1-\beta)^2), \qquad \text{as } \beta \to 1.$$

Thus, $\gamma_{(a,i)} := -\left.\dfrac{\partial \nu_{(a,i)}}{\partial \beta}\right|_{\beta=1}$. As the MP index policy prescribes to route an arriving customer to the queue of the lowest MP index, in the case of constant (first-order) indices the customer is to be routed to the queue of the lowest second-order MP index.

{Input I, λ, μ, β}
{Initialization}
$\zeta := \lambda(1-\mu)$; $\eta := \mu(1-\lambda)$; $\varepsilon := 1-\zeta-\eta$; $A_0 := 0$; $A'_0 := \beta\zeta$;
$B := \beta\mu/(1-\beta+\beta\mu)$; $B' := \beta\zeta B + \beta(\mu-\eta)$; $D_0 := 0$; $D'_0 := -\lambda\mu$;
$\nu_{(1,0)} := -B/\beta$;
{Loop}
for $K = 1$ to $I-1$ do
$A_K := \beta\zeta/[1-\beta+\beta\zeta+\beta\eta(1-A_{K-1})]$; $A'_K := \beta\zeta+\beta(\mu-\eta)A_K$; $Z_K := A_K A'_{K-1}/A'_K$;
if $K = 1$ then $D_1 := -\eta A_1/(\beta\zeta)$; else $D_K := \beta\eta D_{K-1}A_K/(\beta\zeta)$; end {if};
$D'_K := \beta(\mu-\eta)D_K$;

$$g^0 := \frac{\beta\lambda\,(1+B')}{\frac{A'_K}{A_K} + \beta A'_{K-1}B' + \beta\zeta\beta\mu(1-BA_{K-1})}; \quad g^1 := \frac{1+A'_{K-1}}{1+B'}g^0;$$

if $K = 1$ then

$$f^0 := -\frac{-\eta-(\beta D'_0+\mu)B' + \beta\mu\lambda\mu}{\frac{A'_1}{A_1} + \beta A'_0 B' + \beta\zeta\beta\mu}; \quad f^1 := -\frac{-\eta+(1-\beta)D'_0 - \mu A'_{K-1}}{\frac{A'_1}{A_1} + \beta A'_0 B' + \beta\zeta\beta\mu};$$

$$\nu_{(0,0)} := \frac{(1-\zeta)D'_0 - (1-\zeta)A'_0 f^0 - \zeta B' f^1}{\lambda - (1-\zeta)A'_0 g^0 - \zeta B' g^1};$$

else

$$f^0 := -\frac{\frac{\beta\zeta}{A_K}D_K - \beta D'_{K-1}B' + \beta\mu\beta\zeta D_{K-1}B}{\frac{A'_K}{A_K} + \beta A'_{K-1}B' + \beta\zeta\beta\mu(1-BA_{K-1})};$$

$$f^1 := -\frac{\frac{\beta\zeta}{A_K}D_K + (1-\beta)D'_{K-1} + \beta\mu\beta\zeta D_{K-1}}{\frac{A'_K}{A_K} + \beta A'_{K-1}B' + \beta\zeta\beta\mu(1-BA_{K-1})};$$

$$\nu_{(0,K-1)} := \frac{\eta\left[D'_{K-2} + D_{K-1}A'_{K-2}\right] + \varepsilon D'_{K-1} - \left[\eta Z_{K-1}+\varepsilon\right]A'_{K-1}f^0 - \zeta B' f^1}{\lambda - \left[\eta Z_{K-1}+\varepsilon\right]A'_{K-1}g^0 - \zeta B' g^1};$$

end {if};

$$\nu_{(1,K)} := \frac{\mu D'_{K-1} - \mu A'_{K-1}f^0 - (1-\mu)B' f^1}{\lambda - \mu A'_{K-1}g^0 - (1-\mu)B' g^1};$$

end {for};
{Finalization}
$A_I := \beta\zeta/[1-\beta+\beta\zeta+\beta\eta(1-A_{I-1})]$; $A'_I := \beta\zeta+\beta(\mu-\eta)A_I$; $Z_I := A_I A'_{I-1}/A'_I$;
$D_I := \beta\eta D_{I-1}A_I/(\beta\zeta)$;

$$f^0 := -\frac{\frac{\beta\zeta}{A_I}D_I - \beta\mu D'_{I-1}}{\frac{A'_I}{A_I} + \beta\mu A'_{I-1}}; \quad g^0 := \frac{\lambda(1+\beta\mu)}{\frac{A'_I}{A_I} + \beta\mu A'_{I-1}};$$

$$\nu_{(0,I-1)} := \frac{\eta\left[D'_{I-2} + D_{I-1}A'_{I-2}\right] + \varepsilon D'_{I-1} - \left[\eta Z_{I-1}+\varepsilon\right]A'_{I-1}f^0}{(\eta+\varepsilon)\lambda - \left[\eta Z_{I-1}+\varepsilon\right]A'_{I-1}g^0};$$

$$\nu_{(*,I)} := \frac{D'_{I-1} + \frac{\beta\zeta}{A_I}D_I Z_I}{\lambda(1-Z_I)};$$

{Output $\left\{\nu_{(a,i)}\right\}_{(a,i)\in\widetilde{\mathcal{I}}}$}

Figure 4.9: Fast algorithm FA for the calculation of MP indices for the job completions problem.

{Input I, λ, μ}
{Initialization}
$\zeta := \lambda(1-\mu)$; $\eta := \mu(1-\lambda)$; $\varepsilon := 1-\zeta-\eta$;
$A_0 := 0$; $A'_0 := \zeta$; $B := 1$; $B' := \lambda$; $D_0 := 0$; $D'_0 := A'_0 - B'$;
$\nu_{(1,0)} := -1$;
{Loop}
for $K = 1$ to $I-1$ do
$A_K := \zeta/[\zeta + \eta(1-A_{K-1})]$; $A'_K := \zeta + \lambda\mu A_K$; $Z_K := A_K A'_{K-1}/A'_K$;
$D_K := A_K - 1$; $D'_K := A'_K - B'$;
if $K = 1$ then

$$f^0 := \frac{\mu(1-\lambda\mu-\lambda^2)}{\lambda(1+\lambda)+\mu(1-\lambda\mu-\lambda^2)}; \quad f^1 := \frac{\mu(1-\lambda\mu)}{\lambda(1+\lambda)+\mu(1-\lambda\mu-\lambda^2)};$$

else

$$f^0 := \frac{-\mu(1-\lambda\mu-\lambda^2)D_{K-1}}{\lambda(1+\lambda)-\mu(1-\lambda\mu-\lambda^2)D_{K-1}};$$

$$f^1 := \frac{-\mu(1-\lambda\mu)D_{K-1}}{\lambda(1+\lambda)-\mu(1-\lambda\mu-\lambda^2)D_{K-1}};$$

end {if};
$g^0 := 1 - f^0$; $g^1 := 1 - f^1$; $\nu_{(0,K-1)} := -1$; $\nu_{(1,K)} := -1$;
end {for};
{Finalization}
$A_I := \zeta/[\zeta + \eta(1-A_{I-1})]$; $A'_I := \zeta + \lambda\mu A_I$; $Z_I := A_I A'_{I-1}/A'_I$; $D_I := A_I - 1$;

$$f^0 := \frac{-\mu(1-\lambda\mu-\lambda)D_{I-1}}{\lambda(1+\mu)-\mu(1-\lambda\mu-\lambda)D_{I-1}}; \quad g^0 := 1 - f^0;$$

$\nu_{(0,I-1)} := -1$;
$\nu_{(*,I)} := -1$;
{Output $\{\nu_{(a,i)}\}_{(a,i)\in\widetilde{\mathcal{I}}}$}

Figure 4.10: Fast algorithm FA for the calculation of MP indices for the job completions problem under the time-average criterion.

We conjecture that the second-order indices are as follows ($K \geq 1$):

$$\begin{aligned}
\gamma_{(1,0)} &= \frac{1}{\mu} - 1, \\
\gamma_{(0,0)} &= \frac{2}{\mu} - 1 - \frac{\mu - (\mu+\lambda)\zeta}{\mu^2 + \mu\eta\zeta}, \\
\gamma_{(1,1)} &= -\left(2 + \frac{\lambda}{\mu}\right) + \frac{1}{\mu}\left(2 + \frac{\lambda}{\mu}\right), \\
\gamma_{(0,1)} &= -\left(2 + \frac{\lambda}{\mu}\right) + \frac{1}{2\mu}\left(5 + 3\frac{\lambda}{\mu}\right) + \frac{2\lambda - 1}{2\mu(2\zeta+1)}\left(1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{\mu\eta}\right) + \frac{\lambda^2}{2\mu^2\eta(2\zeta+1)}, \\
\gamma_{(1,K+1)} &= \gamma_{(1,K)} + \frac{1}{\mu}\left\{1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{\mu\eta}\left[1 + \frac{\zeta}{\eta} + \cdots + \left(\frac{\zeta}{\eta}\right)^{K-1}\right]\right\}, \\
\gamma_{(0,K+1)} &= \gamma_{(0,K)} + \frac{1}{\mu}\left\{1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{\mu\eta}\left[1 + \frac{\zeta}{\eta} + \cdots + \left(\frac{\zeta}{\eta}\right)^{K-1}\right]\right\} + \frac{\lambda^2(\eta+\lambda)}{\mu^2\eta(2\zeta+1)}\left(\frac{\zeta}{\eta}\right)^K, \\
\gamma_{(*,I)} &= \frac{I}{\mu} - 1 + \frac{\lambda}{\mu\eta}\left[(I-1) + (I-2)\frac{\zeta}{\eta} + \cdots + \left(\frac{\zeta}{\eta}\right)^{I-2}\right].
\end{aligned}$$

The last expressions can be simplified. If $\lambda \neq \mu$, then

$$\begin{aligned}
\gamma_{(1,K+1)} &= \gamma_{(1,K)} + \frac{1}{\mu-\lambda} - \frac{1}{\mu-\lambda}\frac{\lambda^2}{\mu^2}\left(\frac{\zeta}{\eta}\right)^K, \\
\gamma_{(0,K+1)} &= \gamma_{(0,K)} + \frac{1}{\mu-\lambda} + \left[\frac{\eta+\lambda}{\eta(2\zeta+1)} - \frac{1}{\mu-\lambda}\right]\frac{\lambda^2}{\mu^2}\left(\frac{\zeta}{\eta}\right)^K, \\
\gamma_{(*,I)} &= \frac{I}{\mu} - 1 + \frac{\lambda}{\mu}\left[\frac{\zeta\left(\frac{\zeta}{\eta}\right)^{I-1} - \eta + I\,(\mu-\lambda)}{(\mu-\lambda)^2}\right].
\end{aligned}$$

If $\lambda = \mu$, then

$$\begin{aligned}
\gamma_{(1,K+1)} &= \gamma_{(1,K)} + \frac{2}{\mu} + \frac{K}{\eta}, \\
\gamma_{(0,K+1)} &= \gamma_{(0,K)} + \frac{2}{\mu} + \frac{K}{\eta} + \frac{\eta+\lambda}{\eta(2\zeta+1)}, \\
\gamma_{(*,I)} &= \frac{I}{\mu} - 1 + \frac{I(I-1)}{2\eta}.
\end{aligned}$$

4.6 Conclusions

We have presented a restless bandit approach which yielded an efficient exact algorithm for the calculation of the marginal productivity indices and optimal threshold queue lengths of an admission control problem with an action or information delay of

one period. The algorithm draws on and significantly reduces the complexity of the adaptive-greedy algorithm for the calculation of restless bandit optimal index policy.

We propose such indices as building blocks in a heuristic for a harder problem of admission control and/or routing to parallel queues, where the queues can be heterogeneous in buffer lengths, departure probabilities, holding costs, discount factors, and delays. The evaluation of the heuristic is a part of the work in progress.

Our approach seems to be tractable also for the admission control problem with larger delays and, more generally, for arbitrary restless bandits with delays.

Everyone carries a burden.
< J. Nohavica >

Chapter 5

Dynamic Product Promotion and Knapsack Problem for Perishable Items

5.1 Introduction

Senior managers in retail industry make important decisions upon assortment planning, product pricing, and product promotion. Product assortment (collection of products and their shelf space and location) is a strategic decision defining the retail brand image and is taken over a long-term planning period (Kök et al., 2006). The latter two are also strategic, yet in a weaker sense; they can and do be used in practice in day-to-day marketing decisions to dynamically adjust to demand variations. Within the food retail industry, the necessity, frequency, and complexity of pricing and promotion decisions are magnified by *perishability* of products.

An approach routinely applied to the revenue management of perishable products is *dynamic pricing*, i.e. adjusting product price to observed or expected demand variation. The perishability property has deforming implications on product demand, which thus becomes the crucial factor in revenue management models, as documented in Elmaghraby and Keskinocak (2003), who gave an extensive overview of research on dynamic pricing and its adoption in practice.

However, discrete-time decision making, implementation costs, and retail brand image strategy make practitioners not to like changing prices too often or in an "unsystematic" fashion as prescribed by theoretical dynamic pricing models. In addition, the price reduction must usually be done over all units of the product, thus losing possible profit from customers willing to pay the original, higher price. Revenue managers naturally

intend to avoid situations in which the (optimal) price is lower than the costs associated with the inventory, as it leads to a negative net profit. Such a *loss-averse* behavior may then result in conservative product orders, which in turn increase the probability and the length of stockout periods. Based on several studies, Campo and Gijsbrechts (2005) documented that consumers' reaction to stockouts occurrence may have significant negative impacts on retail sales and revenues.

The above suggests that there is a strong need by retail managers for a "softer" marketing tool, which would dynamically allow them to improve sales and revenues, yet not altering product prices.[1] We therefore design a revenue management model in which demand is altered not by price changes, but rather by moving a number of product units to a promotion space, where they are likely to attract extra customers. Examples of the promotion space include shelves close to the cash register, promotion kiosks, or a depot used for selling via the Internet.

Thus, we address the problem of filling the promotion space to maximize the aggregate expected revenue, which we have termed the *Knapsack Problem for Perishable Items*. We assume that the items can be repeatedly reallocated, resulting in strategies allowing for temporary promotions. Nevertheless, given the structural properties we derive, our results also apply to the model, in which backward reallocation of items already promoted is not allowed. Throughout this chapter, we focus on an example of a food retailer, referred as to the food promotion problem and we consider the case when there is a unique unit of each product.

To summarize the chapter, we develop a *dynamic promotion* model addressing the problem faced by merchandiser of choosing a set of products (items) to be reallocated to a promotion location with limited space. Promotion decisions are difficult, because of the combinatorial complexity of allocating a scarce promotion space over multiple periods, and because the demand is uncertain. In fact, we conclude that finding a tractable optimal solution is most likely to be an unreachable goal.

Our approach relies on a decomposition of the problem to single products. Each product is then assigned a *promotion priority index*, which captures the marginal rate of substitution (by properly taking into account both the cost and the opportunity cost of promotion) as a function of its price, salvage value, lifetime, expected demand, and expected promotion power. Such an index, which we derive in closed form, allows us to consider a *promotion-priority-index policy*: Select the products for promotion accordingly to their promotion priority indices. This policy may be suboptimal, however, it possesses important practical characteristics of being interpretable, thus easing managerial

[1] The practical value of such a tool is evident: Capgemini, Intel, Cisco, and Microsoft cooperated on a decision support system which includes Dynamic Promotion Management as one of three key solution areas.

insight, adaptable to additional features, easy-to-implement, and nearly-optimal.

5.1.1 Goals and Contributions

Section 5.2 outlines our model and briefly review the work on related models in the literature. We further describe the milestones in the research on the bandit problem and its extensions, especially the restless bandit problem, which we use to set our model. The celebrated classic result on the bandit problem is the optimality of a dynamic solution defined by certain priority indices. Derivation of such indices for our model is one of the two main goals of this work.

The Knapsack Problem for Perishable Items is formalized in Section 5.3. Since the dynamic programming formulation is most likely to be intractable, we formulate it approximately using a Lagrangian relaxation, which allows us to decompose the problem into single-item case.

An optimal promotion policy for a single item defined via promotion priority indices is derived in Section 5.5. The optimal dynamic promotion policy is shown to be time-monotonous: the efficiency of promoting nondecreases over time. Our single-item model is one of the first finite-horizon bandit problems, in which priority indices are obtained in a closed form.

This leads us to our second goal and contribution, the development of a new index-based heuristic for the Knapsack Problem for Perishable Items in Section 5.6: "Promote the items that are given by an optimal solution to the knapsack subproblem with item's promotion priority indices multiplied by volumes as the objective function price coefficients and item volumes as the knapsack constraint weights". Its superiority to other bandit problem heuristics is suggested by a computational study described in Section 5.7.

Concluding remarks are given in Section 5.8. Proofs are deferred to the Appendix.

5.2 Model Outline and Related Work

A *perishable item* is a product unit with an associated lifetime ending at a *deadline*. At the deadline (e.g., the "best before" date; the moment of replenishment with a fresher product) the product's demand drops to zero and only a *salvage value* is received. An event of selling can happen before the deadline, causing that the standard product *revenue* (i.e., profit margin) is obtained. The probability of selling only depends on whether the item is being promoted or not.

The task is to select regularly a subset of perishable items for a promotion space (knapsack) so that the expected aggregate total discounted revenue is maximized. We

call this model the Knapsack Problem for Perishable Items (KPPI) and give its formal statement in Section 5.3.

We set the model in discrete time as a Markov decision process. We assume that the decisions are made in some regular time moments, say, twice a day, and the problem parameters are adjusted to such time periods. In general, the KPPI defines a stochastic variant of the knapsack problem with multiple units. As time evolves, some items get sold accordingly to a stochastic time-homogeneous demand and some items perish deterministically at their deadlines.

A model related to ours is the so-called *Dynamic and Stochastic Knapsack Problem* (DSKP), which is, however, different in nature (see, e.g. Papastavrou et al., 1996, and further work by the same authors). The DSKP is the problem of finding an online rule to immediately reject or accept arriving items with a random value and/or random weight. In the DSKP, however, the items cannot disappear from the knapsack. In the same vein lies the problem of optimal project selection studied by Lu et al. (1999) and other authors.

The KPPI is further closely related to dynamic pricing problems, and we show that a simple dynamic pricing problem (of a single product) can be formulated in our framework as a dynamic promotion problem when one must pay for promotion. However, our formulation naturally requires a restriction on the volume of promoted items, while such restriction is a nonsense in the dynamic pricing problem.

Accordingly to the food promotion problem, we assume that *no replenishment* occurs. Since the products are perishable, the new delivery usually supplies a fresher product, i.e., perishable at different time moment. As we argue in this chapter, the perishment moment is a crucial factor for optimal promotion strategies, therefore different deliveries can and should be considered as different products. We further assume in our model that the demand is time-homogeneous. Elmaghraby and Keskinocak (2003) remarked that for nondurable goods demand is often independent over time, especially for most necessity items, where consumers make frequent repeat purchases, which is in line with this chapter's focus on food products. Nevertheless, our results can be directly extended to the case of non-homogeneous demand and the heuristic we propose is likely to improve in the case of demand nonincreasing over time.

5.2.1 Bandit Problem Literature

A natural mathematical setting for the KPPI problem is the multi-armed bandit problem (cf. Gittins, 1979), in which one wants to dynamically choose between various *bandits* (reward-yielding processes) one in an optimal fashion. That model captures the fundamental trade-off between *exploitation* of current rewards and *exploration* of possible

future rewards. Gittins (1979) showed that the multi-armed bandit problem can be optimally solved using his indices, which can be calculated in polynomial time.

The bandits are here the perishable items. In our model, however, there are four complications: the bandits are *restless*, because the items can get sold regardless of being in the knapsack or not, the time horizon is *finite* due to perishability, and we are to select *more than one* item for the knapsack, which is allowed to be filled partially, due to the *heterogeneity* of the items. We thus mix up two models: the bandit problem and the knapsack problem.

The multi-armed restless bandit problem (which is a generalization of the multi-armed bandit problem considered in Gittins (1979)) over the infinite horizon was proven to be PSPACE-hard even in its deterministic version (Papadimitriou and Tsitsiklis, 1999). The research focus thus shifts to the design of well-grounded, tractable heuristic policies. For the analysis we use the framework and methodology proposed for restless bandits by Niño-Mora (2002, 2006b). That work provided a sufficient condition for a restless bandit to be *indexable* together with an adaptive-greedy algorithm, which in $\mathcal{O}(n^3)$ operations (where n is the number of states) computes corresponding *marginal productivity (MP) indices* that extend earlier indices of Gittins (classical bandits, 1979) and Whittle (restless bandits, 1988). In our problem, the MP index can be interpreted as the *promotion priority index*.

The indexability property of a single item modeled as a restless bandit means that there exist promotion priority indices such that the optimal solution is to promote the item whenever its promotion priority index is higher than the cost of promotion space occupation (promotion cost). When coupling the bandits back into a multi-armed (non-restless) bandit problem, the promotion priority indices define an optimal policy: At every decision epoch choose the bandit of highest promotion priority index (*promotion-priority-index policy*). Such a priority policy is in general not optimal for the restless case, in which it becomes a well-grounded, efficient and practical heuristic.

Regarding the bandit problems with finite horizon, interesting results of index nature appear very sporadically, because of the intractability of the model, and therefore other methods (such as dynamic programming) are usually used. Even then, the problem is computationally intractable. Nevertheless, there is a tractable instance, the so-called *deteriorating case*, first presented for an infinite-horizon bandit problem by Gittins (1979), which was also successfully applied in a problem with finite-horizon objective (Manor and Kress, 1997). In that setting, the bandits were, however, not restless. The same is the case for the index policies for the finite-horizon multi-armed (non-restless) bandit problem: Niño-Mora (2005) showed that such a problem is indexable and provided an $\mathcal{O}(T^2n^3)$ algorithm (where T is the time horizon and n is the number of states)

to compute the corresponding index. Such an algorithm can be improved to $\mathcal{O}(T^2n^2)$ for certain bandits with special structure.

The bandit problem framework was used to analyze adaptive marketing strategies by various authors. Caro and Gallien (2007) developed a model for dynamic assortment of seasonal goods and proposed an assortment-priority-index policy using approximate indices that they obtained in a closed form. An adaptive model for interactive marketing environment was introduced in Bertsimas and Mersereau (2007). Both works build upon the classical multi-armed bandit problem (in which the bandits are *not* restless) and propose heuristics based on an approximate problem decomposition.

5.3 Knapsack Problem for Perishable Items

In this section we describe formally the Knapsack Problem for Perishable Items (KPPI) arising from the motivation problem of food promotion. Let $s = 0, 1, 2, \ldots$ be the discrete-time *epochs* and let period t be the period between epochs t and $t-1$.

5.3.1 Perishable Items

Throughout the chapter, *item* refers to one unit of one product, and we consider to deal with *perishable items*, defined as follows.

Definition 5.1. An item is called *perishable* at an associated *deadline*, if it possesses the following three features:

(i) there is a stochastic process called *demand* existing before the deadline, which can make the item to be *sold*;

(ii) if the item is sold, a *revenue* (profit margin) is accrued;

(iii) if the item is not sold at its deadline, a *salvage value* is accrued.

Suppose that we have a set $\mathcal{I}$ of $I \geq 2$ perishable items. Let item $i \in \mathcal{I}$ have the deadline $T_i \geq 1$ (integer). Denote the item's revenue $R_i > 0$ and the item's salvage value $\alpha_i R_i$ for some $-\infty < \alpha_i \leq 1$.[2] (The salvage value $\alpha_i R_i$ could be negative in case of high perished-item destroying costs.)

Suppose that at discrete-time *decision epochs* $s = 0, 1, \ldots, T_i - 1$ we can decide between two actions: to *promote* the item, so that the item gets access to a *promotion demand*, or to *not promote* the item, so that it gets access to a *standard demand*. We assume that the

[2]The case of revenue $R_i = 0$ may also be included, but in that case we need to require $\alpha_i R_i < 0$.

promotion demand and the standard demand are independent. Further, a promotion cost ν is paid every time period the item is promoted.

The state space $\mathcal{X}_i$ contains $T_i + 1$ states. State $t \in \mathcal{T}_i := \{1, 2, \ldots, T_i\}$ means that the item is *unsold and t periods remain to its deadline*, and one of the two possible actions must be chosen; therefore the states in $\mathcal{T}_i$ will be called *controllable*. At the deadline, the item moves to the absorbing terminal state 0. The item also moves to the state 0 once it is sold. State 0 will be called *uncontrollable*, because no decision needs to be made.

The time counter s should not be confused with the controllable states of any item i, for which we use variable $t \in \mathcal{T}_i$. At time period $s \in \{0, 1, \ldots, T_i - 1\}$, the item can be either in the controllable state $t = T_i - s$ (if it is still unsold) or in the uncontrollable state 0 (if it has been sold). At the initial time period $s = 0$, the item is in the controllable state $t = T_i$. At the deadline $s = T_i$ and afterwards, the item is in the uncontrollable state 0.

We consider homogeneous standard and promotion demands, which results in homogeneous transition probabilities. The item is sold (moves to state 0) with probability $1 - q_i$ within one time period (and with probability $0 < q_i \leq 1$ remains unsold moving to state $t - 1$), if it is not promoted in state t. Formally,

$$q_i := \mathbb{P}[\text{the not promoted item } i \text{ is not sold in one period}],$$

where the symbol $\mathbb{P}$ denotes probability. Analogously, the item is sold (moves to the state 0) with probability $1 - p_i$ within one time period (and with probability $0 < p_i \leq 1$ remains unsold moving to state $t - 1$), if it is promoted in state t. The difference $q_i - p_i$ will be called *promotion power*, as it captures the increase in the probability of being sold caused by promoting.

We restrict ourselves to the case when future revenues are discounted with a discount factor $0 < \beta < 1$ per period. We can interpret the discount factor as that with probability $1 - \beta$ an event (*bankruptcy*), which implies that there is no need to solve the problem in the next time epoch, happens.[3] At the end of the section we extend the results for the limiting case $\beta = 1$.

To summarize, a perishable item (the subscript i is dropped) is defined as a Markov decision process (MDP) as follows.

- The state space is $\mathcal{X} := \mathcal{T} \cup \{0\}$;
- The action space for controllable states in $\mathcal{T}$ is $\mathcal{A} := \{0, 1\}$: we can either *promote* (active action; 1) or *not promote* (passive action; 0); for uncontrollable state 0 a

[3] The discount factor should reflect both the intrinsic and systematic risk of the company or store in hand.

unique action (say, action 0) is available;

- The transition probability matrix $\boldsymbol{P}^{1|\mathcal{T}}$ for promoting is[4]

$$\boldsymbol{P}^{1|\mathcal{T}} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ T \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \cdots & T-1 & T \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1-p & p & 0 & 0 & 0 & 0 \\ 1-p & 0 & p & 0 & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 & 0 \\ 1-p & 0 & 0 & 0 & p & 0 \end{pmatrix} \end{array},$$

where $\boldsymbol{P}^{1|\mathcal{T}}_{i,j}$ is the probability of moving from state $i \in \mathcal{X}$ to state $j \in \mathcal{X}$ in 1 period if the item is promoted at all states in $\mathcal{T}$ (i.e., including state i). Similarly, the transition probability matrix $\boldsymbol{P}^{1|\emptyset}$ for not promoting is

$$\boldsymbol{P}^{1|\emptyset} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ T \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \cdots & T-1 & T \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1-q & q & 0 & 0 & 0 & 0 \\ 1-q & 0 & q & 0 & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 & 0 \\ 1-q & 0 & 0 & 0 & q & 0 \end{pmatrix} \end{array}.$$

- If the item is not promoted in state $t \in \mathcal{T} \setminus \{1\}$, the one-period expected revenue $R^0_t := R(1-q)$ is incurred; If the item is not promoted in state $t = 1$, the one-period expected revenue $R^0_t := R(1-q) + \beta\alpha Rq$ is incurred; If the item is promoted in state $t \in \mathcal{T} \setminus \{1\}$, the one-period expected revenue $R^1_t := R(1-p)$ minus the promotion cost ν is incurred; If the item is promoted in state $t = 1$, the one-period expected revenue $R^1_t := R(1-p) + \beta\alpha Rp$ minus the promotion cost ν is incurred; In state 0 there is no revenue nor cost, i.e. $R^0_0 := 0$.

[4]Including the row "0" (referring to the uncontrollable state) in the transition probability matrices is correct as long as the transition probabilities are equal under both actions.

5.3.2 KPPI Objective

An item can be either left at its standard shelf (i.e., not promoted) or selected for a promotion knapsack (i.e., promoted) common for all the items and with limited integer capacity $W \geq 1$.[5] Let item i occupy integer space $W_i \geq 1$. To avoid trivial cases we assume that $W_i \leq W$ for all $i \in \mathcal{I}$ and $\sum_{i\in\mathcal{I}} W_i > W$. Decisions are made at time periods $s = 0, 1, \ldots, T-1$, where $T := \max\{T_1, T_2, \ldots, T_I\}$ is the problem's *time horizon*. The revenues are discounted by a one-period *discount factor* $0 < \beta < 1$.

Starting from joint state $\boldsymbol{t} := (t_i)_{i\in\mathcal{I}}$, we define the *expected aggregate total discounted revenue* under policy π as

$$\mathbb{E}^{\pi}_{\boldsymbol{t}}\left[\sum_{i\in\mathcal{I}}\sum_{s=0}^{\infty}\beta^s R^{a_i(s)}_{X_i(s)}\right], \tag{5.1}$$

where $X_i(s)$ is the state of item i at time epoch s and $a_i(s)$ is the action applied in time epoch s to item i, counting promoting as 1 and not promoting as 0. The symbol $\mathbb{E}^{\pi}_{\boldsymbol{t}}$ denotes the expectation under policy π if starting from joint state $\boldsymbol{t}$.

Denote by Π the set of all non-anticipative policies. The goal is to find a policy $\pi^* \in \Pi$ that maximizes (5.1) for $\boldsymbol{t} = \boldsymbol{T} := (T_1, T_2, \ldots, T_I)$ among all such policies, subject to

$$\sum_{i\in\mathcal{I}} W_i \cdot a_i(s) \leq W \text{ at each time } s = 0, 1, \ldots, \infty.$$

5.3.3 Dynamic Programming Formulation

Let us define the following terms for a fixed time epoch $s = 0, 1, \ldots, T$ and product i. The product is *perished*, if its deadline has already passed ($s > T_i$). The product is *perishing*, if its deadline is currently achieved ($s = T_i$). The product is *existing*, if its deadline has not passed yet ($s \leq T_i$). The product is *controllable*, if it is existing and not perishing ($s < T_i$). Let $\mathcal{I}_s = \{i \in \mathcal{I} : T_i \geq s\}$ be the set of existing products, $\mathcal{I}^0_s = \{i \in \mathcal{I} : T_i = s\}$ the set of perishing products, and $\mathcal{I}^+_s = \{i \in \mathcal{I} : T_i > s\}$ the set of controllable products. It should be clear that we have $\mathcal{I}^+_0 = \mathcal{I}$ and $\mathcal{I}^+_T = \emptyset$.

