Markov Decision Processes

the control of high-dimensional systems

Sandjai Bhulai

Markov Decision Processes

the control of high-dimensional systems

Bhulai, Sandjai, 1976 –
Markov Decision Processes: the control of high-dimensional systems
ISBN : 90-9015791-3
NUGI : 811

The figure on the cover is due to M. Dodge and R. Kitchin, "Atlas of Cyberspace".

Printed by Universal Press, The Netherlands.

VRIJE UNIVERSITEIT

Markov Decision Processes
the control of high-dimensional systems

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof. dr. T. Sminia,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Exacte Wetenschappen
op dinsdag 11 juni 2002 om 15.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Sandjai Bhulai

geboren te Paramaribo, Suriname

promotor: prof. dr. G.M. Koole

मेरे माता और पिता के लिये

Preface

This thesis marks an important period in my life during which many people have played an important role. Although the author's name is the only one to appear on the cover, it is the combined effort of these people that helped me in writing this thesis. Therefore, I would like to take this opportunity to thank several individuals for their contributions, realizing that such a list can never be complete.

First and foremost, I would like to express my deepest gratitude to my supervisor Ger Koole. He created an atmosphere with encouragement, faith, freedom, and motivation that made it possible for me to conduct the research in this thesis. I have learned a lot from him, and he has also been a great source of inspiration, even outside the scope of scientific research. Perhaps this is the reason that he succeeded in getting me to do sports; a seemingly impossible task where already many other people have been failed.

I would like to thank the members of my reading committee (Philippe Nain, Jan van Schuppen, Floske Spieksma, Henk Tijms, and Aad van der Vaart) for carefully reading the manuscript. I am also grateful to René Swarttouw for his pleasant discussions on special functions, Michael van Hartskamp for his expert advice on typesetting, Perry Groot for his fine companionship, and Martine Reurings for the necessary tea breaks during work. I am greatly indebted to my colleagues, family, friends, and students; I cherish many good memories, and their interest in my work has been an impetus for going on enthusiastically.

Special thanks go to my brothers Aniel and Rajesh, who put up with the long and weird hours that go into doing mathematics. Finally, I would like to thank my parents for their continuous love and support.

Sandjai Bhulai
June 2002

Table of Contents

Chapter 1

Introduction

1.1 Markov decision problems

In a Markov decision problem we are given a dynamical system whose state may change over time. A decision maker can influence the state by a suitable choice of some of the system's variables, which are called actions or decision variables. The decision maker observes the state of the system at specified points in time, called decision epochs, and gathers information necessary to choose actions in order to control the system. As a result of applying a chosen action to the system, the decision maker incurs an immediate cost. Furthermore, the system changes to a new state according to a probability distribution. In general, the immediate cost and the transition probability function depend on the state and the chosen action.

The outcome of each decision is not fully predictable, but can be anticipated to some extent before the next decision is made through the probability distribution. Also, the actions applied to the system have a long-term consequence. Decisions made at the current epoch have an impact on decisions at the next epoch and so forth. Therefore, decisions cannot be viewed in isolation since one must balance the desire for a low present cost with the undesirability of high future costs. Hence, good decision rules are needed to specify which actions should be chosen at any given epoch and state. A sequence of decision rules will be called a control policy.

Each control policy generates a sequence of costs, and thus measures the system's response to the actions taken. A function defined on the set of control policies that represents the tradeoff the decision maker has to make between present and future costs is called a criterion function. The total discounted cost criterion indicates the importance of present costs, whereas the long-run average cost criterion reflects the importance of future costs. The Markov decision

problem is to find a policy that minimizes the criterion function prior to the first decision epoch. Markov decision theory formally interrelates the set of states, the set of actions, the transition probabilities, and the cost function in order to solve this problem. In Section 2.1 we will show this by formally introducing Markov decision processes. The discounted cost and the average cost criterion will be the subject of study in Sections 2.3 and 2.4, respectively.

The roots of Markov decision theory can be traced back to the pioneering work of Wald [124, 125] on sequential analysis and statistical decision functions. In the late 1940s and in the early 1950s the essential concepts of Markov decision theory were formulated by several researchers working on sequential game models. Bellman [16] identified common ingredients to these problems and is credited for founding the subject through his work on functional equations, dynamic programming, and the principle of optimality.

The discounted Markov decision problem was studied in great detail by Blackwell. Blackwell [28] established many important results, and gave considerable impetus to the research in this area motivating numerous other papers. Howard [65] was the first to study Markov decision problems with an average cost criterion. His introduction of the policy iteration algorithm to find optimal policies can be seen as the first major computational breakthrough in Markov decision theory. Other computational methods were later proposed; Manne [85] gave a linear programming formulation, and White [127] introduced the value iteration, or successive approximations, technique to find optimal policies.

The work discussed so far gives a concise treatment of Markov decision problems with finite state and action spaces. From a theoretical viewpoint existence, uniqueness, and characterization of optimal policies were derived, and from a computational viewpoint algorithms to derive these optimal policies were established. The situation is quite different for problems with denumerably infinite state spaces. In this case, optimal policies need not exist (see, e.g., Maitra [84]). Derman [41] studied the problem with denumerable state spaces, finite action spaces, and bounded cost functions. This paper in conjunction with Derman and Veinott [42] showed that a sufficient condition for the existence of optimal policies is that the expected hitting time of a fixed state under any stationary deterministic policy is bounded uniformly with respect to the choice of the policy and the initial state. In subsequent works many variants of this type of recurrence condition appeared (see, e.g., Federgruen, Hordijk and Tijms [46] and Thomas [120]).

Lippman [80, 82] realized that unbounded cost functions could be handled as well by imposing polynomial bounds on the movement of Markov decision processes in one transition. Moreover, he assumed that there exists a stationary

deterministic policy such that both the mean first passage times and mean first passage costs from any state to a fixed state under the policy are finite. The results could be directly applied to discounted Markov decision problems. This approach is adopted in Section 2.3 to study discounted Markov decision problems. Results for the average cost criterion were obtained by a limiting argument in which the discount factor goes to one. In a series of papers, Sennott [107, 109, 110] relaxed the conditions of Lippman and identified very general conditions.

A direct approach to establish the same results for average cost Markov decision problems is closely related to the ergodic behaviour of the process. To study the ergodic theory, one typically requires some recurrence conditions. Hordijk [63] used a direct approach using a Lyapunov stability condition, which necessarily imposes positive recurrence to a fixed state, the so-called attraction point. Federgruen, Hordijk and Tijms [47] extended Hordijk's result by replacing the single attraction point by a finite set. This work was further extended by Federgruen, Schweitzer and Tijms [48] by dropping the unichain assumption. The conditions used in these papers are more restrictive than Sennott's conditions, but they are easier to verify and provide explicit formulas and convergence results of algorithms.

Dekker and Hordijk studied recurrence conditions in the setting of Blackwell optimality. Their analysis was based on Laurent series expansion techniques. In the case of denumerable state spaces the partial Laurent series expansions are not available without additional assumptions. Dekker and Hordijk [38] derived existence under the uniform geometric ergodicity condition. This ergodicity condition was related to more easily verifiable recurrence conditions in Dekker and Hordijk [39] and Dekker, Hordijk, and Spieksma [40]. In Section 2.2 we shall discuss some of the recurrence conditions appearing in these papers. These conditions will be used in Section 2.4 in order to study average cost Markov decision problems.

Dekker [37] studied the policy iteration algorithm for denumerable Markov decision processes with compact action spaces and unbounded cost functions. He showed convergence of the algorithm using the ergodicity and recurrence conditions in Dekker and Hordijk [38, 39]. An important step in the policy iteration algorithm is solving the Poisson equations, see Equations (2.11) and (2.12), for the long-run expected average cost and the average cost value function. Dekker showed that the solution to the Poisson equations is unique by using partial Laurent series expansions. Bhulai and Spieksma [27] gave a novel probabilistic proof of this fact using relations between ergodicity and recurrence derived in Spieksma [117]. This result is stated in Section 2.4.

1.2 Control of queueing systems

In its simplest form, a queueing system can be described by customers arriving for service, waiting for service if it cannot be provided immediately, and leaving the system after being served. The term customer is used in a general sense here, and does not necessarily refer to human customers. The performance of the system is usually measured on the basis of throughput rates or the average time customers remain in the system. Hence, the average cost criterion is usually the preferred criterion in queueing systems (see, e.g., Puterman [96], and Stidham and Weber [118]).

It is hardly necessary to emphasize the applicability of theory on queueing systems in practice, since many actual queueing situations occur in daily life. During the design and operation of such systems, many decisions have to be made; think of the availability of resources at any moment in time, or routing decisions. This effect is amplified by the advances in the field of data and information processing, and the growth of communication networks (see, e.g., Altman [2]). Hence, there is a need for optimal control of queueing systems.

In general, the control of queueing systems does not fit into the framework of discrete time Markov decision problems, in which the time between state transitions is fixed (as introduced in Section 1.1). When the time spent in a particular state follows an arbitrary probability distribution, it is natural to allow the decision maker to take actions at any point in time. This gives rise to decision models in which the system evolution is described continuously in time, and in which the cost accumulates continuously in time as well. If the cost function is independent of the time spent in a state, such that it only depends on the state and action chosen at the last decision epoch, then we can restrict our attention to models in which the decision epochs coincide with the transition times. Moreover, if the time spent between decision epochs follows an exponential distribution, then the queueing system under study can be reformulated as a discrete-time Markov decision problem. This discretization technique, known as uniformization, is due to Lippman [81], and was later formalized by Serfozo [112]. The uniformization method will be illustrated in Section 2.5.

After uniformization, Markov decision theory can, in principle, be applied to the control of queueing systems, and usually gives rise to infinite state spaces with unbounded cost functions. In this setting, the discussed value iteration and policy iteration algorithms (see pages 20 and 24, respectively) can be used to derive optimal policies. However, these algorithms require memory storage of a real-valued vector of at least the size of the state space. For denumerably infinite state spaces this is not feasible, and one has to rely on appropriate techniques to

bound the state space. But even then, memory may be exhausted by the size of the resulting state space; this holds even more for multi-dimensional models. This phenomenon, known as the curse of dimensionality (see Bellman [17]), gives rise to high-dimensional systems and calls for approximation methods to the optimal policies.

A first approximation method can be distilled from the policy iteration algorithm (see Section 2.4). Suppose that one fixes a policy which, while perhaps not optimal, is not totally unreasonable either. This policy is evaluated through analytically solving the Poisson equations (2.9) and (2.10) induced by the policy in order to obtain the long-run expected average cost and the average cost value function under this policy. Next, using these expressions the policy can be improved by doing one policy improvement step. From Theorem 2.8 we know that this policy is better, i.e., has a lower or equal (if already optimal) average cost. It must be expected that the improved policy is sufficiently complicated to render another policy evaluation step impossible.

The idea of applying one-step policy improvement goes back to Norman [93]. It has been successfully applied by Ott and Krishnan [94] for deriving state-dependent routing schemes for high-dimensional circuit-switched telephone networks. The initial policy in their system was chosen such that the communication lines were independent. In that case, the value function of the system is the sum of the value functions for the independent lines, which are easier to derive (see Theorem 2.11). Since then, this idea has been used in many papers, see, e.g., Hyytiä and Virtamo [66], Rummukainen and Virtamo [102], Sassen, Tijms and Nobel [103], and Towsley, Hwang and Kurose [123]. The initial policy in these papers was chosen such that the queues were independent, reducing the analysis to single queues. For this purpose, Bhulai and Koole [26] presented a unified approach to obtain the value function of the multi-server queue for various cost structures, with the single and infinite server queue as special cases.

In Chapter 3 the result of Bhulai and Koole [26] is generalized to both finite and infinite buffer Markovian birth-death queueing systems with arbitrary cost structures. The Poisson equations of these queueing systems give rise to a set of second-order difference equations for which one homogeneous solution is always known. In Section 3.1 we show that set of equations is reduced to a set of first-order difference equations for which the solution is easy to derive. Hence, knowledge of one homogeneous solution to the set of second-order difference equations enables us to obtain the second homogeneous solution as well as the particular solution for any cost structure.

In Section 3.2 we show that the infinite buffer case does not have a unique

solution to the Poisson equations without further conditions. Here, the ergodicity and recurrence conditions of Chapter 2 play an important role; by constructing a suitable norm, such that the value function is finite with respect to the norm, the unique stable solution among the set of all solutions can be selected. This technique is applied to the multi-server queue in Section 3.3, and the single and infinite server queue in Section 3.4. As an illustration of the one-step policy improvement method, the value function of the multi-server queue will be used to derive a nearly optimal control policy for routing to several finite buffer queues with multiple servers in Section 3.5.

Koole and Spieksma [75] also studied the multi-server queue and obtained explicit expressions for the deviation matrices. The deviation matrix is independent of the cost structure. It enables one to compute the long-run expected average cost and the average cost value function for various cost structures (depending on the state only) by evaluating a sum involving entries of the deviation matrix. However, the expressions they derived for the deviation matrix are cumbersome and difficult to use in practice. Their results can be derived by the method presented in Chapter 3 in an easier way and yields simpler expressions.

Koole and Nain [73] solved the Poisson equations of a two-class priority queue with discounting. Moreover, they were able to explain all the terms appearing in the value function, mostly in terms of busy periods of single server queues. Their work is perhaps the first to study queueing models for which the initial policy is chosen such that there is dependency between the queues. The Poisson equations for such systems give rise to partial difference equations, and make the analysis considerably difficult. This fact is illustrated in Section 4.1 where two independent single server queues are studied. Although the set of solutions to the Poisson equations is easy to derive, it is difficult to formulate a norm such that the unique solution can be selected among this set. In Section 4.3 a tandem model is analyzed. The solution to the Poisson equations turns out to be very hard to obtain, and only partial results for the special case with no arrivals are derived.

The insight obtained in Koole and Nain [73] was used by Groenevelt, Koole, and Nain [55] to derive the value function for the average cost counterpart of the same model. The results for both the discounted and average cost problems was generalized to an arbitrary number of queues by Koole and Nain [74]. Groenevelt, Koole, and Nain did not derive all the solutions to the Poisson equations, and they did not show the correctness of their solution. In Section 4.2 we will complement their results by providing all solutions to the Poisson equations. Furthermore, we shall use the ergodicity and recurrence conditions of Chapter 2 to show the

correctness of their solution.

A second approximation method to circumvent the curse of dimensionality is reinforcement learning; a method that has been primarily developed within the artificial intelligence community several decades ago. Despite the lack of a firm mathematical foundation, early applications of this technique proved successful in a number of situations (see, e.g., Barto, Sutton, and Anderson [13], and Watkins [126]). Bertsekas and Tsitsiklis [23] reintroduced this concept under the name of neuro-dynamic programming, and provided a mathematical foundation for it. Based on this work other authors have been able to apply reinforcement learning successfully to queueing systems (see, e.g., Carlstrom and Nordstrom [31], Marbach, Mihatsch, and Tsitsiklis [86], Nordstrom and Carlstrom [92], and Tong and Brown [122]).

The main idea of this approach is to construct an approximate representation, called the scoring function, of the optimal value function. Typically, constructing a scoring function involves approximating the optimal value function by a certain functional form with free parameters. Next, these parameters are tuned so as to provide the best fit to the optimal value function. The term learning stems from the tuning process, whereas the tuning itself is usually done via an artificial intelligence mechanism, called the reinforcement mechanism.

When the scoring function is fixed, various techniques centered around the optimality equations are used to compute the suboptimal control. However, in order to reduce the dimensionality of the resulting system, it is necessary to describe the scoring function with few free parameters. The choice of the scoring function is essential to the successful application of this method. Hence, this method often requires a considerable amount of trial and error, and a good solid knowledge of the system under study.