On the item level (as opposed to the product level considered in the previous paragraph), we will further need the term *unsold* items. Let $\boldsymbol{z}_s = (z_{s,i})_{i\in\mathcal{I}_s}$ be the *existing inventory* (the binary vector of numbers of unsold items of all existing products) at time epoch s. We will find it useful to define also $\boldsymbol{z}^+_s = (z_{s,i})_{i\in\mathcal{I}^+_s}$, the *controllable inventory* (the binary vector of numbers of unsold items of controllable products).

[5] To avoid unnecessary complications, we set the possibly product-dependent promotion costs $\nu_i := 0$. Their incorporation is straightforward and does not alter our structural results.

Let $\boldsymbol{y}_s^+ = (y_{s,i})_{i\in\mathcal{I}_s^+}$ be the binary vector of decision variables at time epoch s, denoting the number of items of product i chosen at time epoch s to be selected for the knapsack. A dynamic programming formulation of the KPPI follows:

$$D_T^{\text{MAX}}(\boldsymbol{z}_T) = \sum_{i\in\mathcal{I}_T^0} \alpha_i R_i z_{T,i} \tag{DP}$$

$$D_s^{\text{MAX}}(\boldsymbol{z}_s) = \sum_{i\in\mathcal{I}_s^0} \alpha_i R_i z_{s,i} + \max_{\substack{\boldsymbol{y}_s^+\le\boldsymbol{z}_s^+ \\ \sum_{i\in\mathcal{I}_s^+} W_i y_{s,i}\le W}} \left\{ \sum_{i\in\mathcal{I}_s^+} [y_{s,i}R_i(1-p_i) + (z_{s,i}-y_{s,i})R_i(1-q_i)] \right.$$

$$\left. +\beta \sum_{\boldsymbol{m}_s\le\boldsymbol{z}_s^+} \mathbb{P}_{\boldsymbol{z}_s^+}^{\boldsymbol{y}_s^+}[\boldsymbol{M}_s = \boldsymbol{m}_s] \left(D_{s+1}^{\text{MAX}}(\boldsymbol{z}_s^+ - \boldsymbol{m}_s) \right) \right\}$$

for time epochs $s = 0, 1, \ldots, T-1$. Here, $\boldsymbol{M}_s = (M_{s,i})_{i\in\mathcal{I}_s}$ is a nonnegative integer random vector denoting the number of units of each product that get sold at time epoch s. Thus we calculate the probability $\mathbb{P}_{\boldsymbol{z}_s^+}^{\boldsymbol{y}_s^+}[\boldsymbol{M}_s = \boldsymbol{m}_s]$ that items given by vector $\boldsymbol{m}_s$ get sold within one time period, having $\boldsymbol{z}_s^+$ unsold items out of which $\boldsymbol{y}_s^+$ are in the knapsack. The solution to the KPPI, prescribing the number of items selected for the promotion knapsack at the initial time epoch $s = 0$, is given by the vector $\boldsymbol{y}_0^+$ achieving maximum in $D_0^{\text{MAX}}(\mathbf{1})$, where $\mathbf{1} := (1)_{i\in\mathcal{I}}$.

Since selling is assumed to be independent for different products, we have

$$\mathbb{P}_{\boldsymbol{z}_s^+}^{\boldsymbol{y}_s^+}[\boldsymbol{M}_s = \boldsymbol{m}_s] = \prod_{i\in\mathcal{I}_s^+} \mathbb{P}_{z_{s,i}}^{y_{s,i}}[M_{s,i} = m_{s,i}] =$$

$$\prod_{i\in\mathcal{I}_s^+} \left\{ \sum_{m^0=0}^{m_{s,i}} \mathbb{P}_{z_{s,i}}^{y_{s,i}}\left[M_{s,i}^0 = m^0\right] \mathbb{P}_{z_{s,i}}^{y_{s,i}}\left[M_{s,i}^1 = m_{s,i} - m^0\right] \right\}, \tag{5.2}$$

where $M_{s,i}^0$ and $M_{s,i}^1$ are random variables denoting the number of not promoted and promoted units of product i, respectively, that get sold at time epoch s. And further,

$$\mathbb{P}_{z_{s,i}}^{y_{s,i}}\left[M_{s,i}^0 = m^0\right] = \begin{cases} q_i, & \text{if } m^0 = 0 \text{ and } z_{s,i} - y_{s,i} = 1, \\ 1-q_i, & \text{if } m^0 = z_{s,i} - y_{s,i} = 1, \\ 1, & \text{if } m^0 = z_{s,i} - y_{s,i} = 0, \\ 0, & \text{otherwise,} \end{cases} \tag{5.3}$$

$$\mathbb{P}_{z_{s,i}}^{y_{s,i}}\left[M_{s,i}^1 = m^1\right] = \begin{cases} p_i, & \text{if } m^0 = 0 \text{ and } y_{s,i} = 1, \\ 1 - p_i, & \text{if } m^0 = y_{s,i} = 1, \\ 1, & \text{if } m^0 = y_{s,i} = 0, \\ 0, & \text{otherwise,} \end{cases} \tag{5.4}$$

Notice that the time and space complexity of (DP) grows exponentially with the number of products, and thus quickly becomes intractable. The minimization term is somewhat more complex than a classical NP-hard knapsack problem and this must be solved for each possible combinatorial vector $\boldsymbol{z}_s$ of existing inventory for each time epoch s. Further, dynamic programming does not provide any generally good ground for defining an approximate solution. This motivates us to approach the problem from a different perspective, which we explore next.

5.4 Work-Reward Restless Bandit Formulation of KPPI

We next formulate the KPPI as a variant of the multi-armed restless bandit problem, where the restless bandit is replaced by what we call the *work-reward restless bandit*. We set out to obtain a tractable index rule based on the MP indices.

In the multi-armed restless bandit problem (cf. Whittle, 1988), all bandits have the same requirement on the resource. In our model, however, we admit non-uniform resource (i.e., knapsack space) requirements, which is a special case of the model in Niño-Mora (2002). In the restless bandit framework of Whittle (1988), the immediate *work* is assumed to be 1 for the active action and 0 for being passive. Nevertheless, in our case the active action (promoting) requires a non-uniform utilization of the knapsack and we need to reflect this feature in our model. Therefore, we define the immediate (promotion) work of an item i in state $t \in \mathcal{T}_i$ under the active action by its volume, $W_{i,t}^1 := W_i$, and $W_{i,t}^0 := 0$ under the passive action. We further define $W_{i,0}^0 := 0$.

We arrive to the following formulation of the KPPI, equivalent to (DP),

$$\begin{aligned} &\max_{\pi\in\Pi} \mathbb{E}_{\boldsymbol{T}}^{\pi}\left[\sum_{i\in\mathcal{I}}\sum_{s=0}^{\infty}\beta^s R_{i,X_i(s)}^{a_i(s)}\right] \\ &\qquad \text{subject to} \quad \sum_{i\in\mathcal{I}} W_{i,X_i(s)}^{a_i(s)} \le W \text{ at each time } s = 0, 1, \ldots, \infty \end{aligned} \tag{RB}$$

where, as before, $X_i(s)$ denotes the state of item i at time epoch s, starting at state $X_i(0) = T_i$.

5.4.1 Whittle Relaxation and its Interpretation

Whittle (1988) proposed for restless bandits what has become known as the *Whittle relaxation*: replace the infinite set of resource constraints by one constraint requiring to use the full resource *in expectation*. In our case, the Whittle relaxation of the (RB) is the following:

$$\begin{aligned} &\max_{\pi\in\Pi} \mathbb{E}^{\pi}_{\boldsymbol{T}}\left[\sum_{i\in\mathcal{I}}\sum_{s=0}^{\infty}\beta^s R^{a_i(s)}_{i,X_i(s)}\right] \\ \text{subject to}\quad &\mathbb{E}^{\pi}_{\boldsymbol{T}}\left[\sum_{i\in\mathcal{I}}\sum_{s=0}^{\infty}\beta^s W^{a_i(s)}_{i,X_i(s)}\right] = \frac{W}{1-\beta} \end{aligned} \tag{WR}$$

where we have simplified $\sum_{s=0}^{\infty}\beta^s W = \frac{W}{1-\beta}$.

The Whittle relaxation (WR) can be solved by the Lagrangian method. Let κ be a Lagrangian multiplier for the constraint, then the Lagrangian of (WR) is

$$L(\pi,\kappa) = \mathbb{E}^{\pi}_{\boldsymbol{T}}\left[\sum_{i\in\mathcal{I}}\sum_{s=0}^{\infty}\beta^s R^{a_i(s)}_{i,X_i(s)}\right] - \kappa\left(\mathbb{E}^{\pi}_{\boldsymbol{T}}\left[\sum_{i\in\mathcal{I}}\sum_{s=0}^{\infty}\beta^s W^{a_i(s)}_{i,X_i(s)}\right] - \frac{W}{1-\beta}\right)$$

and can be rewritten as

$$L(\pi,\kappa) = \sum_{i\in\mathcal{I}}\left(\mathbb{E}^{\pi}_{\boldsymbol{T}}\left[\sum_{s=0}^{\infty}\beta^s R^{a_i(s)}_{i,X_i(s)}\right] - \kappa\mathbb{E}^{\pi}_{\boldsymbol{T}}\left[\sum_{s=0}^{\infty}\beta^s W^{a_i(s)}_{i,X_i(s)}\right]\right) + \kappa\frac{W}{1-\beta} \tag{L}$$

For a given penalizing parameter κ, (L) can be decomposed and analyzed separately for each item, which we will do in Section 5.5. We interpret the parameter κ as the competitive market price of space, the resource provided by the knapsack. Indeed, the term $\kappa W/(1-\beta)$ can be viewed as the *money budget* allocated for the knapsack space we expect to be using (κW per period). Since we only consider the space utilization in expectation, we in fact assume existence of a space market, where we permit to "buy" some amount of space if necessary or to "sell" some amount of space if it is not used. Then, there is an optimal market price κ^* which balances expected supply (selling free space) and expected demand (buying necessary space). If this price is known, then $\max_{\pi\in\Pi} L(\pi,\kappa^*)$ solves (WR).

This *perfect market assumption* reflected at the Whittle relaxation is sufficient for the KPPI to be solved efficiently. Its solution, however, is not feasible for the original problem (RB), because no such space market is in practice available. The optimal solution is a dynamic adaptive-knapsack policy; in the original problem a dynamic fixed-knapsack

policy is sought. Nevertheless, the optimal solution to the Whittle relaxation yields a tractable bound for the original problem.

The optimal dynamic adaptive-knapsack policy, however, may be relevant in some applications. One could think of adjusting the knapsack's space dynamically as the optimal perfect market solution requires, e.g. by reserving a necessary number of promotion shelves, where the price κ^* must be paid as a promoter's wage for each reserved space unit.

5.5 Optimal Dynamic Promotion of Perishable Item

The aim of this section is to obtain the MP indices for a perishable item in isolation, which will be the building block for the KPPI solution developed in Section 5.6. The MP indices capture the marginal rate of promotion and define an index policy, which furnishes an optimal control of a perishable item by indicating when it is worth promoting. For that end, we introduce a per-period *promotion cost* $\nu_i \geq 0$, which must be paid in every period when the item is being promoted.

Since we are considering item i in isolation, in the following we drop the item's subscript i. Starting from state t, we define the *expected total discounted net revenue* under policy π as

$$\mathbb{E}_t^{\pi}\left[\sum_{s=0}^{\infty}\beta^s R_{X(s)}^{a(s)}\right] - \nu\mathbb{E}_t^{\pi}\left[\sum_{s=0}^{\infty}\beta^s W_{X(s)}^{a(s)}\right], \tag{5.5}$$

where, as before, $X(s)$ is the state at time epoch s and $a(s)$ is the action applied in time epoch s, counting promoting as 1 and not promoting as 0. The symbol $\mathbb{E}_t^{\pi}$ denotes the expectation under policy π if starting from state t.

Denote by Π the set of all non-anticipative policies for such a problem. The goal is to find a policy $\pi^* \in \Pi$ that maximizes (5.5) for $t = T$ among all such policies, and thus optimally resolves the trade-off between the expected total discounted revenue and the expected total discounted promotion cost.

The perishable item as defined above falls to the concept of restless bandit (a binary-action MDP). Under some circumstances, one can identify its optimal control in terms of MP indices. Next we show that such an optimal MP index policy for perishable item exists under a demand regularity condition, and we identify it.

5.5.1 Assumption

We will continue under the following assumption, which requires the promotion power $q - p$ to be positive.

Assumption 5.1. $q - p > 0$.

Assumption 5.1 is a consistency requirement on promotion power which rules out uninteresting items that should never be promoted. Indeed, the optimal action in all the states for an item with promotion power $q-p \leq 0$ is not promoting (as long as $\nu \geq 0$). We will show that promoting is optimal if promoting was optimal in the previous period, i.e., the efficiency of promoting nondecreases over time.

5.5.2 Marginal Productivity Index

Now we examine the economics of promoting the perishable item. In particular, we examine the efficiency of promoting the item in its current state if one must pay for promotion. We will identify circumstances in which it is worth to promote the item, by assigning the *marginal productivity* (MP) index, which captures the marginal rate of promoting, to each controllable state. The optimal MP index policy is: "Promote the item if and only if the MP index of the actual state is greater than the promotion cost ν." We employ the MP indices calculation method as described for restless bandits in Niño-Mora (2002).

We show in Proposition 5.1 that a perishable item is *indexable*, that is, the optimal decisions are prescribed by the MP index policy, using MP indices assigned to controllable states.

Proposition 5.1. *The perishable item is indexable, and the marginal productivity index if t periods remain to the deadline is*

$$\nu_t^* = \frac{R}{W} \frac{(q-p)\left[(1-\beta)\frac{1-(\beta p)^{t-1}}{1-\beta p} + (1-\beta\alpha)(\beta p)^{t-1}\right]}{1-(\beta q-\beta p)\frac{1-(\beta p)^{t-1}}{1-\beta p}}. \tag{5.6}$$

The proof of Proposition 5.1 is presented in the Appendix together with a more detailed description of the work-reward analysis. Next we list the most appealing properties of the MP indices, which have insightful interpretation and define heuristical priorities for promotion if various items compete for a limited promotion space. In Section 5.6 we implement MP index as promotion priority measure: the higher the MP index, the higher the promotion priority.

Proposition 5.2 (MP Index Properties). *For any controllable state t,*

(i) the MP index is nonnegative and proportional to R/W;

(ii) an item with lower probability of being sold when not promoted (i.e., with a higher q), ceteris paribus, has higher MP index;

(iii) (Time Monotonicity) the MP index of an item is nondecreasing as the deadline approaches.

(iv) the MP index of a non-perishable item (i.e., with an infinite deadline) is

$$\nu^*_\infty = \frac{R}{W}\frac{(q-p)(1-\beta)(1-\beta p)}{1-\beta q}. \tag{5.7}$$

MP index resolves the trade-off between immediate and postponed promotion. Proposition 5.2(iii) is a crucial property of MP indices, which demonstrates that the index is nondecreasing as the deadline approaches. Armed with this result, we can look for an *optimal promotion starting time* τ^*,

$$\tau^* := \max\{\tau \in \mathcal{T} : \ \nu^*_t > \nu \text{ for all } t \in \mathcal{T} \text{ such that } t \geq \tau\}. \tag{5.8}$$

In other words, (if τ^* is finite,) τ^* is the threshold time epoch, from which the MP index is larger than the promotion cost ν, i.e. from which it is optimal to start to promote the item, ceasing to promote it at the deadline. If τ^* is not finite, then it is never optimal to promote the item.

Corollary 5.0.1. *The optimal starting time τ^* is finite if and only if*

$$\frac{R}{W}(1-\beta\alpha)(q-p) = \nu^*_1 > \nu.$$

Further,

(i) if τ^* is finite, then promoting is optimal in all time epochs from τ^* to 1 and not promoting is optimal in the remaining time epochs;

(ii) if τ^* is not finite, then not promoting is optimal in all time epochs.

The above result assures that promotion is to be made in a natural way: the item is selected for promotion only once and remains promoted while it is profitable to do it.

5.5.3 Special Cases and Further Remarks

We further give the MP index for certain classes of perishable items.

Proposition 5.3.

(i) [Undiscounted Case] In the case $\beta = 1$, the marginal productivity index if t periods remain to the deadline is

$$\nu_t^* = \frac{R}{W}\frac{(1-\alpha)(q-p)(1-p)p^{t-1}}{1-q+(q-p)p^{t-1}}.$$

(ii) [Reduction to $c\mu$-rule] If $q = 1, W = 1$, and the discount factor $\beta = 1$, then the perishable item is indexable, and the marginal productivity index if t periods remain to the deadline is

$$\nu_t^* = R(1-\alpha)(1-p).$$

Proposition 5.3(ii) tackles the situation in which there is no possibility of selling the item if not promoted. Thus, only promoted item can be sold, and selling happens with probability $1-p$ in every period. Interpreting this probability as a service rate, the MP index reduces to the $c\mu$-index, well-known in the queueing theory (see, e.g., Buyukkoc et al., 1985), where $c := R(1-\alpha)$ is the reduction in revenue if item is not sold during its lifetime. Such an MP index is constant over time and in particular it does not depend on the number of periods before the deadline. This rule is fittingly applied in assortment practice where products are chosen accordingly to their profitability and attractiveness.

5.5.4 Formulation of Dynamic Pricing Problem in our Framework

Suppose that we are given an additional parameter called *discount* (price markdown) $D \geq 0$, so that the revenue is $R - D$ instead of R if the item is promoted. Thus, $1-q$ can be interpreted as the probability of selling the item priced at R, and $1-p$ can be interpreted as the probability of selling the item priced at $R-D$. Let $\widetilde{\nu}$ be the per-period cost of maintaining (or informing about) the lower price. We are thus addressing a simple case of the classic *dynamic pricing problem*.

In particular, we would like to have the following revenues:

$$\begin{aligned}
&\widetilde{R}_t^0 := \beta R(1-q), && \text{for } t \in \mathcal{T}\setminus\{1\};\\
&\widetilde{R}_t^0 := \beta R(1-q) + \beta\alpha R q, && \text{for } t = 1;\\
&\widetilde{R}_t^1 := \beta(R-D)(1-p), && \text{for } t \in \mathcal{T}\setminus\{1\};\\
&\widetilde{R}_t^1 := \beta(R-D)(1-p) + \beta\alpha R p, && \text{for } t = 1;\\
&\widetilde{R}_0^0 := 0,
\end{aligned}$$

Denote by $\widetilde{D} := \beta D(1-p)$. In order to cast the above problem into our framework,

we define the one-period revenue for all $t \in \mathcal{T}$ as follows:

$$R_t^0 := \widetilde{R}_t^0, \qquad R_t^1 := \widetilde{R}_t^1 + \widetilde{D}, \qquad R_0^0 := \widetilde{R}_0^0,$$

and the promotion cost $\nu := \widetilde{\nu} + \widetilde{D}$.

Then, this modified model is effectively the same as the original one, and hence the optimal policy for the dynamic problem is the following: "Price the product at the lower price if and only if the MP index of the actual state is greater than the cost $\widetilde{\nu} + \widetilde{D}$."

5.6 Index-Based Heuristics for KPPI

Together with the relaxation, Whittle (1988) proposed a simple priority policy for the multi-armed restless bandit problem employing the indices once an optimal index policy for each bandit is available: "Promote the bandit of highest index". Heuristics based on this simple idea showed good performance in various problems formulated in the framework of the multi-armed restless bandit problem. In the KPPI problem, however, this heuristic may not be the best proposal, since it assumes the same space requirement of all perishable items ($W_i = 1$).

In a more general model of Niño-Mora (2002), the resource requirements were assumed to be non-uniform and stochastic. Thus, his indices differ from the Whittle indices and it was shown that the latter may be suboptimal in the Niño-Mora (2002)'s model. We applied this approach to the perishable item in Section 5.5, where the item's MP index is volume-adjusted (since it includes a division by the volume W_i), so that the heuristic is: "Promote the item of highest volume-adjusted index".

Since the MP index can be interpreted as measuring the marginal rate of substitution (i.e., the price per unit of space requirement) of promoting the item, we propose the following heuristic construction for the KPPI: "Promote the items that are given by an optimal solution to the knapsack subproblem defined below with item's MP indices multiplied by volumes as the objective function price coefficients and item volumes as the knapsack constraint weights".

Notice that the classic greedy solution to the knapsack problem arising in our heuristic reduces it to "Promote the items of highest volume-adjusted index". It is well known that the greedy solution yields an optimal solution of a knapsack problem when all the weights are uniform; however, in the general case it is suboptimal. Our simulation study presented in Section 5.7 suggests that the latter heuristic reveals an analogous performance: it is inferior and converging to ours.

In the following we assume that $q_i > p_i$ holds for all items i (otherwise not promoting is always optimal for such an item). Recall expression (5.6) for the MP index

calculation, which is to be used in the heuristics via

$$v_i := W_i \nu^*_{i,T_i}. \tag{5.9}$$

Heuristic MPI–OPT: Calculate the prices v_i and then solve the knapsack subproblem optimally.

Heuristic MPI–GRE: Calculate the price/volume ratios v_i/W_i and then select the items for promotion in a greedy manner (highest first).

5.6.1 Knapsack Subproblem

Suppose that the knapsack-problem prices v_i of all items are calculated using expression (5.9). Then we have the following 0-1 knapsack problem to solve:

$$\begin{aligned} &\max_{\boldsymbol{z}} \sum_{i\in\mathcal{I}} z_i v_i \\ \text{subject to} \quad &\sum_{i\in\mathcal{I}} z_i W_i \le W \\ &z_i \in \{0,1\} \text{ for all } i \in \mathcal{I} \end{aligned} \tag{KP}$$

where $\boldsymbol{z} = (z_i : i \in \mathcal{I})$ is the vector of binary decision variables denoting whether the item i is selected for the promotion knapsack or not.

The quality of the solution $\boldsymbol{z}$ is not guaranteed to be optimal. The experimental study in the next section, however, reveals its nearly-optimal behavior, systematically outperforming other considered heuristics.

Finally, the next proposition asserts that the KPPI is a generalization of the knapsack problem.

Proposition 5.4 (KPPI Reduction to KP). *If $T_i = 1, q_i = 1, p_i = 0$ for all $i \in \mathcal{I}$, then any optimal solution $\boldsymbol{z}^*$ of the knapsack problem (KP) is an optimal solution of the KPPI.*

5.7 Experimental Study

In this section we present results of computational experiments, in which we evaluate the performance of heuristics MPI–OPT and MPI–GRE. We further compare their performance to the greedy *Earlier-Deadline-First* policy (EDF–GRE), a benchmark policy often observed in practice.

Heuristic EDF–GRE: Select products in a greedy manner after sorting the items so that product i_1 is preferred to product i_2, if:

(i) $T_{i_1} < T_{i_2}$,

(ii) $T_{i_1} = T_{i_2}$ and $R_{i_1}(1-\alpha_{i_1}) > R_{i_1}(1-\alpha_{i_1})$,

(iii) $T_{i_1} = T_{i_2}$ and $R_{i_1}(1-\alpha_{i_1}) = R_{i_1}(1-\alpha_{i_1})$ and $W_{i_1} < W_{i_2}$.

The following is the worst-case (i.e., minimizing) solution of the knapsack subproblem whenever all the price coefficients are positive, which is our case.

Heuristic MIN: Leave the knapsack empty.

In each experiment we randomly generate 10^4 instances for each fixed pair (I,T), denoting the number of products and the time horizon, respectively, such that $I \in \{2,3,4,\ldots,8\}$ and $T \in \{2,4,6,\ldots,20\}$. For each product i we set $\alpha_i = 0.5$ and we assure that $T_1 := T$. We assume that the standard and the promotion demands are Poisson with the respective means λ_i^0, λ_i^1 and such that $\frac{1}{2}\lambda_i^a T_i \leq 1 < \frac{3}{2}\lambda_i^a T_i$ for both $a \in \{0,1\}$. The last condition assures that each item has a non-extreme probability of being sold before the deadline. Thus, we define $q_i := \exp\{-\lambda_i^0\}$ and $p_i := \exp\{-\lambda_i^1\}$, and assure that $q_i > p_i$. We further generate the following uniformly distributed parameters:

$$W_i \in [10,50]; \qquad R_i \in [10,50]; \qquad T_i \in [2,T]; \qquad \lambda_i^0, \lambda_i^1 \in \left(\frac{2J_i}{3T_i}, \frac{2J_i}{T_i}\right].$$

Finally, a uniformly distributed knapsack volume is generated: $W \in [\max\{W_i\}, 30\% \cdot \sum_i W_i)$.

We focus on the discount factor $\beta = 1$, as this is the case most likely to be implemented in practice. Moreover, our experiments (not reported here) suggest that this is also the hardest case and the performance of index-based heuristics improves as the discount factor diminishes.

The experiments were performed on PC with 2.66 GHz CPU and 1.5 GB RAM working on Windows XP. A Delphi code was developed by the author, implementing a standard enumerative routine for the knapsack subproblem. Finally, the performance evaluation measures (see below) were calculated using Matlab, which also created the figures presented here.

5.7.1 Performance Evaluation Measures

We obtain the maximizing policy solving the (DP) optimally, which also yields the optimal objective value D^{MAX}. The objective values of the other policies are also obtained via the Bellman equations, employing the respective heuristic at each step, denoted D^{π}

for a policy π. We next introduce performance evaluation measures we use to report the experiment results.

The *relative suboptimality gap* of policy π is calculated via

$$\mathrm{rsg}(\pi) = \frac{D^{\mathrm{MAX}} - D^{\pi}}{D^{\mathrm{MAX}}}. \tag{5.10}$$

Clearly, we have $0 \leq \mathrm{rsg}(\pi) \leq 1$, where $\mathrm{rsg}(\pi) = 0$ is obtained by the maximizing policy. However, $\mathrm{rsg}(\pi) = 1$ cannot be achieved unless $\alpha_i \leq 0$ at least for some item i. This motivates us to introduce the *adjusted relative suboptimality gap* of policy π, calculated via

$$\mathrm{arsg}(\pi) = \frac{D^{\mathrm{MAX}} - D^{\pi}}{D^{\mathrm{MAX}} - D^{\mathrm{MIN}}}. \tag{5.11}$$

In our case $0 \leq \mathrm{arsg}(\pi) \leq 1$, and both limiting values can be achieved.

We further introduce a measure to be used to compare the mean performance of an alternative heuristic with respect to Heuristic MPI–OPT, as follows:

$$\mathrm{ratio}(\pi) = \frac{\mathrm{mean}(\mathrm{rsg}(\pi))}{\mathrm{mean}(\mathrm{rsg}(\text{MPI–OPT}))}. \tag{5.12}$$

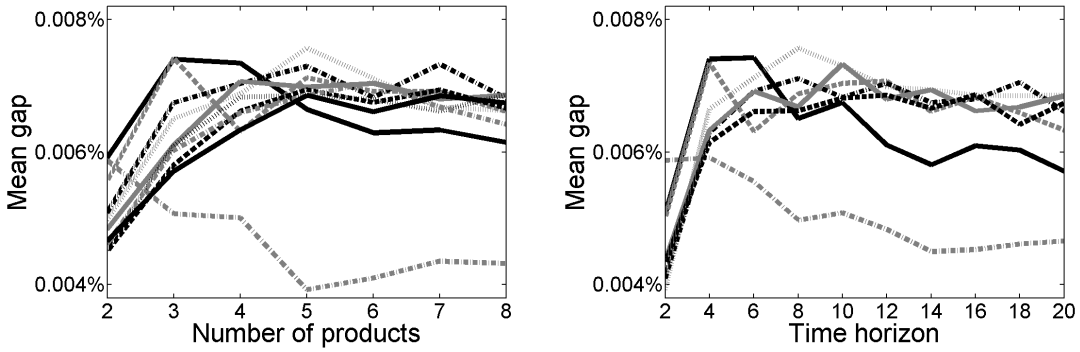
This ratio captures the extent to which the mean absolute gap (i.e., the revenue loss) created by Heuristic MPI–OPT may be expected to be magnified if policy π is implemented instead. Thus, we have $\mathrm{ratio}(\pi) > 1$ if and only if policy π is on average worse than Heuristic MPI–OPT. An analogous ratio is used with the arsg measure.

5.7.2 Results

Figure 5.1 exhibits two projections of the mean $\mathrm{rsg}(\text{MPI–OPT})$ as function of the number of products I and the time horizon T. The figure shows an excellent mean performance of heuristic MPI-OPT well below 0.01%, and further suggests that such a performance can be expected even for higher values of I and T. These strong results are further confirmed in Figure 5.2 considering the arsg measure.

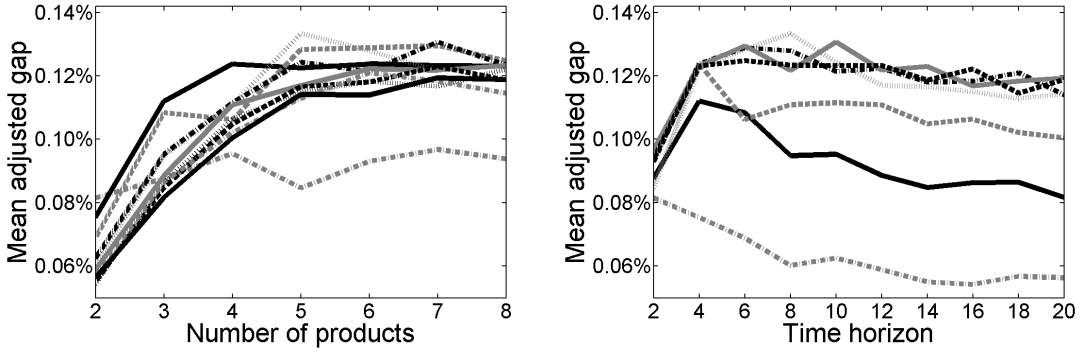
The ratio of the benchmark Heuristic EDF–GRE is presented in Figure 5.3 and Figure 5.4. The benchmark policy's mean gap is in all cases more than 50-times larger than that of Heuristic MPI–OPT, and the ratio grows with the number of items I. Further, in Figure 5.5 and Figure 5.6 we evaluate Heuristic MPI–GRE, whose mean performance is in all cases more than 10-times worse, though improving with higher I once this passes the value 5.

Finally, we remark that the worst-case performance achieved by the maximum rsg (arsg) values of Heuristic MPI–OPT are relatively small, ranging between 0.3% and 15%



(a) for values $T = 2, 4, 6, \ldots, 20$ (b) for values $I = 2, 3, \ldots, 8$

Figure 5.1: Mean relative suboptimality gap of heuristic MPI–OPT.