In Section 3.4 we show that the value function of the infinite buffer single server queue is quadratic. When the buffer is finite then linear and exponential terms appear due to boundary effects. The infinite buffer infinite server queue has a linear value function, and contains hypergeometric terms in case of a finite buffer. The hypergeometric terms are due to the fact that the service rates depend on the state. In Section 3.3 we show that the value function of the multi-server queue has linear, quadratic, exponential, and hypergeometric terms. The value function of the priority queue in Section 4.2 contains linear and quadratic terms. Exponential terms also appear as a function of the number of prioritized customers in the system. Furthermore, cross-terms appear (expressed as the product of the number of prioritized customers and the regular customers) due to dependency between the queues.

It is apparent that for one-step policy improvement as well as reinforcement learning it is necessary to have some insight into the optimality equations. For one-step policy improvement an explicit solution to the optimality equations for a fixed policy is required. For reinforcement learning the structure (e.g., linear, or quadratic in the number of customers in the system) of the solution to the optimality equations for all policies should be known. In Section 2.4 we show that solutions to the Poisson equations are not unique when working in the standard Banach space without additional conditions. When an arbitrary solution is picked among the set of all solutions, convergence of the policy iteration algorithm is not guaranteed, and can be very slow (see, e.g., Chen and Meyn [32], and Meyn [88]). This might lead to bad approximations using one-step policy improvement, just as an inadequate structure for the scoring function in reinforcement learning (see Bertsekas and Tsitsiklis [23, Ch. 8]). Therefore, the results obtained in Chapter 3 and 4 are of practical importance for both one-step policy improvement and reinforcement learning.

The first part of this thesis is concluded with some remarks on the computation of value functions in Section 4.4.

1.3 Partial information models

In a Markov decision problem with partial information the decision maker is faced with a Markov decision problem in which he does not have access to all information. We distinguish two cases: the state the process occupies is not completely observed, and the transition law is not completely known. The former case gives rise to partially observed Markov decision problems and the latter case to Bayesian decision problems.

Drake [43] formulated the first explicit partially observed Markov decision problem. In this problem, the decision maker cannot observe the state of the process. Instead, after every decision he gains some probabilistic information about the state. For a given probability distribution on the initial states, the decision maker can revise this distribution according to Bayes' rule. This approach was proposed by Drake [43], along with Aoki [8], Dynkin [44], and Shiryaev [113]. Furthermore, they showed that by using a Bayesian approach for estimating the state the process occupies, the problem could be transformed to a Markov decision problem with complete information.

Bellman [17] was the first to study Markov decision problems with a transition law that is not completely known. He assumed that the transition probabilities and the cost function depend on a parameter that is not known to the decision

maker. In this problem, the observed states provide information about the value of the unknown parameter. Bellman independently observed that the problem could be transformed into an equivalent Markov decision problem by using the revision scheme based on Bayes' rule.

The transformation to a Markov decision problem with complete information is similar in both partial information problems. The key idea in this transformation is to augment the state space with the set of probability distributions defined on the domain of the unknown quantity (i.e., the unobserved state, or the unknown parameter). The purpose of this construction is to store a probability distribution at every decision epoch that reflects the likely values the unknown quantity can assume. The distribution that one starts with is called the prior distribution. Subsequent probability distributions, called the posterior distributions, are constructed on basis of observations according to Bayes' rule. This procedure results in a process that is Markovian and permits to recast the partial information problem to a Markov decision problem with complete information. The first systematic proof of validity of the transformation was given by Hinderer [62] for the case of denumerable state and action spaces. Sawaragi and Yoshikawa [104] obtained similar results for denumerable state spaces and Borel action spaces, which was extended by Rhenius [97] to Borel state and action spaces. Rieder [98] generalized the results for completely separable metric state and action spaces.

In the literature the partially observed Markov decision problem and the Bayesian decision problem are treated separately as two different problems. However, due to the similar approach for estimating the unknown quantity in both partial information problems, it is possible to handle them simultaneously. In Section 5.1 we formulate the partially observed Markov decision problem, such that it includes the Bayesian decision problem as a special case. Indeed, by specifying that the parameter is part of the state space, its value becomes unknown in the partially observed Markov decision problem. In order to establish this, we need to allow the state space to be a Borel space. In Section 5.2 we show how the Markov decision problem with complete information, as discussed in the literature, can be obtained from the results in Section 5.1.

The transformation of the partial information problem to the Markov decision problem with complete information comes with added computational difficulties. In a denumerable state Markov decision problem a policy can be expressed in simple tabular form, listing optimal actions for each state. However, when dealing with partial information problems the equivalent Markov decision problem with complete information is augmented with the space of probability distributions. Hence, the policies are now defined over a continuum of states, and this is the

fundamental problem in developing algorithms for computing optimal policies.

In Section 5.3 we discuss the computational aspects of the partially observed Markov decision problem. Based on Koole [71], we show how the equivalence between the partially observed Markov decision problem and the Markov decision problem with complete information can be established in an alternative way. This result is very general and allows for different information structures, e.g., the case where the decision rules depend on the complete history of observations, the last observation, or no observations at all. Moreover, this result is not limited to the construction with the augmented state space only. Suppose that one has an alternative description for the state space, such that the information about the states are preserved, then it can be used to show equivalence of the original model with the alternative model. This transformation is worth doing when the state space of the alternative model has lower dimensionality. In Chapter 6 we shall illustrate this approach for a class of routing problems.

For the Bayesian decision problem additional results are available due to the special structure of the problem. DeGroot [36] shows how to construct parameterized families of distributions, known as a conjugate family of distributions, such that the revised distribution belongs to this family if the prior distribution is chosen within this family. Therefore, instead of storing a probability distribution at every decision epoch, it suffices to store the vector of the parameters that determines the distribution. In Section 5.4 we shall show how to construct such a family of distributions for various distributions of the transition law.

Applications that can be modeled as partially observed Markov decision problems can be found in a lot of different areas, e.g., inventory management, learning theory, manufacturing systems, medical sciences, optimal stopping, and quality control (see Chapter 6 in Bertsekas [21], van Hee [58], Kumar [77], and Monahan [90]). The earliest applications, however, deal with inventory management systems with unknown demand. Scarf [105, 106] studied an inventory control model with linear purchasing, holding, and shortage costs. In this model the demand distribution was assumed to be a member of an exponential family with one unknown parameter. Scarf constructed a conjugate family for this model, which was extended to a model with more general cost functions by Karlin [69] and Iglehart [67]. Azoury [11] generalized this result for a broad class of continuous demand distributions, including the Gamma, normal, uniform, and Weibull distributions.

Azoury and Miller [12] studied the flexibility of ordering policies in Bayesian inventory models, and showed that in most cases more flexible ordering policies were obtained in Bayesian decision models than in non-Bayesian models. In this

setting, an action is considered more flexible compared with a different action, if the feasible action set at the next decision epoch is larger under the first action than under the second. A similar result was obtained in Strijbosch and Moors [119]. They studied a periodic review inventory control model with an order-up-to-level policy. For common criteria, such as the fill-rate and the stock-out probability, they showed that replacing the unknown demand with an estimate can lead to substantially higher safety stocks. Moreover, they showed that even exponential smoothing may fall short in some cases. Therefore, they identified the need for adaptive policies in such inventory systems.

Freedman, Parmar, and Spiegelharter [50] discuss the use of partially observed decision models in medical sciences. They argue that the Bayesian approach to estimation allows a formal basis for using external evidence (by formulating an adequate prior distribution), and in addition it provides a rational way of dealing with accumulating data and the prediction of the consequence of continuing a study. The paper contains a discussion in which the opinion of over 40 people, working in the field of statistics and medical sciences, is presented. The main criticism was focused on how to obtain an adequate prior distribution, and the role of this distribution. However, many discussants seem to agree that partially observed decision models can play an important role in medical sciences, in particular in clinical trials.

In Chapters 6 and 7 we shall discuss two applications of partially observed Markov decision problems. Chapter 6 discusses open-loop routing problems. Here, a queueing system with several parallel servers with no waiting room is considered. Upon arrival of a customer, the decision maker must route this customer to one of the servers in the system. However, the decision maker does not have knowledge about the state of the servers (i.e., idle or busy). When a customer is routed to a busy server, the customer in service is lost. The problem is to minimize the number of lost customers in the system. Similar problems have already been studied by Coffman, Liu and Weber [34], Combé and Boxma [35], Koole [72], and Rosberg and Towsley [99]. However, exact results derived using these models depend on the assumption that the interarrival times of the arrival process are independent identically distributed.

In general, structural results for the optimal policies in partially observed Markov decision process are difficult to obtain, due to the potential error in determining the underlying state of the process (see, e.g., Monahan [90] and White [128]). Koole [72] showed that for the special case of symmetrical servers and arrival processes, the optimality of the round-robin policy was established. For the case of 2 servers, it was shown that a periodic policy whose period has the

form $(1, 2, 2, 2, \ldots, 2)$ is optimal, where 2 means sending a packet to the faster server. Similar results have been obtained in a dual model of server assignment in Coffman, Liu, and Weber [34] with a somewhat different cost. Altman, Bhulai, Gaujal, and Hordijk [3] also showed that the optimal policy is periodic under general stationary arrival processes. This result was later generalized to general convex cost functions in Altman, Bhulai, Gaujal, and Hordijk [4]. The results of this work are presented in Sections 6.2–6.5.

Hajek [57] introduced a novel approach, based on multimodularity, to solve such problems in the context of queueing systems. Elementary properties of multimodular functions and their relations with open-loop optimal control problems were further developed in Altman, Gaujal and Hordijk [6]. It was shown that the class of control problems with a multimodular structure is rich and wide. In Altman, Gaujal, Hordijk and Koole [7] the problem under consideration was solved by using this concept under the assumption that there is only one server, and that the process of interarrival times is stationary. Koole and van der Sluis [76] showed that for the optimization of multimodular functions a local search algorithm could be developed that is guaranteed to converge to the optimal value. In Section 6.3 some properties of multimodular functions are explored. We show that the cost function in the model is multimodular. Moreover, we derive computational results by showing that the local search algorithm of Koole and van der Sluis holds for any convex subset of the search space.

In Section 6.4 we consider three explicit arrival processes: the Poisson process, the Markov modulated Poisson process, and the Markovian arrival process. We show that for all these arrival processes, it is possible to obtain an explicit expression for the cost function. Note that the Markovian arrival process is a process that can model both bursty and regular input. Moreover, every marked point process is the limit of a sequence of Markovian arrival processes. As such, the Markovian arrival process is a versatile family of arrival processes. The chapter is concluded by applying the results to the problem of robot scheduling for web search engines in Section 6.5.

Chapter 7 considers multi-armed bandit problems. This class of problems is motivated by the design of clinical trials (see Bellman [15]). In a general setting the problem can be formulated as follows. Suppose that one has to select a project to work on from a given set of projects. Each project is finished successfully with a fixed unknown probability, and provides the decision maker with a reward of one unit, and zero otherwise. The problem in this setting is to choose a project at each epoch such that the long term discounted rewards are maximized. Indeed, the problem models clinical trials when the projects are seen as different treatments,

and a reward is seen as a healed patient.

The multi-armed bandit problem is a classical problem (see Berry and Fristedt [18], Gittins [52], and Kumar [77] for an overview of the work done in this field). The decision maker has to value obtaining a high reward against acquiring information that can be used to determine which project is profitable in the future. It is this possible tension that makes the problem difficult to solve (see, e.g., Kelly [70], and Kumar and Seidman [78]). The quantification of the amount of information the decision maker has is very important; it helps to decide whether the decision maker should select a less rewarding but more informative project over one that is more rewarding but less informative.

Berry and Kertz [19] have studied the worth of perfect information for multi-armed bandits; they compared the reward of the decision maker with the reward of the decision maker who has perfect information. Gittins and Wang [53] investigated the relationship between the importance of acquiring additional information and the amount of information that is already available. The methods for quantifying the value of information studied in Berry and Kertz [19], and Gittins and Wang [53] are cumbersome in practice, since the computational complexity is too large.

Bhulai and Koole [24] have adopted a direct approach by considering two extreme situations when a fixed number of information samples have been acquired: the situation where the decision maker stops learning, and the situation where the decision maker obtains full information about the project. They showed that the difference in rewards between this lower and upper bound goes to zero as the number of information samples grows large. This approach is illustrated in Section 7.2, and numerical results are given in Section 7.3.

Chapter 2

Markov Decision Processes

Markov decision processes have been applied to a wide range of stochastic control problems. Particularly, inspection-maintenance-replacement systems, inventory management, and economic planning and consumption models were the earliest applications of this theory. Typically, these systems and models have a finite state space with bounded cost functions. A great deal of literature is devoted to such models, see, e.g., Bertsekas [21, 22], Kumar and Varaiya [79], Puterman [96], Ross [101], and Tijms [121]. More general models with Borel state and action spaces are treated in, e.g., Dynkin and Yushkevich [45], Hinderer [62], and Hernández-Lerma and Lasserre [60, 61].

In this chapter we study Markov decision processes with denumerable state spaces, finite action spaces, and unbounded cost functions. After some preliminaries on weighted-norm spaces, both the discounted and the average cost Markov decision problems are discussed. For the average cost case extra emphasis is put on the uniqueness of solutions to the Poisson equations induced by a fixed policy. We derive a novel probabilistic proof to obtain unicity results that are partly based on Bhulai and Spieksma [27]. The material discussed in this chapter is fundamental for both parts of the thesis.

2.1 The canonical construction

Markov decision processes are completely determined by the tuple

$$\big(\mathcal{X}, \{\mathcal{A}_x \mid x \in \mathcal{X}\}, p, c\big).$$

The denumerable set $\mathcal{X}$ is called the state space. The finite set $\mathcal{A}_x$ for $x \in \mathcal{X}$ denotes the set of feasible actions available to the decision maker when the system is in state x. The set $\mathcal{A} = \cup_{x\in\mathcal{X}} \mathcal{A}_x$ is referred to as the action space. Define the set of

feasible state-action pairs $\mathcal{K}$ by

$$\mathcal{K} = \{(x, a) \mid x \in \mathcal{X}, a \in \mathcal{A}_x\}. \tag{2.1}$$

The transition probability function p defined on $\mathcal{X}$ given $\mathcal{K}$ is called the transition law. Finally, the function $c : \mathcal{K} \to \mathbb{R}$ is known as the cost function. Note that for general state spaces one needs measurable selection theorems to ensure that $\mathcal{K}$ contains the graph of a measurable function from $\mathcal{X}$ to $\mathcal{A}$ (see Assumption 2.2.2 in Hernández-Lerma and Lasserre [60]). In the case of denumerable state spaces (with the discrete topology) this requirement is always fulfilled.

Define the set of admissible histories up to epoch t by $\mathcal{H}_t = \mathcal{K}^t \times \mathcal{X}$ for $t \in \mathbb{N}_0$. A generic element $h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t) \in \mathcal{H}_t$ is called a history up to epoch t. A randomized decision rule for epoch t is a non-negative function φ_t on the action set $\mathcal{A}$ given $\mathcal{H}_t$ satisfying the constraint

$$\sum_{a \in \mathcal{A}_{x_t}} \varphi_t(a \mid h_t) = 1,$$

for all $h_t \in \mathcal{H}_t$. A sequence of decision rules $\pi = \{\varphi_t\}_{t \in \mathbb{N}_0}$ is called a randomized control policy. When all φ_t are independent of t, the policy π is called stationary. The policy π is called a Markov policy, when φ_t given h_t depends only on x_t for all $t \in \mathbb{N}_0$. Finally, when all φ_t are functions from $\mathcal{H}_t$ to $\mathcal{A}$ such that $\varphi_t(h_t) \in \mathcal{A}_{x_t}$ (or, equivalently $\varphi_t(a \mid h_t) = 1$ for an $a \in \mathcal{A}_{x_t}$), the policy π is called deterministic. The combinations randomized, randomized Markov, randomized stationary, deterministic, deterministic Markov, and deterministic stationary will be denoted by R, RM, RS, D, DM, and DS respectively. The set of all policies with property K will be denoted by Π_K where $K \in \{R, RM, RS, D, DM, DS\}$. Observe that $\Pi_{RS} \subset \Pi_{RM} \subset \Pi_R$ and $\Pi_{DS} \subset \Pi_{DM} \subset \Pi_D \subset \Pi_R$.

The Markov decision process can now be formally formulated as a stochastic process on a suitable probability space. Define $\Omega = (\mathcal{X} \times \mathcal{A})^\infty$, and let $\mathcal{F}$ be the product σ-algebra on Ω. Define the $\mathcal{X}$-valued random variable X_t by $X_t(w) = x_t$ for $\omega = (x_0, a_0, \ldots) \in \Omega$ and $t \in \mathbb{N}_0$. Similarly, define the $\mathcal{A}$-valued random variable A_t by $A_t(\omega) = a_t$. The history of the process is defined as the random variable H_t by $H_0(\omega) = x_0$, $H_t(\omega) = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t)$ for $t \in \mathbb{N}$ and $H_\infty(\omega) = (x_0, a_0, \ldots)$. Let $x_0 \in \mathcal{X}$ be given, and let $\pi \in \Pi_R$ be an arbitrary control policy. By the theorem of Ionescu-Tulcea (see Proposition C.10 of Hernández-Lerma and Lasserre [60]) there exists a unique probability measure $\mathbb{P}^\pi_{x_0}$ on the measurable space $(\Omega, \mathcal{F})$ determined by x_0 and the policy π, given the transition law p, such that $\mathbb{P}^\pi_{x_0}(\mathcal{H}_\infty) = 1$, and

$$\mathbb{P}^\pi_{x_0}(A_t = a_t \mid H_t = h_t) = \varphi_t(a_t \mid h_t), \tag{2.2}$$

$$\mathbb{P}^{\pi}_{x_0}(X_{t+1} = x_{t+1} \,|\, H_t = h_t, A_t = a_t) = p(x_{t+1} \,|\, x_t, a_t). \tag{2.3}$$

The expectation with respect to this probability measure will be denoted by $\mathbb{E}^{\pi}_{x_0}$.

The stochastic process $\{X_t\}_{t\in\mathbb{N}_0}$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P}^{\pi}_{x_0})$ is called a Markov decision process. The stochastic process $\{X_t\}_{t\in\mathbb{N}_0}$ depends on the particular policy π being used and on the given initial state x_0. This dependence will not be reflected in the notation when it is clear from the context which π and x_0 are used. The collection $\{X_t\}_{t\in\mathbb{N}_0}$ and $\{(\Omega, \mathcal{F}, \mathbb{P}^{\pi}_{x_0}) \,|\, \pi \in \Pi_R\}$ together with a given criterion function to be optimized is called a Markov decision problem.

Note that the construction of the probability space for $\{X_t\}_{t\in\mathbb{N}_0}$ is the same as the construction in the case of finite state spaces. In the latter case, existence and uniqueness of optimal policies can be obtained rather easily. In the former case, additional conditions have to be imposed on the Markov chain induced by the policies. Before moving on to Markov decision problems, we first study Markov chains on weighted-norm spaces, since this will be related to the conditions one needs to impose on the Markov chain.

2.2 Weighted-norm spaces

Consider a discrete-time uncontrolled Markov chain determined by the tuple $(\mathcal{X}, p, c)$. In this setting, $\{X_t\}_{t\in\mathbb{N}_0}$ is the uncontrolled Markov chain having values in the denumerable set $\mathcal{X}$, p is the transition probability function, and c is the cost function. This stochastic process is better known as a Markov reward chain in the literature, where the corresponding reward function is given by $-c$; we shall refer to it as a Markov cost chain. Let P denote the matrix of transition probabilities determined by p. Let $\mathbb{B}(\mathcal{X})$ denote the Banach space of bounded real-valued functions u on $\mathcal{X}$ with the supremum norm, i.e., the norm $\|\cdot\|$ is defined by

$$\|u\| = \sup_{x\in\mathcal{X}} |u(x)|.$$

We want to study Markov chains with unbounded cost functions. However, these cost functions are not contained in the Banach space $\mathbb{B}(\mathcal{X})$. A remedy to this situation is to consider suitable larger Banach spaces instead of the space $\mathbb{B}(\mathcal{X})$. In order to construct such a space, consider a function $w : \mathcal{X} \to [1, \infty)$, which we will refer to as a weight function. The w-norm is defined as

$$\|u\|_w = \left\|\tfrac{u}{w}\right\| = \sup_{x\in\mathcal{X}} \frac{|u(x)|}{w(x)}.$$

Let the matrix A be an operator on $\mathcal{X} \times \mathcal{X}$. The operator w-norm is defined by $\|A\|_w = \sup\{\|Au\|_w : \|u\|_w \leq 1\}$. This norm can be rewritten in the following equivalent form (see Expression (7.2.8) in Hernández-Lerma and Lasserre [61])

$$\|A\|_w = \sup_{i \in \mathcal{X}} \frac{1}{w(i)} \sum_{j \in \mathcal{X}} |A_{ij}|\, w(j).$$

A function u is said to be bounded if $\|u\| < \infty$, and w-bounded if $\|u\|_w < \infty$. Let $\mathbb{B}_w(\mathcal{X})$ denote the normed linear space of w-bounded functions on $\mathcal{X}$. Note that for every function u which is bounded, $\|u\|_w \leq \|u\| < \infty$. Furthermore, it is easy to see that every Cauchy-sequence $\{u_n\}$ in w-norm has a w-limit. Hence, $\mathbb{B}_w(\mathcal{X})$ is a Banach space that contains $\mathbb{B}(\mathcal{X})$.

Consider the transition matrix P of the Markov chain. The 0^{th} iterate P^0 is set equal to the identity matrix. A state that communicates with every state it leads to is called essential; otherwise inessential. The set of essential states is partitioned in a number of closed classes. In each closed class all states communicate with each other. The number of closed classes in the Markov chain will be denoted with κ. A set $B \subset \mathcal{X}$ will be called a set of reference states if it contains exactly one state from each closed class and no other states. Since the stationary distribution of the Markov chain depends on the initial distribution of the chain, we will consider the stationary matrix P^* given by

$$P^*_{ij} = \lim_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n-1} P^t_{ij},$$

for a transition to state $j \in \mathcal{X}$ given that the Markov chain is in state $i \in \mathcal{X}$. The stationary matrix P^* is well-defined, since the limit always exists (see the corollary to Theorem 6.4 in Chung [33]). The Markov chain is said to be stable if P^* is a stochastic matrix. Define the taboo transition matrix for $i \in \mathcal{X}$ and $M \subset \mathcal{X}$ by

$${}_M P_{ij} = \begin{cases} P_{ij}, & j \notin M, \\ 0, & j \in M. \end{cases}$$

We write ${}_M P^n$ for the n^{th} iterate of ${}_M P$, and set ${}_M P^0$ equal to the identity matrix. Let $F^{(n)}_{iM}$ denote the probability that the system, starting in state i, reaches set M for the first time after $n \geq 1$ steps. Then

$$F^{(n)}_{iM} = \sum_{m \in M} ({}_M P^{n-1} \cdot P)_{im}, \quad n \in \mathbb{N}.$$

The probability that set M is eventually reached is given by $F_{iM} = \sum_{n\in\mathbb{N}} F_{iM}^{(n)}$. For a stable Markov chain one has $F_{iB} = 1$ for all sets B of reference states, and all classes are positive recurrent (see Theorems 6.3 to 7.4 of Chung [33]).

Let $M \subset \mathcal{X}$ be finite, and let w be a weight function. The following properties of P will be used in the sequel (cf. Dekker and Hordijk [38, 39], Dekker, Hordijk and Spieksma [40], Hordijk and Spieksma [64]). A Markov chain is called

- w-geometrically ergodic [w-GE] if there are constants $r > 0$ and $\beta < 1$ such that $\|P\|_w < \infty$ and $\|P^n - P^*\|_w \leq r\beta^n$ for $n \in \mathbb{N}_0$;
- w-geometrically recurrent with respect to M [w-GR(M)] if there exists an $\varepsilon > 0$ such that $\|_M P\|_w \leq 1 - \varepsilon$;
- w-weak geometrically recurrent with respect to M [w-WGR(M)] if there are constants $r > 0$ and $\beta < 1$ such that $\|_M P^n\|_w \leq r\beta^n$ for $n \in \mathbb{N}$;
- w-weak geometrically recurrent with respect to reference states [w-WGRRS(M)] if there is a set of reference states $B \subset M$ such that the Markov chain is w-WGR(B).

The relation between these ergodicity and recurrence concepts is summarized in the following lemma.

Lemma 2.1 (Ch. 2 in [117]): Let $M \subset \mathcal{X}$ be finite, and let w be a weight function.

- The condition that P is the transition matrix of an aperiodic Markov chain and that w-WGR(M) holds for M is equivalent to w-GE with $\kappa < \infty$;
- w-GR(M) implies w-WGR(M);
- w-WGR(M) is equivalent to w-WGRRS(M).

In the sequel, w-geometric ergodicity will play an important role. However, it is not easy to check if a Markov chain is w-GE. Lemma 2.1 shows that w-GR(M) with the condition that P is the transition matrix of an aperiodic Markov chain implies w-GE. This condition is easier to check, since w-GR(M) is a property which only concerns the transition matrix P, instead of the power matrix P^n for $n \in \mathbb{N}$. We will use this fact in Chapters 3 and 4 to show that the queueing models under study are indeed geometrically ergodic.

If a Markov chain is geometrically ergodic, then the Markov chain is also geometrically recurrent (see Theorem 1.1 of Spieksma [117]). However, this does not necessarily hold for the same weight function. When the weight function is bounded, then w-geometric ergodicity is equivalent to w-geometric recurrence.

2.3 Discounted Markov decision problems

Let $(\mathcal{X}, \{\mathcal{A}_x \mid x \in \mathcal{X}\}, p, c)$ determine a Markov decision process. Let the discount factor $\alpha \in (0,1)$ be fixed, then the discounted cost criterion is given by

$$V(\pi, x) = \mathbb{E}_x^\pi \sum_{t=0}^{\infty} \alpha^t \, c(X_t, A_t), \quad \pi \in \Pi_R, \; x \in \mathcal{X}.$$

So far, this expression may be undefined; conditions guaranteeing its existence will be given later. The corresponding α-discount value function is

$$V^*(x) = \inf_{\pi \in \Pi_R} V(\pi, x), \quad x \in \mathcal{X}.$$

The Markov decision problem is to find a control policy $\pi^* \in \Pi_R$ such that $V(\pi^*, x) = V^*(x)$ for $x \in \mathcal{X}$. An important operator in the analysis of this problem is the operator T_α defined by

$$T_\alpha u(x) = \min_{a \in \mathcal{A}_x} \left[c(x,a) + \alpha \sum_{y \in \mathcal{X}} p(y \mid x, a) u(y) \right],$$

for $x \in \mathcal{X}$. If the state space is finite, and hence the costs are bounded, then one can easily show that the operator T_α is a contraction mapping on $\mathbb{B}(\mathcal{X})$ with contraction factor α (see Proposition 6.2.4 of Puterman [96]), thus

$$\|T_\alpha u - T_\alpha u'\| \leq \alpha \, \|u - u'\|,$$

for all $u, u' \in \mathbb{B}(\mathcal{X})$. Therefore, by Banach's Fixed-Point Theorem, it follows that there exists a unique function $v^* \in \mathbb{B}(\mathcal{X})$ such that it satisfies the optimality equation $T_\alpha v^* = v^*$, and v^* is the limit of $v_n = T_\alpha v_{n-1} = T_\alpha^n v_0$ with $v_0 = 0$. Finally, one shows that v^* equals the α-discount value function V^*, so that V^* is the unique bounded solution in $\mathbb{B}(\mathcal{X})$ to the optimality equation. The algorithm of obtaining v^* through iterative computation of v_n is known as value iteration.

Lippman [82] was the first to realize that value iteration could be extended to denumerable state spaces with possibly unbounded cost functions (that satisfy a growth condition) when the supremum norm is replaced by a suitable weighted supremum norm. The weighted supremum norm was chosen such that the mapping T_α remained a contraction mapping so that Banach's Fixed-Point Theorem could be applied again. The conditions used by Lippman were relaxed by Sennott [111]. The following theorem summarizes the conditions needed on the Markov chain in order to obtain existence and uniqueness results.

Theorem 2.2 (Ch. 8 in [61]): Suppose that there exist constants $1 \leq \beta < \frac{1}{\alpha}$ and $\overline{c} \geq 0$, and a weight function w such that for any $\pi = \varphi^{\infty} = (\varphi, \varphi, \ldots) \in \Pi_{DS}$

$$\big\|c\big(\cdot, \varphi(\cdot)\big)\big\|_w \leq \overline{c}, \quad \text{and} \quad \|P(\varphi)\|_w \leq \beta, \tag{2.4}$$

with $P(\varphi)$ the transition probability matrix induced by decision rule φ. Then

- The α-discount value function V^* is the unique solution to the optimality equation $V(x) = T_\alpha V(x)$ in the space $\mathbb{B}_w(\mathcal{X})$, and

$$\|v_n - V^*\|_w \leq \frac{\overline{c}\,(\alpha\,\beta)^n}{1 - \alpha\,\beta}, \quad n \in \mathbb{N}.$$

- There exists a deterministic Markovian decision rule φ such that

$$V^*(x) = c\big(x, \varphi(x)\big) + \alpha \sum_{y \in \mathcal{X}} p\big(y \,|\, x, \varphi(x)\big) V^*(y), \tag{2.5}$$

  for all x. Furthermore, the policy $\pi = \varphi^{\infty} \in \Pi_{DS}$ is α-discount optimal. Conversely, if $\pi = \varphi^{\infty} \in \Pi_{DS}$ is α-discount optimal, then it satisfies (2.5).

Note that the growth conditions in Expression (2.4) can be interpreted as follows. The condition that the cost is w-bounded implies that the cost increases by at most rate $w(x)$ for any $a \in \mathcal{A}_x$. The second condition can be rewritten as

$$\sup_{a \in \mathcal{A}_x} \mathbb{E}\left[w(X_{n+1}) \,|\, X_n = x, A_n = a\right] \leq \beta\, w(x), \quad n \in \mathbb{N}_0.$$

Hence, the second condition limits the growth of the weight function itself. Puterman [96, Assumption 6.10.2] considers weaker conditions for Theorem 2.2 by studying J-stage contraction mappings in the Banach space $\mathbb{B}_w(\mathcal{X})$, i.e.,

$$\|T_\alpha^J u - T_\alpha^J u'\|_w \leq \alpha' \, \|u - u'\|_w,$$

for some $J \in \mathbb{N}$ and $\alpha' \in (0, 1)$. The J-stage contraction mappings inherit many of the important properties of contraction mappings. The conditions on the weight function w are different from the conditions in Expression (2.4), however, they are harder to verify in practice.

Theorem 2.2 resembles Theorem 4.2.3 in Hernández-Lerma and Lasserre [60]. However, the result is derived in a different way with different assumptions on the Markov decision chain. The latter result holds for quite general action spaces, and non-negative cost functions which are virtually without restriction in their

growth rate. The former holds for unbounded cost functions, but can only be generalized to compact action spaces.