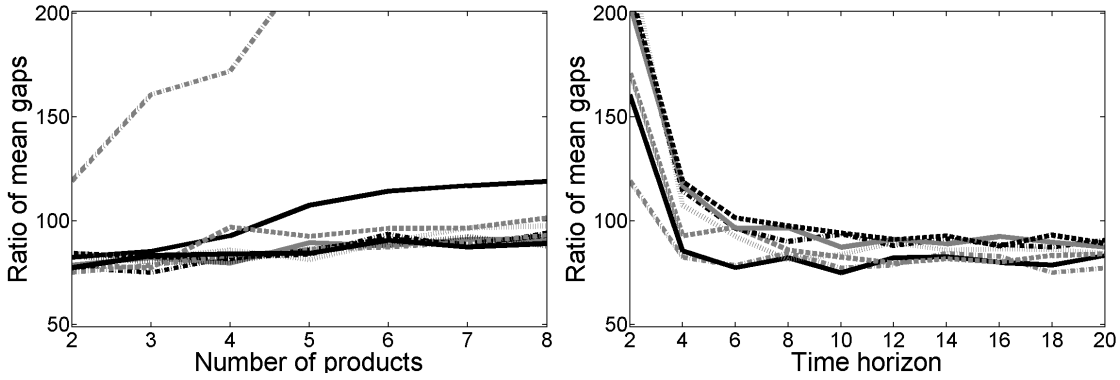
(a) for values $T = 2, 4, 6, \ldots, 20$ (b) for values $I = 2, 3, \ldots, 8$

Figure 5.2: Mean adjusted relative suboptimality gap of heuristic MPI–OPT.

(4% and 14%). The maximum rsg (arsg) values of Heuristic MPI–GRE range between 1% and 8% (22% and 72%), that is, its worst-case performance is good in absolute terms, but is especially bad in the problems where promotion has small effect on total revenues. The worst-case performance of Heuristic EDF–GRE ranges between 3% and 8% (51% and 100%).

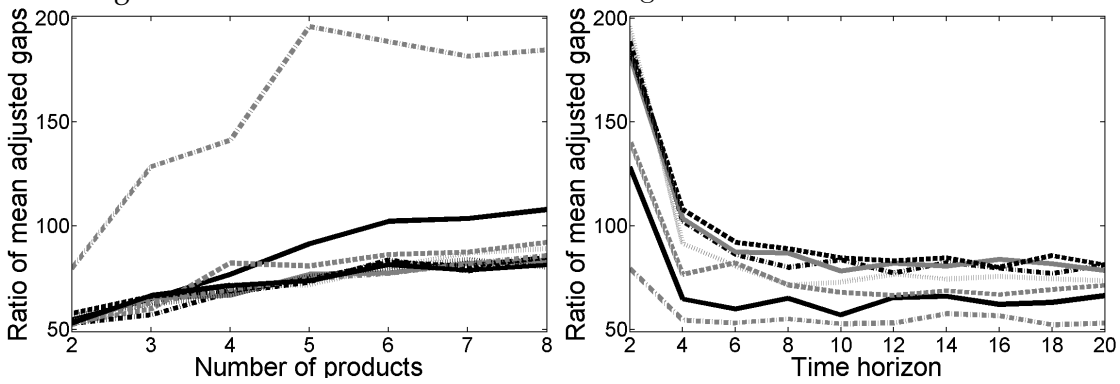
5.8 Conclusions

We have developed a dynamic and stochastic model of dynamic promotion and proposed a policy that has a natural economic interpretation and suggests itself to be easily implementable in practice. These advantages come at the cost of possible suboptimality of such a dynamic solution, which was, however, shown to be negligible and smaller than the cost of implementing a naïve marketing solution. The model has an appealing property of being extensible to a variety of ad-hoc requirements that managers or



(a) for values $T = 2, 4, 6, \ldots, 20$ (b) for values $I = 2, 3, \ldots, 8$

Figure 5.3: Performance ratio in terms of rsg of EDF–GRE over MPI–OPT.



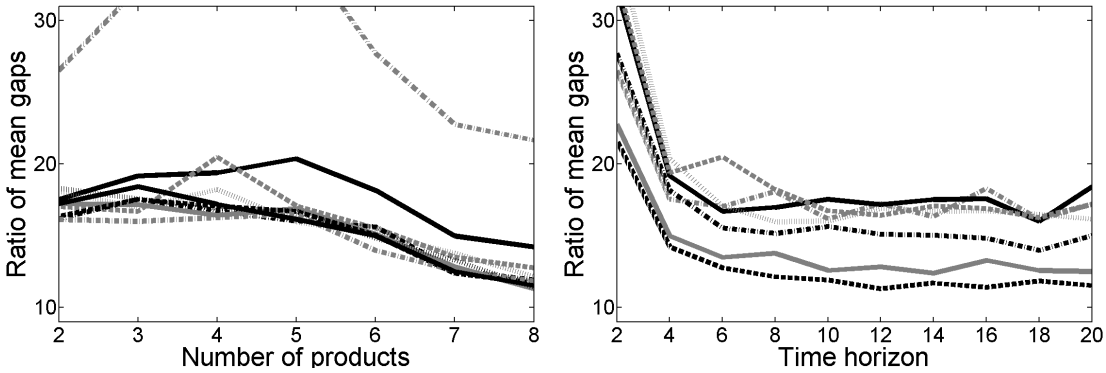
(a) for values $T = 2, 4, 6, \ldots, 20$ (b) for values $I = 2, 3, \ldots, 8$

Figure 5.4: Performance ratio in terms of arsg of EDF–GRE over MPI–OPT.

certain circumstances may impose.

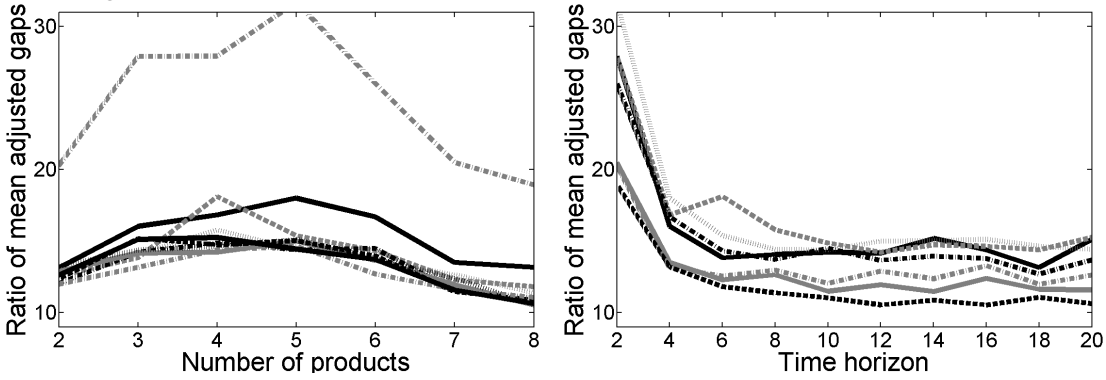
A challenge showing itself is to extend the model presented in this chapter to account for price changes, inventories with dependent demands and product assortment, and obtain an appealing index-based solution. The analysis of that problem is, however, more complex and the theoretical background must be extended in order to tackle such problems.

Our model offers a comprehensive modeling framework that may be used in other applications, since the items considered in knapsack problems are often perishable, either naturally or due to special restrictions. An application, for example, arises in surgery, when only a limited number of patients may be chosen to undertake an alternative treatment (e.g., a transplantation). Further, the task management problem, in which tasks have associated deadlines and one can work only on a subset of them at a time, also falls to the general KPPI setting.



(a) for values $T = 2, 4, 6, \ldots, 20$ (b) for values $I = 2, 3, \ldots, 8$

Figure 5.5: Performance ratio in terms of rsg of MPI–GRE over MPI–OPT.



(a) for values $T = 2, 4, 6, \ldots, 20$ (b) for values $I = 2, 3, \ldots, 8$

Figure 5.6: Performance ratio in terms of arsg of MPI–GRE over MPI–OPT.

Sometimes everybody hurts.
< R.E.M. >

Chapter 6

Congestion Control in Routers with Future-Path Information

6.1 Introduction

With the growth of traffic volume and traffic heterogeneity in best-effort networks, congestion control has been getting more importance. The queue tail drop policy showed to be prone to creating various serious problems including bias against bursty traffic and global synchronization, eventually resulting in congestion collapse (cf. Braden et al., 1998). Such a *reactive* congestion control has a significant negative impact on the efficiency of scarce resources (bandwidth and buffer space) allocation in networks.

Alternative proposals focused on *preventive* congestion control developing *congestion avoidance mechanisms*, such as RED (Floyd and Jacobson, 1993), BLUE (Feng et al., 2002), and a palette of their variants, which try to detect local congestion in its early stage and warn by random packet dropping the traffic sources expecting that they decrease their transmission rates.

Nevertheless, packet losses in the Internet are still high and Quality of Service strongly suffers from this fact. Packets drops will persist even in the networks where *explicit congestion notification* (ECN) is deployed. Explicitly marked (instead of dropped) packets should be understood by users as a warning about possible dropping of packets in a very near future. However, since users are let to decide how to react, non-cooperative flows exist, which together with the bursty nature of traffic are the main causes of packet drops.

This chapter deals with the congestion control at routers with future-path information. By congestion control problem at router with future-path information we mean to implement particular control methods using the actual network congestion information

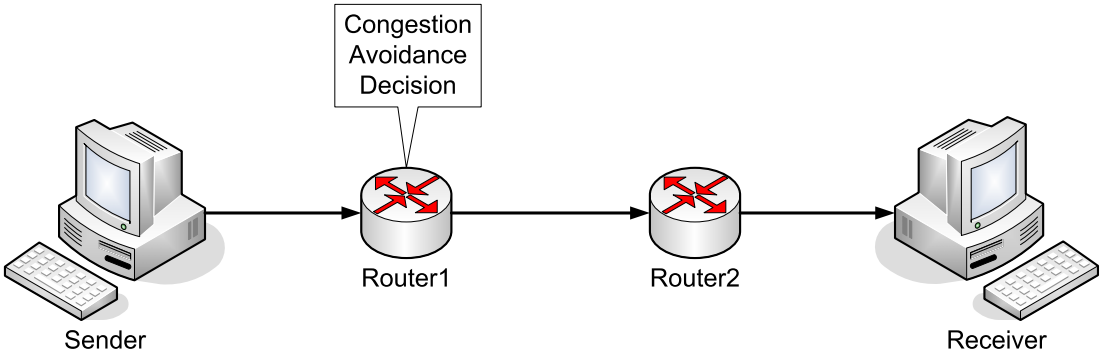


Figure 6.1: A design of an end-to-end connection.

to reduce overall congestion for the route. For instance, in this work we specifically deal with dropping policy, intelligent marking, and admission control.

Our setting differs from the preventive congestion control mainly by trying to exploit available information in order to know the whole network congestion state, not only the congestion at the node where this control is implemented. Gathering of such information inside the network may be costly, however, increasing ECN deployment improves the possibilities for routers to have fresh news from other network nodes (see, e.g., Molle and Xu, 2005).

More information at network nodes may result in a more efficient resource allocation. Notice that the information about network congestion gathered by intermediate nodes experiences a significantly lower delay than the one users obtain from packet acknowledgements. Similarly, the delay in the response to congestion is also lower for nodes than for users. Thus, the ECN may not only decrease packet losses directly by marking instead of dropping packets, but may also provide useful information for a more efficient resource allocation at network nodes.

As Floyd and Fall (1999) pointed out, traffic lacking end-to-end congestion control may cause congested links sending packets that will only be dropped later in the network. Since dropping a packet on its route implies that all the scarce resources it has consumed so far are wasted, congestion control that uses future-path information may be highly valuable for the network performance in periods of congestion. It is then intuitively appealing that when a scarce resource is to be allocated to a packet, the possibility of getting that packet lost in the remainder of its route should be taken into account.

To illustrate the idea on a simple example, consider an end-to-end connection that includes two bottleneck routers, as in Figure 6.1. If Router2 is busy (yet still has some free capacity in the buffer) and Router1 is able to anticipate it, then congestion avoidance decisions at Router1 should take into account the transmission rate of an incoming

flow. If the rate is small, so that Router2 would be able to transmit it, the flow should be transmitted at Router1. On the other hand, if the transmission rate is too high, so that Router2 is very likely to drop it, the flow should be dropped already at Router1; or, it could be a strong candidate for the congestion warning policy implemented in the congestion avoidance mechanism at Router1.

Exploiting the network congestion information leads to a novel concept that we coin *network-capability fairness*. Put simply, flows are treated fairly according to what the network *can* transmit, and not according to what the flows *want* to transmit as it is usually assumed in the existing concepts of fairness. For instance, we put in doubt whether two flows arriving at the same rate should be treated equally if we know that one of them is routed to a congested link and the other one to a congestion-free link. The network-capability fairness arises by making the routers maximize the expected time-average network goodput, which relies on flow's future-path information.

We model the congestion control problem in the framework of Markov decision processes (MDPs), which leads to a formulation as the multi-armed restless bandit problem with an additional feature of random arrivals. After decomposing the problem into single-flow subproblems, we deploy the restless bandit indexation methodology. As we show in this chapter, the network-capability fairness can be achieved by implementing *transmission indices* which evaluate the usefulness of flow transmission as a function of its current transmission rate and current network congestion state. The transmission index is the marginal productivity (MP) index arising in the context of this problem.

We consider a bottleneck router with a scarce resource that is given by the bandwidth available, for which several flows compete. Each flow generates certain *goodput* reward for its receiver, if it is delivered, which can be achieved only if it is transmitted by the router. The difficulty is that these flows are dynamically changing their transmission rate, so the rewards may increase or decrease over time. Thus, the question is whether to exploit the present rewards by transmitting at the arrival rate, or to take a locally-suboptimal action which may yield higher rewards further downstream or to the following packets arriving at the router.

In this work we assume that routers have estimates on congestion probabilities on downstream links, i.e., on the future paths of all the flows, and that these probabilities are available for every possible sending rate of each flow. Continuous gathering of such information would be desirable, but it is hardly implementable due to overwhelming amount of data transmission and data processing by the routers. Thus, we assume that this information is gathered in certain time intervals that are bigger than the existence of flows (for instance, once per hour). This allows us to assume that each flow finds these congestion probabilities constant.

6.1.1 Related Congestion Control Protocols

In this chapter we present a stochastic model of a single bottleneck router at which multiple flows compete for available bandwidth. Models simpler in network topology assumptions allow to incorporate more complex and more realistic dynamics. For instance, such a problem was studied in order to design and develop congestion control schemes obeying max-min fairness, which is hard to analyze for general network topologies. A deterministic fluid-flows model of a single bottleneck router led to several congestion control protocol proposals, including the Explicit Control Protocol in Katabi et al. (2002), the Rate Control Protocol in Dukkipati et al. (2005), and the Adaptive Control Protocol in Lestas et al. (2008). They, however, differ in the packet-level implementation significantly, as we outline next.

Explicit Control Protocol (XCP). XCP was designed to work well in networks with high bandwidth-delay products. This protocol assumes that users fill in the header of each packet their current sending rate (i.e., the current TCP window) and their current estimate of the round-trip time (RTT) for a given flow, which provides a *state* information for the routers. Moreover, the packet headers carry a feedback-rate information on a possible increase (positive or negative) in the current sending rate which is initially set by the sender to its desired increase, and then modified by the routers encountered on the packet's path. At every packet acknowledgment the users are supposed to modify their sending rate as indicated by the feedback-rate information, which carries the minimum of all the routers' required or allowed increases in the sending rate. Each XCP router maintains a per-link estimation of average RTT to control the feedback delay. Note that XCP does not drop packets, since it operates on top of a router's dropping policy such as tail dropping or implicit congestion notification (e.g., RED).

Adaptive Control Protocol (ACP). This protocol assumes that users fill in the header of each packet their current estimate of the round-trip time (RTT) for a given flow (as in XCP), and that the header carries a feedback-rate information on a desired sending rate modified by the routers on the path. Moreover, it assumes that there is an ECN bit set to 1 by each of the routers on the path, whose current input rate is above 95% of the router's bandwidth capacity. The users are supposed to smoothly (and even less aggressively if the ECN bit equals 1) modify their sending rate after every packet acknowledgment as indicated by the feedback-rate information, which carries the minimum of all the routers' required sending rate. As in XCP, each ACP router maintains a per-link estimation of average RTT to control the feedback delay. It was shown by simulations that ACP corrects the problem of XCP in achieving max-min fairness in the

presence of multiple bottleneck routers.

Rate Control Protocol (RCP). This simple protocol tries to minimize the duration of flows by emulating processor sharing. It assigns an equal input rate to all the flows, i.e., it admits the same amount of data from each flow. As in ACP, it is assumed that users fill in the header of each packet their current estimate of the round-trip time (RTT) for a given flow, and that the header carries a feedback-rate information on a desired sending rate modified by the routers on the path. The users are supposed to modify their sending rate after every packet acknowledgment as indicated by the feedback-rate information. This policy is shown to improve over TCP and XCP in the settings when new flows arrive randomly and are finite-length, since the latter two protocols make adapt the users' sending rates over several RTTs, which works well only when all flows are long-lived. Note that the RCP router does not keep flows-state and does no per-packet calculations.

6.1.2 Congestion Control Problem of Multiple Flows at Bottleneck Router

In this subsection we present a general formulation of the congestion control problem. While earlier formulations were based on deterministic fluid models, we develop here a stochastic extension using the MDP framework.

Consider the time slotted into discrete time epochs $t = 0, 1, 2, \dots$ Let us denote the router parameters by

- W the bandwidth, i.e., the deterministic "server capacity" (in packets per period);
- $\overline{W}$ the target time-average router throughput, or "virtual capacity" (in packets per period); $\overline{W} < W$;
- B the buffer size, possibly infinite (in packets); $B \geq W$;
- $B(t)$ the backlog process (in packets) at epochs t.

Suppose further that flow $m \in \mathcal{M} := \{1, 2, \dots\}$ appears at a random time epoch T_m in the (deterministic) initial state n_m and that its transmission lasts for a random number of periods, given by the probability $0 < 1 - \beta_m < 1$ that the flow is terminated by the sender before the next time epoch. That is, the length of the flow-m existence (i.e., the flow's time between starting and terminating) follows a geometric distribution with mean $1/(1-\beta_m)$. Denote by $\mathcal{M}^{\text{started}}(t) \subset \mathcal{M}$ the random process of the set of flows that have started by time epoch t. The distributions of T_m can be arbitrary; we only require

that the number of flows that have started, $|\mathcal{M}^{\text{started}}(t)|$, grows linearly with t, i.e.,

$$0 < L := \lim_{t\to\infty} \frac{|\mathcal{M}^{\text{started}}(t)|}{t} < \infty. \tag{6.1}$$

Denote by $\mathcal{M}(t) \subset \mathcal{M}$ the random process of the set of existing (i.e., started and not terminated) flows at time epochs t. The above linearity condition assures that $\mathbb{P}[T_m < t \text{ for all } m \in \mathcal{M}] < 1$ for any $t = 0, 1, \ldots$, and therefore $\mathcal{M}(t) \subset \mathcal{M}^{\text{started}}(t)$ is finite at all epochs. Then, we define the following parameters for any existing flow $m \in \mathcal{M}(t)$:

- N_m the number of states of flow m; $\mathcal{N}_m := \{0, 1, \ldots, N_m - 1\}$;
- $A_m := |\mathcal{A}_m|$ the number of available actions by the router for flow m;
- $X_m(t) \in \mathcal{N}_m$ the state process of flow m at epochs t;
- $a_m(t) \in \mathcal{A}_m$ the action process of flow m at epochs t;
- $W^{\text{sent}}_{m,n}$ the workload (in number of packets) sent by the sender of flow m at state n;
- $W^{a}_{m,n}$ the one-period expected bandwidth used (in number of packets) of flow m at state n if action a is applied at the router;
- $R^{a}_{m,n}$ the one-period expected goodput (or reward) of flow m at state n if action a is applied at the router.

The flows dynamics is as follows (see Figure 6.2). At epoch t, the sender of each existing flow $m \in \mathcal{M}(t)$ sets its state $X_m(t)$ (that depends on whether the previous-epoch workload was transmitted without congestion, given by the complete acknowledgements of the receiver of flow m sent back to the sender in the previous period) and sends the workload of $W^{\text{sent}}_{m,X_m(t)}$ packets to the bottleneck router. However, only $0 \le W^{a_m(t)}_{m,X_m(t)} \le W^{\text{sent}}_{m,X_m(t)}$ packets are allowed to queue in the buffer for being transmitted. The transmitted packets with possible losses arrive to the receiver of flow m, who obtains the goodput (or reward) $R^{a_m(t)}_{m,X_m(t)}$. If the router transmitted the flow without any congestion warning and if there was no congestion downstream, then the receiver sends complete acknowledgements back to the sender; otherwise only partial acknowledgments are sent. Next, with probability $1 - \beta_m$ the flow m terminates so that there are no more packets sent by the sender in the future time epochs. This may well be due to finalizing the file transmission (i.e., sending all the planned packets), an impatience of the sender, or external factors such as broken connections (see, e.g., Massoulié and Roberts, 1999). If not terminated, then the sender sets its next-epoch state $X_m(t+1)$ and repeats the process.

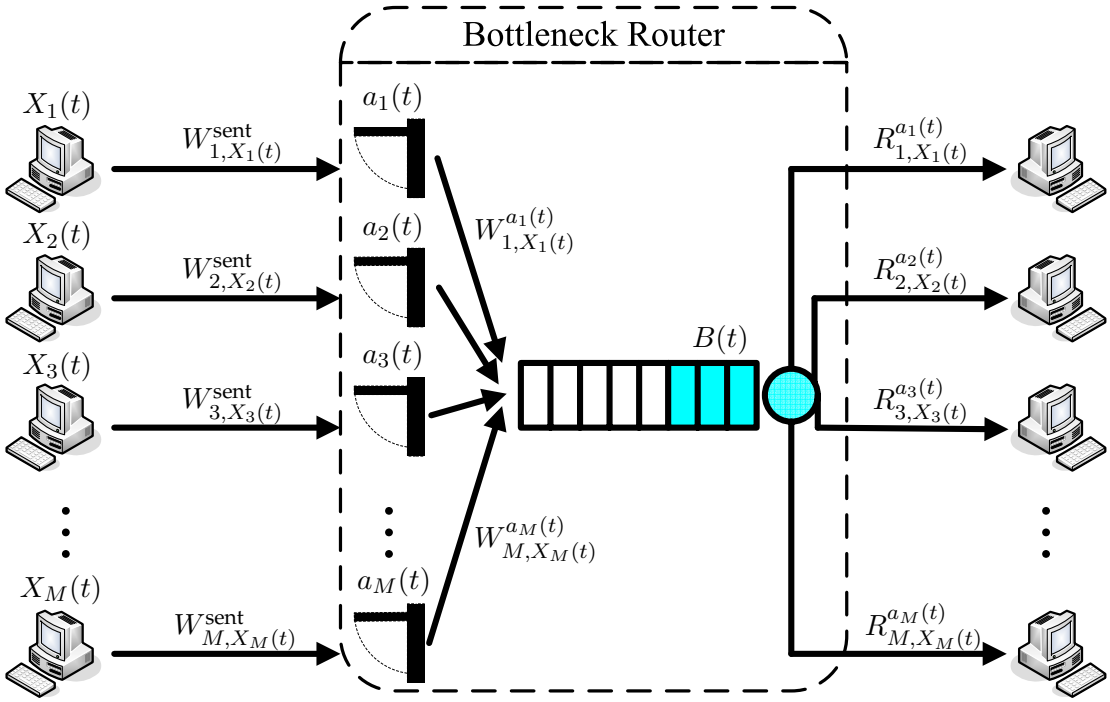


Figure 6.2: A scheme of $M := |\mathcal{M}(t)|$ flows sharing a bottleneck router.

The congestion avoidance decision at the router is taken in the following way. At epoch t the router controller observes the backlog $B(t)$ and the flows states $X_m(t)$ of all existing flows $m \in \mathcal{M}(t)$. Based on that she decides the flow actions $a_m(t)$ (which may be viewed to be taken in virtual gates, as illustrated in Figure 6.2), instantaneously appends (in FIFO order) $W^{a_m(t)}_{m,X_m(t)}$ packets of each flow m to the buffer, and transmits (in FIFO order) W packets (or all the packets if there are less than W packets in the buffer) during the period. Thus, at the next epoch there is the backlog

$$B(t+1) := \max\left\{B(t) + \sum_{m\in\mathcal{M}(t)} W^{a_m(t)}_{m,X_m(t)} - W; 0\right\}.$$

The above description implies that $B(t) + \sum_{m\in\mathcal{M}(t)} W^{a_m(t)}_{m,X_m(t)} \le B$, so we have $B(t) \le B - W$ at all epochs t.

To summarize, the senders make no decisions and therefore the flows dynamics can be modeled as a Markov chain. Section 6.3 presents a binary-action MDP (i.e., a restless bandit) model for the router-based control of a single flow, where the dynamics and the parameters are defined in more detail. At this moment we present a generic formulation of the congestion control problem.

Let Π be the set of all history-dependent randomized policies. Denote by the symbol $\mathbb{E}^{\pi}_{\boldsymbol{n},B_0}$ the conditional expectation given that the initial conditions are $\boldsymbol{n} := (n_m)_{m\in\mathcal{M}}$, $B_0 = B(0) \leq B - W$ and the policy applied is $\pi \in \Pi$. The router controller's problem to solve under the time-average criterion (which is well-defined due to bounded operands) is

$$\max_{\pi\in\Pi} \lim_{T\to\infty} \frac{1}{T} \mathbb{E}^{\pi}_{\boldsymbol{n},B_0} \left[\sum_{t=0}^{T-1} \sum_{m\in\mathcal{M}(t)} R^{a_m(t)}_{m,X_m(t)} \right] \tag{6.2}$$

$$\text{subject to} \quad \lim_{T\to\infty} \frac{1}{T} \mathbb{E}^{\pi}_{\boldsymbol{n},B_0} \left[\sum_{t=0}^{T-1} \sum_{m\in\mathcal{M}(t)} W^{a_m(t)}_{m,X_m(t)} \right] \leq \overline{W} \tag{6.3}$$

$$B(t) + \sum_{m\in\mathcal{M}(t)} W^{a_m(t)}_{m,X_m(t)} \leq B, \text{ for all } t = 0, 1, 2, \ldots \tag{6.4}$$

The virtual capacity seen as the target time-average router throughput $\overline{W}$ is in fact a delay- and stability-controlling parameter, since the higher the $\overline{W}$, the higher the probability of non-zero backlogs $B(t)$, and therefore the higher the router's time-average contribution to end-to-end propagation delays of the flows. We note that an analogous constraint (6.3) formulation was used in Ma et al. (2008). To avoid the trivial problem of underloaded router, we assume that

$$\lim_{T\to\infty} \frac{1}{T} \mathbb{E}^{\pi}_{\boldsymbol{n},B_0} \left[\sum_{t=0}^{T-1} \sum_{m\in\mathcal{M}(t)} W^{\text{sent}}_{m,X_m(t)} \right] > \overline{W}. \tag{6.5}$$

6.1.3 Related Models

Closely related to our work is a wide stream of literature on network economics and pricing (see, e.g., Courcoubetis and Weber, 2003) aiming at improving network resource allocation. Based on an auction model, MacKie-Mason and Varian (1995) proposed that each user sets a *bid* in the packet headers. When such a packet arrives to a router, it is accepted if and only if its bid is higher than an actual congestion threshold. If the packet is blocked, the packet with an increased bid can be retransmitted by the user, or it must wait until the threshold decreases. This however, leads to complex practical problems of setting the bids and therefore such threshold policies are difficult to implement in practice (see Shenker et al., 1996). In our model, such a bid can be seen as given by the flow transmission rate (the higher the rate, the smaller the bid) and clearly having analogous consequences. Nevertheless, we do not restrict ourselves to such strict accepting/blocking decisions, yet let the routers self-calculate the *price* of the

entire flow based on the bid (transmission rate) and the current congestion information.

Kelly (1997) developed decompositions of a deterministic network model, in which a Lagrange multiplier mediates between subproblems and leads to a social optimum. He further explained how different fairness criteria arise from different utility functions. However, Stidham (2004) showed that if users are heterogeneous, such models may be difficult to solve to global optimality due to the existence of various local optima. The book by Srikant (2004) and the surveys in Lestas et al. (2008); Low and Srikant (2004); Low et al. (2002); Low and Lapsley (1999) provide summaries of the most important mathematical models and results on congestion control for whole networks.

The above papers are alike in optimizing the sum of user's utility functions, and showing that such a problem can be decomposed into per-user problems of setting transmission rates in an adaptive (reactive to network congestion signals) manner. The typical result is that a particular type of fairness (such as proportional fairness, max-min fairness, etc.) is achieved if all the users have utility functions of same type. Moreover, the flows are assumed to be persistent and their number be constant. However, in the current and future Internet these are not appropriate assumptions, due to variability of traffic flow types, such as FTP, VoIP, video, mice etc., that appear randomly and have a finite duration. Such assumptions are dropped in this work.

Continuous information gathering at network nodes were considered for congestion control problems in Neely et al. (2008), Paganini (2006) and Molle and Xu (2005). They showed that communication across network nodes, especially the neighboring ones, is beneficial and the latter two discussed also practical implementation of such information gathering. It is crucial for our approach to the congestion control problem to assume that routers have certain information about actual congestion downstream. However, as noted at the beginning of this chapter, we assume that this information is gathered in certain time intervals (and not continuously) that are bigger than the flows lifetimes (for instance, once per hour). This greatly reduces the implementability problem of overwhelming amount of data transmission and data processing by the routers.

6.1.4 Goals and Contributions

We do not use utility functions existing in the previous work (though the model is general enough to incorporate them), nor we want to punish users who experience congestion by charging them in monetary terms. Rather, we focus on finding simple nearly-optimal network strategies that make the network truly *best-effort*: transmitting all the packets that can be transmitted, so that the expected goodput be maximized. Since the expected goodput can be viewed as a particular utility function, it gives rise to a particular fairness criterion, which we term *network-capability fairness*. We will show that it

Multi-Armed Restless Bandit Problem		*Congestion Control Problem*
bandits	$\longleftrightarrow$	flows
server	$\longleftrightarrow$	router's bandwidth
work	$\longleftrightarrow$	flow transmission
reward	$\longleftrightarrow$	flow delivery
marginal productivity index	$\longleftrightarrow$	transmission index

Table 6.1: Analogy between the multi-armed restless bandit problem and the congestion control problem.

possesses several desirable properties. Moreover, it is analytically fruitful, since the expected goodput does not suffer from non-concavity of some utilities functions described in Shenker et al. (1996).

Our multiple-flow problem formulation targets the trade-off between throughput and delay in networks with finite-length flows. Given its complexity, the problem is relaxed and decomposed in Section 6.2 so that individual flows are treated in isolation, building on an analogy with the multi-armed restless bandit problem introduced by Whittle (1988) (see Table 6.1). Moreover, we show in Section 6.3 that the decisions upon dropping and marking of flow packets can be transformed into a flow admission control problem. This decomposition thus shows that congestion control at a router can be decomposed into a collection of per-flow admission control problems. Similar observations were made in Ferragut and Paganini (2008) and Paganini (2006), who studied stability of the problem using a classic fluid-flow model. They illustrated that properly designed admission control can be shown to be sufficient for both an efficient congestion control and an efficient routing.

Section 6.3 presents an MDP (restless bandit) model of congestion control of an individual flow. We build on an improved idea of Wischik (1999), who concluded that a fair marking of a flow (in explicit congestion avoidance mechanisms) should reflect (i) how much of the capacity it uses, and (ii) the congestion state at the router. We extend the latter to (ii') the congestion state at the router *and the remainder of the route*. We believe that such a modification may significantly improve the efficiency of resource allocation across the network.

Section 6.4 employs the restless bandit indexation methodology surveyed in Niño-Mora (2007b) to obtain an optimal solution for the congestion control of an individual flow via the *marginal productivity index*, called the transmission index in our context. This index identifies locally optimal actions for the decentralized problem, and captures the value of network services to users. If multiplied by the packet size, it can be

seen as a list of certain internal *prices*, and can be implemented in congestion avoidance mechanisms in order to resolve the fairness problem between different flows.

We use these general results in Section 6.5 to obtain closed-form expressions of the transmission index under three basic router variants with the following network congestion control functions:

(i) *TD router*: congestion control based on tail dropping (buffer overflow),

(ii) *ICN router*: congestion avoidance with implicit congestion notification (packet dropping), and

(iii) *ECN router*: congestion avoidance with explicit congestion notification (packet marking).

Section 6.6 presents two simple settings in which the transmission index defines an optimal transmission priority policy for a multiple-flow problem at a bottleneck router.

Finally, Section 6.7 discusses heuristic proposals for the implementation of the transmission index in existing congestion avoidance mechanisms in order to improve the performance of the whole network. Our proposal is very flexible and a uniform implementation across the network is not necessary.

6.2 Decomposition of the Multiple-Flows Problem

The problem (6.2)–(6.4) is difficult to solve due to the sample path constraint (6.4). One possibility for relaxing the problem is to assume that the buffer space B is infinite, so that the constraint (6.4) is trivially fulfilled. Another possibility is to relax that constraint as did Whittle (1988), by requiring it only *on time-average*, i.e.,

$$\lim_{T\to\infty} \frac{1}{T}\, \mathbb{E}^{\pi}_{\boldsymbol{n},B_0} \left[\sum_{t=0}^{T-1} \sum_{m\in\mathcal{M}(t)} W^{a_m(t)}_{m,X_m(t)} \right] + B^{\pi}_{\boldsymbol{n},B_0} \leq B,$$

where $B^{\pi}_{\boldsymbol{n},B_0}$ is the time-average of the backlog process $B(t)$ under policy π and initial conditions $\boldsymbol{n}, B_0$. However, such a constraint is weaker than (6.3), because $B - B^{\pi}_{\boldsymbol{n},B_0} \geq W > \overline{W}$ under any $\pi, \boldsymbol{n}, B_0$.

Either of these two relaxation possibilities results in omitting the constraint (6.4). Note also that the initial backlog B_0 is then irrelevant and therefore can also be omitted. Thus, we end up with a problem formulation (6.2)–(6.3), which is precisely the Whittle relaxation of the multi-armed restless bandit problem (Whittle, 1988), except for our generalization into time-variant number of flows.

The standard solution of such a formulation is by solving the Lagrangian relaxation of (6.2)–(6.3), which is

$$\max_{\pi\in\Pi}\lim_{T\to\infty}\frac{1}{T}\,\mathbb{E}^{\pi}_{\boldsymbol{n}}\left[\sum_{t=0}^{T-1}\sum_{m\in\mathcal{M}(t)}\left(R^{a_m(t)}_{m,X_m(t)}-\nu W^{a_m(t)}_{m,X_m(t)}\right)\right]+\nu\overline{W}, \tag{6.6}$$

where ν is the Lagrangian parameter that can be interpreted as a per-packet *transmission cost*. The Lagrangian theory assures that there exists ν^*, for which the Lagrangian relaxation (6.6) achieves optimum of (6.2)–(6.3). Since for any fixed ν the flows are independent, we can decompose (6.6) into an infinite number of individual-flow problems, as detailed in the following.[1]

Proposition 6.1. *Let Π_m be the set of all history-dependent randomized policies for flow m, and individual-flow policies $\pi^*_m \in \Pi_m$ such that they form the joint policy $\pi^* \in \Pi$. If for a given parameter ν, each policy π^*_m for $m \in \mathcal{M}$ optimizes the individual-flow problem*

$$\max_{\pi_m\in\Pi_m}\mathbb{E}^{\pi_m}_{n_m}\left[\sum_{t=0}^{\infty}\beta^t_m\left(R^{a_m(t)}_{m,X_m(t)}-\nu W^{a_m(t)}_{m,X_m(t)}\right)\right], \tag{6.7}$$

then π^* optimizes the multi-flow problem (6.6).

In Section 6.3 we will find under certain natural conditions an optimal solution to such a ν-parameter problem (if $A_m = 2$ for all $m \in \mathcal{M}$) in terms of flow- and state-dependent *marginal productivity indices* $\nu_{m,n}$, which in our setting can be interpreted as *transmission indices*. If the optimal transmission cost ν^* is known, then these indices define the following optimal policy for problem (6.2)–(6.3): "At each time epoch transmit all the flows of actual-state transmission index greater than the transmission cost ν^* and warn the remaining flows".

Since in practice ν^* is typically unknown, the buffer space is finite, and it is desirable to have work-conserving transmission in order to increase bandwidth utilization, we will use the transmission indices to define practically feasible and desirable heuristical policies in Section 6.7.

6.3 Individual-Flow Congestion Control Problem

In this section we consider an individual flow requiring router resources (buffer space and bandwidth capacity). We present an MDP model for flows that behave under the any-increase/multiplicative-decrease policy (such as TCP connections). A similar

[1] I am grateful to Bernardo D'Auria for helping to clear the proof of this proposition.

model could be developed for other flow types, but these have been left out of the scope of this work. Before modeling a flow with multiplicative decrease, we develop an MDP model of a flow with any-increase/restart behavior, henceforth called the *restarting flow*.

We consider that router is of one out of three variants described in Section 6.1: TD router, ICN router, ECN router. In order to model these router variants, we assume that the router has two available actions to choose from to control an individual flow. One of these actions is *transmitting the flow without warning* in all the variants. The other action depends on the router variant, and it is: (i) blocking the flow for TD router, (ii) dropping some packets of the flow for ICN router, and (iii) marking some packets of the flow for ECN router.

In order to define the problem of congestion control of an individual flow, we will use information about the router variant (to know what to decide upon), the current flow transmission rate (to know the flow's bandwidth requirements), and the probability of losses downstream the flow's route (to calculate flow's goodput). In the following subsections we present a formal model of the restarting flow and define the problem. An optimal congestion control via the transmission index is derived in Section 6.4. Since we focus on a single flow, we drop the flow subscript m.

6.3.1 Markov Decision Process Model of Restarting Flow

The restarting flow in every period increases its transmission rate defined by the *actualWindow* variable by a discrete increment up to a certain *maximumWindow* constant unless the flow sender is warned about congestion, when it restarts by setting the transmission rate to a certain *minimumWindow* constant. We assume that the constants $0 < \mathit{minimumWindow} \leq \mathit{maximumWindow}$ are given in advance, as they are typically set by the users' operation systems. We set the MDP model in discrete time, defining one time period as one round-trip time (RTT). We assume that all packets are of the same size, which we further define to be one bandwidth capacity unit.

The states $n \in \mathcal{N} := \{0, 1, \ldots, N-1\}$ ($N \geq 1$) denote possible levels of the sending rate, i.e., of the *actualWindow* variable. The 0-th state represents $\mathit{actualWindow} = \mathit{minimumWindow}$, and the $(N-1)$-th state represents $\mathit{actualWindow} = \mathit{maximumWindow}$. The *actualWindow* variable in state n assumes the value W_n^{sent} (in packets/RTT), which can therefore be interpreted as the bandwidth capacity the flow requires for complete transmission at the current period. Hence, in the following we assume that $0 < W_0^{\text{sent}} := \mathit{minimumWindow} < W_1^{\text{sent}} < \cdots < W_{N-2}^{\text{sent}} < W_{N-1}^{\text{sent}} := \mathit{maximumWindow}$. The schematic behavior of the restarting flow as a Markov chain is shown in Figure 6.3, where "OK" represents a congestion-free reception of the flow by the receiver (complete acknowledgments) and "NO" represents a congestion-experienced transmission (incomplete ac-

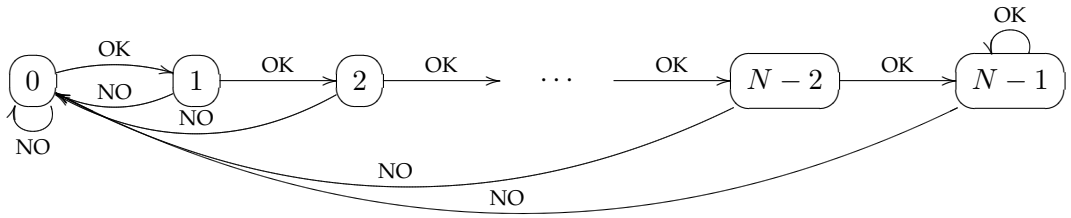


Figure 6.3: A model of the restarting flow as a Markov chain. The arrows represent one-period transitions among the states $0, 1, \ldots, N-1$ after a congestion-free (OK) and a congestion-experienced (NO) transmission.

knowledgments).

While no decisions are taken by the sender, congestion control decisions must be taken by the routers. At a particular router on the connection route we want to decide whether the incoming flow in state n should be transmitted without any congestion warning (achieved by action $a(t) = 1$ of transmitting $W_n^1 := W_n^{\text{sent}}$ packets), or warned by employing a congestion control function (action $a(t) = 0$ of transmitting $0 \leq W_n^0 \leq W_n^{\text{sent}}$ packets) depending on the router variant.

Thus, the parameter W_n^0 gives us the flexibility to consider router variants with different congestion warning. In particular, the warning action corresponds to

(i) *blocking* of entire flow ($W_n^0 := 0$) in TD router,

(ii) *dropping* some of the packets ($0 < W_n^0 < W_n^{\text{sent}}$) in ICN router, or

(iii) *marking* some of the packets ($W_n^0 := W_n^{\text{sent}}$) in ECN router.

If the flow is warned at the router, the flow restarts to state 0. Notice, however, that if the flow is transmitted without warning, then the flow being in state n restarts to state 0 with the probability $0 \leq 1 - p_n < 1$, reflecting the probability of experiencing congestion in the remainder of its route. On the other hand, with the probability p_n of congestion-free transmission it increases the transmission rate from W_n^{sent} to W_{n+1}^{sent} (or remains at the *maximumWindow* rate if being in state $N-1$). The assumption $p_n \neq 1$ is what makes our model suitable for exploiting the future-path congestion information in the congestion control problem.

Given the actions interpretation given above, R_n^1 is the expected one-period goodput (receiver reward) from a transmitted flow and R_n^0 is the expected one-period goodput from a warned flow. We will find it convenient to further decompose R_n^1 into the congestion-free reward R_n^{1+} and the congestion-experienced reward R_n^{1-}, so that $R_n^1 := p_n R_n^{1+} + (1 - p_n) R_n^{1-}$. Thus, R_n^0 and R_n^1 can capture the receiver sensitivity to

dropped or marked packets, and should also depend on W_n^0 and W_n^1, respectively. The difference of receiver rewards and transmission costs (i.e., $R_n^0 - \nu W_n^0$ and $R_n^1 - \nu W_n^1$) will be henceforth called the *net reward* under transmission cost ν.

In summary, our MDP model for the router-based control of the restarting flow is defined as follows:

- *State space* is $\mathcal{N} := \{0, 1, \ldots, N-1\}$.
- *Actions*: active action (*transmitting*) and passive action (*warning*) are available in each state.
- *Dynamics if active*: If the flow is in state n and the flow is transmitted at a given period, then during that period
  - with probability $p_n > 0$: it generates net reward $R_n^{1+} - \nu W_n^1$ and the flow moves to state $n+1$ for the next period (or remains in $N-1$, if $n = N-1$).
  - with probability $1 - p_n$: it generates net reward $R_n^{1-} - \nu W_n^1$ and the flow moves to state 0 for the next period.
- *Dynamics if passive*: If the flow is in state n and the flow is warned at a given period, then during that period it generates net reward $R_n^0 - \nu W_n^0$ and the flow moves to state 0 for the next period.

6.3.2 The Three Router Variants: Definitions

Since $W_n^1 = W_n^{\text{sent}}$ is the number of packets sent if the flow is in state n (i.e., the actual flow transmission rate), the congestion-free one-period receiver reward is $R_n^{1+} := W_n^{\text{sent}}$ in all the mechanisms we consider below.

We suppose that the non-warned flow continuously increases the transmission rate, $W_0^{\text{sent}} < W_1^{\text{sent}} < \cdots < W_{N-1}^{\text{sent}}$, and that the probabilities of congestion-free transmission downstream, p_n's, are nonincreasing in n. We narrow our focus to networks in which the congestion downstream the route is treated in the same way as at the router where our policy is implemented.

The three mechanisms below differ by defining the congestion-warned transmission rate W_n^0, the expected one-period receiver reward of the congestion experienced downstream R_n^{1-}, and the congestion-warned expected one-period receiver reward R_n^0.

In the light of the model in Section 6.3, we assume that the flow reacts by restarting to the minimal transmission rate $W_0^1 = 1$ in all the mechanisms.

TD Router. The action 0 at the router refers to *blocking* the entire flow, i.e., $W_n^0 = R_n^0 := 0$. Naturally, the reward from the blocked flow is $R_n^{1-} := 0$.

ICN Router. While in practice there are many ways of implicit congestion notification, as an illustration we focus on a simple one. Suppose that the action 0 at the router refers to *dropping one packet* of the flow, i.e., $W_n^0 := W_n^1 - 1$. Therefore, we define $R_n^0 = R_n^{1-} := W_n^1 - 1$.

ECN Router. Suppose that the passive action at the router refers to *marking* the packets of the flow (dropping none of them), using explicit congestion notification (ECN). Then, $W_n^0 := W_n^1$ and $R_n^0 := R_n^1$. If all the routers use ECN, dropping downstream only occurs when the buffers overflow, which we consider very harmful, so that $R_n^{1-} := 0$.

6.3.3 Optimization Problem

Consider now the individual-flow problem in (6.7). Thus, we look for a policy maximizing the expected total net rewards under the β-discounted criterion. From the above interpretations we conclude that our model captures the trade-off between high throughput and long queues on one hand, and low throughput and short queues on the other hand for finite flows. This is known as network optimization under the *throughput/delay criterion*.

To evaluate a policy π under the β-discounted criterion, we consider the following two measures. Let $g_i^\pi := \mathbb{E}_i^\pi \left[\sum_{t=0}^{\infty} \beta^t W_{X(t)}^{a(t)} \right]$ be the *expected total β-discounted bandwidth utilization* if starting from state i under policy π. For convenience, we will also call g_i^π the expected total β-discounted work, since the bandwidth utilization can be seen as the work performed by the router in order to transmit the flow. Analogously we denote by $f_i^\pi := \mathbb{E}_i^\pi \left[\sum_{t=0}^{\infty} \beta^t R_{X(t)}^{a(t)} \right]$ the *expected total β-discounted reward* if starting from state i under policy π.

The objective (6.7)[2] is for each transmission cost ν,

$$\max_{\pi \in \Pi} f_i^\pi - \nu g_i^\pi. \tag{6.8}$$

6.4 Optimal Solution via the Marginal Productivity Index

In this section we solve the problem (6.8) using the restless bandit indexation methodology surveyed in Niño-Mora (2007b). First we reduce the optimization problem by *normalizing* the net reward parameters and narrow the focus to stationary policies. Then,

[2] Similarly could be formulated the problem under the time-average criterion. The marginal productivity index under the latter are obtained in the limit $\beta \to 1$ from the marginal productivity index under the discounted criterion (Niño-Mora, 2002).

we carry out a work-reward analysis in order to establish structural results of the optimal control policy. Finally, we obtain closed formulae of the transmission index of the restarting flow, and of the flow with fast recovery.

The MDP model for any of the router variants aims at finding the *transmission index* that will be assigned to a particular flow on a particular router. The transmission index captures the actual value of network services to the flow receiver, and falls into the concept of the *marginal productivity* (MP) index developed in Niño-Mora (2001, 2002, 2006b). If ν denotes the *transmission cost* (or, the cost of providing bandwidth capacity) paid for each unit of router's bandwidth required by the flow, then the transmission index is defined as a set of state-dependent values so that the following is an optimal policy: "Transmit the flow without congestion warning if and only if the actual transmission index is higher than the transmission cost ν".

Niño-Mora (2002, p. 383) showed that the above problem can be normalized so that no passive-action net rewards are present. This is achieved by normalizing the active-action net rewards (see also the Appendix, Section D.3) for all $n \in \mathcal{N}$ by

$$R_n := (R_n^1 - R_n^0) + \beta p_n (R_{n+1}^0 - R_0^0), \quad W_n := (W_n^1 - W_n^0) + \beta p_n (W_{n+1}^0 - W_0^0), \quad (6.9)$$

where we have defined $R_N^0 := R_{N-1}^0$ and $W_N^0 := W_{N-1}^0$.

Thus, the above-defined dynamics of the system is modified in the following way:

- *Dynamics if active*: If the flow is in state n and the flow is warned at a given period, then during that period it generates net reward $R_n - \nu W_n$ and
  - with probability $p_n > 0$: the flow moves to state $n+1$ for the next period (or remains in $N-1$, if it is already there).
  - with probability $1 - p_n$: the flow moves to state 0 for the next period.
- *Dynamics if passive*: If the flow is in state n and the flow is warned at a given period, then during that period it generates no net reward and the flow moves to state 0 for the next period.

Thus, we end up with an admission control problem, pioneered by Naor (1969) for Poisson arrivals. For the same arrival process, Niño-Mora (2002) identified an optimal policy via MP indices. However, the TCP-like behavior described in the previous section does not result in a Poisson process for the number of arriving packets, therefore we analyze it next.

6.4.1 Reduction to Stationary Policies

Since for MDPs with finite state space and finite action space there exists an optimal stationary policy independent of the initial state, we focus only on stationary policies and represent them via *active sets* $\mathcal{S} \subseteq \mathcal{N}$. In other words, a policy $\mathcal{S}$ prescribes to be active (to transmit) in states in $\mathcal{S}$ and passive (to warn) in states in $\mathcal{N} \setminus \mathcal{S}$.

In the remainder of this section we consider the normalized problem under the set of all active sets. The above-defined total bandwidth utilization (total work) g_i^π and total reward f_i^π can readily be defined for this problem as $g_i^{\mathcal{S}}$ and $f_i^{\mathcal{S}}$, respectively. Moreover, they satisfy the following balance equations, to be used in the later analysis.

Lemma 6.1. *For any state n and any active set $\mathcal{S}$ we have:*
If state $n \in \mathcal{S}$ and $n \neq N-1$, then

$$f_n^{\mathcal{S}} = R_n + \beta\left(p_n f_{n+1}^{\mathcal{S}} + (1-p_n) f_0^{\mathcal{S}}\right), \qquad g_n^{\mathcal{S}} = W_n + \beta\left(p_n g_{n+1}^{\mathcal{S}} + (1-p_n) g_0^{\mathcal{S}}\right).$$

If state $N-1 \in \mathcal{S}$, then

$$f_{N-1}^{\mathcal{S}} = \frac{R_{N-1} + \beta(1-p_{N-1}) f_0^{\mathcal{S}}}{1-\beta p_{N-1}}, \qquad g_{N-1}^{\mathcal{S}} = \frac{W_{N-1} + \beta(1-p_{N-1}) g_0^{\mathcal{S}}}{1-\beta p_{N-1}}.$$

If state $n \notin \mathcal{S}$, then $f_n^{\mathcal{S}} = \beta f_0^{\mathcal{S}}, g_n^{\mathcal{S}} = \beta g_0^{\mathcal{S}}$. If state $0 \notin \mathcal{S}$, then $f_0^{\mathcal{S}} = 0, g_0^{\mathcal{S}} = 0$.

6.4.2 Marginal Reward and Marginal Bandwidth Utilization

In this subsection we define measures of marginal reward and marginal bandwidth utilization, and present results which will be used later to derive the transmission index of problem (6.8).

Let $\langle a, \mathcal{S}\rangle$ be the policy that implements action a in the initial period and policy $\mathcal{S}$ proceeds. We consider the $(n,\mathcal{S})$-*marginal reward* defined as $r_n^{\mathcal{S}} := f_n^{\langle 1,\mathcal{S}\rangle} - f_n^{\langle 0,\mathcal{S}\rangle}$, and the $(n,\mathcal{S})$-*marginal bandwidth utilization* (or, $(n,\mathcal{S})$-marginal work) defined as $w_n^{\mathcal{S}} := g_n^{\langle 1,\mathcal{S}\rangle} - g_n^{\langle 0,\mathcal{S}\rangle}$. Finally, the $(n,\mathcal{S})$-*marginal transmission rate* is denoted by $\nu_n^{\mathcal{S}} := r_n^{\mathcal{S}}/w_n^{\mathcal{S}}$. The following results are analogous to the balance equations in Lemma 6.1.

Lemma 6.2. *For any state $n \neq N-1$ and any active set $\mathcal{S}$ we have*

$$f_n^{\langle 1,\mathcal{S}\rangle} = R_n + \beta\left(p_n f_{n+1}^{\mathcal{S}} + (1-p_n) f_0^{\mathcal{S}}\right), \qquad g_n^{\langle 1,\mathcal{S}\rangle} = W_n + \beta\left(p_n g_{n+1}^{\mathcal{S}} + (1-p_n) g_0^{\mathcal{S}}\right),$$
$$f_{N-1}^{\langle 1,\mathcal{S}\rangle} = R_{N-1} + \beta\left(p_{N-1} f_{N-1}^{\mathcal{S}} + (1-p_{N-1}) f_0^{\mathcal{S}}\right), \qquad g_{N-1}^{\langle 1,\mathcal{S}\rangle} = W_{N-1} + \beta\left(p_{N-1} g_{N-1}^{\mathcal{S}} + (1-p_{N-1}) g_0^{\mathcal{S}}\right),$$
$$f_n^{\langle 0,\mathcal{S}\rangle} = \beta f_0^{\mathcal{S}}, \qquad g_n^{\langle 0,\mathcal{S}\rangle} = \beta g_0^{\mathcal{S}},$$
$$f_{N-1}^{\langle 0,\mathcal{S}\rangle} = \beta f_0^{\mathcal{S}}, \qquad g_{N-1}^{\langle 0,\mathcal{S}\rangle} = \beta g_0^{\mathcal{S}}.$$

Using the above lemma and the definition of the marginal reward and marginal work, respectively, we obtain the following identities.

Proposition 6.2. *For any state $n \neq N-1$ and any active set $\mathcal{S}$ we have*

$$r_n^{\mathcal{S}} = R_n + \beta p_n (f_{n+1}^{\mathcal{S}} - f_0^{\mathcal{S}}) \qquad w_n^{\mathcal{S}} = W_n + \beta p_n (g_{n+1}^{\mathcal{S}} - g_0^{\mathcal{S}})$$
$$r_{N-1}^{\mathcal{S}} = R_{N-1} + \beta p_{N-1} (f_{N-1}^{\mathcal{S}} - f_0^{\mathcal{S}}) \qquad w_{N-1}^{\mathcal{S}} = W_{N-1} + \beta p_{N-1} (g_{N-1}^{\mathcal{S}} - g_0^{\mathcal{S}})$$

Finally, the next result is obtained by employing Lemma 6.1 into Proposition 6.2.

Proposition 6.3. *If state $n \in \mathcal{S}$, then $r_n^{\mathcal{S}} = f_n^{\mathcal{S}} - \beta f_0^{\mathcal{S}}$ and $w_n^{\mathcal{S}} = g_n^{\mathcal{S}} - \beta g_0^{\mathcal{S}}$.*

For a given active set $\mathcal{S}$, the above propositions allow us to narrow the focus to quantities $f_n^{\mathcal{S}}$ and $g_n^{\mathcal{S}}$ for $n \in \mathcal{S}$ that can be obtained from the recursion in Lemma 6.1, as we will see in the following subsection.

6.4.3 Structure of Optimal Policies

Now we prepare the ground for establishing the structure of the optimal active sets (policies), under the following two conditions. First, we will use a natural monotonicity assumption upon congestion probabilities, termed *deteriorating QoS* (Quality of service).

Assumption 6.1 (Deteriorating QoS). The one-period works W_n are positive and we have $\dfrac{W_m}{p_m} \leq \dfrac{W_n}{p_n}$ for all $m, n \in \mathcal{N}$ such that $m < n$.

Further, we will concentrate on the case, in which the one-period reward R_n possesses a sort of concavity in W_n. The concavity behavior is a natural one for the expected goodput in communications networks. In fact, one expects the network to increment losses as the required bandwidth per connection is increased.

Assumption 6.2 (Concave Adjusted Rewards). There is a real-valued function R with $R(0) \geq 0$, which is concave on the domain $\{0, W_0/p_0, \ldots, W_{N-1}/p_{N-1}\}$ and satisfies $R_n/p_n = R(W_n/p_n)$.

We will prove that, under these two conditions, the optimal policy for any transmission cost ν is a *threshold policy* belonging to family

$$\mathcal{F} := \{\mathcal{N}_{k-1} : k \in \mathcal{N} \cup \{N\}\}, \qquad \text{where } \mathcal{N}_{k-1} := \{0, 1, \ldots, k-1\}. \tag{6.10}$$

That is, for any transmission cost ν there is a threshold state k such that it is optimal to transmit the flow if and only if it is in any state smaller than k. We will accomplish this by establishing PCL($\mathcal{F}$)-indexability, introduced in Niño-Mora (2001, 2002), of our problem for the family of active sets $\mathcal{F}$ defined in (6.10).

Definition 6.1. Problem (6.8) is called *PCL($\mathcal{F}$)-indexable* if the following two conditions hold:

(i) the marginal work $w_n^{\mathcal{N}_{k-1}}$ is positive for all $n \in \mathcal{N}$ and $k \in \mathcal{N} \cup \{N\}$.

(ii) the marginal transmission rates $\nu_n^{\mathcal{N}_{n-1}}$ are nonincreasing in $n \in \mathcal{N}$.

The two assumptions given above will be shown sufficient for conditions (i) and (ii) of PCL($\mathcal{F}$)-indexability. They are, however, not necessary, and thus the threshold policies defined in (6.10) may remain optimal even if our assumptions are not valid.

6.4.4 Work-Reward Analysis

To achieve the above goal, we first characterize the marginal rewards and works. For all $n \in \mathcal{N}$, let $q_n := \prod_{m=0}^{n-1} \beta p_m$, $Q_n := \sum_{m=0}^{n} q_m$, $Q_n^{(R)} := \sum_{m=0}^{n} q_m R_m$, and $Q_n^{(W)} := \sum_{m=0}^{n} q_m W_m$.

Lemma 6.3. *For $k, n \in \mathcal{N}$ with $n \geq k-1$, the marginal reward and work, respectively, are*

$$r_n^{\mathcal{N}_{k-1}} = R_n - \beta p_n \frac{Q_{k-1}^{(R)}}{Q_k}, \qquad w_n^{\mathcal{N}_{k-1}} = W_n - \beta p_n \frac{Q_{k-1}^{(W)}}{Q_k}. \tag{6.11}$$

Proof. For $k = 0$ consider $\mathcal{N}_{-1} = \emptyset$. Then for any state $n \in \mathcal{N}$ we have $f_n^{\mathcal{N}_{-1}} = g_n^{\mathcal{N}_{-1}} = 0$, and hence $r_n^{\mathcal{N}_{-1}} = R_n$ and $w_n^{\mathcal{N}_{-1}} = W_n$ by Proposition 6.2.

Now for $k \in \mathcal{N} \setminus \{0\}$ consider $\mathcal{N}_{k-1}$. Using Lemma 6.1, for $n \leq k-1$ we have

$$f_n^{\mathcal{N}_{k-1}} = R_n + \beta \left(p_n f_{n+1}^{\mathcal{N}_{k-1}} + (1-p_n) f_0^{\mathcal{N}_{k-1}} \right), \quad g_n^{\mathcal{N}_{k-1}} = W_n + \beta \left(p_n g_{n+1}^{\mathcal{N}_{k-1}} + (1-p_n) g_0^{\mathcal{N}_{k-1}} \right),$$

and for $n \geq k$ we have $f_n^{\mathcal{N}_{k-1}} = \beta f_0^{\mathcal{N}_{k-1}}$ and $g_n^{\mathcal{N}_{k-1}} = \beta g_0^{\mathcal{N}_{k-1}}$.

The solution of the two above linear-equation systems gives

$$(1-\beta) f_0^{\mathcal{N}_{k-1}} = \frac{Q_{k-1}^{(R)}}{Q_k}, \qquad (1-\beta) g_0^{\mathcal{N}_{k-1}} = \frac{Q_{k-1}^{(W)}}{Q_k}. \tag{6.12}$$

Therefore, we have $f_n^{\mathcal{N}_{k-1}} - f_0^{\mathcal{N}_{k-1}} = -(1-\beta) f_0^{\mathcal{N}_{k-1}}$ and $g_n^{\mathcal{N}_{k-1}} - g_0^{\mathcal{N}_{k-1}} = -(1-\beta) g_0^{\mathcal{N}_{k-1}}$ for $n \geq k$. Then Proposition 6.2 yields (6.11) for $k, n \in \mathcal{N}$ with $n \geq k-1$. □

In the following lemma we establish positivity of the marginal works as required in condition (i) of PCL($\mathcal{F}$)-indexability.

Lemma 6.4. *Under deteriorating QoS, the marginal work $w_n^{\mathcal{N}_{k-1}}$ is positive for all $n \in \mathcal{N}$ and $k \in \mathcal{N} \cup \{N\}$, i.e., condition (i) of PCL($\mathcal{F}$)-indexability holds.*

Proof. If $k \neq N$, then the marginal work $w_n^{\mathcal{N}_{k-1}}$ is positive for $n \geq k-1$, because from (6.11) it can be rewritten as

$$w_n^{\mathcal{N}_{k-1}} = \frac{1}{Q_k}\left(W_n + \sum_{m=0}^{k-1} q_{m+1}\left(W_n - \frac{p_n}{p_m}W_m\right)\right), \tag{6.13}$$

which is positive due to deteriorating QoS.

For $n \leq k-1 \leq N-1$ we proceed as follows. Lemma 6.1 after rearranging gives, for $n \geq 1$,

$$g_n^{\mathcal{N}_{k-1}} = \left(g_{n-1}^{\mathcal{N}_{k-1}} - \beta(1-p_{n-1})g_0^{\mathcal{N}_{k-1}} - W_{n-1}\right)/\beta p_{n-1}, \tag{6.14}$$

whose recursive implementation yields

$$g_n^{\mathcal{N}_{k-1}} = \frac{Q_n - \beta Q_{n-1}}{q_n} g_0^{\mathcal{N}_{k-1}} - \frac{Q_{n-1}^{(W)}}{q_n}. \tag{6.15}$$

Due to Proposition 6.3, we want to prove $g_n^{\mathcal{N}_{k-1}} > \beta g_0^{\mathcal{N}_{k-1}}$, i.e.,

$$(1-\beta)g_0^{\mathcal{N}_{k-1}} > \frac{Q_{n-1}^{(W)}}{Q_n}. \tag{6.16}$$

For $k = N$, (6.15) evaluated for $n = N-1$ together with Lemma 6.1 give

$$(1-\beta)g_0^{\mathcal{N}_{N-1}} = \frac{Q_{N-2}^{(W)} + W_{N-1}\frac{q_{N-1}}{1-\beta p_{N-1}}}{Q_{N-1} + \beta p_{N-1}\frac{q_{N-1}}{1-\beta p_{N-1}}}. \tag{6.17}$$

For $k \neq N$, we use (6.12). In both cases, deteriorating QoS and D.1 imply (6.16). □

Next we characterize and bound the marginal transmission rates $\nu_k^{\mathcal{N}_{k-1}}$, which will be crucial for characterization of the transmission index in the next subsection.