In general, the solution to the optimality equation is not unique. In addition to the w-bounded solution, there can be other unbounded solutions. However, when one has a weight function w such that the conditions in Expression (2.4) are satisfied, the correct solution can be selected from the set of all solutions by requiring that the solution has a finite norm with respect to the weight function. The following example due to Sennott [108] illustrates this technique.

Example 2.3: Consider the discounted Markov decision problem with state space $\mathcal{X} = \{1, 2, \ldots\}$, and feasible action spaces $\mathcal{A}_x = \{1\}$ for all $x \in \mathcal{X}$. The transition probabilities are given by

$$p(1 \mid x, 1) = \frac{2x}{3\,(2x-1)} \quad \text{and} \quad p(2x \mid x, 1) = \frac{4x-3}{3\,(2x-1)},$$

for all $x \in \mathcal{X}$. The cost function is given by $c(x, a) = 0$ for all $x \in \mathcal{X}$, $a \in \mathcal{A}$.

Let $\alpha = \frac{3}{4}$, then the optimality equation is given by

$$V(x) = \frac{3}{4}\left[\frac{2x}{3\,(2x-1)}\,V(1) + \frac{4x-3}{3\,(2x-1)}\,V(2x)\right].$$

Clearly, $V(x) = 0$ for all $x \in \mathcal{X}$ is a solution. However, $V(x) = x$ is also a solution to the optimality equation, since $x = \frac{1}{2}$ is never attained. Now consider the constant weight function $w(x) = K$ for some $K > 1$. Then $\|P\|_w = 1$, where P is the transition probability matrix independent of $a \in \mathcal{A}$. The conditions in Expression (2.4) are satisfied, thus the unique solution must have a finite norm with respect to the weight function w. Hence, the α-discount value function cannot grow faster than a constant, and must therefore be a constant function itself. Thus, we finally obtain that $V(x) = 0$ is the unique solution in $\mathbb{B}_w(\mathcal{X})$.

2.4 Average cost Markov decision problems

In discounted Markov decision problems the discount factor $\alpha \in (0, 1)$ is crucial for the analysis. It ensures that the mapping T_α is a contraction mapping on $\mathbb{B}_w(\mathcal{X})$ for some suitable weighted norm w, from which uniqueness of solutions to the optimality equations can be derived. Since this will not hold for average cost Markov decision problems, we have to use a different method of analysis.

Observe that in the previous section we ignored the chain structure of the transition matrices of the Markov chains induced by the policies in Π_S. This was

not necessary due to discounting; the long-run behaviour of the system did not have a substantial contribution to the α-discount value function. However, in average cost Markov decision problems results and analyses explicitly depend on communicating state patterns of the Markov chain corresponding to policies in Π_S. We distinguish two classes of chain structures; a Markov decision problem is called

- Unichain if the transition matrix corresponding to every $\pi \in \Pi_{DS}$ consists of a single recurrent class plus a possibly empty set of transient states;
- Multichain if the transition matrix corresponding to at least one $\pi \in \Pi_S$ contains two or more closed irreducible recurrent classes.

In the analysis we shall allow models with a multichain structure, as this does not essentially complicate the analysis. We start with the discussion of multichain models, and afterwards give specialized results for the unichain case.

Consider the average cost criterion given by

$$g(\pi, x) = \lim_{n\to\infty} \mathbb{E}_x^{\pi} \frac{1}{n} \sum_{t=0}^{n-1} c(X_t, A_t), \quad \pi \in \Pi_R, x \in \mathcal{X}. \tag{2.6}$$

This expression may be undefined when the limit does not exist; conditions guaranteeing the existence will be given later. The corresponding long-run expected average cost is given by

$$g^*(x) = \inf_{\pi\in\Pi_R} g(\pi, x), \quad x \in \mathcal{X}.$$

The Markov decision problem is to find a control policy $\pi^* \in \Pi_R$ such that $g(\pi^*, x) = g^*(x)$ for $x \in \mathcal{X}$. Let V denote the average cost value function. From arguments put forward in Section 8.2 of Puterman [96] it follows that the value function $V(x)$ can be seen as the difference in total cost between starting in state x and starting in a reference state, usually taken to be 0, when the induced Markov chain is aperiodic. This interpretation can be helpful in deriving or estimating the structure of the value function in queueing systems (see Chapters 3 and 4).

The optimality equations for $x \in \mathcal{X}$ are given by

$$g(x) = \min_{a\in\mathcal{A}_x} \left[\sum_{y\in\mathcal{X}} p(y \mid x, a) g(y) \right], \tag{2.7}$$

and

$$g(x) + V(x) = \min_{a\in\mathcal{B}_x} \left[c(x, a) + \sum_{y\in\mathcal{X}} p(y \mid x, a) V(y) \right], \tag{2.8}$$

with $\mathcal{B}_x = \{a \in \mathcal{A}_x \mid g(x) = \sum_{y\in\mathcal{X}} p(y \mid x, a)g(y)\}$. Observe that this system of equations is nested, because the set of actions over which the minimum is sought in the second equation depends on the set of actions that attain the minimum in the first equation. The optimality equations can be solved by an iterative procedure called policy iteration. The procedure starts with an arbitrary policy, and yields a sequence of improved policies. This procedure can be implemented under the following assumption.

Assumption 2.4: There exists a weight function w on $\mathcal{X}$ such that the cost function is bounded and the Markov chain is stable, w-GE with $\kappa < \infty$ uniformly and continuous over policies in Π_{DS}, i.e., for any $\pi = \varphi^\infty \in \Pi_{DS}$ the cost function satisfies $\left\|c\big(\cdot, \varphi(\cdot)\big)\right\|_w < \infty$, and there are constants $r > 0$ and $\beta < 1$ such that

$$\begin{cases} \|P(\varphi)\|_w < \infty, \\ \|P^n(\varphi) - P^*(\varphi)\|_w \le r\beta^n, \quad n \in \mathbb{N}_0, \end{cases}$$

with $\kappa(\varphi) < \infty$ continuous on Π_{DS}, where $P(\varphi)$, $P^*(\varphi)$ are stochastic matrices, and $\kappa(\varphi)$ is the number of closed classes induced by decision rule φ.

Policy Iteration

1. Let w be such that the conditions in Assumption 2.4 are met. Choose any policy $\pi_0 = \varphi_0^\infty \in \Pi_{DS}$, and set $n = 0$.
2. Solve g_n and v_n in the Banach space $\mathbb{B}_w(\mathcal{X})$ from

$$g_n(x) = \sum_{y\in\mathcal{X}} p\big(y \mid x, \varphi_n(x)\big) g_n(y), \tag{2.9}$$

   and

$$g_n(x) + v_n(x) = c\big(x, \varphi_n(x)\big) + \sum_{y\in\mathcal{X}} p\big(y \mid x, \varphi_n(x)\big) v_n(y). \tag{2.10}$$

3. Determine a decision rule φ_{n+1} from the set of minimizing actions in the optimality equations (2.7) and (2.8) for g_n and v_n. If it is possible to select $\varphi_{n+1} = \varphi_n$, then the algorithm ends, else step 2 is repeated.

The policy iteration algorithm thus consists of two steps. In the policy evaluation step the long-run expected average cost and the average cost value function are determined for a fixed policy by solving Equations (2.9) and (2.10), known as

the Poisson equations. Then a policy improvement step follows which yields a decision rule providing an improvement through the optimality equations.

In case of finite state spaces, the analysis of the policy iteration algorithm is done via Laurent series expansion techniques. This technique allows one to establish existence and uniqueness of solutions to the Poisson equations. However, in case of denumerable state spaces partial Laurent series expansions are not available without additional assumptions. Dekker [37] provides conditions under which the partial Laurent series expansions exist and are unique. We shall give an alternative probabilistic proof of Dekker's results using relations between ergodicity and recurrence that will clarify the conditions used in Assumption 2.4.

Note that for a fixed policy $\pi = \varphi^\infty \in \Pi_{DS}$ the Markov decision process can be seen as a discrete-time uncontrolled Markov cost chain determined by the tuple $(\mathcal{X}, p, c)$. Here $\{X_t\}_{t\in\mathbb{N}_0}$ is the $\mathcal{X}$-valued uncontrolled Markov chain, p the transition probability function induced by the decision rule φ, and c the cost function. We will not reflect the dependency on the policy π in the notation as this will not play an essential role in the sequel. Denote with P the matrix of transition probabilities determined by the transition law p. Let $\mathbb{B}_w(\mathcal{X})$ be a normed vector space of w-bounded real-valued functions on $\mathcal{X}$ for some weight function w such that $c \in \mathbb{B}_w(\mathcal{X})$.

The Poisson equations (2.9) and (2.10) can then be rewritten as

$$g = Pg, \tag{2.11}$$

$$g + v = c + Pv, \tag{2.12}$$

for some functions g and v in $\mathbb{B}_w(\mathcal{X})$. By virtue of this definition we will call g and v a solution to the Poisson equations when they have a finite norm with respect to w. The following lemma displays the assumptions needed to guarantee the existence of solutions to the Poisson equations.

Lemma 2.5 ([64], Prop 5.1 in [117]): Assume that the induced Markov chain is w-GE with $\kappa < \infty$. Then solutions (g, v) to the Poisson equations exist in the space of w-bounded functions $\mathbb{B}_w(\mathcal{X})$.

Note that by Lemma 2.1 the requirement that P is the matrix of an aperiodic Markov chain with w-GR(M) for some M, instead of w-GE and $\kappa < \infty$ also suffices. The ergodicity conditions are hard to verify in practice, whereas the recurrence conditions as a property of the one-step transition matrices are relatively easy to verify. We prefer to work with the ergodicity condition in the sequel, as this is the more general concept.

As a first step towards proving uniqueness of the solution to the Poisson equations, we study the average cost first. The average cost is uniquely determined by the transition matrix P and the cost function c. The following theorem summarizes this result.

Theorem 2.6: Suppose that the induced Markov chain is w-GE, and that $\kappa < \infty$. Let the functions g and v in $\mathbb{B}_w(\mathcal{X})$ be a solution to the Poisson equations. Then

$$g = \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} P^t g = \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} P^t c = P^* c. \tag{2.13}$$

Proof: By iterating Equation (2.11) we find that $g = P^t g$ for $t \in \mathbb{N}_0$. Taking a finite sum over t from 0 to $n-1$ yields

$$ng = \sum_{t=0}^{n-1} P^t g, \quad n \in \mathbb{N}, \tag{2.14}$$

which implies the first equality. The second equality is obtained by writing Equation (2.12) as $v = (c - g) + Pv$. By iteration of this expression we derive

$$v = \sum_{t=0}^{n-1} P^t (c - g) + P^n v, \quad n \in \mathbb{N}. \tag{2.15}$$

Since the Markov chain is w-GE and $\kappa < \infty$, there exist non-negative constants r and $\beta < 1$ such that

$$\|P^n - P^*\|_w \le r\beta^n, \quad n \in \mathbb{N}.$$

As a consequence there exists a constant C such that $\|P^n\|_w \le C$ for all $n \in \mathbb{N}$. Hence

$$\left\| \tfrac{1}{n} P^n v \right\|_w \le \tfrac{1}{n} \|P^n\|_w \|v\|_w \to 0, \quad n \to \infty.$$

The last equality directly follows from w-geometric ergodicity. □

Note that in Theorem 2.6 the w-GE condition is only used for the second equality. The results of Theorem 2.6 are also stated in Corollary 7.5.6(a) and (b) of Hernández-Lerma and Lasserre [61]. They do not impose that the Markov chain should be geometrically ergodic, but merely assume that $P^n v/n \to 0$ as n grows to infinity. We have shown that this condition is fulfilled when the Markov chain is geometrically ergodic. Hence, this also holds under the more easily verifiable geometric recurrence condition.

Next, we move on to study the properties of the value function v. Since a Markov chain with a multichain structure can have classes with different state classifications, one cannot expect the value function v to be uniquely determined. However, when the Markov chain is stable, then more can be said about the value function. The following result can be found in Dekker [37, Ch. 3, Lemma 2.2] in the setting of Blackwell optimality equations using Laurent series expansion techniques. As mentioned before, we give an alternative probabilistic proof using the relations between ergodicity and recurrence.

Theorem 2.7: Suppose that the induced Markov chain is stable, w-GE, and that $\kappa < \infty$. Then a solution to the Poisson equations is given by

$$g = P^* c, \quad v = \sum_{n=0}^{\infty} (P^n - P^*)\, c = D\, c,$$

with D the deviation matrix. Furthermore, if both (g, v) and (g', v') are solutions to the Poisson Equations, then $g = g'$ and

$$v - v' = \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} P^t (v - v'). \tag{2.16}$$

Moreover, the functions v and v' only differ by a constant on each closed class.

Proof: From Theorem 2.6 it immediately follows that $g = P^* c$ and $g = g'$. The fact that $v = Dc$ is a solution follows from direct insertion into Equation (2.12), and using w-boundedness. Equation (2.12) yields $v - v' = P(v - v')$. Consequently, the derivation of Expression (2.16) proceeds along the same lines as the derivation of Expression (2.13).

For the last statement, write Equation (2.12) as $v = (c - g) + Pv$. Since the Markov chain is w-GE, it is also w-WGRRS(M) for some finite set $M \subset \mathcal{X}$. Hence, there exist non-negative constants r and $\beta < 1$, and a set $B = \{b_1, \ldots, b_\kappa\} \subset M$ such that $\|{}_B P^n\|_w \le r\beta^n$ for $n \in \mathbb{N}$. Let i be an essential state, and observe that

$$v_i = (c - g)_i + (Pv)_i = (c - g)_i + ({}_B P v)_i + \sum_{k=1}^{\kappa} P_{ib_k} v_{b_k}.$$

By iteration of this expression we derive

$$v_i = \left[\sum_{t=0}^{n-1} {}_B P^t (c - g)\right]_i + ({}_B P^n v)_i + \sum_{k=1}^{\kappa} \sum_{t=0}^{n-1} ({}_B P^t \cdot P)_{ib_k} v_{b_k}, \quad n \in \mathbb{N}. \tag{2.17}$$

Note that $({}_BP^t \cdot P)_{ib} = F_{ib}^{(t+1)}$; the probability that the system, starting in state i, enters state b for the first time after $t+1$ steps. A similar expression also holds for v'. Hence, by substituting $g' = g$ in the expression of v', we find

$$(v - v')_i = ({}_BP^n v)_i - ({}_BP^n v')_i + \sum_{k=1}^{\kappa} \sum_{t=0}^{n-1} F_{ib_k}^{(t+1)} (v_{b_k} - v'_{b_k}), \quad n \in \mathbb{N}.$$

The first two terms converge geometrically fast to zero as $n \to \infty$. Therefore, by taking the limit we find

$$(v - v')_i = \lim_{n\to\infty} \sum_{k=1}^{\kappa} \sum_{t=0}^{n-1} F_{ib_k}^{(t+1)} (v_{b_k} - v'_{b_k}) = \sum_{k=1}^{\kappa} F_{ib_k} (v_{b_k} - v'_{b_k}).$$

Since the Markov chain is stable, all closed classes $\mathcal{C}_{b_1}, \ldots, \mathcal{C}_{b_\kappa}$ of the Markov chain are positive recurrent. Hence, there is exactly one state $j = j(i) \in \{1, \ldots, \kappa\}$ such that $i \in \mathcal{C}_{b_j}$. Furthermore, for all $i \in \mathcal{C}_{b_j}$ we have $F_{ib_j} = 1$ and zero otherwise. Therefore,

$$(v - v')_i = v_{b_j} - v'_{b_j}, \quad i \in \mathcal{C}_{b_j}.$$

Thus, v and v' differ by only a constant on each closed class. □

Observe that the conditions used in Assumption 2.4 resemble the conditions of Theorem 2.7. Basically, the conditions are the same with the sole difference that in Assumption 2.4 they must hold uniformly for all deterministic stationary policies. In view of these results, the equivalent of Theorem 2.2 for average cost Markov decision problems is given by the following theorem.