Lemma 6.5. *Under concave adjusted rewards and deteriorating QoS, for $k \in \mathcal{N}$, we have*

$$\frac{R_k - \frac{p_k}{p_{k-1}}R_{k-1}}{W_k - \frac{p_k}{p_{k-1}}W_{k-1}} \leq \nu_k^{\mathcal{N}_{k-1}} = \frac{R_k + \sum_{m=0}^{k-1} q_{m+1}\left(R_k - \frac{p_k}{p_m}R_m\right)}{W_k + \sum_{m=0}^{k-1} q_{m+1}\left(W_k - \frac{p_k}{p_m}W_m\right)} \leq \frac{R_k}{W_k}. \tag{6.18}$$

Proof. The expression for $\nu_k^{\mathcal{N}_{k-1}}$ is a simple reformulation of (6.11), as in (6.13). Having $R_m/W_m \geq R_k/W_k$ for all $m < k$ from D.4(ii), we can rewrite it as

$$\frac{R_k - \frac{p_k}{p_m}R_m}{W_k - \frac{p_k}{p_m}W_m} \leq \frac{R_k}{W_k}. \tag{6.19}$$

Then, D.2 yields the upper bound for $\nu_k^{\mathcal{N}_{k-1}}$.

We prove the lower bound next. D.2 together with the upper bound implies

$$\min\left\{\frac{R_k-\frac{p_k}{p_m}R_m}{W_k-\frac{p_k}{p_m}W_m} \text{ for all } 0\le m\le k-1\right\}\le \nu_k^{\mathcal{N}_{k-1}}. \tag{6.20}$$

Further, we have

$$\frac{R_k-\frac{p_k}{p_m}R_m}{W_k-\frac{p_k}{p_m}W_m}\ge \frac{R_k-\frac{p_k}{p_{k-1}}R_{k-1}}{W_k-\frac{p_k}{p_{k-1}}W_{k-1}} \text{ for all } 0\le m\le k-1 \tag{6.21}$$

by D.3(iii), which gives the lower bound. □

Further, we will need the following monotonicity result for transmission rates.

Lemma 6.6. *Under concave adjusted rewards and deteriorating QoS, $\nu_k^{\mathcal{N}_{k-1}}\ge \nu_n^{\mathcal{N}_{k-1}}$ for all $k,n\in\mathcal{N}$ with $n\ge k+1$.*

Proof. Denote by

$$a_k:=\frac{\beta Q_{k-1}^{(R)}}{Q_k}, \qquad\qquad b_k:=\frac{\beta Q_{k-1}^{(W)}}{Q_k}.$$

Using the identities in (6.11), we are to show that, for all $N-1\ge n\ge k+1$,

$$\nu_k^{\mathcal{N}_{k-1}}=\frac{\frac{R_k}{p_k}-a_k}{\frac{W_k}{p_k}-b_k}\ge \frac{\left(\frac{R_k}{p_k}-a_k\right)+\left(\frac{R_n}{p_n}-\frac{R_k}{p_k}\right)}{\left(\frac{W_k}{p_k}-b_k\right)+\left(\frac{W_n}{p_n}-\frac{W_k}{p_k}\right)}=\frac{\frac{R_n}{p_n}-a_k}{\frac{W_n}{p_n}-b_k}=\nu_n^{\mathcal{N}_{k-1}}. \tag{6.22}$$

Having

$$\frac{\frac{R_k}{p_k}-a_k}{\frac{W_k}{p_k}-b_k}\ge \frac{\frac{R_n}{p_n}-\frac{R_k}{p_k}}{\frac{W_n}{p_n}-\frac{W_k}{p_k}} \qquad \text{for all } N-1\ge n\ge k+1$$

is equivalent to

$$\frac{\frac{R_k}{p_k}-a_k}{\frac{W_k}{p_k}-b_k}\ge \frac{\frac{R_{k+1}}{p_{k+1}}-\frac{R_k}{p_k}}{\frac{W_{k+1}}{p_{k+1}}-\frac{W_k}{p_k}}. \tag{6.23}$$

since the right-hand side is nonincreasing in n, due to concave adjusted rewards with D.3(ii). The last inequality holds due to the lower bound in Lemma 6.5 and, again, concave adjusted rewards with D.3(iv). Therefore, by D.1(ii), (6.22) holds. □

Next we establish condition (ii) of PCL($\mathcal{F}$)-indexability.

Lemma 6.7. *Under concave adjusted rewards and deteriorating QoS, marginal transmission rates $\nu_n^{\mathcal{N}_{n-1}}$ are nonincreasing in $n \in \mathcal{N}$, i.e., condition (ii) of PCL($\mathcal{F}$)-indexability holds.*

Proof. Niño-Mora (2002, Proposition 6.4(c)) showed that under positive marginal works $\nu_n^{\mathcal{N}_{n-1}} \geq \nu_{n+1}^{\mathcal{N}_n}$ is equivalent to $\nu_n^{\mathcal{N}_{n-1}} \geq \nu_{n+1}^{\mathcal{N}_{n-1}}$, which is satisfied due to Lemma 6.6. □

To summarize, in this subsection we have proved the following.

Proposition 6.4. *Under concave adjusted rewards and deteriorating QoS, problem (6.8) is PCL($\mathcal{F}$)-indexable for the family $\mathcal{F}$ defined in (6.10).*

6.4.5 Transmission Indices

Now we are ready to give a complete characterization of the transmission index, which is the MP index interpreted in the setting of the congestion control problem.

Proposition 6.5. *Under concave adjusted rewards and deteriorating QoS, the optimal active sets for problem (6.8) are $\mathcal{N}_{k-1}$, for $k \in \mathcal{N} \cup \{N\}$ and the transmission index of state n under the β-discounted criterion is*

$$\nu_n = \frac{R_n + \sum_{m=0}^{n-1} q_{m+1} \left(R_n - \frac{p_n}{p_m} R_m \right)}{W_n + \sum_{m=0}^{n-1} q_{m+1} \left(W_n - \frac{p_n}{p_m} W_m \right)}. \tag{6.24}$$

Proof. Since the problem (6.8) is PCL($\mathcal{F}$)-indexable by Proposition 6.4, the transmission index exists and is defined by $\nu_n^{\mathcal{N}_{n-1}}$ for $n \in \mathcal{N}$. □

The transmission index can be bounded by quantities independent of the discount factor β. Given the interpretation of the discount factor as the probability that the flow continues in the next period, these bounds are valid for flows of any length.

Proposition 6.6. *Under concave adjusted rewards and deteriorating QoS, for any β,*

$$\frac{R_n - \frac{p_n}{p_{n-1}} R_{n-1}}{W_n - \frac{p_n}{p_{n-1}} W_{n-1}} \leq \nu_n \leq \frac{R_n}{W_n}. \tag{6.25}$$

Proof. By Lemma 6.5, because the transmission index is defined by $\nu_n^{\mathcal{N}_{n-1}}$ for $n \in \mathcal{N}$. □

We further show that the transmission index is higher for shorter flows.

Proposition 6.7. *The transmission index ν_n is nonincreasing in β. Further, $\nu_n \to \frac{R_n}{W_n}$ as $\beta \to 0$.*

Proof. Notice first in the transmission index expression in Proposition 6.5 that q_{m+1} depends multiplicatively on β. Further, we have

$$\frac{R_n - \frac{p_n}{p_m} R_m}{W_n - \frac{p_n}{p_m} W_m} \geq \frac{R_n - \frac{p_n}{p_{m+1}} R_{m+1}}{W_n - \frac{p_n}{p_{m+1}} W_{m+1}} \text{ for all } 0 \leq m \leq n-2 \tag{6.26}$$

by D.3(iii). Therefore, D.5 implies that the expression for the transmission index in Proposition 6.5 nondecreases when β is decreased. The convergence as $\beta \to 0$ is obtained directly since q_{m+1} depends multiplicatively on β. □

6.4.6 Flows with Fast Recovery

Now we consider a flow which restarts to state $i \in \mathcal{N}$ instead of state 0, which resembles the *Fast Recovery* (i.e., multiplicative decrease) mechanism commonly implemented in several TCP variants. Let $q_{i,n} := \prod_{m=i}^{n-1} \beta p_m$ and $Q_{i,n} := \sum_{m=i}^{n} q_{i,m}$ for all $i \leq n$.

Proposition 6.8. *Under concave adjusted rewards and deteriorating QoS, if the flow restarts to state $i \in \mathcal{N}$, the transmission index of state $n \geq i$ under the discounted criterion is*

$$\nu_{i,n} = \frac{R_n + \sum_{m=i}^{n-1} q_{i,m+1} \left(R_n - \frac{p_n}{p_m} R_m \right)}{W_n + \sum_{m=i}^{n-1} q_{i,m+1} \left(W_n - \frac{p_n}{p_m} W_m \right)}. \tag{6.27}$$

The transmission index of flows restarting at different minimum transmission rates (i.e., at different states) seem to be quantitatively very akin. In fact, the numerator and denominator of the above index formulae are dominated by terms $Q_{i,n} R_n$ and $Q_{i,n} W_n$, respectively, and hence the parameters of other states, apart from being smaller, become negligible for larger values of n. This is especially true if the flow uses *slow start*.

We believe that the above formulae can remain sound even for more complex protocols, like those that can react both by restarting and by fast recovery following a more complex rule. The reason being that, from the probabilistic perspective, these protocols can be approximated as randomizations of the two extreme variants we explored above. Their indices thus should rather closely follow the above formulae and the results presented here could be considered quite robust. Nevertheless, only a profound analysis of such protocols may verify this intuition.

6.5 The Three Router Variants: Solutions

In this section we employ the above general results to obtain the transmission index under three common router variants, maximizing the throughput/delay criterion. For notational convenience, we define $p_N := p_{N-1}$, $W_N^1 := W_{N-1}^1$ and $W_{-1}^1 := 0$.

The requirements needed in the propositions below should be commonly satisfied in practice. For instance, the concavity of the throughput functions $R(\cdot)$ holds if the congestion probability downstream is proportional to the transmission rate. Note also that in a congestion-free network (i.e., $p_n = 1$ for all $n \in \mathcal{N}$), the transmission index equals 1, independently of the router variant and of the actual flow transmission rate.

6.5.1 TD Router

Proposition 6.9. *Suppose that $R(0) := 0$ and $R(W_n^1/p_n) := W_n^1$ for all $n \in \mathcal{N}$ is concave on the domain $\{0, W_0^1/p_0, W_1^1/p_1, \ldots, W_{N-1}^1/p_{N-1}\}$, i.e.,*

$$\frac{W_n^1 - W_{n-1}^1}{\frac{W_n^1}{p_n} - \frac{W_{n-1}^1}{p_{n-1}}} \geq \frac{W_{n+1}^1 - W_n^1}{\frac{W_{n+1}^1}{p_{n+1}} - \frac{W_n^1}{p_n}}, \qquad \text{for all } n \in \mathcal{N} \setminus \{0, N-1\}. \tag{6.28}$$

Then, the transmission index of state $n \in \mathcal{N}$ is

$$\nu_n = \frac{W_n^1 + \sum_{m=0}^{n-1} q_{m+1} \left(W_n^1 - W_m^1\right)}{\frac{W_n^1}{p_n} + \sum_{m=0}^{n-1} q_{m+1} \left(\frac{W_n^1}{p_n} - \frac{W_m^1}{p_m}\right)}. \tag{6.29}$$

Proof. By (6.9), $W_n := W_n^1$ and $R_n = p_n W_n^1$ for all $n \in \mathcal{N}$. It is easy to check that deteriorating QoS holds. The result then follows from Proposition 6.5. □

6.5.2 ICN Router

Proposition 6.10. *Suppose that $R(0) := 0$ and $R(1/p_n + \beta(W_{n+1}^1 - 1)) := 1 + \beta(W_{n+1}^1 - 1)$ for all $n \in \mathcal{N}$ is concave on the domain $\{0, 1/p_0 + \beta(W_1^1 - 1), 1/p_1 + \beta(W_2^1 - 1), \ldots, 1/p_{N-1} + \beta(W_N^1 - 1)\}$, i.e.,*

$$\frac{W_2^1}{W_1^1} \geq \frac{p_0}{p_1} \quad \text{and} \quad \frac{W_{n+2}^1 - W_{n+1}^1}{W_{n+1}^1 - W_n^1} \leq \frac{\frac{1}{p_{n+1}} - \frac{1}{p_n}}{\frac{1}{p_n} - \frac{1}{p_{n-1}}}, \quad \text{for all } n \in \mathcal{N} \setminus \{0, N-1\}. \tag{6.30}$$

Then, the transmission index of state $n \in \mathcal{N}$ is

$$\nu_n = \frac{1 + \beta(W^1_{n+1} - 1) + \sum_{m=0}^{n-1} q_{m+1}\beta\left(W^1_{n+1} - W^1_{m+1}\right)}{\frac{1}{p_n} + \beta(W^1_{n+1} - 1) + \sum_{m=0}^{n-1} q_{m+1}\left(\frac{1}{p_n} - \frac{1}{p_m} + \beta\left(W^1_{n+1} - W^1_{m+1}\right)\right)}. \tag{6.31}$$

Proof. By (6.9) and using $W^0_0 = R^0_0 = 0$, $W_n = 1 + \beta p_n(W^1_{n+1} - 1)$ and $R_n = p_n + \beta p_n(W^1_{n+1} - 1)$ for all $n \in \mathcal{N}$. It is easy to check that deteriorating QoS holds. The result then follows from Proposition 6.5. □

6.5.3 ECN Router

Proposition 6.11. *Suppose that $R(0) := 0$ and $R(\beta(W^1_{n+1} - 1)) := \beta(p_{n+1}W^1_{n+1} - p_0)$ for all $n \in \mathcal{N}$ is concave on the domain $\{0, \beta(W^1_1 - 1), \beta(W^1_2 - 1), \ldots, \beta(W^1_{N-1} - 1)\}$, i.e.,*

$$\frac{p_{n+2}W^1_{n+2} - p_{n+1}W^1_{n+1}}{p_{n+1}W^1_{n+1} - p_nW^1_n} \leq \frac{W^1_{n+2} - W^1_{n+1}}{W^1_{n+1} - W^1_n}, \qquad \text{for all } n \in \mathcal{N} \setminus \{N-2, N-1\}. \tag{6.32}$$

Then, the transmission index of state $n \in \mathcal{N}$ is

$$\nu_n = \frac{(p_{n+1}W^1_{n+1} - p_0) + \sum_{m=0}^{n-1} q_{m+1}\left(p_{n+1}W^1_{n+1} - p_{m+1}W^1_{m+1}\right)}{(W^1_{n+1} - 1) + \sum_{m=0}^{n-1} q_{m+1}\left(W^1_{n+1} - W^1_{m+1}\right)}. \tag{6.33}$$

Proof. By (6.9) and using $W^0_0 = R^0_0 = 1$, $R_n = \beta p_n(p_{n+1}W^1_{n+1} - p_0)$ and $R_n = W_n = \beta p_n(W^1_{n+1} - 1)$ for all $n \in \mathcal{N}$. It is easy to check that deteriorating QoS holds. The result then follows from Proposition 6.5. □

6.6 Transmission Index Priority Policies for Bottleneck Problem

We now have all the necessary results for the individual flows to tackle the multi-flow problem in the following sections. In this section we consider the *bottleneck problem*, in which several flows compete for router's resources. We show how transmission index policies can be optimal in two particular bottleneck problems, which are special cases of the real-life situations. Let $\mathcal{M} := \{1, 2, \ldots, M\}$ be the set of flows and write (m, n) if flow m is in state n. Thus, e.g., $\nu_{m,n}$ is the transmission index of flow m in state n.

6.6.1 Optimality in Single-Transmitted-Flow Problem

In this subsection we treat the problem of transmitting a single flow, while all the remaining flows are warned. We show that the optimal policy for such a case is to transmit the flow with the highest MP (transmission) index.

Suppose that the router has enough resources to transmit flows of any arrival rate. We prove optimality of the transmission priority policy using a similarity of the above problem with the classic bandit problem, where warned flows are not allowed to change state. In our problem, a warned flow is allowed to change state, however, the transition is made to state 0, where it remains until transmitted next time. The idea of our proof below draws on Weber (1992), who observed that a priority policy in multiple-flow problem is optimal if it is a *holding policy*: If flow m of index $\nu_{m,n}$ is transmitted in the initial time period, then it must be transmitted in all subsequent periods while its index is greater than $\nu_{m,n}$; if its index equals or drops below $\nu_{m,n}$ at certain time period, then that time period is considered initial and the requirement is repeated.

Proposition 6.12. *Suppose that each flow in isolation satisfies concave adjusted rewards and deteriorating QoS after its normalization. If each flow m is initially in state $(m, 0)$, then the following is an optimal policy for the above problem: "At each time period transmit the flow of highest actual transmission index and warn the remaining flows".*

Proof. Suppose that in the initial period flow m's index $\nu_{m,0}$ is higher than or equal to other flows' indices, and the flow m is transmitted. Then, in the next period the flow m will move to state 1, whose index is by Lemma 6.7 not greater than $\nu_{m,n}$. Therefore, we can consider that period as a new initial period and repeat the argument, which implies that the transmission priority policy is a holding policy, and thus optimal. □

6.6.2 Optimality in Expected-Queue-Length Problem

Some congestion avoidance mechanisms proposed in the literature or currently implemented in the Internet try to control the average buffer queue length. Suppose that our objective is to find a policy that maximizes the aggregate sum of expected total discounted rewards obtained from all the flows, given that queue length equals to a given target value W at each time period *in expectation*. This requirement can be seen as a relaxation of having the queue length equal to W at each time period, and it became known as the *Whittle relaxation* (see Whittle, 1988; Niño-Mora, 2001). The Whittle relaxation, using a Lagrangian approach, can be perfectly decomposed into parametric problems (6.8), where each flow is considered in isolation.

Proposition 6.13 (Whittle (1988), Proposition 3). *Suppose that each flow in isolation satisfies concave adjusted rewards and deteriorating QoS after its normalization. For every target*

queue length W there exists a value $\nu(W)$ such that the following is an optimal policy for the above problem: "At each time period transmit all the flows of actual transmission index higher than $\nu(W)$ and warn the remaining flows".

Note that the value $\nu(W)$ depends on the number of flows and their characteristics.

6.7 Practical Implementation

While the indices have usually been used as an ordinal measure in priority policies, results in Chapter 5 suggest that their pricing interpretation can give rise to novel policies. Next we develop a packet-pricing policy using the MP (transmission) indices.

6.7.1 Transmission Indices Implementation in Congestion Avoidance Mechanisms

Consider any congestion avoidance mechanism (e.g., RED, REM, etc.) with the following property: on an arrival or while a packet is queued in the buffer, the mechanism calculates the probability of dropping/marking it, generates a random event, and eventually drops/marks the packet. We will now discuss how such a mechanism may be modified so that it takes into account the *transmission index* of the packet during the dropping/marking decision stage.

The price of dropping/marking a packet can be evaluated via the transmission index multiplied by its size in Bytes, say $\nu_n s_n$. Let the dropping/marking probability calculated by the congestion avoidance mechanism for this packet at a given time moment be π_n. We next set out to calculate what should be the dropping/marking probability π_m if a packet with price $\nu_m s_m$ arrived instead.

For a fair admission control we may want to impose that the *expected loss* of dropped/marked packets be equal: $\pi_m \nu_m s_m = \pi_n \nu_n s_n$. Hence,

$$\pi_m = \frac{\nu_n s_n}{\nu_m s_m} \pi_n. \tag{6.34}$$

Therefore, either all or none of the incoming packets have zero dropping/warning probability.

When the router is heavily congested and the dropping/marking probability $\pi_n = 1$, then all other packets should experience the same dropping/marking probability. Yet in that case, according to (6.34), the dropping/marking probability π_m will be larger than 1 for lower priced packets, and smaller than 1 for higher priced packets. An alternative

formula satisfying both $\pi_n = 0 \iff \pi_m = 0$ and $\pi_n = 1 \iff \pi_m = 1$ is

$$\pi_m = 1 - (1 - \pi_n)^{\frac{\nu_n s_n}{\nu_m s_m}}. \tag{6.35}$$

Note that the two formulae are roughly equivalent for small values of dropping/-marking probability π_n. Any of them can hence be implemented in the congestion avoidance mechanisms, in which the dropping/marking probability is maintained at very low levels (except when the buffer overflows and all incoming packets are being dropped). For congestion avoidance mechanisms, in which the dropping/marking probability smoothly increases to 1, the latter might be the preferred formula.

6.8 Conclusions and Further Work

We have presented a decentralized approach to model congestion control at network nodes with future-path information. We have focused on the most representative type of traffic with finite or infinite length and additive-increase/multiplicative-decrease mechanism. Other flow types are likely to be tractable in our framework with moderate modifications. Our analytical approach is flexible and it directly applies to both wired and wireless networks. By the means of the transmission index we have obtained locally optimal solution that can be interpreted as the packet price (when multiplied by the packet size) and easily implemented in congestion avoidance mechanisms to improve their resource allocation decisions.

From the modeling perspective, several challenges remain open. For instance, our model assumes that the congestion information is instantaneously received by network nodes. A more realistic model would take into account information delays and censoring due to unobserved dropped packets. Existence of finite buffers and non-cooperative flows further raises the necessity of consideration of three actions at the router: transmitting, warning, and dropping. However, index policies are only known in the literature for optimization problems with two actions such as transmitting and warning, as we have considered here.

From the practical point of view, however, we believe that the outcome of our model is roughly preserved also in more complicated mechanisms. The reason being that the transmission index depends on the actual transmission rate much more strongly than on other aspects of the dynamics of the mechanism.

Apart from the congestion avoidance mechanisms on which we have focused in this work, the transmission index could be also implemented in scheduling algorithms. The use of transmission indices as weights in weighted fair queueing would imply that each flow is transmitted accordingly to its actual network price.

We could also take advantage of our approach turning it upside-down. Instead of improving congestion avoidance mechanisms for a given protocol, we could set out to develop a protocol that is economically sound given a congestion avoidance mechanism. The task would then be to formulate a protocol for which the transmission index would be constant in transmission rate. Then, the fairness considered accordingly to the packet size would be correct.

References

Altman, E. and Nain, P. (1992). Closed-loop control with delayed information. *Performance Evaluation Review*, 20(1):193–204.

Artiges, D. (1995). Optimal routing into two heterogeneous service stations with delayed information. *IEEE Transactions on Automatic Control*, 40(7):1234–1236.

Bertsimas, D. and Mersereau, A. J. (2007). A learning approach for interactive marketing to a customer segment. *Operations Research*, 55(6):1120–1135.

Bertsimas, D. and Niño-Mora, J. (1996). Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. *Mathematics of Operations Research*, 21(2):257–306.

Braden, B., Clark, D., Crowcroft, J., Davie, B., Deering, S., Estrin, D., Floyd, S., Jacobson, V., Minshall, G., Partridge, C., Peterson, L., Ramakrishnan, K., Shenker, S., Wroclawski, J., and Zhang, L. (1998). RFC 2309: Recommendations on queue management and congestion avoidance in the Internet. Available as RFC 2309. ftp://ftp.isi.edu/in-notes/rfc2309.txt.

Brooks, D. M. and Leondes, C. T. (1972). Markov decision processes with state-information lag. *Operations Research*, 20(4):904–907.

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

Campo, K. and Gijsbrechts, E. (2005). Retail assortment, shelf and stockout management: Issues, interplay and future challenges. *Applied Stochastic Models in Business and Industry*, 21:383–392.

Caro, F. and Gallien, J. (2007). Dynamic assortment with demand learning for seasonal consumer goods. *Management Science*, 53(2):276–292.

Coffman, E. G. J. and Mitrani, I. (1980). A characterization of waiting time performance realizable by single-server queues. *Operations Research*, 28(3):810–821.

Courcoubetis, C. and Weber, R. (2003). *Pricing Communication Networks: Economics, Technology and Modelling*. John Wiley & Sons Ltd, Chichester, England.

Dacre, M., Glazebrook, K., and Niño-Mora, J. (1999). The achievable region approach to the optimal control of stochastic systems. *Journal of the Royal Statistical Society, Series B*, 61(4):747–791.

Dukkipati, N., Kobayashi, M., Zhang-Shen, R., and McKeown, N. (2005). Processor sharing flows in the Internet. In *Proceedings of IWQoS '05*, Lecture Notes in Computer Science 3552, pages 271–285.

Elmaghraby, W. and Keskinocak, P. (2003). Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Management Science*, 49(10):1287–1309.

Federgruen, A. and Groenevelt, H. (1988). Characterization and optimization of achievable performance in general queueing systems. *Operations Research*, 36(5):733–741.

Feng, W., Shin, K. G., Kandlur, D. D., and Saha, D. (2002). The BLUE active queue management algorithms. *IEEE/ACM Transactions on Networking*, 10(4):513–528.

Ferragut, A. and Paganini, F. (2008). Achieving network stability and user fairness through admission control of TCP connections. In *Proceedings of 42nd Annual Conference on Information Sciences and Systems*, pages 1195–1200.

Floyd, S. and Fall, K. (1999). Promoting the use of end-to-end congestion control in the Internet. *IEEE/ACM Transactions on Networking*, 7(4):458–472.

Floyd, S. and Jacobson, V. (1993). Random early detection gateways for congestion avoidance. *IEEE/ACM Transactions on Networking*, 1(4):397–413.

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

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

Guignard, M. (2003). Lagrangean relaxation. *TOP*, 11(2):151–228.

Hajek, B. (1984). Optimal control of two interacting service stations. *IEEE Transactions on Automatic Control*, 29:491–499.

Hordijk, A. and Koole, G. (1990). On the optimality of the generalized shortest queue policy. *Probability in the Engineering and Informational Sciences*, 4:477–487.

Houck, D. J. (1987). Comparison of policies for routing customers to parallel queueing sytems. *Operations Research*, 35:306–310.

Jacko, P. and Niño-Mora, J. (2007). Time-constrained restless bandits and the knapsack problem for perishable items (extended abstract). *Electronic Notes in Discrete Mathematics*, 28:145–152.

Jacko, P. and Niño-Mora, J. (2008). Admission control and routing to parallel queues with delayed information via marginal productivity indices. In *Proceedings of the Third International Conference on Performance Evaluation Methodologies and Tools (ValueTools)*, ACM International Conference Proceeding Series. ICST Gent.

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

Kalman, D. and Mena, R. (2003). The Fibonacci numbers—exposed. *Mathematics Magazine*, 76(3):167–181.

Katabi, D., Handley, M., and Rohrs, C. (2002). Congestion control for high bandwidth-delay product networks. In *Proceedings of SIGCOMM '02*.

Kelly, F. (1997). Charging and rate control for elastic traffic. *European Transactions on Telecommunications*, 8:33–37.

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

Kök, A. G., Fisher, M., and Vaidyanathan, R. (2006). *Retail Supply Chain Management*, chapter Assortment Planning: Review of Literature and Industry Practice. Kluwer Publishers. (to appear).

Kuri, J. and Kumar, A. (1995). Optimal control of arrivals to queues with delayed queue length information. *IEEE Transactions on Automatic Control*, 40(8):1444–1450.

Lestas, M., Pitsillides, A., and Ioannou, P. (2008). Congestion control in computer networks. In *Modeling and Control of Complex Systems*. CRC Press.

Low, S. H. and Lapsley, D. E. (1999). Optimization flow control—I: Basic algorithm and convergence. *IEEE/ACM Transactions on Networking*, 7(6):861–874.

Low, S. H., Paganini, F., and Doyle, J. C. (2002). Internet congestion control. *IEEE Control Systems Magazine*, 22(1):28–43.

Low, S. H. and Srikant, R. (2004). A mathematical framework for designing a low-loss, low-delay Internet. *Networks and Spatial Economics*, 4:75–101.

Lu, L. L., Chiu, S. Y., and Cox, L. A. J. (1999). Optimal project selection: Stochastic knapsack with finite time horizon. *Journal of the Operational Research Society*, 50:645–650.

Lucas, E. (1878). Théorie des fonctions numériques simplement périodiques. *American Journal of Mathematics*, 1:184–240, 289–321.

Ma, K., Mazumdar, R., and Luo, J. (2008). On the performance of primal/dual schemes for congestion control in networks with dynamic flows. In *Proceedings of The 27th Conference on Computer Communications (IEEE INFOCOM)*, pages 960–967.

MacKie-Mason, J. K. and Varian, H. R. (1995). Pricing the internet. In Kahin, B. and Keller, J., editors, *Public Access to the Internet*. Prentice-Hall, Englewood Cliffs, New Jersey.

Manor, G. and Kress, M. (1997). Optimality of the greedy shooting strategy in the presence of incomplete damage information. *Naval Research Logistics*, 44:613–622.

Massoulié, L. and Roberts, J. (1999). Arguments in favour of admission control for TCP flows. In *Teletraffic Engineering in a Competitive World, Proceedings of ITC 16*, pages 33–44.

Molle, M. and Xu, Z. (2005). Short-circuiting the congestion signaling path for AQM algorithms using reverse flow matching. *Computer Communications*, 28:2082–2093.

Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica*, 37(1):15–24.

Neely, M. J., Modiano, E., and Li, C.-P. (2008). Fairness and optimal stochastic control for heterogeneous networks. *IEEE/ACM Transactions on Networking*, 16(2):396–409.

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

Niño-Mora, J. (2002). Dynamic allocation indices for restless projects and queueing admission control: A polyhedral approach. *Mathematical Programming, Series A*, 93(3):361–413.

Niño-Mora, J. (2005). Marginal productivity index policies for the finite-horizon multiarmed bandit problem. In *Proceedings of the 44th IEEE Conference on Decision and Control and European Control Conference ECC 2005 (CDC-ECC '05)*, pages 1718–1722.

Niño-Mora, J. (2006a). Marginal productivity index policies for scheduling a multiclass delay-/loss-sensitive queue. *Queueing Systems*, 54:281–312.

Niño-Mora, J. (2006b). 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(1):50–84.

Niño-Mora, J. (2007a). Characterization and computation of restless bandit marginal productivity indices. In *Proceedings of the 2nd International Conference on Performance Evaluation Methodologies and Tools*. ICST, Brussels, Belgium.

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

Niño-Mora, J. (2007c). Marginal productivity index policies for admission control and routing to parallel multi-server loss queues with reneging. *Lecture Notes in Computer Science*, 4465:138–149.