Theorem 2.8 (Thm. 2.7, Ch. 3 in [37]): Suppose there exists a weight function w on $\mathcal{X}$ such that Assumption 2.4 holds. Then

- The optimal long-run expected average cost g^* together with the corresponding average cost value function V are the unique, in the sense of Theorem 2.7, solution to the optimality equations (2.7) and (2.8) in the space $\mathbb{B}_w(\mathcal{X})$.
- The policy iteration algorithm yields a sequence of improving policies $\{\pi_n\}_{n\in\mathbb{N}_0}$. The pair (g_n, v_n) converges to the optimal long-run expected average cost g^* and to a corresponding average cost value function V satisfying the optimality equations.

– There exists a deterministic Markovian decision rule φ such that Equations (2.9) and (2.10) are satisfied. Furthermore, the policy $\pi = \varphi^\infty \in \Pi_{DS}$ is average cost optimal. Conversely, if $\pi = \varphi^\infty \in \Pi_{DS}$ is average cost optimal, then it satisfies Equations (2.9) and (2.10).

The results of Theorem 2.8 show that g^* is uniquely determined, whereas the average cost value function V is not. In view of Theorem 2.7 this cannot be expected either. The chain structure of the Markov decision chain is too complicated to guarantee a unique value function. More can be said when a unichain structure is assumed, thus $\kappa(\varphi) = 1$ for all deterministic Markovian decision rules φ; there is only one recurrent class with a possibly empty set of transient states. We will therefore reformulate the derived theorems for the unichain case, such that they convey more information and utilize the unichain structure of the Markov decision chain.

Theorem 2.9: Suppose that the Markov chain, induced by policies in Π_{DS}, is unichain, stable, and w-GE. Let the functions g and v in $\mathbb{B}_w(\mathcal{X})$ be a solution to the Poisson equations. Then P^* has equal rows, say ψ, and g is a constant given by

$$g = \sum_{i \in \mathcal{X}} \psi_i \, c_i.$$

Proof: Since the Markov chain is stable, P^* is a stochastic matrix. Due to the unichain structure the rows of the matrix P^* are identical (Theorem A.2 and Expression (A.6) of Puterman [96]). Therefore, the result follows directly by Theorem 2.6. □

Hernández-Lerma and Lasserre [61] work with Borel state spaces. In order to derive the equivalent result of Theorem 2.9, they have to work with the invariant probability measure, and would have to state that the result holds almost everywhere (in probability). Since this would not result in an "explicit" expression for g, they have chosen to consider Markov processes for which the result holds for every $x \in \mathcal{X}$. Note that in the case of denumerable state spaces, this is not necessary.

Observe that due to the fact that g is a constant, Equation (2.11) is always satisfied since P is a stochastic matrix. Hence, the set $\mathcal{B}_x$ in Equation (2.8) is equal to the set $\mathcal{A}_x$. It follows that for unichain models the optimal policies and their average costs are characterized through a single optimality equation. The average cost value functions also possess more structure. The specialized result of Theorem 2.7 is as follows.

Theorem 2.10: Suppose that the induced Markov chain, induced by policies in Π_{DS}, is unichain, stable, and w-GE. Let both (g, v) and (g', v') be solutions to the Poisson equations. Then $g = g'$, and v, v' only differ by a constant. Furthermore, for any positive recurrent state m

$$v = \sum_{t=0}^{\infty} {}_m P^t (c - g) + v_m.$$

Proof: Due to the unichain structure, there is only one closed class. The possibly empty set of transient states all reach states in the positive recurrent class with probability one. Therefore, the first two statements follow from Theorem 2.7. The last statement follows by taking the limit $n \to \infty$ in Expression (2.17). □

By Theorem 2.10 we have that for any two solutions (g, v) and (g', v') to the Poisson equations $v = v' + k$ for some constant k. If we wish to guarantee a unique solution, it suffices to impose that

$$\sum_{i \in \mathcal{X}} \psi_i \, v(i) = 0.$$

Then $v - v' = 0$ and from Theorem 2.7 it follows that $v = D\,c$, where the matrix $D = \sum_{n=0}^{\infty} (P^n - P^*)$ is the deviation matrix.

2.5 Continuous-time Markov decision problems

In the Markov decision problems discussed so far, the decision maker could choose actions only at a discrete equidistant set of epochs. If the time spent in a particular state follows an arbitrary probability distribution, then it is natural to allow actions to be made at any point in time. This gives rise to decision models in which the system evolution is described continuously in time, and in which cost is accumulated continuously in time. We shall assume that the cost is independent of the time spent in a state, and that it only depends on the state and action chosen at the last decision epoch. In this case we can restrict our attention to models in which the decision epochs coincide with the transition times (see Section 11.5 of Puterman [96]). These models are called semi-Markov decision models.

Continuous-time Markov decision problems may be regarded as a class of semi-Markov decision problems in which the time between decision epochs follows an exponential distribution. In general, the parameter of this distribution

will depend on the state and the action chosen at the decision epoch. It is possible to convert this class of problems into an equivalent class of problems that can be studied in discrete time. This discretization method, due to Lippman [81] and later formalized by Serfozo [112], is called uniformization. We shall illustrate this method for both the continuous-time discounted and average cost Markov decision problems.

Consider a continuous-time Markov decision process determined by the tuple $(\mathcal{X}, \{\mathcal{A}_x \mid x \in \mathcal{X}\}, p, \nu, c)$. Here, the state space $\mathcal{X}$ and the feasible action set $\mathcal{A}_x$ for $x \in \mathcal{X}$ are defined as usual. If the system is in state $x \in \mathcal{X}$ and action $a \in \mathcal{A}_x$ is chosen, the next state will be $y \in \mathcal{X}$ with probability $p(y \mid x, a)$. The time interval τ between the transition from state x to state y is exponentially distributed with parameter $\nu(x, a)$. Furthermore, the duration τ is independent of earlier transition times, states, and actions. We assume that the parameters $\nu(x, a)$ are uniformly bounded, i.e.,

$$\sup_{(x,a)\in\mathcal{K}} \nu(x, a) < \infty,$$

with $\mathcal{K}$ the set of feasible state-action pairs (see Expression 2.1). During a transition period τ, the cost, determined by the cost function c, is accumulated continuously. Recall that we assumed that the cost function does not depend on the time spent in a state.

The probabilistic behaviour of the process is described by the infinitesimal generator Q, i.e., the matrix with components

$$Q(y \mid x, a) = \begin{cases} -[1 - p(x \mid x, a)]\nu(x, a), & y = x \\ p(y \mid x, a)\nu(x, a), & y \neq x, \end{cases}$$

for $a \in \mathcal{A}_x$. Observe that in most natural formulations $p(x \mid x, a) = 0$, meaning that at the end of a sojourn time in state x, the system will jump to a different state. In order to analyze the continuous-time model in discrete-time, it is essential to allow the system to occupy the same state before and after a jump. The infinitesimal generator determines the probability distribution of the system through the differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t} P^t(y \mid x, a) = \sum_{z\in\mathcal{X}} Q(y \mid z, a) P^t(z \mid x, a),$$

or

$$\frac{\mathrm{d}}{\mathrm{d}t} P^t(y \mid x, a) = \sum_{z\in\mathcal{X}} P^t(y \mid z, a) Q(z \mid x, a),$$

which are referred to as the forward and backward Kolmogorov equations, respectively. Therefore, for a fixed policy, processes with the same infinitesimal generator have identical finite-dimensional distributions (provided that both systems start in the same state). Hence, modifying the process into one that is easier to analyze such that the infinitesimal generator does not change, does not alter the probabilistic structure.

Let us construct an equivalent stochastic process such that the sojourn time in each state has an exponential distribution with a fixed parameter. Since the parameter ν is uniformly bounded, there exists a constant $\widetilde{\nu}$, such that

$$\sup_{(x,a)\in\mathcal{K}} [1 - p(x \,|\, x, a)]\nu(x, a) \leq \widetilde{\nu} < \infty,$$

Define the transition probabilities $\widetilde{p}$ by

$$\widetilde{p}(y \,|\, x, a) = \begin{cases} 1 - \frac{[1-p(x \,|\, x,a)]\nu(x,a)}{\widetilde{\nu}}, & y = x \\ \frac{p(y \,|\, x,a)\nu(x,a)}{\widetilde{\nu}}, & y \neq x. \end{cases} \tag{2.18}$$

The infinitesimal generator $\widetilde{Q}$ of the continuous-time Markov decision process determined by $\big(\mathcal{X}, \{\mathcal{A}_x \,|\, x \in \mathcal{X}\}, \widetilde{p}, \widetilde{\nu}, c\big)$ equals Q. Thus, the constructed process is an equivalent process, in which the system is observed at exponentially distributed times with parameter $\widetilde{\nu}$. Therefore, this process is called the uniformization of $\{X_t\}$, because it has an identical sojourn time distribution in every state. Observe that due to the choice of $\widetilde{\nu}$ the observations in the process are more frequent than in the original system. As a result the probability that the system occupies the same state at different observation times is also higher. Alternatively, the uniformization can be viewed as adding fictitious transitions from a state to itself.

Since the duration in every state has an exponential distribution with the same parameter $\widetilde{\nu}$, we can study the embedded process at the transition times without making a distinction between the different states and actions. However, it is important to know the cost incurred during a transition period τ. When the discounted cost criterion with discount factor α is used, this quantity is given by

$$\widetilde{c}(x, a) = c(x, a)\, \mathbb{E} \int_0^\tau e^{-\alpha t} \mathrm{d}t = c(x, a)\, \mathbb{E} \left[\frac{1 - e^{-\alpha\tau}}{\alpha} \right] = \frac{c(x, a)}{\alpha + \widetilde{\nu}}.$$

Note that the proper discount factor in this situation is given by $\widetilde{\alpha} = \mathbb{E}[e^{-\alpha\tau}] = \widetilde{\nu}/(\alpha + \widetilde{\nu})$. Therefore, the optimality equation for the equivalent discrete-time

discounted Markov decision problem becomes

$$V(x) = \min_{a \in \mathcal{A}_x} \left[\widetilde{c}(x,a) + \widetilde{\alpha} \sum_{y \in \mathcal{X}} \widetilde{p}(y \,|\, x, a) V(y) \right].$$

This formulation enables us to study the continuous-time process in discrete time as described in Section 2.3. Moreover, the value function in both formulations coincide due to equivalence of both processes (see also Proposition 11.5.1 of Puterman [96]).

Let us now consider the average cost Markov decision problem under the assumption that the Markov chain induced by policies in Π_{DS} is unichain. In that case, the cost incurred during a transition period is given by

$$\widetilde{c}(x,a) = c(x,a)\,\mathbb{E}[\tau] = \frac{c(x,a)}{\widetilde{\nu}}.$$

The optimality equation for the equivalent discrete-time average-cost Markov decision problem becomes

$$\frac{g}{\widetilde{\nu}} + V(x) = \min_{a \in \mathcal{A}_x} \left[\widetilde{c}(x,a) + \sum_{y \in \mathcal{X}} \widetilde{p}(y \,|\, x, a) V(y) \right]. \tag{2.19}$$

The validity of this optimality equation can be found in Proposition 3.3 of Bertsekas [22]. Observe that in this case the long-run expected average cost g has to be corrected for the change in time as well.

Now consider a continuous-time Markov cost chain determined by the tuple $(\mathcal{X}, p, \nu, c)$. The system dynamics of this process is the same as in a continuous-time Markov decision process with the sole difference that no actions are taken. Thus, we still assume that the Markov chain has a unichain structure, and that $\sup_{x \in \mathcal{X}} \nu(x) < \infty$. Alternatively, one can view this as a continuous-time Markov decision process where a fixed policy is applied. The equivalent to Equation (2.19) for the Markov cost chain now becomes

$$g + \nu(x) V(x) = c(x) + \sum_{y \in \mathcal{X}} p(y \,|\, x) \nu(x) V(y).$$

In the sequel we shall study several independent Markov cost chains simultaneously. It seems intuitively clear that the value function in this case is the sum of the value functions of the independent Markov cost chains. The following theorem shows that this is indeed true.

Theorem 2.11: Consider n independent continuous-time Markov cost chains determined by $(\mathcal{X}_i, p_i, \nu_i, c_i)$ for chain $i = 1, \ldots, n$, with (g_i, V_i) the solution to the Poisson equations. The solution to the Poisson equations of the Markov chain that considers the n independent Markov cost chains simultaneously is given by $(g_1 + \cdots + g_n, V_1 + \cdots + V_n)$.

Proof: Assume that the Poisson equations for chain $i = 1, \ldots, n$, after uniformization, are given by

$$g_i + \nu_i(x_i)V_i(x_i) = c_i(x_i) + \sum_{y_i \in \mathcal{X}_i} p_i(y_i \,|\, x_i)\nu_i(x_i)V_i(y_i). \tag{2.20}$$

Let $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. Define $g = \sum_{i=1}^n g_i$, $c(x) = \sum_{i=1}^n c_i(x_i)$, $\nu(x) = \sum_{i=1}^n \nu_i(x_i)$, and $V(x) = \sum_{i=1}^n V_i(x_i)$ for $x = (x_1, \ldots, x_n) \in \mathcal{X}$. By taking a finite sum over i from 1 to n in Expression (2.20), we derive

$$g + \sum_{i=1}^n \nu_i(x_i)V_i(x_i) = c(x) + \sum_{i=1}^n \sum_{y_i \in \mathcal{X}_i} p_i(y_i \,|\, x_i)\nu_i(x_i)V_i(y_i).$$

Adding both sides $\sum_{i=1}^n \sum_{j \neq i} \nu_i(x_i)V_j(x_j)$ yields

$$g + \nu(x)V(x) = c(x) + \sum_{i=1}^n \sum_{y_i \in \mathcal{X}_i} p_i(y_i \,|\, x_i)\nu_i(x_i)V_i(y_i) + \sum_{i=1}^n \sum_{j \neq i} \nu_i(x_i)V_j(x_j).$$

Note that the last term is equal to $\sum_{i=1}^n \sum_{y_i \in \mathcal{X}_i} \sum_{j \neq i} p_i(y_i \,|\, x_i)\nu_i(x_i)V_j(x_j)$. Define $p(y \,|\, x) = p_i(y_i \,|x_i)\nu_i(x_i)/\nu(x)$ if y and x differ exclusively in the i-th component for $x, y \in \mathcal{X}$; otherwise $p(y \,|\, x) = 0$. Thus, the transition probabilities are chosen such that only one event can occur at transitions. Consequently, we have

$$g + \nu(x)V(x) = c(x) + \sum_{y \in \mathcal{X}^n} p(y \,|\, x)\nu(x)V(y).$$

Observe that we have derived the Poisson equations of the Markov chain that considers all n independent Markov cost chains simultaneously. Moreover, the pair $(g_1 + \cdots + g_n, V_1 + \cdots + V_n)$ satisfies the Poisson equations.