Niño-Mora, J. (2007d). Marginal productivity index policies for scheduling multiclass delay/loss-sensitive traffic with delayed state observation. In *Proceedings of the 3rd EuroNGI Conference on Next Generation Internet Networks - Design and Engineering for Heterogeneity*, pages 209–217.

Niño-Mora, J. (2009). Conservation laws and related approaches. In *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons.

Paganini, F. (2006). Congestion control with adaptive multipath routing based on optimization. In *Proceedings of the Conference on Information Sciences and Systems*, pages 333–338.

Papadimitriou, C. H. and Tsitsiklis, J. N. (1999). The complexity of optimal queueing network. *Mathematics of Operations Research*, 24(2):293–305.

Papastavrou, J. D., Rajagopalan, S., and Kleywegt, A. J. (1996). The dynamic and stochastic knapsack problem with deadlines. *Management Science*, 42(12):1706–1718.

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

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

Saaty, T. and Gass, S. (1954). Parametric objective function (part 1). *Journal of the Operations Research Society of America*, 2(3):316–319.

Shanthikumar, J. G. and Yao, D. D. (1992). Multiclass queueing systems: Polymatroidal structure and optimal scheduling control. *Operations Research*, 40(Supplement 2: Stochastic Processes):S293–S299.

Shenker, S., Clark, D., Estrin, D., and Herzog, S. (1996). Pricing in computer networks: Reshaping the research agenda. *Telecommunications Policy*, 20(3):183–201.

Smith, W. E. (1956). Various optimizers for single-stage production. *Naval Research Logistics Quarterly*, 3(1-2):59–66.

Srikant, R. (2004). *The Mathematics of Internet Congestion Control*. Birkhäuser Boston, Cambridge, Massachusetts.

Stidham, S. J. (2002). Analysis, design, and control of queueing systems. *Operations Research*, 50(1):197–216.

Stidham, S. J. (2004). Pricing and congestion management in a network with heterogeneous users. *IEEE Transactions on Automatic Control*, 49(6):976–981.

Tsoucas, P. (1991). The region of achievable performance in a model of klimov. Technical Report RC16543, IBM T. J. Watson Research Center, Yorktown Heights, New York.

Visweswaran, V. (2009). Decomposition techniques for milp: Lagrangian relaxation. In Floudas, C. and Pardalos, P., editors, *Encyclopedia of Optimization*, page 38183824. Springer, New York, 2nd edition edition.

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

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

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

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

Winston, W. (1977). Optimality of the shortest line discipline. *Journal of Applied Probability*, 14:181–189.

Wischik, D. (1999). How to mark fairly. manuscript.

Appendix A

Appendix to Chapter 3

A.1 Pure-Active-Action Normalization

In this section we show that the ν-wage problem can be transformed into an equivalent problem with normalized active-action parametric rewards and zero passive-action (parametric) rewards. Such a normalization typically leads to a simpler solution procedure for the problem.

The ν-wage problem (3.7) starting from initial state $i \in \mathcal{N}$ admits a standard LP formulation arising as the dual of the LP formulation of its Bellman equations (cf. Niño-Mora, 2002, Section 6). The LP formulation (under the β-discounted criterion) is the following

$$\begin{aligned} v^{\mathrm{LP}}(\nu) := \max_{\boldsymbol{x}^0, \boldsymbol{x}^1} \quad & \boldsymbol{x}^0(\boldsymbol{R}^0 - \nu \boldsymbol{W}^0) + \boldsymbol{x}^1(\boldsymbol{R}^1 - \nu \boldsymbol{W}^1) \\ \text{subject to} \quad & \boldsymbol{x}^0(\boldsymbol{I} - \beta \boldsymbol{P}^0) + \boldsymbol{x}^1(\boldsymbol{I} - \beta \boldsymbol{P}^1) = \boldsymbol{e}_i \\ & \boldsymbol{x}^0, \boldsymbol{x}^1 \geq \boldsymbol{0} \end{aligned} \tag{LP}$$

where all vectors' subscripts run over $\mathcal{N}$ and those of matrices run over $\mathcal{N} \times \mathcal{N}$. Here, $\boldsymbol{e}_i$ is the i-th unit coordinate vector defining the initial state and $\boldsymbol{I}$ is an identity matrix. The variable x_n^a has the interpretation as the *state-action frequency measure* denoting the expected total β-discounted amount of time when action a is taken in state n.

If the matrix $\boldsymbol{I} - \beta \boldsymbol{P}^0$ is invertible, let us define the following pure-active-action

normalized LP formulation:

$$\begin{aligned} v^{\text{NLP}}(\nu) := \max_{\boldsymbol{x}} \quad & \boldsymbol{x}(\boldsymbol{R} - \nu \boldsymbol{W}) \\ \text{subject to} \quad & \boldsymbol{e}_i(\boldsymbol{I} - \beta \boldsymbol{P}^0)^{-1} - \boldsymbol{x}(\boldsymbol{I} - \beta \boldsymbol{P}^1)(\boldsymbol{I} - \beta \boldsymbol{P}^0)^{-1} \geq \boldsymbol{0} \\ & \boldsymbol{x} \geq \boldsymbol{0} \end{aligned} \tag{NLP}$$

where

$$\boldsymbol{R} := \boldsymbol{R}^1 - (\boldsymbol{I} - \beta \boldsymbol{P}^1)(\boldsymbol{I} - \beta \boldsymbol{P}^0)^{-1}\boldsymbol{R}^0, \quad \boldsymbol{W} := \boldsymbol{W}^1 - (\boldsymbol{I} - \beta \boldsymbol{P}^1)(\boldsymbol{I} - \beta \boldsymbol{P}^0)^{-1}\boldsymbol{W}^0. \tag{A.1}$$

It was shown in Niño-Mora (2001, 2002) that the two formulations are equivalent, as stated in the following proposition. Let $v^{\emptyset}(\nu) := \boldsymbol{e}_i(\boldsymbol{I} - \beta \boldsymbol{P}^0)^{-1}(\boldsymbol{R}^0 - \nu \boldsymbol{W}^0)$ be the (LP) objective value of the *pure-passive-action policy*, i.e., the policy taking the passive action in all states, for which by definition $\boldsymbol{x}^1 = \boldsymbol{0}$.

Proposition A.1. *Let the wage parameter ν be fixed and suppose that $\boldsymbol{I} - \beta \boldsymbol{P}^0$ is invertible. A solution $\boldsymbol{x}$ is an optimal solution to* (NLP) *if and only if*

$$(\boldsymbol{x}^0, \boldsymbol{x}^1) := \big(\boldsymbol{e}_i(\boldsymbol{I} - \beta \boldsymbol{P}^0)^{-1} - \boldsymbol{x}(\boldsymbol{I} - \beta \boldsymbol{P}^1)(\boldsymbol{I} - \beta \boldsymbol{P}^0)^{-1}, \boldsymbol{x}\big)$$

is an optimal solution to (LP). *We further have $v^{LP}(\nu) = v^{NLP}(\nu) + v^{\emptyset}(\nu)$.*

Appendix B

Appendix to Chapter 4

B.1 Admission Control Problem with Delay: Marginal Work Analysis

In this section we set out to obtain closed formulae for marginal works in the admission control problem with delay and present results in terms of action-differences in total work that will facilitate the proof of their positivity.

B.1.1 Preliminaries

Recall the definition of the expected total β-discounted work (briefly, *total work*) in (4.11). The following is the work balance equation for a fixed active set $\mathcal{S}$:

$$g^{\mathcal{S}}_{(a,i)} = \begin{cases} W_{(a,i)} + \beta \sum_{j\in\mathcal{I}} p^a_{ij} g^{\mathcal{S}}_{(1,j)} & \text{if } (a,i) \in \mathcal{S} \\ W_{(a,i)} + \beta \sum_{j\in\mathcal{I}} p^a_{ij} g^{\mathcal{S}}_{(0,j)} & \text{if } (a,i) \notin \mathcal{S} \end{cases} \tag{B.1}$$

The following two implications of the work balance equation were found useful in the problem analysis. First, we give a characterization of total works $g^{\mathcal{S}}_{(a,i)}$'s in terms of their action-differences $\Delta_1 g^{\mathcal{S}}_{(1,i)}$'s and state-differences $\Delta_2 g^{\mathcal{S}}_{(a,i)}$'s.

Lemma B.1. *For a fixed active set $\mathcal{S}$,*

$$(1-\beta)g^{\mathcal{S}}_{(1,i)} = \lambda - \beta\mu\Delta_2 g^{\mathcal{S}}_{(1,i)} \qquad \text{if } (1,i)\in\mathcal{S} \quad \text{(B.2)}$$

$$(1-\beta)g^{\mathcal{S}}_{(1,i)} = \lambda - \beta\mu\Delta_2 g^{\mathcal{S}}_{(0,i)} - \beta\Delta_1 g^{\mathcal{S}}_{(1,i)} \qquad \text{if } (1,i)\notin\mathcal{S} \quad \text{(B.3)}$$

$$(1-\beta)g^{\mathcal{S}}_{(0,i)} = \begin{cases} \lambda - \beta\mu\Delta_2 g^{\mathcal{S}}_{(1,I)} + \beta\Delta_1 g^{\mathcal{S}}_{(1,I)} & \text{if } i = I \\ \beta\zeta\Delta_2 g^{\mathcal{S}}_{(1,i+1)} - \beta\eta\Delta_2 g^{\mathcal{S}}_{(1,i)} + \beta\Delta_1 g^{\mathcal{S}}_{(1,i)} & \text{otherwise} \end{cases} \qquad \text{if } (0,i)\in\mathcal{S} \quad \text{(B.4)}$$

$$(1-\beta)g^{\mathcal{S}}_{(0,i)} = \begin{cases} \lambda - \beta\mu\Delta_2 g^{\mathcal{S}}_{(0,I)} & \text{if } i = I \\ \beta\zeta\Delta_2 g^{\mathcal{S}}_{(0,i+1)} - \beta\eta\Delta_2 g^{\mathcal{S}}_{(0,i)} & \text{otherwise} \end{cases} \qquad \text{if } (0,i)\notin\mathcal{S} \quad \text{(B.5)}$$

Proof. Suppose first that $(1,i)\in\mathcal{S}$. By adding $-\beta g^{\mathcal{S}}_{(1,i)}$ at both sides of identity (B.1) for $(1,i)$, we obtain

$$(1-\beta)g^{\mathcal{S}}_{(1,i)} = W_{(1,i)} - \beta\sum_{j\in\mathcal{I}} p^1_{ij}\left(g^{\mathcal{S}}_{(1,i)} - g^{\mathcal{S}}_{(1,j)}\right),$$

which simplifies to (B.2) after plugging the definition of $W_{(1,i)}$ in (4.10) and that of p^1_{ij} in (4.5)-(4.6), and finally using (4.23).

The remaining identities are obtained analogously by adding $-\beta g^{\mathcal{S}}_{(a,i)}$ at both sides of identity (B.1), then plugging the definition of $W_{(a,i)}$ in (4.10) and that of p^a_{ij} in (4.5)-(4.6), and finally using (4.22)-(4.23). □

The following lemma characterizes action-differences $\Delta_1 g^{\mathcal{S}}_{(1,i)}$'s in terms of state-differences $\Delta_2 g^{\mathcal{S}}_{(a,i)}$'s.

Lemma B.2. *For a fixed active set $\mathcal{S}$ and any state $0\le i\le I-1$,*

$$\Delta_1 g^{\mathcal{S}}_{(1,i)} = \lambda - \beta\zeta\Delta_2 g^{\mathcal{S}}_{(1,i+1)} - \beta(\mu-\eta)\Delta_2 g^{\mathcal{S}}_{(1,i)} \qquad \text{if } (0,i),(1,i)\in\mathcal{S} \quad \text{(B.6)}$$

$$\Delta_1 g^{\mathcal{S}}_{(1,i)} = \lambda - \beta\zeta\Delta_2 g^{\mathcal{S}}_{(0,i+1)} - \beta(\mu-\eta)\Delta_2 g^{\mathcal{S}}_{(0,i)} \qquad \text{if } (0,i),(1,i)\notin\mathcal{S} \quad \text{(B.7)}$$

$$(1+\beta)\Delta_1 g^{\mathcal{S}}_{(1,i)} = \lambda - \beta\zeta\Delta_2 g^{\mathcal{S}}_{(1,i+1)} - \beta\mu\Delta_2 g^{\mathcal{S}}_{(0,i)} + \beta\eta\Delta_2 g^{\mathcal{S}}_{(1,i)} \qquad \text{if } (0,i)\in\mathcal{S}, (1,i)\notin\mathcal{S} \quad \text{(B.8)}$$

$$(1-\beta)\Delta_1 g^{\mathcal{S}}_{(1,i)} = \lambda - \beta\zeta\Delta_2 g^{\mathcal{S}}_{(0,i+1)} - \beta\mu\Delta_2 g^{\mathcal{S}}_{(1,i)} + \beta\eta\Delta_2 g^{\mathcal{S}}_{(0,i)} \qquad \text{if } (0,i)\notin\mathcal{S}, (1,i)\in\mathcal{S} \quad \text{(B.9)}$$

$$\Delta_1 g^{\mathcal{S}}_{(1,I)} = 0. \quad \text{(B.10)}$$

Proof. If $(0,i),(1,i)\in\mathcal{S}$, identity (B.6) is obtained by subtracting (B.4) from (B.2), and using (4.22). The remaining identities are obtained analogously. □

Recall the definition of marginal works $w^{\mathcal{S}}_{(a,i)}$ in (4.16). In order to obtain a character-

ization of marginal works in terms of $\Delta_1 g_{(1,i)}$'s, we specialize the work balance equation in (B.1) for policies $\langle 0, \mathcal{S}\rangle$ and $\langle 1, \mathcal{S}\rangle$.

Lemma B.3. *For a fixed active set $\mathcal{S}$,*

$$g^{\langle 1,\mathcal{S}\rangle}_{(a,i)} = W_{(a,i)} + \beta \sum_{j\in\mathcal{I}} p^a_{ij} g^{\mathcal{S}}_{(1,j)}$$
$$g^{\langle 0,\mathcal{S}\rangle}_{(a,i)} = W_{(a,i)} + \beta \sum_{j\in\mathcal{I}} p^a_{ij} g^{\mathcal{S}}_{(0,j)}$$

Now we are ready to express marginal works $w^{\mathcal{S}}_{(a,i)}$ in terms of action-differences $\Delta_1 g^{\mathcal{S}}_{(1,i)}$'s.

Lemma B.4. *For a fixed active set $\mathcal{S}$,*

$$w^{\mathcal{S}}_{(1,i)} = \begin{cases} \beta \Delta_1 g^{\mathcal{S}}_{(1,0)} & \text{if } i = 0 \\ \beta\mu \Delta_1 g^{\mathcal{S}}_{(1,i-1)} + \beta(1-\mu)\Delta_1 g^{\mathcal{S}}_{(1,i)} & \text{otherwise} \end{cases} \tag{B.11}$$

$$w^{\mathcal{S}}_{(0,i)} = \begin{cases} \beta(1-\zeta)\Delta_1 g^{\mathcal{S}}_{(1,0)} + \beta\zeta \Delta_1 g^{\mathcal{S}}_{(1,1)} & \text{if } i = 0 \\ \beta\eta \Delta_1 g^{\mathcal{S}}_{(1,i-1)} + \beta\varepsilon \Delta_1 g^{\mathcal{S}}_{(1,i)} + \beta\zeta \Delta_1 g^{\mathcal{S}}_{(1,i+1)} & \text{otherwise} \\ w^{\mathcal{S}}_{(1,I)} & \text{if } i = I \end{cases} \tag{B.12}$$

Proof. From plugging the identities in Lemma B.3 into the definition of $w^{\mathcal{S}}_{(a,i)}$, we obtain

$$w^{\mathcal{S}}_{(a,i)} = \beta \sum_{j\in\mathcal{I}} p^a_{ij} \Delta_1 g^{\mathcal{S}}_{(1,j)}.$$

Then, using the definition of p^a_{ij} in (4.5)-(4.6) gives the result. □

The last lemma shows that marginal work is equal to the *expected next-period β-discounted* increment in total work if starting from state $(1, i)$ instead of $(0, i)$, i.e.,

$$w^{\mathcal{S}}_{(a,i)} = \mathbb{E}^a_i \left[\Delta_1 g^{\mathcal{S}}\right], \tag{B.13}$$

where the random variable $\Delta_1 g^{\mathcal{S}}$ has value $\Delta_1 g^{\mathcal{S}}_{(1,j)}$ with probability p^a_{ij}. This suggests a way for establishing positivity of marginal works needed in Definition 4.1(i) by establishing positivity of $\Delta_1 g^{\mathcal{S}}_{(1,i)}$'s for all states $0 \le i \le I-1$ (recall that by (B.10), $\Delta_1 g^{\mathcal{S}}_{(1,i)} = 0$ for state I).

Before calculating the action-differences needed above, we prepare notation and state an auxiliary result. These quantities will appear in the action-differences recur-

sion developed in the following subsection. For $j \geq 0$ we define

$$A_0 := 0, \qquad A_{j+1} := \frac{\beta\zeta}{1-\beta+\beta\zeta+\beta\eta(1-A_j)}, \qquad B := \frac{\beta\mu}{1-\beta+\beta\mu}. \tag{B.14}$$

and

$$A'_j := \beta\zeta + \beta(\mu-\eta)A_j, \qquad Z_{j+1} := A_{j+1}\frac{A'_j}{A'_{j+1}}, \qquad B' := \beta\zeta B + \beta(\mu-\eta). \tag{B.15}$$

Lemma B.5.

(i) $0 < B < \beta$*;*

(ii) $0 < A_{j+1} < \beta$ *and* $A_j \leq A_{j+1}$ *for all* $j \geq 0$*;*

(iii) $0 < Z_{j+1} < \beta$ *for all* $j \geq 0$*.*

Proof. (i) The positivity is straightforward from the definition of B in (B.14), because $\beta\mu > 0$ and $1-\beta > 0$ by the model parameter assumptions given in (4.7). On the other hand, the same assumptions imply $B < \beta$. □

(ii) We proceed by induction. Since $\beta\eta(1-A_0) \geq 0$, we have $A_1 \leq \frac{\beta\zeta}{1-\beta+\beta\zeta} < \beta$, where the last inequality is due to the model parameter assumptions. Hence, assuming inductively $\beta\eta(1-A_j) \geq 0$, we have $A_{j+1} \leq \frac{\beta\zeta}{1-\beta+\beta\zeta} < \beta$. The positivity of A_{j+1} follows from $\beta\zeta > 0$, $1-\beta > 0$ and $\beta\eta(1-A_j) \geq 0$.

Similarly by induction we prove the monotonicity. As the first step, the above implies $0 = A_0 < A_1$. Hence we have $\beta\eta(1-A_0) \geq \beta\eta(1-A_1)$, which implies $A_1 \leq A_2$. Inductively, assuming $A_{j-1} \leq A_j$ analogously implies $A_j \leq A_{j+1}$. □

(iii) The model parameter assumptions given in (4.7) imply that $A'_j > 0$, since $\mu > \eta$ for all $j \geq 0$. Therefore, the definition of Z_{j+1} in (B.15) implies $Z_{j+1} > 0$. On the other hand, using (B.15) we can write

$$Z_{j+1} = \frac{\beta\zeta+\beta(\mu-\eta)A_j}{\frac{\beta\zeta}{A_{j+1}}+\beta(\mu-\eta)} < \frac{\beta\zeta+\beta(\mu-\eta)\beta}{\frac{\beta\zeta}{\beta}+\beta(\mu-\eta)} = \beta,$$

where the inequality is due to $A_j, A_{j+1} < \beta$ by (ii). □

B.1.2 Calculation of Action-Differences in Total Work

Since Lemma B.4 characterizes marginal works $w^{\mathcal{S}}_{(a,i)}$'s as weighted averages of action-differences $\Delta_1 g^{\mathcal{S}}_{(1,i)}$'s, in this subsection we focus on the calculation of the latter. The

ultimate goal of proving positivity of marginal works under $\mathcal{F}$ in order to establish condition Definition 4.1(i), will be reached by proving positivity of action-differences in subsection B.1.3 and subsection B.1.4 for active sets $\widetilde{\mathcal{I}}_{K,K}$ and $\widetilde{\mathcal{I}}_{K,K+1}$, respectively.

In the following we show that all the relevant state-differences can be obtained by recursion from two *pivot state-differences* associated to the two thresholds K_0, K_1 under any policy $\widetilde{\mathcal{I}}_{K_0,K_1}$. Denote by $K_0' := \min\{K_0, I\}$ and note that the relevant state-differences needed in the subsequent analysis are $\Delta_2 g^{\mathcal{S}}_{(0,i)}$'s for $1 \le i \le K_0' - 1$, and $\Delta_2 g^{\mathcal{S}}_{(1,i)}$ for $K_1 + 1 \le i \le I$.

Lemma B.6. *For a fixed active set $\mathcal{S} = \widetilde{\mathcal{I}}_{K_0,K_1}$,*

$$\Delta_2 g^{\mathcal{S}}_{(0,i)} = \Delta_2 g^{\mathcal{S}}_{(0,K_0')} \prod_{j=i}^{K_0'-1} A_j, \qquad \text{for } 1 \le i \le K_0' - 1, \tag{B.16}$$

$$\Delta_2 g^{\mathcal{S}}_{(1,i)} = \Delta_2 g^{\mathcal{S}}_{(1,K_1)} B^{i-K_1}, \qquad \text{for } K_1 + 1 \le i \le I. \tag{B.17}$$

Proof. For $1 \le i \le K_0' - 1$, augmented states $(0,i), (0,i-1) \notin \mathcal{S}$, so taking the difference of (B.5) for i and for $i-1$ gives

$$(1 - \beta + \beta\eta + \beta\zeta)\Delta_2 g^{\mathcal{S}}_{(0,i)} = \beta\eta \Delta_2 g^{\mathcal{S}}_{(0,i-1)} + \beta\zeta \Delta_2 g^{\mathcal{S}}_{(0,i+1)}.$$

Expressed for $i = 1$ and divided by $1 - \beta + \beta\eta + \beta\zeta$, we have $\Delta_2 g^{\mathcal{S}}_{(0,1)} = \Delta_2 g^{\mathcal{S}}_{(0,2)} A_1$, since, by definition,

$$\Delta_2 g^{\mathcal{S}}_{(0,0)} = 0 \qquad \text{and} \qquad A_1 = \frac{\beta\zeta}{1 - \beta + \beta\zeta + \beta\eta}.$$

Inductively, if $\Delta_2 g^{\mathcal{S}}_{(0,i-1)} = \Delta_2 g^{\mathcal{S}}_{(0,i)} A_{i-1}$, then

$$(1 - \beta + \beta\eta + \beta\zeta(1 - A_{i-1}))\Delta_2 g^{\mathcal{S}}_{(0,i)} = \beta\zeta \Delta_2 g^{\mathcal{S}}_{(0,i+1)}.$$

which is the same as $\Delta_2 g^{\mathcal{S}}_{(0,i)} = \Delta_2 g^{\mathcal{S}}_{(0,i+1)} A_i$ for all $1 \le i \le K_0' - 1$. This recursion gives (B.16).

Similarly for $K_1 + 1 \le i \le I$, augmented states $(1,i), (1,i-1) \in \mathcal{S}$, so taking the difference of (B.2) for i and for $i-1$ gives $\Delta_2 g^{\mathcal{S}}_{(1,i)} = \Delta_2 g^{\mathcal{S}}_{(1,i-1)} B$. This recursion gives (B.17). □

Further we identify a recursion to calculate action-differences $\Delta_1 g^{\mathcal{S}}_{(1,i)}$'s in terms of the two pivot state-differences $\Delta_2 g^{\mathcal{S}}_{(0,K_0')}$ and $\Delta_2 g^{\mathcal{S}}_{(1,K_1)}$. Thus, this is a simplification of Lemma B.2.

Proposition B.1. *For a fixed active set* $\mathcal{S} = \widetilde{\mathcal{I}}_{K_0,K_1}$,

$$\Delta_1 g^{\mathcal{S}}_{(1,K_0'-1)} = \lambda - A'_{K_0'-1}\Delta_2 g^{\mathcal{S}}_{(0,K_0')}, \qquad \text{if } 1 \le K_0', \tag{B.18}$$

$$\Delta_1 g^{\mathcal{S}}_{(1,i)} = \lambda(1 - Z_{i+1}) + \Delta_1 g^{\mathcal{S}}_{(1,i+1)} Z_{i+1}, \qquad \text{for } 0 \le i \le K_0' - 2, \tag{B.19}$$

$$\Delta_1 g^{\mathcal{S}}_{(1,K_1)} = \lambda - B'\Delta_2 g^{\mathcal{S}}_{(1,K_1)}, \qquad \text{if } K_1 \le I - 1, \tag{B.20}$$

$$\Delta_1 g^{\mathcal{S}}_{(1,i)} = \lambda(1 - B) + \Delta_1 g^{\mathcal{S}}_{(1,i-1)} B, \qquad \text{for } K_1 + 1 \le i \le I - 1. \tag{B.21}$$

Proof. As a consequence of plugging (B.16) into (B.7) we have

$$\Delta_1 g^{\mathcal{S}}_{(1,i)} = \lambda - A'_i \Delta_2 g^{\mathcal{S}}_{(0,K_0')} \prod_{j=i+1}^{K_0'-1} A_j, \qquad \text{for } 0 \le i \le K_0' - 1,$$

This identity expressed for $i = K_0' - 1$ gives (B.18), and expressed for i and $i+1$ implies (B.19).

Similarly, by plugging (B.17) into (B.6) we have

$$\Delta_1 g^{\mathcal{S}}_{(1,i)} = \lambda - B'\Delta_2 g^{\mathcal{S}}_{(1,K_1)} B^{i-K_1}, \qquad \text{for } K_1 \le i \le I - 1.$$

This identity expressed for $i = K_1$ gives (B.20), and expressed for i and $i-1$ implies (B.21). □

The above results help significantly simplify the subsequent analysis, which we present in separate subsections for optimal active sets $\widetilde{\mathcal{I}}_{K,K}$ and $\widetilde{\mathcal{I}}_{K,K+1}$. In each subsection we first present expressions for the pivot state-differences in a lemma and then establish positivity of $\Delta_1 g^{\mathcal{S}}_{(1,i)} > 0$ for all $0 \le i \le I - 1$.

B.1.3 Positivity of Action-Differences in Total Work under Active Set $\widetilde{\mathcal{I}}_{K,K}$

Lemma B.7. *Under active set* $\mathcal{S} = \widetilde{\mathcal{I}}_{K,K}$,

(i) if $K = 0$, *then*

$$\Delta_2 g^{\mathcal{S}}_{(1,0)} = 0. \tag{B.22}$$

(ii) if $1 \le K \le I - 1$, *then*

$$\Delta_2 g^{\mathcal{S}}_{(0,K)} = \frac{\beta\lambda\,(1 + B')}{\frac{A'_K}{A_K} + \beta A'_{K-1} B' + \beta\zeta\beta\mu(1 - BA_{K-1})}, \tag{B.23}$$

$$\Delta_2 g^{\mathcal{S}}_{(1,K)} = \frac{1 + A'_{K-1}}{1 + B'} \Delta_2 g^{\mathcal{S}}_{(0,K)}. \tag{B.24}$$

(iii) if $K = I$, then

$$\Delta_2 g^{\mathcal{S}}_{(0,I)} = \frac{\lambda(1+\beta\mu)}{\frac{A'_I}{A_I} + \beta\mu A'_{I-1}}. \tag{B.25}$$

(iv) if $K = I+1$, then

$$\Delta_2 g^{\mathcal{S}}_{(0,I)} = \frac{\lambda A_I}{A'_I}. \tag{B.26}$$

Proof. (i) The identity is by definition. □

(ii) Taking the difference of (B.2) for $i = K$ and (B.3) for $i = K-1$, and plugging (B.7) for $i = K-1$, gives

$$(1-\beta+\beta\mu)\Delta_2 g^{\mathcal{S}}_{(1,K)} = \beta\lambda + \beta\left[\mu - \beta(\mu-\eta)\right]\Delta_2 g^{\mathcal{S}}_{(0,K-1)} - \beta^2\zeta\Delta_2 g^{\mathcal{S}}_{(0,K)}. \tag{B.27}$$

Similarly, taking the difference of (B.4) for $i = K$ and (B.5) for $i = K-1$, and plugging (B.6) for $i = K$, gives

$$\begin{aligned}(1-\beta+\beta\zeta)\Delta_2 g^{\mathcal{S}}_{(0,K)} &= \beta\lambda + (1-\beta)\beta\zeta\Delta_2 g^{\mathcal{S}}_{(1,K+1)} \\ &\quad - \beta\left[\eta + \beta(\mu-\eta)\right]\Delta_2 g^{\mathcal{S}}_{(1,K)} + \beta\eta\Delta_2 g^{\mathcal{S}}_{(0,K-1)}.\end{aligned} \tag{B.28}$$

Using (B.16) for $i = K-1$ and (B.17) for $i = K+1$, and solving the above system of two equations yields the results. □

(iii) The proof goes along the same lines as in the previous case, yet the latter identity becomes

$$(1-\beta+\beta\zeta)\Delta_2 g^{\mathcal{S}}_{(0,I)} = \lambda - \beta\mu\Delta_2 g^{\mathcal{S}}_{(1,I)} + \beta\eta\Delta_2 g^{\mathcal{S}}_{(0,I-1)}. \tag{B.29}$$

□

(iv) Taking the difference of (B.5) for $i = I$ and for $i = I-1$, and plugging (B.16) for $i = I-1$, yields the result. □

Proposition B.2. *Under active set $\mathcal{S} = \widetilde{\mathcal{I}}_{K,K}$ with $0 \le K \le I+1$, action-differences $\Delta_1 g^{\mathcal{S}}_{(1,i)} > 0$ for all $0 \le i \le I-1$ and $\Delta_1 g^{\mathcal{S}}_{(1,I)} = 0$.*

Proof. We will divide the proof into three steps, in which we prove the following:

(i) if $1 \le K' := \min\{K, I\}$, then action-difference $\Delta_1 g^{\mathcal{S}}_{(1,K'-1)} > 0$;

(ii) if $K \leq I-1$, then action-difference $\Delta_1 g^{\mathcal{S}}_{(1,K)} > 0$;

(iii) action-differences $\Delta_1 g^{\mathcal{S}}_{(1,i)} > 0$ for all $0 \leq i \leq I-1$ and $\Delta_1 g^{\mathcal{S}}_{(1,I)} = 0$.

(i) Suppose first that $1 \leq K \leq I-1$. Identity (B.18) gives

$$\Delta_1 g^{\mathcal{S}}_{(1,K-1)} = \lambda - A'_{K-1} \Delta_2 g^{\mathcal{S}}_{(0,K)}.$$

In order to show $\Delta_1 g^{\mathcal{S}}_{(1,K-1)} > 0$, using (B.23) we need to have

$$\lambda > \frac{\beta\lambda\,(1+B')\,A'_{K-1}}{\frac{A'_K}{A_K} + \beta A'_{K-1} B' + \beta\zeta\beta\mu(1-BA_{K-1})}.$$

Since the denominator is positive due to Lemma B.5, this is equivalent to

$$\beta A'_{K-1} < \frac{A'_K}{A_K} + \beta\zeta\beta\mu(1-BA_{K-1}).$$

This is true, because Lemma B.5(iii) with $\beta < 1$ implies $\beta A'_{K-1} < \frac{A'_K}{A_K}$, and Lemma B.5(i)-(ii) implies $\beta\zeta\beta\mu(1-BA_{K-1}) > 0$.