Due to the fact that $c(x) = \sum_{i=1}^n c_i(x_i)$ in the Markov chain that considers all n independent Markov cost chains simultaneously, it follows from Equation (2.6) that $g_1 + \cdots + g_n$ is indeed the average cost. Consequently, from Equation (2.12) it follows that the value function is uniquely determined, since g is now fixed. Hence, $(g_1 + \cdots + g_n, V_1 + \cdots + V_n)$ is the unique solution for the Markov chain that considers all n independent Markov cost chains simultaneously. □

Chapter 3

The Multi-Server Queue

In this chapter we study the multi-server queue with Poisson arrivals and identical independent servers with exponentially distributed service times. Customers arriving to the system are admitted or rejected according to a fixed threshold policy. The motivation for studying threshold policies stems from the fact that threshold policies are optimal or close to optimal in many queueing systems (see, e.g., Bhulai and Koole [25], and Stidham and Weber [118]). Hence, one can expect to obtain good approximations through one-step policy improvement or reinforcement learning when initially starting with a threshold policy or the structure of its induced value function.

Additionally, the system is subject to holding, waiting, and rejection costs. Note that this model can also be viewed as a multi-server queue with a finite buffer where no control is applied (the standard $M/M/s/N$ queue). We shall derive a closed-form expression for the long-run expected average cost and the value function. Also the standard $M/M/s$ queue with an infinite buffer will be analyzed. Since this model results in a denumerable state space, we need to construct a suitable weighted norm in order to provide uniqueness of the value function. The result will then be used in a single step of policy improvement in a model where a controller has to route to several parallel finite buffer queues with multiple servers. Numerical experiments show that the improved policy has a close to optimal value.

Koole and Spieksma [75] obtain expressions for deviation matrices for birth-death processes, and in particular to both the $M/M/s/N$ and $M/M/s$ queue. The deviation matrix is independent of the cost structure. Hence, it enables one to compute the average cost and the value function for various cost structures (depending on the state only) by evaluating a sum involving entries of the deviation matrix. However, the expressions they derive for the deviation matrix are very complicated. Therefore, evaluating the sum is not easy in many situations. The

method adopted in this chapter shows the benefit of working with costs integrated into the problem formulation. In the same general setting, we derive expressions that are simpler and easier to obtain in contrast to either working with deviation matrices or equilibrium probabilities.

In the literature one usually tries to derive the value function for a specific policy (see, e.g., Koole and Nain [73], Ott and Krishnan [94]). However, the results of Theorems 3.4–3.6 concern a parameterized class of policies in contrast to a specific policy. The results can therefore be used for finding the best threshold policy within the class as well. This optimization problem with respect to the threshold level is not difficult, and can be done analytically.

We start this chapter with some theory on difference calculus for second-order difference equations. This will be essential for analyzing general birth-death queueing processes, and thus also the multi-server queue. These results greatly simplify the proofs given in Bhulai and Koole [26], and study the problem in greater generality. The value function in the finite buffer case has a unique solution due to extra conditions on the boundaries. In the infinite buffer case results from the previous chapter provide the unique solution to the Poisson equations. After obtaining the value function we move on to discussing the single and infinite server queues, which are special cases of the multi-server queue. We conclude the chapter with applying the derived results to a routing problem.

3.1 Difference calculus

Many queueing models, although being a stochastic process in continuous time, can be analyzed in discrete time (see Section 2.5). The Poisson equations of most of these models give rise to linear difference equations. Due to the nature of the Poisson equations, the difference equations have a lot of structure when the state description is one-dimensional. Therefore, it is worthwhile to study difference equations prior to the analysis of the multi-server queue. We shall restrict attention to second-order difference equations, since this is general enough to model birth-death queueing processes which encapsulate the multi-server queue.

Let $V(x)$ be an arbitrary function defined on $\mathbb{N}_0$. Define the backward difference operator Δ as

$$\Delta V(x) = V(x) - V(x-1),$$

for $x \in \mathbb{N}$. Note that the value of $V(x)$ can be expressed as

$$V(x) = V(k-1) + \sum_{i=k}^{x} \Delta V(i), \tag{3.1}$$

for every $k \in \mathbb{N}$ such that $k \leq x$. This observation is the key to solving first-order difference equations, and second-order difference equations when one solution to the homogeneous equation, i.e., Equation (3.4), is known. We first state the result for first-order difference equations. The result can be found in Chapter 2 of Mickens [89], but is stated less precise there.

Lemma 3.1: Let $f(x)$ be a function defined on $\mathbb{N}$ satisfying the relation

$$f(x+1) - \gamma(x) f(x) = r(x), \tag{3.2}$$

with γ and r arbitrary functions defined on $\mathbb{N}$ such that $\gamma(x) \neq 0$ for all $x \in \mathbb{N}$. With the convention that an empty product equals one, $f(x)$ is given by

$$f(x) = f(1)\, Q(x) + Q(x) \sum_{i=1}^{x-1} \frac{r(i)}{Q(i+1)}, \quad \text{with} \quad Q(x) = \prod_{i=1}^{x-1} \gamma(i).$$

Proof: Let the function $Q(x)$ be as stated in the theorem. By assumption we have that $Q(x) \neq 0$ for all $x \in \mathbb{N}$. Dividing Equation (3.2) by $Q(x+1)$ yields

$$\Delta \left[\frac{f(x+1)}{Q(x+1)} \right] = \frac{f(x+1)}{Q(x+1)} - \frac{f(x)}{Q(x)} = \frac{r(x)}{Q(x+1)}.$$

From Expression (3.1) with $k = 2$ it follows that

$$f(x) = Q(x) \frac{f(1)}{Q(1)} + Q(x) \sum_{i=2}^{x} \frac{r(i-1)}{Q(i)} = f(1)\, Q(x) + Q(x) \sum_{i=1}^{x-1} \frac{r(i)}{Q(i+1)}.$$

□

Note that the condition $\gamma(x) \neq 0$ for all $x \in \mathbb{N}$ is not very restrictive in practice. If there is a state $y \in \mathbb{N}$ for which $\gamma(y) = 0$, then the analysis can be reduced to two other first-order difference equations, namely the part for states $x < y$, and the part for states $x > y$ for which $\gamma(y) = 0$ is the boundary condition.

The solution to first-order difference equations plays an important part in solving second-order difference equations when one solution to the homogeneous equation is known. In that case, the second-order difference equation can be reduced to a first-order difference equation expressed in the homogeneous solution. Application of Lemma 3.1 then gives the solution to the second-order difference equation. The following theorem summarizes this result.

Theorem 3.2: Let $V(x)$ be a function defined on $\mathbb{N}_0$ satisfying the relation

$$V(x+1) + \alpha(x)V(x) + \beta(x)V(x-1) = q(x), \tag{3.3}$$

with α, β, and q arbitrary functions defined on $\mathbb{N}$ such that $\beta(x) \neq 0$ for all $x \in \mathbb{N}$. Suppose that one homogeneous solution is known, say $V_1^h(x)$, such that $V_1^h(x) \neq 0$ for all $x \in \mathbb{N}_0$. Then, with the convention that an empty product equals one, $V(x)$ is given by

$$\frac{V(x)}{V_1^h(x)} = \frac{V(0)}{V_1^h(0)} + \left[\Delta\left[\frac{V(1)}{V_1^h(1)}\right]\right]\sum_{i=1}^{x} Q(i) + \sum_{i=1}^{x} Q(i)\sum_{j=1}^{i-1}\frac{q(j)}{V_1^h(j+1)\,Q(j+1)},$$

where $Q(x) = \prod_{i=1}^{x-1}\beta(i)\,V_1^h(i-1)/V_1^h(i+1)$.

Proof: Note that V_1^h always exists, since a second-order difference equation has exactly two linearly independent homogeneous solutions (see Theorem 3.11 of Mickens [89]). The known homogeneous solution $V_1^h(x)$ satisfies

$$V_1^h(x+1) + \alpha(x)V_1^h(x) + \beta(x)V_1^h(x-1) = 0. \tag{3.4}$$

Set $V(x) = V_1^h(x)\,u(x)$ for an arbitrary function u defined on $\mathbb{N}_0$. Substitution into Equation (3.3) yields

$$V_1^h(x+1)\,u(x+1) + \alpha(x)V_1^h(x)\,u(x) + \beta(x)V_1^h(x-1)\,u(x-1) = q(x).$$

By subtracting Equation (3.4) $u(x)$ times from this expression, and rearranging the terms, we derive

$$\Delta u(x+1) - \gamma(x)\Delta u(x) = r(x),$$

with

$$\gamma(x) = \frac{V_1^h(x-1)}{V_1^h(x+1)}\,\beta(x), \quad \text{and} \quad r(x) = \frac{q(x)}{V_1^h(x+1)}.$$

From Lemma 3.1 it follows that

$$\Delta u(x) = [\Delta u(1)]\,Q(x) + Q(x)\sum_{j=1}^{x-1}\frac{r(j)}{Q(j+1)}, \quad \text{with} \quad Q(x) = \prod_{i=1}^{x-1}\gamma(i).$$

From Expression (3.1) it finally follows that

$$u(x) = u(0) + [\Delta u(1)]\sum_{i=1}^{x} Q(i) + \sum_{i=1}^{x} Q(i)\sum_{j=1}^{i-1}\frac{r(j)}{Q(j+1)}.$$

Since $V(x) = V_1^h(x)\,u(x)$ it follows that $u(x) = V(x)/V_1^h(x)$ for $x = 1, 2$. □

From Theorem 3.11 of Mickens [89] it follows that a second-order difference equation has exactly two homogeneous solutions that are also linearly independent. In Theorem 3.2 it is assumed that one homogeneous solution V_1^h is known. From the same theorem it follows that the second homogeneous solution is determined by $V_2^h(x) = V_1^h(x) \sum_{i=1}^{x} Q(i)$. This can be easily checked as follows.

$$
\begin{aligned}
&V_2^h(x+1) + \alpha(x)V_2^h(x) + \beta(x)V_2^h(x-1) = \\
&V_1^h(x+1) \sum_{i=1}^{x+1} Q(i) + \alpha(x)V_1^h(x) \sum_{i=1}^{x} Q(i) + \beta(x)V_1^h(x-1) \sum_{i=1}^{x-1} Q(i) = \\
&\sum_{i=1}^{x} Q(i) \left[V_1^h(x+1) + \alpha(x)V_1^h(x) + \beta(x)V_1^h(x-1) \right] + \\
&V_1^h(x+1)\, Q(x+1) - \beta(x)V_1^h(x-1)\, Q(x) = 0.
\end{aligned}
$$

The last equality follows from the fact that

$$
Q(x+1) = \frac{V_1^h(x-1)}{V_1^h(x+1)} \beta(x)\, Q(x) = \frac{V_1^h(0)\, V_1^h(1)}{V_1^h(x)\, V_1^h(x+1)} \prod_{i=1}^{x} \beta(i), \quad x \in \mathbb{N}.
$$

Note that the homogeneous solutions V_1^h and V_2^h are also linearly independent, since their Casorati determinant $C(x) = V_1^h(x)\, V_1^h(x+1)\, Q(x+1)$ is non-zero for all $x \in \mathbb{N}_0$ (see Sections 3.2 and 3.3 of Mickens [89]).

3.2 Markovian birth-death queueing models

In this section we study birth-death queueing systems. The name birth-death stems from the fact that arrivals (birth) and departures (death) of customers occur only in sizes of one, i.e., batch arrivals and batch services are not allowed. Let state $x \in \mathbb{N}_0$ denote the number of customers in the system. When the system is in state x, customers arrive according to a Poisson process with rate $\lambda(x)$. At the same time, customers receive service with an exponentially distributed duration at rate $\mu(x)$, such that $\mu(0) = 0$. Moreover, the system is subject to costs $c(x)$ in state x, where $c(x)$ is a polynomial function of x.

For stability of the system, we assume that

$$
0 < \liminf_{x \to \infty} \frac{\lambda(x)}{\mu(x)} \leq \limsup_{x \to \infty} \frac{\lambda(x)}{\mu(x)} < 1. \tag{3.5}
$$

Moreover, we assume that the transition rates satisfy

$$0 < \inf_{x\in\mathbb{N}_0} \big(\lambda(x)+\mu(x)\big) \le \sup_{x\in\mathbb{N}_0} \big(\lambda(x)+\mu(x)\big) < \infty. \tag{3.6}$$

Without loss of generality we can assume that $\sup_{x\in\mathbb{N}_0} \big(\lambda(x)+\mu(x)\big) < 1$. This can always be obtained after a suitable renormalization without changing the long-run system behaviour. After uniformization, see Expression (2.18), the resulting birth-death process has the following transition rate matrix.

$$\begin{aligned} P_{0,0} &= 1-\lambda(0), & P_{0,1} &= \lambda(0), \\ P_{x,x-1} &= \mu(x), \quad P_{x,x} = 1-\lambda(x)-\mu(x), & P_{x,x+1} &= \lambda(x), \qquad x = 1,2,\ldots. \end{aligned}$$

Inherent to the model is that the Markov chain has a unichain structure. Hence, by Theorem 2.10 we know that the long-run expected average cost g is constant. Furthermore, the value function V is unique up to a constant. As a result we can use this degree of freedom to set $V(0) = 0$. By rearranging the terms in Equation (2.19), we see that the Poisson equations for this system are given by

$$g + \big(\lambda(x)+\mu(x)\big)V(x) = \lambda(x)V(x+1) + \mu(x)V\big([x-1]^+\big) + c(x), \tag{3.7}$$

for $x \in \mathbb{N}_0$, where $[x]^+ = \max\{0, x\}$. The relation with the previous section becomes clear when the Poisson equations are written in the form of Equation (3.3), with

$$\alpha(x) = -\frac{\lambda(x)+\mu(x)}{\lambda(x)}, \quad \beta(x) = \frac{\mu(x)}{\lambda(x)}, \quad \text{and} \quad q(x) = \frac{g-c(x)}{\lambda(x)}.$$

Observe that $\beta(x) \neq 0$ for $x \in \mathbb{N}$ due to the stability assumption. Furthermore, note that for any set of Poisson equations, the constant function is always a solution to the homogeneous equations. This fact follows directly from Equation (2.12), since P is a transition probability matrix. Hence, by applying Theorem 3.2 with $V_1^h(x) = 1$ for $x \in \mathbb{N}_0$, it follows that the solution to Equation (3.7) is given by

$$V(x) = \frac{g}{\lambda(0)} \sum_{i=1}^{x} Q(i) + \sum_{i=1}^{x} Q(i) \sum_{j=1}^{i-1} \frac{q(j)}{Q(j+1)}, \quad \text{with} \quad Q(x) = \prod_{i=1}^{x-1} \frac{\mu(i)}{\lambda(i)}. \tag{3.8}$$

From the previous chapter we know that this expression is not the unique solution to the Poisson equations. Hence, we need to construct a weighted norm w such

that the Markov chain is w-GE. By Lemma 2.1 it suffices to construct a weighted norm w such that the Markov chain is w-GR(M) for some finite set $M \subset \mathbb{N}_0$. Due to stability of the Markov chain there exists a constant $K \in \mathbb{N}_0$ such that $\lambda(x)/\mu(x) < 1$ for all $x > K$. Let $M = \{0, \ldots, K\}$, and assume that $w(x) = z^x$ for some $z > 1$. Now consider

$$\sum_{y \notin M} \frac{P_{xy}\, w(y)}{w(x)} = \begin{cases} \lambda(K)\, z, & x = K, \\ \lambda(K+1)\, z + \big(1 - \lambda(K+1) - \mu(K+1)\big), & x = K+1, \\ \lambda(x)\, z + \big(1 - \lambda(x) - \mu(x)\big) + \frac{\mu(x)}{z}, & x > K+1. \end{cases}$$

We need to choose z such that all expressions are strictly less than 1. The first expression immediately gives $z < 1/\lambda(K)$. The second expression gives that $z < 1 + \mu(K+1)/\lambda(K+1)$. Solving the third expression shows that $1 < z < \mu(x)/\lambda(x)$ for $x > K+1$. Define $z^* = \min\{1/\lambda(K), \inf_{x \in \mathbb{N} \setminus M} \mu(x)/\lambda(x)\}$, and note that the choice of M ensures us that $z^* > 1$. Then, the three expressions are strictly less than 1 for all $z \in (1, z^*)$. Thus, we have shown that for $w(x) = z^x$ with $1 < z < z^*$, there exists an $\varepsilon > 0$ such that $\|{}_M P\|_w \leq 1 - \varepsilon$. Hence, the Markov chain is w-GR(M), and therefore by Lemma 2.1 also w-GE.

Note that $c \in \mathbb{B}_w(\mathbb{N}_0)$, since $c(x)$ is a polynomial in x by assumption. From Section 2.4 we know that the unique value function V is bounded with respect to the norm w. Therefore, the value function cannot contain terms θ^x with $\theta > 1$, since the weight function can be chosen such that $z \in (1, \min\{\theta, z^*\})$. Consequently, $\|V\|_w = \infty$, hence the value function cannot grow exponentially with a growth factor greater than 1. Thus, we have the following corollary.

Corollary 3.3: The value function of a stable birth-death queueing process cannot grow exponentially with a growth factor greater than 1.

3.3 The multi-server queue

Consider a queueing system with one queue and s identical independent servers. The arrivals are determined by a Poisson process with parameter λ. The service times are exponentially distributed with parameter μ. Let state x denote the number of customers in the system. A controller decides to admit or reject arriving customers to the system according to a threshold policy with threshold level $\tau \in \mathbb{N}_0$. Thus, when at arrival $x < \tau$, the controller decides to admit the customer, whereas the customer is rejected when $x \geq \tau$. Hence, when starting with an empty system, the states are limited to $x \in \{0, \ldots, \tau\}$. Note that a threshold

level of $\tau = 0$ rejects every customer, and the limiting case $\tau \to \infty$ admits every customer.

Observe that this system is a special case of a birth-death queueing process. This is easily seen by setting the arrival rate to $\lambda(x) = \lambda \mathbb{1}_{\{x<\tau\}}$, and the departure rate to $\mu(x) = \min\{x, s\}\mu$ for $x \in \{0, \ldots, \tau\}$. Hence, the Markov chain satisfies the unichain assumption. Thus, under holding, waiting, and rejection costs the long-run expected average cost g_τ is independent of the initial state. The same result holds for the limiting case $\tau \to \infty$ under the assumption that $\rho = \lambda/s\mu < 1$, see Expression (3.5). The Poisson equations for the system are given by

$$\begin{aligned}
g_\tau + \lambda V_\tau(0) &= \lambda V_\tau(1), \\
g_\tau + (\lambda + x\mu)V_\tau(x) &= hx + \lambda V_\tau(x+1) + x\mu V_\tau(x-1), && x = 1, \ldots, s-1, \\
g_\tau + (\lambda + s\mu)V_\tau(x) &= hx + \lambda w(x-s+1) + \lambda V_\tau(x+1) + \\
&\quad s\mu V_\tau(x-1), && x = s, \ldots, \tau-1, \\
g_\tau + s\mu V_\tau(\tau) &= h\tau + \lambda r + s\mu V_\tau(\tau-1).
\end{aligned}$$

In this set of equations the constants h, w, and r denote the holding, waiting, and rejection costs, respectively. The function $V_\tau(x)$ is the value function depending on the threshold level τ. Observe that due to linearity the value function can be decomposed into $V_\tau(x) = V_\tau^h(x) + V_\tau^w(x) + V_\tau^r(x)$, which are due to holding, waiting, and rejection costs respectively. In the same way the long-run expected average cost $g_\tau = g_\tau^h + g_\tau^w + g_\tau^r$.

We adopt the following approach to solve the Poisson equations. We first consider the set of equations for $x = 0, \ldots, s-1$. By taking $V_\tau(0) = 0$ as in the previous section, the equations have a unique solution expressed in g_τ. The solution also holds for $x = s$ when considering holding and rejection costs. Then the set of equations for $x = s, \ldots, \tau-1$ is solved. In this case $V_\tau(s-1)$ or $V_\tau(s)$ is known and again guarantees a unique solution expressed in g_τ. Finally, the equation for $x = \tau$ is considered. This equation provides an expression for g_τ, which solves the complete system explicitly. Before solving the Poisson equations, first define the hypergeometric function $F(x)$ by

$$F(x) = \sum_{k=0}^{x-1} \frac{\Gamma(x)}{\Gamma(x-k)} \left(\frac{\mu}{\lambda}\right)^k,$$

with $\Gamma(x) = (x-1)!$ when x is integer. Then the first step of our approach is given by the following theorem.

Theorem 3.4: Let $k \in \{h, w, r\}$, and consider the Poisson equations for states $x = 0, \ldots, s$ when $k \in \{h, r\}$, and $x = 0, \ldots, s-1$ when $k = w$. The unique solution to this set of equations is given by

$$V_\tau^k(x) = \frac{g_\tau^k}{\lambda} \sum_{i=1}^{x} F(i) - \frac{\alpha(k)}{\lambda} \sum_{i=1}^{x} (i-1) F(i-1),$$

with $\alpha(h) = h$, $\alpha(w) = 0$ and $\alpha(r) = 0$.

Proof: This follows directly from evaluating Expression (3.8). □

The first term in the value function is the particular solution to g_τ^k in the Poisson equations. Therefore this term appears in $V_\tau^k(x)$ for all $k \in \{h, w, r\}$. The second term is the particular solution to the costs. Since in this case no waiting or rejections occur, the second term is zero in $V_\tau^w(x)$ and $V_\tau^r(x)$. The terms are rather complicated due to the fact that the rates in the Poisson equations are dependent on the state. This does not occur when the rates are constant. The following theorem shows this for the solution to the Poisson equations for $x = s, \ldots, \tau-1$.

Theorem 3.5: Consider the Poisson equations for $x = s, \ldots, \tau-1$. Let $\rho = \lambda/s\mu$ and $\Delta V_\tau^k(x) = V_\tau^k(x) - V_\tau^k(x-1)$. The unique solution to this set of equations is given by

$$\begin{aligned} V_\tau^k(x) = &-\frac{\big(x-\sigma(k)\big)\rho}{1-\rho} \frac{g_\tau^k}{\lambda} + V_\tau^k\big(\sigma(k)\big) \\ &+ \left[\frac{\big(x-\sigma(k)\big)\big(x-\sigma(k)+1\big)\rho}{2\,(1-\rho)} + \frac{\big(x-\sigma(k)\big)\big(\rho+\gamma(k)\big)\rho}{(1-\rho)^2} \right] \frac{\beta(k)}{\lambda} \\ &+ \frac{\left(\frac{1}{\rho}\right)^{x-\sigma(k)} - 1}{1-\rho} \left[\frac{\rho}{1-\rho} \frac{g_\tau^k}{\lambda} + \frac{\sigma(k)}{s} \Delta V_\tau^k\big(\sigma(k)\big) - \frac{\big(\rho+\gamma(k)\big)\rho}{(1-\rho)^2} \frac{\beta(k)}{\lambda} \right], \end{aligned}$$

for $k \in \{h, w, r\}$ with $\sigma(k) = s - \mathbb{1}_{\{k=w\}}$, $\gamma(k) = s(1-\rho)\,\mathbb{1}_{\{k\neq w\}}$, $\beta(h) = h$, $\beta(w) = \lambda w$ and $\beta(r) = 0$.

Proof: Let $k \in \{h, w, r\}$. Note that the multi-server queue constitutes a birth-death process with $\lambda(x) = \lambda$ and $\mu(x) = s\mu$ for states $x = \sigma(k), \ldots, \tau-1$. Hence, $V_\tau^k(x) - V_\tau^k\big(\sigma(k)\big)$ is given by Expression (3.8) after shifting the expression by $\sigma(k)$. □

Theorem 3.4 and Theorem 3.5 fully characterize the solution to the Poisson equations expressed in g_τ^k. The equation for state $x = \tau$ can now be used to explicitly determine g_τ^k. This will also explicitly determine the solution to the complete set of Poisson equations. The results are given by the following theorem.

Theorem 3.6: The average cost g_τ^k for $k \in \{h, w, r\}$ are given by

$$g_\tau^h = \left[\frac{1}{\rho} F(s) + \frac{1-\rho^{\tau-s+1}}{1-\rho}\right]^{-1} \times \left[s\, F(s) + \tau\rho^{\tau-s} + \frac{s\rho}{1-\rho} + \frac{\rho}{(1-\rho)^2} - \frac{\rho^{\tau-s+1}}{(1-\rho)^2} - \frac{\tau\rho^{\tau-s}}{1-\rho}\right] h,$$

$$g_\tau^w = \left[\frac{s-1}{s\rho} F(s-1) + \frac{1-\rho^{\tau-s+2}}{1-\rho}\right]^{-1} \times \left[\frac{\rho}{(1-\rho)^2} - \frac{(\tau-s+1)\rho^{\tau-s+1}}{1-\rho} - \frac{\rho^{\tau-s+2}}{(1-\rho)^2}\right] \lambda w,$$

$$g_\tau^r = \left[\frac{1}{\rho} F(s) + \frac{1-\rho^{\tau-s+1}}{1-\rho}\right]^{-1} \cdot \rho^{\tau-s}\, \lambda r.$$

Proof: The Poisson equation for state $x = \tau$ and $k \in \{h, w, r\}$ can be written as

$$g_\tau^k + s\mu\, \Delta V_\tau^k(\tau) - \mathbb{1}_{\{k=h\}}\, h\tau - \mathbb{1}_{\{k=r\}}\, \lambda r = 0.$$

After substitution of $V_\tau^k(\tau)$ from Theorem 3.5 one gets an equation in g_τ^k only. After some tedious calculus one shows that the solution is indeed as stated in the theorem. □

The set of Poisson equations is now solved, and we have derived an explicit solution expressed in the parameters λ, μ, s and τ. Note that we did not require any restriction on the parameters. However, when we consider the limiting case $\tau = \infty$, we require stability of the queueing system, i.e., $\rho = \lambda/s\mu < 1$. Assuming that the stability condition holds, one can directly obtain that

- $g_\infty^r = 0$,
- $g_\infty^h/h = g_\infty^w/(ws\mu) + \lambda/\mu$.

The first line directly follows from Theorem 3.6 when taking the limit, since we assumed that $\rho < 1$. Indeed, when all customers are admitted the costs due to

rejection are zero. The second line is more involved and is explained as follows. The mean time spent waiting in the queue is obtained when $w = 1/\lambda\, s\mu$. Adding the mean service time $1/\mu$ to it gives the mean sojourn time. Applying Little's Law gives the mean queue length, and thus yields the second result.

Let us now compute the expression for g^h_∞. For simplicity we assume that $h = 1$. From Theorem 3.6 it follows that

$$\begin{aligned} g^h_\infty = \lim_{\tau\to\infty} g^h_\tau &= \left[\frac{1}{\rho}\,F(s) + \frac{1}{1-\rho}\right]^{-1} \cdot \left[s\,F(s) + \frac{s\rho}{1-\rho} + \frac{\rho}{(1-\rho)^2}\right] \\ &= \frac{\rho}{(1-\rho)^2}\left[\frac{1}{\rho}\,F(s) + \frac{1}{1-\rho}\right]^{-1} + s\rho \\ &= \frac{(s\rho)^s\rho}{s!\,(1-\rho)^2}\left[\frac{(s\rho)^{s-1}}{\Gamma(s)}\,F(s) + \frac{(s\rho)^s}{s!\,(1-\rho)}\right]^{-1} + s\rho \\ &= \frac{(s\rho)^s\rho}{s!\,(1-\rho)^2}\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!} + \frac{(s\rho)^s}{s!\,(1-\rho)}\right]^{-1} + s\rho. \end{aligned}$$

Note that we have derived in an alternative way the well-known expression for the average queue length in a multi-server queueing system with an infinite buffer (see, e.g., Section 2.3 of Gross and Harris [56]).

It seems obvious that g^h_∞ is indeed given by $\lim_{\tau\to\infty} g^h_\tau$, and similarly for $V^h_\infty = \lim_{\tau\to\infty} V^h_\tau$. However, it is not evident that one would get the same result for a different Poisson equation for state $x = \tau$. One might choose the equation such that the Poisson equations do not have a solution for $\tau < \infty$, whereas the limiting case does. Hence, we need a formal proof to show validity of the expression derived for g^h_∞.

The solutions for the limiting case $\tau = \infty$ are given in Theorems 3.4 and 3.5. By Corollary 3.3 we know that the value function cannot contain terms θ^x with $\theta > 1$. If it does contain them, then the norm $w(x) = z^x$ where $z \in (1, s\mu/\lambda)$ (constructed in Section 3.2) shows that $\|V\|_w = \infty$. Therefore, V cannot be a solution to the Poisson equations. Thus, it follows that g must be chosen such that the exponential terms cancel. Hence, to meet the final requirement that $V \in \mathbb{B}_w(\mathbb{N}_0)$, i.e., $\|V\|_w < \infty$, we need that

$$\frac{\rho}{1-\rho}\,\frac{g^h_\infty}{\lambda} + \Delta V^h_\infty(s) - \frac{\big(\rho + s(1-\rho)\big)\rho}{(1-\rho)^2}\,\frac{h}{\lambda} = 0.$$

Since $\Delta V_\infty^h(s) = (g_\infty^h/\lambda)\cdot F(s) - (h/\lambda)\cdot(s-1)F(s-1)$, we find that the desired quantity is given by

$$g_\infty^h = \left[\frac{1}{\rho}F(s) + \frac{1}{1-\rho}\right]^{-1} \cdot \left[\frac{s-1}{\rho}\,F(s-1) + \frac{\rho + s\,(1-\rho)}{(1-\rho)^2}\right] h$$

$$= \left[\frac{1}{\rho}F(s) + \frac{1}{1-\rho}\right]^{-1} \cdot \left[s\,F(s) + \frac{s\rho}{1-\rho} + \frac{\rho}{(1-\rho)^2}\right] h.$$

The results explicitly depict the structure of the value function as well. This information can be fruitfully used in reinforcement learning to guess the structure or even the values of value functions for other queueing models. As will become clear in Section 3.4, the value function of the infinite buffer single server queue is a sum of linear and quadratic terms. However, the finite buffer single server queue also contains exponential terms due to boundary effects. When the rates in the Poisson equations depend on the states (as in the infinite server queue), then also hypergeometric terms appear.

The results can be directly applied to one-step policy improvement. In this setting one chooses a policy which has the property that the value function and the average cost can be computed. Next, this policy will be used in one step of the policy iteration algorithm (see page 24) resulting in an improved policy. The improved policy will in general be sufficiently complicated to render another step of policy iteration impossible. In Section 3.5 we shall apply this technique for routing to several parallel finite buffer queues. Before doing that, we first consider special cases of the multi-server queue.

3.4 Special cases

In this section we consider special cases of the multi-server queue. The case where $s = 1$ results in the single server queue. The case where $s = \tau$ results in the infinite server queue. We will discuss both the finite buffer (τ finite) and the infinite buffer ($\tau = \infty$ and $\rho < 1$) model. We first start with the treatment of the single server queue.