For $K = I$ and $K = I+1$, we have $\Delta_1 g^{\mathcal{S}}_{(1,I-1)} = \lambda - A'_{I-1}\Delta_2 g^{\mathcal{S}}_{(0,I)}$ as above. In order to show $\Delta_1 g^{\mathcal{S}}_{(1,I-1)} > 0$, using (B.25) for $K = I$ we need to have

$$\lambda > \frac{\lambda(1+\beta\mu)A'_{I-1}}{\frac{A'_I}{A_I} + \beta\mu A'_{I-1}}, \qquad \text{which is equivalent to} \qquad 1 > \frac{A'_{I-1}}{\frac{A'_I}{A_I}},$$

which is true by Lemma B.5(iii).

Finally, using (B.26) for $K = I+1$ we need to have

$$\lambda > \frac{\lambda A_I A'_{I-1}}{A'_I}, \qquad \text{which is equivalent to} \qquad 1 > \frac{A_I A'_{I-1}}{A'_I},$$

which is again true by Lemma B.5(iii). □

(ii) Similarly, for $1 \leq K \leq I-1$ identity (B.20) gives

$$\Delta_1 g^{\mathcal{S}}_{(1,K)} = \lambda - B' \Delta_2 g^{\mathcal{S}}_{(1,K)}.$$

In order to show $\Delta_1 g^{\mathcal{S}}_{(1,K)} > 0$, using (B.24) we need to have

$$\lambda > \frac{\beta\lambda\left(1+A'_{K-1}\right)B'}{\frac{A'_K}{A_K} + \beta A'_{K-1} B' + \beta\zeta\beta\mu(1-BA_{K-1})},$$

which is equivalent to

$$\beta B' < \frac{A'_K}{A_K} + \beta\zeta\beta\mu(1 - BA_{K-1}).$$

This is true, because $\beta B' < \frac{A'_K}{A_K}$ using the definitions of B' and A'_K in (B.15) and properties from Lemma B.5.

Finally, for $K = 0$, plugging (B.22) into (B.6) gives $\Delta_1 g^{\mathcal{S}}_{(1,0)} = \lambda > 0$. □

(iii) We will show that positivity of action-difference $\Delta_1 g^{\mathcal{S}}_{(1,K'-1)}$ implies $\Delta_1 g^{\mathcal{S}}_{(1,i)} > 0$ for all $0 \leq i \leq K'-1$, and positivity of action-difference $\Delta_1 g^{\mathcal{S}}_{(1,K)}$ implies $\Delta_1 g^{\mathcal{S}}_{(1,i)} > 0$ for all $K \leq i \leq I-1$.

Recursion (B.19) shows that $\Delta_1 g^{\mathcal{S}}_{(1,i)}$ is a weighted average of $\lambda > 0$ and $\Delta_1 g^{\mathcal{S}}_{(1,i+1)}$ for all $0 \leq i \leq K' - 2$ (the weights are between 0 and 1 due to Lemma B.5). Since state-difference $\Delta_1 g^{\mathcal{S}}_{(1,K'-1)} > 0$ by (i), by induction we obtain $\Delta_1 g^{\mathcal{S}}_{(1,i)} > 0$ for all $0 \leq i \leq K' - 1$.

Similarly, recursion (B.21) shows that $\Delta_1 g^{\mathcal{S}}_{(1,i)}$ is a weighted average of $\lambda > 0$ and $\Delta_1 g^{\mathcal{S}}_{(1,i-1)}$ for all $K + 1 \leq i \leq I - 1$ (the weights are between 0 and 1 due to Lemma B.5). Since state-difference $\Delta_1 g^{\mathcal{S}}_{(1,K)} > 0$ by (ii), by induction we obtain $\Delta_1 g^{\mathcal{S}}_{(1,i)} > 0$ for all $K \leq i \leq I - 1$.

In summary, we have shown that $\Delta_1 g^{\mathcal{S}}_{(1,i)} > 0$ for all $0 \leq i \leq I - 1$. Finally, $\Delta_1 g^{\mathcal{S}}_{(1,I)} = 0$ by (B.10). □

B.1.4 Positivity of Action-Differences in Total Work under Active Set $\widetilde{\mathcal{I}}_{K,K+1}$

Lemma B.8. *Under active set $\mathcal{S} = \widetilde{\mathcal{I}}_{K,K+1}$,*

(i) if $1 \leq K \leq I - 1$, then

$$\Delta_2 g^{\mathcal{S}}_{(0,K)} = \frac{\beta\lambda\left[1 - \beta(1 - \zeta - \mu)\right]}{\left[1 - \beta^2(1 - \zeta - \mu)\right]\frac{A'_K}{A_K} + \beta\eta\beta(1 - \mu)\left[1 - \beta + \beta\mu(1 - A_{K-1})\right]}, \tag{B.30}$$

$$\Delta_2 g^{\mathcal{S}}_{(1,K+1)} = \frac{\left[1 - \beta(1 - \zeta - \mu)\right] + \beta\eta\left[\beta - A_{K-1} - \beta\mu(1 - A_{K-1})\right]}{1 - \beta(1 - \zeta - \mu)}\Delta_2 g^{\mathcal{S}}_{(0,K)}, \tag{B.31}$$

$$\Delta_1 g^{\mathcal{S}}_{(1,K)} = \lambda - \frac{A'_K}{A_K}\Delta_2 g^{\mathcal{S}}_{(0,K)}. \tag{B.32}$$

(ii) if $K = 0$, then

$$\Delta_1 g^{\mathcal{S}}_{(1,0)} = \lambda \frac{1-\beta+\beta\mu}{1-\beta^2(1-\zeta-\mu)+\beta\mu}, \tag{B.33}$$

$$\Delta_1 g^{\mathcal{S}}_{(1,1)} = \lambda \frac{1-\beta^2(1-\zeta-\mu)+\beta\mu-\beta B'}{1-\beta^2(1-\zeta-\mu)+\beta\mu}. \tag{B.34}$$

Proof. (i) Taking the difference of (B.2) for $i = K+1$ and (B.3) for $i = K$, gives

$$\Delta_2 g^{\mathcal{S}}_{(1,K+1)} = B\Delta_2 g^{\mathcal{S}}_{(0,K)} + \frac{B}{\mu}\Delta_1 g^{\mathcal{S}}_{(1,K)}. \tag{B.35}$$

Taking the difference of (B.3) for $i = K$ and for $i = K-1$, employing the identity (4.24) together with (B.16) for $i = K-1$ and the expression of A_K in (B.14) gives

$$\Delta_2 g^{\mathcal{S}}_{(1,K)} = \left[\beta - \beta\mu(1-A_{K-1})\right]\Delta_2 g^{\mathcal{S}}_{(0,K)}. \tag{B.36}$$

Taking the difference of (B.4) for $i = K$ and (B.5) for $i = K-1$, and plugging again (B.16) for $i = K-1$ and (B.14), gives

$$\frac{\beta\zeta}{A_K}\Delta_2 g^{\mathcal{S}}_{(0,K)} = \beta\Delta_1 g^{\mathcal{S}}_{(1,K)} - \beta\eta\Delta_2 g^{\mathcal{S}}_{(1,K)} + \beta\eta\Delta_2 g^{\mathcal{S}}_{(0,K)} + \beta\zeta\Delta_2 g^{\mathcal{S}}_{(1,K+1)}. \tag{B.37}$$

By (B.8), the right-hand side of the above identity is equal to $\lambda - \Delta_1 g^{\mathcal{S}}_{(1,K)} - \beta\mu\Delta_2 g^{\mathcal{S}}_{(0,K)}$, yielding (B.32).

Using (B.32), (B.35) can be reformulated as

$$\Delta_2 g^{\mathcal{S}}_{(1,K+1)} = \frac{B\lambda}{\mu} - \frac{B}{\mu}\left[\frac{A'_K}{A_K} - \mu\right]\Delta_2 g^{\mathcal{S}}_{(0,K)}. \tag{B.38}$$

Further, (B.17), (B.35) and (B.36) can be used to reformulate (B.37) as (B.31). Finally, (B.31) and (B.38) after some algebra yield (B.30). □

(ii) (B.35) holds as before and simplifies to

$$\Delta_2 g^{\mathcal{S}}_{(1,1)} = \frac{B}{\mu}\Delta_1 g^{\mathcal{S}}_{(1,0)}. \tag{B.39}$$

By (B.8),

$$(1+\beta)\Delta_1 g^{\mathcal{S}}_{(1,0)} = \lambda - \beta\zeta\Delta_2 g^{\mathcal{S}}_{(1,1)}. \tag{B.40}$$

Solving and rearranging yields (B.33).

Further, (B.20) together with (B.39) and (B.33) give (B.34). □

Proposition B.3. *Under active set* $\mathcal{S} = \widetilde{\mathcal{I}}_{K,K+1}$ *with* $0 \leq K \leq I-1$, *action-differences* $\Delta_1 g^{\mathcal{S}}_{(1,i)} > 0$ *for all* $0 \leq i \leq I-1$ *and* $\Delta_1 g^{\mathcal{S}}_{(1,I)} = 0$.

Proof. We will divide the proof into four steps, in which we prove the following:

(i) if $1 \leq K$, then action-difference $\Delta_1 g^{\mathcal{S}}_{(1,K-1)} > 0$;

(ii) action-difference $\Delta_1 g^{\mathcal{S}}_{(1,K)} > 0$;

(iii) if $K \leq I-2$, then action-difference $\Delta_1 g^{\mathcal{S}}_{(1,K+1)} > 0$;

(iv) action-differences $\Delta_1 g^{\mathcal{S}}_{(1,i)} > 0$ for all $0 \leq i \leq I-1$ and $\Delta_1 g^{\mathcal{S}}_{(1,I)} = 0$.

(i) Suppose first that $1 \leq K$. Identity (B.18) gives

$$\Delta_1 g^{\mathcal{S}}_{(1,K-1)} = \lambda - A'_{K-1} \Delta_2 g^{\mathcal{S}}_{(0,K)}$$

and identity (B.32) gives

$$\Delta_1 g^{\mathcal{S}}_{(1,K)} = \lambda - \frac{A'_K}{A_K} \Delta_2 g^{\mathcal{S}}_{(0,K)}.$$

Because of Lemma B.5(iii), we have $\Delta_1 g^{\mathcal{S}}_{(1,K-1)} > \Delta_1 g^{\mathcal{S}}_{(1,K)}$. The positivity of the latter is proved in (ii).

(ii) In order to show $\Delta_1 g^{\mathcal{S}}_{(1,K)} > 0$ for $1 \leq K \leq I-1$, using (B.30) we need to have

$$\lambda > \frac{\beta\lambda\left[1-\beta(1-\zeta-\mu)\right]\frac{A'_K}{A_K}}{\left[1-\beta^2(1-\zeta-\mu)\right]\frac{A'_K}{A_K} + \beta\eta\beta(1-\mu)\left[1-\beta+\beta\mu(1-A_{K-1})\right]}.$$

Since the denominator is positive due to Lemma B.5, this is equivalent to

$$\beta\frac{A'_K}{A_K} < \frac{A'_K}{A_K} + \beta\eta\beta(1-\mu)\left[1-\beta+\beta\mu(1-A_{K-1})\right].$$

This is true, because of the properties in Lemma B.5(iii).

For $K = 0$, positivity of action-difference $\Delta_1 g^{\mathcal{S}}_{(1,0)}$ can be seen directly in (B.33).

(iii) Similarly, if $K \leq I-2$ identity (B.20) gives

$$\Delta_1 g^{\mathcal{S}}_{(1,K+1)} = \lambda - B' \Delta_2 g^{\mathcal{S}}_{(1,K+1)}.$$

In order to show $\Delta_1 g^{\mathcal{S}}_{(1,K+1)} > 0$ for $1 \leq K$, using (B.31) where we have plugged (B.30), we need to have

$$\lambda > \frac{B'\beta\lambda\left\{[1-\beta(1-\zeta-\mu)]+\beta\eta\left[\beta-A_{K-1}-\beta\mu(1-A_{K-1})\right]\right\}}{\left[1-\beta^2(1-\zeta-\mu)\right]\frac{A'_K}{A_K}+\beta\eta\beta(1-\mu)\left[1-\beta+\beta\mu(1-A_{K-1})\right]},$$

which is equivalent to

$$\begin{aligned}&\left[1-\beta^2(1-\zeta-\mu)\right]\frac{A'_K}{A_K}+\beta\eta\beta(1-\mu)\left[1-\beta+\beta\mu(1-A_{K-1})\right]\\ &> B'\beta\left\{[1-\beta(1-\zeta-\mu)]+\beta\eta\left[1-A_{K-1}-1+\beta-\beta\mu(1-A_{K-1})\right]\right\}.\end{aligned}$$

This can be further reformulated as

$$\begin{aligned}&\beta^2(\zeta+\mu)\frac{A'_K}{A_K}+\left(1-\beta^2\right)\frac{A'_K}{A_K}+\beta\eta\beta(1-\mu)\left[1-\beta+\beta\mu(1-A_{K-1})\right]\\ &> B'\beta\left\{[1-\beta(1-\zeta-\mu)]+\beta\eta(1-A_{K-1})\right\}-B'\beta\left\{\beta\eta\left[1-\beta+\beta\mu(1-A_{K-1})\right]\right\}.\end{aligned}$$

This is true, if the first terms of both sides (divided by $\beta > 0$) satisfy

$$\beta(\zeta+\mu)\frac{A'_K}{A_K} > B'\left\{[1-\beta(1-\zeta-\mu)]+\beta\eta(1-A_{K-1})\right\},$$

because the remaining two terms on the left-hand side are non-negative, and the last term on the right-hand side is non-positive. We further reformulate the last inequality as

$$\left(\frac{A'_K}{A_K}-B'\right)\left[1-\beta(1-\zeta-\mu)+\beta\eta(1-A_{K-1})\right] > \left[1-\beta+\beta\eta(1-A_{K-1})\right]\frac{A'_K}{A_K}.$$

Now, definitions in (B.14)-(B.15) imply the following identities:

$$\begin{aligned}\frac{A'_K}{A_K}-B' &= \frac{\beta\zeta}{A_K}-\beta\zeta B\\ 1-\beta(1-\zeta-\mu)+\beta\eta(1-A_{K-1}) &= \frac{A'_K}{A_K}+\beta\eta\\ 1-\beta+\beta\eta(1-A_{K-1}) &= \frac{\beta\zeta}{A_K}-\beta\zeta.\end{aligned}$$

The above inequality is therefore equivalent to

$$\left(\frac{\beta\zeta}{A_K}-\beta\zeta B\right)\left(\frac{A'_K}{A_K}+\beta\eta\right) > \left(\frac{\beta\zeta}{A_K}-\beta\zeta\right)\frac{A'_K}{A_K},$$

which is true because $B < 1$ and $\beta\eta \geq 0$.

For $K = 0$, positivity of action-difference $\Delta_1 g^{\mathcal{S}}_{(1,1)}$ given in (B.34) is straightforward after substituting for B' and using Lemma B.5(i). □

(iv) As in the proof of Proposition B.2(iii), one can show that positivity of action-difference $\Delta_1 g^{\mathcal{S}}_{(1,K-1)}$ implies $\Delta_1 g^{\mathcal{S}}_{(1,i)} > 0$ for all $0 \leq i \leq K-1$, and positivity of action-difference $\Delta_1 g^{\mathcal{S}}_{(1,K+1)}$ implies $\Delta_1 g^{\mathcal{S}}_{(1,i)} > 0$ for all $K+1 \leq i \leq I-1$.

Therefore, $\Delta_1 g^{\mathcal{S}}_{(1,i)} > 0$ for all $0 \leq i \leq I-1$. Finally, $\Delta_1 g^{\mathcal{S}}_{(1,I)} = 0$ by (B.10). □

B.2 Admission Control Problem with Delay: Marginal Reward Analysis

General case.

Analogously to Section B.1, in this section we set out to obtain closed formulae for marginal rewards in the admission control problem with delay. The proofs are similar to the formers and therefore they are omitted.

B.2.1 Preliminaries

Next we state analogies of Lemma B.1, Lemma B.2, and Lemma B.4 for reward measures in the admission control problem with delay. First we give a characterization of total rewards $f^{\mathcal{S}}_{(a,i)}$'s in terms of their action-differences $\Delta_1 f^{\mathcal{S}}_{(1,i)}$'s and state-differences $\Delta_2 f^{\mathcal{S}}_{(a,i)}$'s.

Lemma B.9. *For a fixed active set $\mathcal{S}$,*

$$(1-\beta) f^{\mathcal{S}}_{(1,i)} = R_{(1,i)} - \beta\mu\Delta_2 f^{\mathcal{S}}_{(1,i)} \qquad \text{if } (1,i) \in \mathcal{S} \tag{B.41}$$

$$(1-\beta) f^{\mathcal{S}}_{(1,i)} = R_{(1,i)} - \beta\mu\Delta_2 f^{\mathcal{S}}_{(0,i)} - \beta\Delta_1 f^{\mathcal{S}}_{(1,i)} \qquad \text{if } (1,i) \notin \mathcal{S} \tag{B.42}$$

$$(1-\beta) f^{\mathcal{S}}_{(0,i)} = \begin{cases} R_{(0,I)} - \beta\mu\Delta_2 f^{\mathcal{S}}_{(1,I)} + \beta\Delta_1 f^{\mathcal{S}}_{(1,I)} & \text{if } i = I \\ R_{(0,i)} + \beta\zeta\Delta_2 f^{\mathcal{S}}_{(1,i+1)} - \beta\eta\Delta_2 f^{\mathcal{S}}_{(1,i)} + \beta\Delta_1 f^{\mathcal{S}}_{(1,i)} & \text{otherwise} \end{cases} \qquad \text{if } (0,i) \in \mathcal{S} \tag{B.43}$$

$$(1-\beta) f^{\mathcal{S}}_{(0,i)} = \begin{cases} R_{(0,I)} - \beta\mu\Delta_2 f^{\mathcal{S}}_{(0,I)} & \text{if } i = I \\ R_{(0,i)} + \beta\zeta\Delta_2 f^{\mathcal{S}}_{(0,i+1)} - \beta\eta\Delta_2 f^{\mathcal{S}}_{(0,i)} & \text{otherwise} \end{cases} \qquad \text{if } (0,i) \notin \mathcal{S} \tag{B.44}$$

The following lemma characterizes action-differences $\Delta_1 f^{\mathcal{S}}_{(1,i)}$'s in terms of state-differences $\Delta_2 f^{\mathcal{S}}_{(a,i)}$'s.

Lemma B.10. *For a fixed active set $\mathcal{S}$ and any state $0 \le i \le I-1$,*

$$\Delta_1 f^{\mathcal{S}}_{(1,i)} = \Delta_1 R_{(1,i)} - \beta\zeta\Delta_2 f^{\mathcal{S}}_{(1,i+1)} - \beta(\mu-\eta)\Delta_2 f^{\mathcal{S}}_{(1,i)} \quad \text{if } (0,i),(1,i) \in \mathcal{S} \tag{B.45}$$

$$\Delta_1 f^{\mathcal{S}}_{(1,i)} = \Delta_1 R_{(1,i)} - \beta\zeta\Delta_2 f^{\mathcal{S}}_{(0,i+1)} - \beta(\mu-\eta)\Delta_2 f^{\mathcal{S}}_{(0,i)} \quad \text{if } (0,i),(1,i) \notin \mathcal{S} \tag{B.46}$$

$$(1+\beta)\Delta_1 f^{\mathcal{S}}_{(1,i)} = \Delta_1 R_{(1,i)} - \beta\zeta\Delta_2 f^{\mathcal{S}}_{(1,i+1)} - \beta\mu\Delta_2 f^{\mathcal{S}}_{(0,i)} + \beta\eta\Delta_2 f^{\mathcal{S}}_{(1,i)} \quad \text{if } (0,i)\in\mathcal{S}, (1,i)\notin\mathcal{S} \tag{B.47}$$

$$(1-\beta)\Delta_1 f^{\mathcal{S}}_{(1,i)} = \Delta_1 R_{(1,i)} - \beta\zeta\Delta_2 f^{\mathcal{S}}_{(0,i+1)} - \beta\mu\Delta_2 f^{\mathcal{S}}_{(1,i)} + \beta\eta\Delta_2 f^{\mathcal{S}}_{(0,i)} \quad \text{if } (0,i)\notin\mathcal{S}, (1,i)\in\mathcal{S} \tag{B.48}$$

$$\Delta_1 f^{\mathcal{S}}_{(1,I)} = 0. \tag{B.49}$$

Now we are ready to express marginal rewards $r^{\mathcal{S}}_{(a,i)}$'s in terms of action-differences $\Delta_1 f^{\mathcal{S}}_{(1,i)}$'s.

Lemma B.11. *For a fixed active set $\mathcal{S}$,*

$$r^{\mathcal{S}}_{(1,i)} = \begin{cases} \beta\Delta_1 f^{\mathcal{S}}_{(1,0)} & \text{if } i = 0 \\ \beta\mu\Delta_1 f^{\mathcal{S}}_{(1,i-1)} + \beta(1-\mu)\Delta_1 f^{\mathcal{S}}_{(1,i)} & \text{otherwise} \end{cases} \tag{B.50}$$

$$r^{\mathcal{S}}_{(0,i)} = \begin{cases} \beta(1-\zeta)\Delta_1 f^{\mathcal{S}}_{(1,0)} + \beta\zeta\Delta_1 f^{\mathcal{S}}_{(1,1)} & \text{if } i = 0 \\ \beta\eta\Delta_1 f^{\mathcal{S}}_{(1,i-1)} + \beta\varepsilon\Delta_1 f^{\mathcal{S}}_{(1,i)} + \beta\zeta\Delta_1 f^{\mathcal{S}}_{(1,i+1)} & \text{otherwise} \\ r^{\mathcal{S}}_{(1,I)} & \text{if } i = I \end{cases} \tag{B.51}$$

Let, for $j \ge i \ge 0$,

$$C_{i,i} := 0, \qquad C_{i,j+1} := \left[C_{i,j} - \frac{\Delta_2 R_{(1,j+1)}}{\beta\mu}\right] B, \qquad C''_{i,j} := \Delta_1 R_{(1,j)} + \beta\zeta C_{i,j+1}, \tag{B.52}$$

$$D_0 := 0, \quad D_{j+1} := \left[\beta\eta D_j - \Delta_2 R_{(1,j+1)}\right] \frac{A_{j+1}}{\beta\zeta}, \quad D'_j := \Delta_1 R_{(1,j)} + \beta(\mu-\eta) D_j. \tag{B.53}$$

B.2.2 Calculation of Action-Differences in Total Reward

Since Lemma B.11 characterizes marginal rewards $r^{\mathcal{S}}_{(a,i)}$'s in terms of action-differences $\Delta_1 f^{\mathcal{S}}_{(1,i)}$'s, in this subsection we focus on the calculation of the latter. In the following we show that all the relevant state-differences can be obtained by recursion from two pivot state-differences associated to the two thresholds K_0, K_1 under any policy $\widetilde{\mathcal{I}}_{K_0,K_1}$.

Lemma B.12. *For a fixed active set* $\mathcal{S} = \widetilde{\mathcal{I}}_{K_0,K_1}$,

$$\Delta_2 f^{\mathcal{S}}_{(0,i)} = \Delta_2 f^{\mathcal{S}}_{(0,i+1)} A_i - D_i, \qquad \text{for } 1 \le i \le K_0' - 1, \tag{B.54}$$

$$\Delta_2 f^{\mathcal{S}}_{(1,i)} = \Delta_2 f^{\mathcal{S}}_{(1,K_1)} B^{i-K_1} - C_{K_1,i}, \qquad \text{for } K_1 + 1 \le i \le I. \tag{B.55}$$

Further we identify a recursion to calculate action-differences $\Delta_1 f^{\mathcal{S}}_{(1,i)}$'s in terms of the two pivot state-differences $\Delta_2 f^{\mathcal{S}}_{(0,K_0')}$ and $\Delta_2 f^{\mathcal{S}}_{(1,K_1)}$. Thus, this is a simplification of Lemma B.10.

Proposition B.4. *For a fixed active set* $\mathcal{S} = \widetilde{\mathcal{I}}_{K_0,K_1}$,

$$\Delta_1 f^{\mathcal{S}}_{(1,K_0'-1)} = D'_{K_0'-1} - \Delta_2 f^{\mathcal{S}}_{(0,K_0')} A'_{K_0'-1}, \qquad \text{if } 1 \le K_0', \tag{B.56}$$

$$\Delta_1 f^{\mathcal{S}}_{(1,i)} = D'_i(1 - Z_{i+1}) + \left[\Delta_1 f^{\mathcal{S}}_{(1,i+1)} + \beta\mu D_i - \Delta_2 R_{(1,i+1)}\right] Z_{i+1}, \quad \text{for } 0 \le i \le K_0' - 2, \tag{B.57}$$

$$\Delta_1 f^{\mathcal{S}}_{(1,K_1)} = C''_{K_1,K_1} - \Delta_2 f^{\mathcal{S}}_{(1,K_1)} B', \qquad \text{if } K_1 \le I - 1, \tag{B.58}$$

$$\Delta_1 f^{\mathcal{S}}_{(1,i)} = \left[C''_{K_1,i} + \beta(\mu - \eta) C_{K_1,i}\right] - \left[C''_{K_1,i-1} + \beta(\mu - \eta) C_{K_1,i-1}\right] B + \Delta_1 f^{\mathcal{S}}_{(1,i-1)} B, \quad \text{for } K_1 + 1 \le i \le I - 1. \tag{B.59}$$

The above results help significantly simplify the subsequent analysis, which we present in the next subsection for optimal active set $\widetilde{\mathcal{I}}_{K,K}$, and where we identify closed-form expressions for pivot state-differences in total reward. The results under the active set $\widetilde{\mathcal{I}}_{K,K+1}$ are not necessary, since they are not implemented in the algorithm FA.

B.2.3 Pivot State-Differences under Active Set $\widetilde{\mathcal{I}}_{K,K}$

Lemma B.13. *Under active set* $\mathcal{S} = \widetilde{\mathcal{I}}_{K,K}$

(i) if $K = 0$, *then*

$$\Delta_2 f^{\mathcal{S}}_{(1,0)} = 0. \tag{B.60}$$

(ii) if $1 \le K \le I - 1$,

$$\Delta_2 f^{\mathcal{S}}_{(0,K)} = -\frac{\frac{\beta\zeta}{A_K} D_K - \Delta_1 R_{(1,K)} + (1-\beta) C''_{K,K+1} - (\beta D'_{K-1} + \Delta_2 R_{(1,K)}) B' + \beta\mu(\beta\zeta D_{K-1} B + C''_{K,K+1} - \Delta_1 R_{(1,K-1)})}{\frac{A'_K}{A_K} + \beta A'_{K-1} B' + \beta\zeta\beta\mu(1 - BA_{K-1})}, \tag{B.61}$$

$$\Delta_2 f^{\mathcal{S}}_{(1,K)} = -\frac{\frac{\beta\zeta}{A_K} D_K - \Delta_1 R_{(1,K)} + (1-\beta) D'_{K-1} - (\beta C''_{K,K+1} + \Delta_2 R_{(1,K)}) A'_{K-1} + \beta\mu(\beta\zeta D_{K-1} + (C''_{K,K+1} - \Delta_1 R_{(1,K-1)}) A_{K-1})}{\frac{A'_K}{A_K} + \beta A'_{K-1} B' + \beta\zeta\beta\mu(1 - BA_{K-1})}. \tag{B.62}$$

(iii) if $K = I$*,*

$$\Delta_2 f^{\mathcal{S}}_{(0,I)} = -\frac{\frac{\beta\zeta}{A_I} D_I - \beta\mu D'_{I-1}}{\frac{A'_I}{A_I} + \beta\mu A'_{I-1}}. \tag{B.63}$$

(iv) if $K = I + 1$*,*

$$\Delta_2 f^{\mathcal{S}}_{(0,I)} = -\frac{\beta\zeta D_I}{A'_I}. \tag{B.64}$$

Appendix C

Appendix to Chapter 5

C.1 Work-Reward Analysis

In order to prove Proposition 5.1, we describe some crucial points from the restless bandit framework in more detail. For a survey on this methodology refer to Niño-Mora (2007b). Note that any set $\mathcal{S} \subseteq \mathcal{T}$ can represent a stationary policy, by being active in all the states belonging to $\mathcal{S}$ and being passive in all the states belonging to $\mathcal{T} \setminus \mathcal{S}$. We will call such a policy an *$\mathcal{S}$-active policy*, and $\mathcal{S}$ an *active set*. Note also that we can restrict our attention to stationary deterministic policies, since it is well-known from the MDP theory that there exists an optimal policy which is stationary, deterministic, and independent of the initial state.

Let $\mathcal{S} \subseteq \mathcal{T}$ be an active set. We can reformulate (5.5) as

$$f_t^{\mathcal{S}} - \nu g_t^{\mathcal{S}} := \mathbb{E}_t^{\mathcal{S}} \left[\sum_{s=0}^{t-1} \beta^s \boldsymbol{P}_{t,t-s}^{s|\mathcal{S}} R_{t-s}^{I_{\mathcal{S}}(t-s)} \right] - \nu \mathbb{E}_t^{\mathcal{S}} \left[\sum_{s=0}^{t-1} \beta^s \boldsymbol{P}_{t,t-s}^{s|\mathcal{S}} W_{t-s}^{I_{\mathcal{S}}(t-s)} \right], \tag{C.1}$$

where $\boldsymbol{P}_{i,j}^{j-i|\mathcal{S}}$ is the probability of moving from state $i \in \mathcal{X}$ to state $j \in \mathcal{X}$ in exactly $j-i$ periods under policy $\mathcal{S}$ and $I_{\mathcal{S}}(s)$ is the indicator function $I_{\mathcal{S}}(s) = \begin{cases} 1, & \text{if } s \in \mathcal{S}, \\ 0, & \text{if } s \notin \mathcal{S}. \end{cases}$ We will call $f_t^{\mathcal{S}}$ the *expected total discounted revenue* under policy $\mathcal{S}$ if starting from state t, and we will write it in a more convenient way as

$$f_t^{\mathcal{S}} = \mathbb{E}_t^{\mathcal{S}} \left[\sum_{s=1}^{t} \beta^{t-s} \boldsymbol{P}_{t,s}^{t-s|\mathcal{S}} R_s^{I_{\mathcal{S}}(s)} \right]. \tag{C.2}$$

Similarly, we will call $g_t^{\mathcal{S}}$ the *expected total discounted promotion work* under policy $\mathcal{S}$

if starting from state t, and we will write it in a more convenient way as

$$g_t^{\mathcal{S}} = \mathbb{E}_t^{\mathcal{S}} \left[\sum_{s=1}^{t} \beta^{t-s} \boldsymbol{P}_{t,s}^{t-s|\mathcal{S}} W_s^{I_{\mathcal{S}}(s)} \right]. \tag{C.3}$$

Let, further, $\langle a, \mathcal{S} \rangle$ be the policy which takes action $a \in \mathcal{A}$ in the current time epoch and adopts an $\mathcal{S}$-active policy thereafter. For any state $t \in \mathcal{T}$ and an $\mathcal{S}$-active policy, the $(t, \mathcal{S})$-*marginal revenue* is defined as

$$r_t^{\mathcal{S}} := f_t^{\langle 1, \mathcal{S} \rangle} - f_t^{\langle 0, \mathcal{S} \rangle}, \tag{C.4}$$

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

$$w_t^{\mathcal{S}} := g_t^{\langle 1, \mathcal{S} \rangle} - g_t^{\langle 0, \mathcal{S} \rangle}. \tag{C.5}$$

These marginal revenue and marginal promotion work capture the change in the expected total discounted revenue and promotion work, respectively, which results from being active instead of passive in the first time epoch and following the $\mathcal{S}$-active policy afterwards. Finally, if $w_t^{\mathcal{S}} \neq 0$, we define the $(t, \mathcal{S})$-*marginal productivity rate* as

$$\nu_t^{\mathcal{S}} := \frac{r_t^{\mathcal{S}}}{w_t^{\mathcal{S}}}. \tag{C.6}$$

In order to verify that a perishable item satisfies PCL-indexability (which implies existence of MP indices), we need to postulate a family $\mathcal{F}$ of optimal active sets. We define it as a family of nested sets $\mathcal{F} := \{\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_T\}$, where $\mathcal{S}_k := \{1, 2, \ldots, k\}$. The following assumption must be verified.

Assumption C.1 (Positive Marginal Works). The marginal promotion work $w_t^{\mathcal{S}} > 0$ for all states $t \in \mathcal{T}$ under any $\mathcal{S}$-active policy from a feasible family of active sets $\mathcal{F} \subseteq 2^{\mathcal{T}}$.

If Assumption C.1 holds and the quantities $\widehat{\nu}_{\tau_{k+1}}$ computed in the *adaptive-greedy algorithm* presented in Figure C.1 are nonincreasing in k, then the marginal productivity indices exist and equal $\nu_t^* := \widehat{\nu}_t$ and $\mathcal{F}$ contains an optimal active set for any ν.

C.2 Proofs

C.2.1 Proof of Proposition 5.1

We next show that Assumption C.1 holds and derive a closed-form expression for the MP index given in Proposition 5.1. Plugging (C.2) and (C.3) into (C.4) and (C.5), respec-

```
set  $\widehat{\mathcal{S}}_0 := \emptyset$;
for  $k := 0$  to  $T+1$  do
   choose  $\tau_{k+1} \in \{\tau \in \mathcal{T} \setminus \widehat{\mathcal{S}}_k;\ \nu_\tau^{\widehat{\mathcal{S}}_k} \geq \nu_t^{\widehat{\mathcal{S}}_k}$  for all  $t \in \mathcal{T} \setminus \widehat{\mathcal{S}}_k\}$;
   set  $\widehat{\mathcal{S}}_{k+1} := \widehat{\mathcal{S}}_k \cup \{\tau_{k+1}\}$;
   set  $\widehat{\nu}_{\tau_{k+1}} := \nu_{\tau_{k+1}}^{\widehat{\mathcal{S}}_k}$;
end {for};
```

Figure C.1: Adaptive-greedy algorithm for calculation of MP indices.

tively, we obtain two expressions that will be used in the following analysis:

$$r_t^{\mathcal{S}} = \left(R_t^1 - R_t^0\right) - (\beta q - \beta p) \sum_{s=1}^{t-1} \beta^{t-s-1} \boldsymbol{P}_{t-1,s}^{t-s-1|\mathcal{S}} R_s^{I_{\mathcal{S}}(s)}, \tag{C.7}$$

$$w_t^{\mathcal{S}} = \left(W_t^1 - W_t^0\right) - (\beta q - \beta p) \sum_{s=1}^{t-1} \beta^{t-s-1} \boldsymbol{P}_{t-1,s}^{t-s-1|\mathcal{S}} W_s^{I_{\mathcal{S}}(s)}. \tag{C.8}$$

It is well known from the MDP theory that the transition probability matrix for multiple periods is obtained by multiplication of transition probability matrices for subperiods. Hence, given an active set $\mathcal{S} \subseteq \mathcal{T}$, we have

$$\boldsymbol{P}^{t-s|\mathcal{S}} = \left(\boldsymbol{P}^{1|\mathcal{S}}\right)^{t-s}, \tag{C.9}$$

where the matrix $\boldsymbol{P}^{1|\mathcal{S}}$ is an $(T+1) \times (T+1)$-matrix constructed so that its row $x \in \mathcal{X}$ is the row x of the matrix $\boldsymbol{P}^{1|\mathcal{T}}$ if $x \in \mathcal{S}$, and is the row x of the matrix $\boldsymbol{P}^{1|\emptyset}$ otherwise. For definiteness, we remark that $\boldsymbol{P}^{0|\mathcal{S}}$ is an identity matrix.

Lemma C.1. *Let $t \in \mathcal{T}$ and consider any integer $0 \leq k \leq T$. Then,*

$$r_t^{\mathcal{S}_k} = \begin{cases} R(q-p)\left[(1-\beta)\dfrac{1-(\beta p)^{t-1}}{1-\beta p}\right. \\ \left.+(1-\beta\alpha)\,(\beta p)^{t-1}\right], & \text{if } k \geq t-1 \geq 0, \\ R(q-p)\left[(1-\beta)\dfrac{1-(\beta q)^{t-k-1}}{1-\beta q}\right. \\ +(1-\beta)\,(\beta q)^{t-k-1}\,\dfrac{1-(\beta p)^{k}}{1-\beta p} \\ \left.+(1-\beta\alpha)\,(\beta q)^{t-k-1}\,(\beta p)^{k}\right], & \text{if } T-1 \geq t-1 \geq k. \end{cases} \tag{C.10}$$

$$w_t^{\mathcal{S}_k} = \begin{cases} W\left[1-(\beta q-\beta p)\dfrac{1-(\beta p)^{t-1}}{1-\beta p}\right], & \text{if } k \geq t-1 \geq 0, \\ W\left[1-(\beta q-\beta p)\,(\beta q)^{t-k-1}\,\dfrac{1-(\beta p)^{k}}{1-\beta p}\right], & \text{if } T-1 \geq t-1 \geq k. \end{cases} \tag{C.11}$$

Proof. Under an active set $\mathcal{S}_k$, from (C.9) we get for $T \geq t-1 \geq s \geq 1$,

$$\boldsymbol{P}_{t-1,s}^{t-s-1|\mathcal{S}_k} = \begin{cases} p^{t-s-1}, & \text{if } k \geq t-1 \geq s \geq 0, \\ q^{t-s-1}, & \text{if } T \geq t-1 \geq s \geq k, \\ q^{t-k-1}p^{k-s}, & \text{if } T \geq t-1 \geq k \geq s \geq 0, \end{cases}$$

These expressions together with the definitions of R_t^a and W_t^a plugged into (C.7)–(C.8) after simplifying conclude the proof. □

Lemma C.2. *For any integer $k \geq 0$ we have*

(i) $w_k^{\mathcal{S}_k} > 0$;

(ii) $w_t^{\mathcal{S}_k} > 0$ *for all* $t \in \mathcal{T}$.

Proof. Denote by

$$h(k) := (\beta q-\beta p)\frac{1-(\beta p)^k}{1-\beta p}, \tag{C.12}$$

so that, using (C.11), $w_k^{\mathcal{S}_k} = W\,[1-h(k)]$.

(i) For $k=0$, we have $h(0)=0$ by definition. For $k \geq 1$, we have

$$h(k) = (1-(\beta p)^k)\frac{\beta q-\beta p}{1-\beta p} < 1.$$

□

(ii) Implied by (i) and (C.11). □

Lemma C.3. *In each step $k = 0, 1, \ldots, T-1$ of the adaptive-greedy algorithm presented in Figure C.1, Assumption C.1 is satisfied and the algorithm sets*

$$\tau_{k+1} = k+1; \qquad \widehat{\mathcal{S}}_{k+1} = \{1, \ldots, k+1\};$$
$$\widehat{\nu}_{k+1} = \frac{R(q-p)\left[(1-\beta)\dfrac{1-(\beta p)^k}{1-\beta p} + (1-\beta\alpha)\,(\beta p)^k\right]}{W\left[1-(\beta q-\beta p)\dfrac{1-(\beta p)^k}{1-\beta p}\right]}$$

Proof. Consider the step k. Having $\widehat{\mathcal{S}}_k = \{1, 2, \ldots, k\}$, for $t \in \mathcal{T} \setminus \widehat{\mathcal{S}}_k$, by Lemma C.1 we have

$$\begin{aligned} r_t^{\widehat{\mathcal{S}}_k} &= R(q-p)\left[(1-\beta)\frac{1-(\beta q)^{t-k-1}}{1-\beta q}\right. \\ &\quad + (1-\beta)\,(\beta q)^{t-k-1}\,\frac{1-(\beta p)^k}{1-\beta p} \\ &\quad \left.+(1-\beta\alpha)\,(\beta q)^{t-k-1}\,(\beta p)^k\right], \\ w_t^{\widehat{\mathcal{S}}_k} &= W\left[1-(\beta q-\beta p)\,(\beta q)^{t-k-1}\,\frac{1-(\beta p)^k}{1-\beta p}\right] > 0, \end{aligned}$$

where the positivity is due to Lemma C.2 (which also holds for $t \in \widehat{\mathcal{S}}_k$), implying that Assumption C.1 is satisfied. Furthermore, then $r_t^{\widehat{\mathcal{S}}_k}$ is nondecreasing and $w_t^{\widehat{\mathcal{S}}_k}$ is nonincreasing as t diminishes (i.e. as t gets closer to the deadline). Hence the maximum $\nu_t^{\widehat{\mathcal{S}}_k}$ is attained at $t = k+1$ and the algorithm sets what is stated. □

Lemma C.3 is a crucial result in the proof of Proposition 5.1. It verifies that the family $\mathcal{F}$ required in Assumption C.1 is the family of nested active sets $\mathcal{F} = \{\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_T\}$. Finally, Proposition 5.2(iii) assures that the algorithm's output $\widehat{\nu}_t$ is nondecreasing as t diminishes. This concludes the proof of Proposition 5.1. □

C.2.2 Proof of Proposition 5.2

(i) Immediate from (5.6). □

(ii) Formally, we are to prove the following statement: If the probability q is replaced by $q' \leq q$, then $\nu_t^{*\prime} \leq \nu_t^*$ for any $t \in \mathcal{T}$. It is straightforward to see that (5.6) is nondecreasing in q. □

(iii) In order to see that the MP index is nondecreasing as t diminishes, it is enough to compare the MP indices for $t \in \mathcal{T} \setminus \{T\}$ and $t+1$. Niño-Mora (2007b, p. 172)

showed that under positive marginal works, $\nu_t^* \geq \nu_{t+1}^*$ is equivalent to $\nu_t^{\mathcal{S}_{t+1}} \geq \nu_{t+1}^{\mathcal{S}_{t+1}}$, which is satisfied as shown in the proof of Lemma C.3. □

(iv) The expression is obtained readily by letting $T \to \infty$. □

C.2.3 Proof of Corollary 5.0.1

In order the set $\{\tau \in \mathcal{T} : \nu_t^* > \nu \text{ for all } t \in \mathcal{T} \text{ such that } t \geq \tau\}$ to be nonempty, due to Proposition 5.2(iii) we need to have $\nu_1^* > \nu$, that is, $\frac{R}{W}(1-\beta\alpha)(q-p) > \nu$. □

C.2.4 Proof of Proposition 5.3

(i) The MP index for $\beta = 1$ is given by the limit of the discounted MP index (5.6), if it exists. The limit exists and is equal to the stated expression. □

(ii) Straightforward from (5.6) after setting $q := 1$ and $\beta := 1$. □

C.2.5 Proof of Proposition 5.4

Under the above assumptions, all the products perish within one period, and promoting is equivalent to avoiding the deadline cost. The problem thus reduces to a combinatory problem of choosing a subset of items not to be promoted that minimizes the aggregate cost of not promoted items while the remaining items do not occupy more than W. Since the aggregate cost of all items is constant, this problem is equivalent to choosing a subset of items to be promoted that maximizes the aggregate cost of promoted items while their aggregate volume is not greater than W, which is the knapsack problem. □

Appendix D

Appendix to Chapter 6

D.1 Proofs

D.1.1 Proof of Proposition 6.1

Note first that the additive term $\nu\overline{W}$ is constant and can be ignored. Further,

$$\lim_{T\to\infty}\frac{1}{T}\mathbb{E}^{\pi}_{\boldsymbol{n}}\left[\sum_{t=0}^{T-1}\sum_{m\in\mathcal{M}(t)}\left(R^{a_m(t)}_{m,X_m(t)}-\nu W^{a_m(t)}_{m,X_m(t)}\right)\right]$$

$$=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}^{\pi}_{\boldsymbol{n}}\left[\sum_{m\in\mathcal{M}(t)}\left(R^{a_m(t)}_{m,X_m(t)}-\nu W^{a_m(t)}_{m,X_m(t)}\right)\right]$$

$$=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{m\in\mathcal{M}^{\text{started}}(T-1)}\mathbb{P}\left[m\in\mathcal{M}(t)\right]\mathbb{E}^{\pi_m}_{n_m}\left[\left(R^{a_m(t)}_{m,X_m(t)}-\nu W^{a_m(t)}_{m,X_m(t)}\right)\right]$$

$$=\lim_{T\to\infty}\frac{L}{|\mathcal{M}^{\text{started}}(T-1)|}\sum_{m\in\mathcal{M}^{\text{started}}(T-1)}\sum_{t=0}^{T-1}\mathbb{P}\left[m\in\mathcal{M}(t)\right]\mathbb{E}^{\pi_m}_{n_m}\left[\left(R^{a_m(t)}_{m,X_m(t)}-\nu W^{a_m(t)}_{m,X_m(t)}\right)\right],$$

where we have used (6.1). Further, using

$$\lim_{T\to\infty}\sum_{t=0}^{T-1}\mathbb{P}\left[m\in\mathcal{M}(t)\right]\mathbb{E}_{n_m}^{\pi_m}\left[\left(R_{m,X_m(t)}^{a_m(t)}-\nu W_{m,X_m(t)}^{a_m(t)}\right)\right]$$
$$=\mathbb{E}_{n_m}^{\pi_m}\left[\sum_{t=T_m}^{\infty}\beta_m^{t-T_m}\left(R_{m,X_m(t)}^{a_m(t)}-\nu W_{m,X_m(t)}^{a_m(t)}\right)\right]<\infty,$$

the Lagrangian relaxation (6.6) (without the additive constant) can be written as

$$\max_{\pi\in\Pi}\lim_{T\to\infty}\frac{L}{|\mathcal{M}^{\text{started}}(T-1)|}\sum_{m\in\mathcal{M}^{\text{started}}(T-1)}\mathbb{E}_{n_m}^{\pi_m}\left[\sum_{t=T_m}^{\infty}\beta_m^{t-T_m}\left(R_{m,X_m(t)}^{a_m(t)}-\nu W_{m,X_m(t)}^{a_m(t)}\right)\right],$$

which we can rewrite due to flow independence as

$$\lim_{T\to\infty}\frac{L}{|\mathcal{M}^{\text{started}}(T-1)|}\sum_{m\in\mathcal{M}^{\text{started}}(T-1)}\max_{\pi_m\in\Pi_m}\mathbb{E}_{n_m}^{\pi_m}\left[\sum_{t=T_m}^{\infty}\beta_m^{t-T_m}\left(R_{m,X_m(t)}^{a_m(t)}-\nu W_{m,X_m(t)}^{a_m(t)}\right)\right].$$

Notice that the operator $\lim_{T\to\infty}\frac{L}{|\mathcal{M}^{\text{started}}(T-1)|}\sum_{m\in\mathcal{M}^{\text{started}}(T-1)}$ is equivalent to the operator $\lim_{M\to\infty}\frac{L}{M}\sum_{m=1}^{M}$, so the Lagrangian relaxation (6.6) (without the additive constant) is the same as

$$\lim_{M\to\infty}\frac{L}{M}\sum_{m=1}^{M}\max_{\pi_m\in\Pi_m}\mathbb{E}_{n_m}^{\pi_m}\left[\sum_{t=T_m}^{\infty}\beta_m^{t-T_m}\left(R_{m,X_m(t)}^{a_m(t)}-\nu W_{m,X_m(t)}^{a_m(t)}\right)\right].$$

Therefore, the maximizing policy π^* is the one that decomposes into π_m^* for each $m\in\mathcal{M}$ that maximize the individual-flow subproblems

$$\max_{\pi_m\in\Pi_m}\mathbb{E}_{n_m}^{\pi_m}\left[\sum_{t=T_m}^{\infty}\beta_m^{t-T_m}\left(R_{m,X_m(t)}^{a_m(t)}-\nu W_{m,X_m(t)}^{a_m(t)}\right)\right]\text{ for each } m\in\mathcal{M}.$$

Note that the flow-m subproblem is independent of T_m whenever $X_m(T_m)=n_m$, and therefore it is enough to solve it for $T_m=0$. □

D.2 Auxiliary Results

Lemma D.1. *Let $b_0,b_1,\alpha_1>0$. Then the following statements are equivalent:*

(i) $\frac{a_0}{b_0}\geq\frac{a_1}{b_1}$;

(ii) $\dfrac{a_0}{b_0} \geq \dfrac{a_0 + \alpha_1 a_1}{b_0 + \alpha_1 b_1}$;

(iii) $\dfrac{a_0 + \alpha_1 a_1}{b_0 + \alpha_1 b_1} \geq \dfrac{a_1}{b_1}$.

Further, the above equivalence holds also if all inequalities are strict.

Proof. A simple algebraic exercise. □

Lemma D.2. *Let for a positive integer K be $\dfrac{a_0}{b_0} \geq \dfrac{a_1}{b_1} \geq \cdots \geq \dfrac{a_K}{b_K}$ such that $b_0, b_1, \ldots, b_K > 0$ and $\alpha_1, \ldots, \alpha_K \geq 0$. Then*

$$\frac{a_0}{b_0} \geq \frac{a_0 + \alpha_1 a_1 + \cdots + \alpha_K a_K}{b_0 + \alpha_1 b_1 + \cdots + \alpha_K b_K} \geq \frac{a_K}{b_K}. \tag{D.1}$$

Proof. For $k = 1$ see Lemma D.1. To proceed by induction, suppose that the claim holds for all $k = 1, 2, \ldots, K-1$. If $\alpha_1 = 0$, then we are in the case $k = K-1$, so suppose $\alpha_1 > 0$. Then, by the induction assumption for $k = K-1$, we have

$$\frac{a_1}{b_1} \geq \frac{a_1 + \frac{\alpha_2}{\alpha_1} a_2 + \cdots + \frac{\alpha_K}{\alpha_1} a_K}{b_1 + \frac{\alpha_2}{\alpha_1} b_2 + \cdots + \frac{\alpha_K}{\alpha_1} b_K}. \tag{D.2}$$

Since $\dfrac{a_0}{b_0} \geq \dfrac{a_1}{b_1}$, we have $\dfrac{a_0}{b_0}$ larger than or equal to the right-hand side of (D.2). Applying Lemma D.1 gives

$$\frac{a_0}{b_0} \geq \frac{a_0 + \alpha_1 a_1 + \cdots + \alpha_K a_K}{b_0 + \alpha_1 b_1 + \cdots + \alpha_K b_K} \geq \frac{a_1 + \frac{\alpha_2}{\alpha_1} a_2 + \cdots + \frac{\alpha_K}{\alpha_1} a_K}{b_1 + \frac{\alpha_2}{\alpha_1} b_2 + \cdots + \frac{\alpha_K}{\alpha_1} b_K}, \tag{D.3}$$

where the right-hand side expression is, again by the induction assumption for $k = K-1$, larger than or equal to $\dfrac{a_K}{b_K}$. This completes the proof for $k = K$. □

Lemma D.3 (Equivalent Definitions of Concavity). *Consider a real-valued function $a(\cdot)$, a positive integer K, and a set $\mathcal{B} := \{0, b_0, b_1, \ldots, b_K\}$ such that $0 < b_0 < b_1 < \cdots < b_K$. Denote by $a_k := a(b_k)$ for any $k = 0, 1, \ldots, K$. Then the following statements are equivalent:*

(i) the function $a(\cdot)$ is concave on $\mathcal{B}$;

(ii) $\dfrac{a_{k_1} - a_{k_0}}{b_{k_1} - b_{k_0}} \geq \dfrac{a_{k_2} - a_{k_0}}{b_{k_2} - b_{k_0}}$ *for any $b_{k_0}, b_{k_1}, b_{k_2} \in \mathcal{B}$ with $b_{k_0} < b_{k_1} < b_{k_2}$;*

(iii) $\dfrac{a_{k_2} - a_{k_0}}{b_{k_2} - b_{k_0}} \geq \dfrac{a_{k_2} - a_{k_1}}{b_{k_2} - b_{k_1}}$ *for any $b_{k_0}, b_{k_1}, b_{k_2} \in \mathcal{B}$ with $b_{k_0} < b_{k_1} < b_{k_2}$;*

(iv) $\dfrac{a_{k_1} - a_{k_0}}{b_{k_1} - b_{k_0}} \geq \dfrac{a_{k_2} - a_{k_1}}{b_{k_2} - b_{k_1}}$ *for any $b_{k_0}, b_{k_1}, b_{k_2} \in \mathcal{B}$ with $b_{k_0} < b_{k_1} < b_{k_2}$.*

Lemma D.4. *Let a be a concave real-valued function with $a(0) \geq 0$. Consider a positive integer K and a set $\mathcal{B} := \{0, b_0, b_1, \dots, b_K\}$ such that $0 < b_0 < b_1 < \cdots < b_K$. Denote by $a_k := a(b_k)$ for any $k = 0, 1, \dots, K$. Then*

(i) $\dfrac{a_{k_2}}{b_{k_2}} \geq \dfrac{a_{k_2} - a_{k_1}}{b_{k_2} - b_{k_1}}$ *for any $b_{k_1}, b_{k_2} \in \mathcal{B}$ with $0 < b_{k_1} < b_{k_2}$;*

(ii) $\dfrac{a_0}{b_0} \geq \dfrac{a_1}{b_1} \geq \cdots \geq \dfrac{a_K}{b_K}$;

(iii) $\dfrac{a_{k_1}}{b_{k_1}} \geq \dfrac{a_{k_2} - a_{k_1}}{b_{k_2} - b_{k_1}}$ *for any $b_{k_1}, b_{k_2} \in \mathcal{B}$ with $0 < b_{k_1} < b_{k_2}$;*

Proof. (i) By setting $b_{k_0} = 0$ in Lemma D.3(iii), for any $k_1 \in \{0, 1, \dots, K-1\}$,

$$\frac{a_{k_2}}{b_{k_2}} \geq \frac{a_{k_2} - a(0)}{b_{k_2}} \geq \frac{a_{k_2} - a_{k_1}}{b_{k_2} - b_{k_1}},$$

where the first inequality is due to $a(0) \geq 0$. □

(ii) By rearranging the terms in (i) we get $a_{k_1}/b_{k_1} \geq a_{k_2}/b_{k_2}$ for any $k_1 \in \{0, 1, \dots, K-1\}$ and $k_2 > k_1$. □

(iii) Since (i) holds and by (ii), $a_{k_1}/b_{k_1} \geq a_{k_2}/b_{k_2}$, the result is immediate. □

Lemma D.5. *Let for a positive integer K be $\dfrac{a_0}{b_0} \geq \dfrac{a_1}{b_1} \geq \cdots \geq \dfrac{a_K}{b_K}$ such that $b_0, b_1, \dots, b_K > 0$. Let $\alpha_1, \dots, \alpha_K \geq 0$ and $1 \geq \gamma_1 > \gamma_2 > \cdots > \gamma_K \geq 0$. Then*

$$\frac{a_0 + \alpha_1 a_1 + \cdots + \alpha_K a_K}{b_0 + \alpha_1 b_1 + \cdots + \alpha_K b_K} \leq \frac{a_0 + \gamma_1\alpha_1 a_1 + \cdots + \gamma_K\alpha_K a_K}{b_0 + \gamma_1\alpha_1 b_1 + \cdots + \gamma_K\alpha_K b_K}. \tag{D.4}$$

Proof. Consider $k = 1$ and $\gamma_1 < 1$ (for $\gamma_1 = 1$ it trivially holds). Hence, we can multiply $\dfrac{a_0}{b_0} \geq \dfrac{a_1}{b_1}$ by the positive expression $\alpha_1(1 - \gamma_1)b_0 b_1$, and add $a_0 b_0 + \gamma_1\alpha_1^2 a_1 b_1$, which after rearranging gives the desired result. To proceed by induction, suppose that the claim holds for all $k = 1, 2, \dots, K-1$, and suppose $\alpha_1 > 0$ (otherwise it is true by the induction assumption). Denote

$$d_a := a_1 + \frac{\gamma_2}{\gamma_1}\frac{\alpha_2}{\alpha_1}a_2 + \cdots + \frac{\gamma_K}{\gamma_1}\frac{\alpha_K}{\alpha_1}a_K, \qquad d_b := b_1 + \frac{\gamma_2}{\gamma_1}\frac{\alpha_2}{\alpha_1}b_2 + \cdots + \frac{\gamma_K}{\gamma_1}\frac{\alpha_K}{\alpha_1}b_K,$$

and further denote by c_a and c_b the expressions obtained from d_a and d_b by omitting γ's. Note that we want to show

$$\frac{a_0 + \alpha_1 c_a}{b_0 + \alpha_1 c_b} \leq \frac{a_0 + \gamma_1\alpha_1 d_a}{b_0 + \gamma_1\alpha_1 d_b}. \tag{D.5}$$

Using Lemma D.2 and multiplying both numerator and denominator by $\alpha_1(1-\gamma_1)$ we obtain

$$\frac{a_1}{b_1} \geq \frac{\alpha_1(1-\gamma_1)a_1 + \alpha_2(1-\gamma_2)a_2 + \cdots + \alpha_K(1-\gamma_K)a_K}{\alpha_1(1-\gamma_1)b_1 + \alpha_2(1-\gamma_2)b_2 + \cdots + \alpha_K(1-\gamma_K)b_K},$$

and therefore we can write

$$\frac{a_0}{b_0} \geq \frac{\alpha_1 c_a - \gamma_1\alpha_1 d_a}{\alpha_1 c_b - \gamma_1\alpha_1 d_b},$$

which is the same as $a_0(\alpha_1 c_b - \gamma_1\alpha_1 d_b + b_0) \geq b_0(\alpha_1 c_a - \gamma_1\alpha_1 d_a + a_0)$.

On the other hand, the induction assumption implies $\frac{c_a}{c_b} \leq \frac{d_a}{d_b}$, hence $0 \geq \alpha_1^2\gamma_1(c_a d_b - d_a c_b)$. Adding up the last two inequalities and rearranging yields (D.5). □

D.3 Normalization of the Optimization Problem

Next we develop the matrices needed for normalization of the above general model setting. We have

$$\boldsymbol{P}^0 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N-1 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \cdots & N-1 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & \ddots & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array},$$

and therefore

$$\boldsymbol{I} - \beta\boldsymbol{P}^0 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N-1 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \cdots & N-1 \end{array} \\ \begin{pmatrix} 1-\beta & 0 & 0 & 0 & 0 \\ -\beta & 1 & 0 & 0 & 0 \\ -\beta & 0 & 1 & 0 & 0 \\ -\beta & 0 & 0 & \ddots & 0 \\ -\beta & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Is is easy to see that $\boldsymbol{I} - \beta\boldsymbol{P}^0$ is invertible if and only if $\beta \neq 1$, and the inverse is

$$(\boldsymbol{I} - \beta\boldsymbol{P}^0)^{-1} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N-1 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \cdots & N-1 \end{array} \\ \begin{pmatrix} \frac{1}{1-\beta} & 0 & 0 & 0 & 0 \\ \frac{\beta}{1-\beta} & 1 & 0 & 0 & 0 \\ \frac{\beta}{1-\beta} & 0 & 1 & 0 & 0 \\ \frac{\beta}{1-\beta} & 0 & 0 & \ddots & 0 \\ \frac{\beta}{1-\beta} & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Further,

$$\boldsymbol{P}^1 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N-2 \\ N-1 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \cdots & N-2 & N-1 \end{array} \\ \begin{pmatrix} 1-p_0 & p_0 & 0 & 0 & 0 & 0 \\ 1-p_1 & 0 & p_1 & 0 & 0 & 0 \\ 1-p_2 & 0 & 0 & p_2 & 0 & 0 \\ \vdots & 0 & 0 & 0 & \ddots & 0 \\ 1-p_{N-2} & 0 & 0 & 0 & 0 & p_{N-2} \\ 1-p_{N-1} & 0 & 0 & 0 & 0 & p_{N-1} \end{pmatrix} \end{array},$$

and therefore

$$\boldsymbol{I} - \beta\boldsymbol{P}^1 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N-2 \\ N-1 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \cdots & N-2 & N-1 \end{array} \\ \begin{pmatrix} 1-\beta(1-p_0) & -\beta p_0 & 0 & 0 & 0 & 0 \\ -\beta(1-p_1) & 1 & -\beta p_1 & 0 & 0 & 0 \\ -\beta(1-p_2) & 0 & 1 & \ddots & 0 & 0 \\ \vdots & 0 & 0 & \ddots & \ddots & 0 \\ -\beta(1-p_{N-2}) & 0 & 0 & 0 & 1 & -\beta p_{N-2} \\ -\beta(1-p_{N-1}) & 0 & 0 & 0 & 0 & 1-\beta p_{N-1} \end{pmatrix} \end{array}.$$

Finally,

$$(\boldsymbol{I} - \beta\boldsymbol{P}^1)(\boldsymbol{I} - \beta\boldsymbol{P}^0)^{-1} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N-2 \\ N-1 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \cdots & N-2 & N-1 \end{array} \\ \begin{pmatrix} 1+\beta p_0 & -\beta p_0 & 0 & 0 & 0 & 0 \\ \beta p_1 & 1 & -\beta p_1 & 0 & 0 & 0 \\ \beta p_2 & 0 & 1 & \ddots & 0 & 0 \\ \vdots & 0 & 0 & \ddots & \ddots & 0 \\ \beta p_{N-2} & 0 & 0 & 0 & 1 & -\beta p_{N-2} \\ \beta p_{N-1} & 0 & 0 & 0 & 0 & 1-\beta p_{N-1} \end{pmatrix} \end{array}.$$

Thus, for the normalized LP formulation we set (cf. (A.1)), for all $n \in \mathcal{N}$,

$$R_n := (R_n^1 - R_n^0) + \beta p_n (R_{n+1}^0 - R_0^0), \qquad W_n := (W_n^1 - W_n^0) + \beta p_n (W_{n+1}^0 - W_0^0).$$

where we have defined $R_N^0 := R_{N-1}^0$ and $W_N^0 := W_{N-1}^0$.