The single server queue

The single server queue can be obtained by considering the multi-server queue with one server only, i.e., $s = 1$. In this case the Poisson equations become

simpler. These equations are now given by

$$\begin{aligned} g_\tau + \lambda V_\tau(0) &= \lambda V_\tau(1), \\ g_\tau + (\lambda+\mu)V_\tau(x) &= hx + \lambda w x + \lambda V_\tau(x+1) + \mu V_\tau(x-1), \qquad x = 1,\ldots,\tau-1, \\ g_\tau + \mu V_\tau(\tau) &= h\tau + \lambda r + \mu V_\tau(\tau-1). \end{aligned}$$

The solution to this set of equations is given by the expression in Theorem 3.5 and Theorem 3.6 with $s = 1$. After some simple calculus one derives that the value function is given by

$$V_\tau^k(x) = a_1^k \frac{\left(\frac{1}{\rho}\right)^x - 1}{1-\rho} + \frac{x(x+1)}{2\,\mu(1-\rho)}\,\beta(k) - a_1^k\, x,$$

for $k \in \{h, w, r\}$ with $\beta(h) = h$, $\beta(w) = \lambda w$ and $\beta(r) = 0$. The constants a_1^k for $k \in \{h, w, r\}$ and the average costs are given by

$$a_1^h = -\frac{(\tau+1)\rho}{\mu\left(\left(\frac{1}{\rho}\right)^\tau - \rho\right)(1-\rho)}\,h, \qquad g_\tau^h = \left[1 - \frac{(\tau+1)(1-\rho)}{\left(\frac{1}{\rho}\right)^\tau - \rho}\right]\frac{\rho}{1-\rho}\,h,$$

$$a_1^w = -\frac{(\tau+\rho)\rho}{\left(\left(\frac{1}{\rho}\right)^\tau - \rho\right)(1-\rho)}\,w, \qquad g_\tau^w = \left[1 - \frac{(\tau+\rho)(1-\rho)}{\left(\frac{1}{\rho}\right)^{\tau-1} - \rho^2}\right]\frac{\rho}{1-\rho}\,\lambda w,$$

$$a_1^r = \frac{\rho}{\left(\frac{1}{\rho}\right)^\tau - \rho}\,r, \qquad g_\tau^r = \frac{1-\rho}{\left(\frac{1}{\rho}\right)^\tau - \rho}\,\lambda r.$$

The average costs and the value function for the single server queue with an infinite buffer, or equivalently with no rejections, are given by

$$g_\infty^k = \lim_{\tau\to\infty} g_\tau^k = \frac{\rho}{1-\rho}\,\beta(k), \quad \text{and} \quad V_\infty^k(x) = \lim_{\tau\to\infty} V_\tau^k(x) = \frac{x(x+1)}{2\,\mu(1-\rho)}\,\beta(k).$$

The value function of the infinite buffer single server queue is thus the sum of linear and quadratic terms. However, exponential terms appear in the value function when working with a finite buffer.

Note that for the simple case of the infinite buffer single server queue the value function can be easily interpreted. The value function can be seen as the difference in total cost between starting in state x and starting in the reference state 0. Focus on the holding costs, and assume that there are x customers in the

system initially. The time until one customer has left the system is the same as the duration of a busy period in an $M/M/1$ queue with arrival rate λ and service rate μ, which is equal to $1/(\mu-\lambda)$ (see Gross and Harris [56]). During the first busy period the cost incurred by customers present in the system is hx, and reduces to $h(x-1)$ during the second busy period. Hence, it is expected that the holding cost is given by $h\sum_{i=0}^{x-1}(x-i)/(\mu-\lambda) = x(x+1)h/\big(2\,(\mu-\lambda)\big)$.

The infinite server queue

The infinite server queue is obtained by considering the multi-server queue with $s=\tau$. The Poisson equations then become

$$\begin{aligned} g_\tau + \lambda V_\tau(0) &= \lambda V_\tau(1), \\ g_\tau + (\lambda + x\mu)V_\tau(x) &= hx + \lambda V_\tau(x+1) + x\mu V_\tau(x-1), \qquad x = 1,\dots,\tau-1, \\ g_\tau + \tau\mu V_\tau(\tau) &= h\tau + \lambda r + \tau\mu V_\tau(\tau-1). \end{aligned}$$

The solution to these equations is in fact already given in Theorem 3.4. However, the average costs do differ and are given by

$$g_\tau^h = \left[1 - \frac{1}{F(c+1)}\right]\rho h, \qquad g_\tau^w = 0, \qquad g_\tau^r = \frac{1}{F(c+1)}\,\lambda r,$$

where now $\rho = \lambda/\mu$. The average costs and the value function of the infinite server queue with no rejections are given by

$$g_\infty^k = \lim_{\tau\to\infty} g_\tau^k = \rho\,\alpha(k), \quad \text{and} \quad V_\infty^k(x) = \lim_{\tau\to\infty} V_\tau^k(x) = \frac{x}{\mu}\,\alpha(k),$$

with $\alpha(h)=h$, $\alpha(w)=0$ and $\alpha(r)=0$. In this case we observe that the infinite server model with no rejections has a linear value function. However, rejections cause hypergeometric terms to appear in the value function. This is due to the fact that the rates in the Poisson equations depend on the state.

It is not surprising that the value function of the infinite server model with no blocking is linear in the number of customers. Recall that the value function represents the difference in cost between starting in state x and starting in the reference state 0. Again, focus on the holdings costs in the system, and assume that there are x customers in the system initially. Since there are infinitely many servers, every customer is being served immediately. Hence, the expected duration to empty the system is $1/\mu$. Therefore, the value function is given by xh/μ.

3.5 Application to routing problems

In this section we illustrate the one-step policy improvement method by studying a routing problem to parallel queues. The general idea is to start with a policy such that each queue behaves as a multi-server queue. In this way, the value function and the average costs can be determined from the results of the previous sections. Finally, one step of policy improvement can be applied to obtain a better policy without having to compute the value function in an iterative way.

Consider two parallel finite buffer queues. Queue i has a buffersize of τ_i customers and has its own set of s_i dedicated servers, each working at rate μ_i for $i = 1, 2$. Furthermore, queue i has holding, waiting, and rejection costs of h_i, w_i, and r_i, respectively. Arriving customers, determined by a Poisson process with rate λ, can either be sent to queue 1 or to queue 2. The objective is to minimize the average cost. The optimality equations for this system are given by

$$\begin{aligned} g + \big(\lambda + (x \wedge s_1)\mu_1 + (y \wedge s_2)\mu_2\big)\, V(x,y) &= h_1 x_1 + h_2 x_2 + \\ &\lambda \min\Big\{ \mathbb{1}_{\{x<\tau_1\}}[x - s_1 + 1]^+ w_1 + \mathbb{1}_{\{x=\tau_1\}} r_1 + V\big((x+1 \wedge \tau_1), y\big), \\ &\mathbb{1}_{\{y<\tau_2\}}[y - s_2 + 1]^+ w_2 + \mathbb{1}_{\{y=\tau_2\}} r_2 + V\big(x, (y+1 \wedge \tau_2)\big)\Big\} + \\ &(x \wedge s_1)\mu_1\, V\big([x-1]^+, y\big) + (y \wedge s_2)\mu_2\, V\big(x, [y-1]^+\big), \end{aligned}$$

with $[k]^+ = \max\{k, 0\}$, $(k \wedge l) = \min\{k, l\}$, and x, $y \in \mathbb{N}_0$ the number of customers in queue 1 and 2 respectively.

Consider the policy that splits the arrival stream in two streams, such that there are arrivals to queue 1 at rate $\eta\lambda$ and to queue 2 at rate $(1 - \eta)\lambda$ with $\eta \in [0, 1]$ independent of the arrival process. We call this policy a Bernoulli policy with parameter η. Let $F(x, y)$ and $G(x, y)$ denote the two expressions in the minimization in the optimality equation. Then the optimality equations under the Bernoulli policy are obtained by changing $\lambda \min\{F(x, y), G(x, y)\}$ into $\eta\lambda\, F(x, y) + (1 - \eta)\lambda\, G(x, y)$. Hence, we can see that the two queues behave independently as a multi-server queue. Therefore, from Theorem 2.11 it follows that the corresponding value function becomes

$$V_B(x, y) = V^{\mathrm{MS}}{}_{(\eta\lambda,\mu_1,s_1,\tau_1,h_1,w_1,r_1)}(x) + V^{\mathrm{MS}}{}_{((1-\eta)\lambda,\mu_2,s_2,\tau_2,h_2,w_2,r_2)}(y),$$

with V^{MS} the value function of the multi-server queue of Section 3.3 with the corresponding parameters. Similarly, the average cost is expressed as

$$g_B = g^{\mathrm{MS}}{}_{(\eta\lambda,\mu_1,s_1,\tau_1,h_1,w_1,r_1)} + g^{\mathrm{MS}}{}_{((1-\eta)\lambda,\mu_2,s_2,\tau_2,h_2,w_2,r_2)},$$

y/x	0	1	2	3	4	5	6	7	8	9
9	2	2	2	2	2	2	2	2	2	1
8	1	1	1	1	1	2	2	2	1	1
7	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	2	1	1
4	1	1	1	1	1	2	2	2	1	1
3	1	1	1	1	2	2	2	2	2	1
2	1	1	1	2	2	2	2	2	2	1
1	2	2	2	2	2	2	2	2	2	1
0	2	2	2	2	2	2	2	2	2	1

one-step improved policy

y/x	0	1	2	3	4	5	6	7	8	9
9	2	2	2	2	2	2	2	2	2	1
8	1	1	1	2	2	2	2	2	1	1
7	1	1	1	1	1	2	2	1	1	1
6	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	2	2	1	1	1
3	1	1	1	1	2	2	2	2	1	1
2	1	1	1	2	2	2	2	2	2	1
1	2	2	2	2	2	2	2	2	2	1
0	2	2	2	2	2	2	2	2	2	1

optimal policy

Table 3.1: Relevance of rejection costs.

with g^{MS} the average cost for the multi-server queue. From numerical experiments it follows that not all parameters of the Bernoulli policy result in an improved policy which is close to the optimal value. Therefore we will use the optimal Bernoulli policy for deriving the improved policy in the sequel. The one-step policy improvement step now follows from the minimizing action in $\min\{F(x,y), G(x,y)\}$.

First, we display the relevance of rejection costs when working with finite buffer queues. Set $h_1 = h_2 = 1$, $w_1 = w_2 = r_1 = r_2 = 0$, $\lambda = 5$, $\mu_1 = 2$, $\mu_2 = 3$, $s_1 = 3$, $s_2 = 2$, and $\tau_1 = \tau_2 = 9$. Thus, we study two parallel finite buffer queues with holding costs only. The first queue has more dedicated servers than the second, but they work at a lower rate.

The optimal Bernoulli policy yields a value of $g_B = 2.351414$, the one-step improved policy $g' = 1.993648$, and the optimal policy $g^* = 1.993563$. Table 3.1 shows the routing policy for the values of x and y under the one-step improved policy and the optimal policy. One would expect an increasing switching curve. However, when one of the queues becomes congested, lack of rejection costs results in routing to that queue such that rejections occur.

In the previous example the one-step improved policy had a value close to the optimal value. Table 3.2 shows that this also holds for other parameter values. Note that this method can easily be used for more than two queues. In this section we restricted ourselves to two queues, since the computation of the optimal policy becomes numerically difficult for more than two queues. Hence, it becomes hard

λ	μ_1	μ_2	s_1	s_2	τ_1	τ_2	h_1	h_2	w_1	w_2	g_B	g'	g^*
10	2	2	3	3	10	10	0	0	0	0	0.390401	0.082642	0.082642
10	2	2	3	3	10	5	0	0	0	0	0.836706	0.253959	0.226499
10	3	2	2	3	10	10	0	0	0	0	0.367001	0.072194	0.071396
8	2	2	3	3	10	10	0	0	1	1	8.807790	3.595779	3.531940
8	2	2	3	3	10	5	0	0	1	1	4.662343	1.917528	1.911727
8	3	2	2	3	10	10	0	0	1	1	9.945102	4.081310	3.921034
8	2	2	3	3	10	10	1	1	0	0	5.491495	4.606377	4.599034
8	2	2	3	3	10	5	1	1	0	0	4.999463	4.454041	4.425574
8	3	2	2	3	10	10	1	1	0	0	5.024346	3.950910	3.914964
8	2	2	3	3	10	10	1	1	1	1	14.228695	8.182282	8.092028
8	4	2	2	3	10	5	1	1	1	1	7.654585	4.386521	4.200002

Table 3.2: Numerical results for $r_1 = r_2 = 1$.

to assess the quality of the one-step improved policy. For N stations of $M/M/s/\tau$ queues the number of states is equal to $(\tau+1)^N$; thus the complexity is exponential in the number of queues. However, a single step of policy iteration has linear complexity.

Chapter 4

Multi-Dimensional Models

In this chapter we study queueing models for which the state space cannot be described by one variable only. These, so-called, multi-dimensional models differ considerably in analysis from the one-dimensional models. The Poisson equations give rise to inhomogeneous partial difference equations. Since it is a challenging task to solve these partial difference equations analytically, it is difficult to obtain the value function for multi-dimensional models. Moreover, in order to derive the unique solution to the Poisson equations, one needs to construct a weighted norm such that the Markov chain is geometrically ergodic. This can be a problem as well due to multiple boundary effects of the system dynamics.

The standard techniques to solve partial difference equations, such as the symbolic, LaGrange, and Laplace methods, usually focus on power solutions of the form $\alpha_1^{x_1} \cdots \alpha_n^{x_n}$ for n-dimensional models (see, e.g., Chapter 5 of Mickens [89]). In the previous chapter we have seen that the value functions have a linear, quadratic, or hypergeometric structure with perhaps a term θ^x where $\theta \in (0, 1)$. Therefore, these techniques do not work well for deriving the value function. Hence, one has to rely on different techniques.

Adan [1] discusses a compensation approach for queueing problems in the context of solving difference equations for equilibrium probabilities. The idea is to solve the equations for the interior of the state space. Then, a compensation term for a boundary is added such that the equation on that boundary is satisfied. This has an impact on a different boundary, and a compensation term has to be added for that boundary too. This procedure repeats itself and results in an infinite series of compensation terms which is convergent under particular conditions.

The set of equations for the equilibrium probabilities are always homogeneous, and usually give rise to power solutions or solutions with that structure. Therefore, it is a natural choice to use power solutions in the compensation step. This does not hold anymore for Poisson equations and one has to use different

compensation terms. This is very complicated due to the inhomogeneous nature of the equations. Therefore, it still remains unclear what structure these terms should possess.

Boxma and Cohen [30] analyze random walks and queueing systems with a two-dimensional state space in the context of boundary value problems. In the analysis the focus is again on solving the equilibrium probabilities from a set of equations. The idea is to write these equations in terms of their generating function. Note that whereas difference equations with constant coefficients necessarily lead to rational generating functions, this is no longer true for partial difference equations (see, e.g., Bousquet-Mélou and Petkovšek [29]). Next, if the kernel of the resulting equations is simple enough, the problem can be formulated as a Riemann or Riemann-Hilbert boundary value problem. The problem can then be tackled by standard techniques from boundary value theory (see, e.g., Gakhov [51]).

Note that when the equilibrium probabilities are summed over all states, the sum is equal to 1. This yields an extra equation that can be used to reduce the complexity of the expressions derived when writing them in terms of generating functions. This cannot be done for the Poisson equations, since the extra condition is given by the weighted norm. The norm cannot be used to reduce the complexity of the expressions in an early stage. Due to this and the non-homogeneity of the equations the kernel can be very complex. This forms a serious bottleneck, since the solution to the boundary value problem, if possible to formulate at all, is very hard to derive.

Due to the fact that the equilibrium probabilities have had a lot of attention in the past, there are various methods for solving them from the set of equilibrium equations. The previous discussion shows that these techniques are not easy to adapt in order to solve the Poisson equations. Although there are no straightforward techniques to solve partial difference equations for queueing models, results can be derived to some extent for specific queueing models. In this chapter we shall study three such queueing models; two independent single-server queues, a priority queue with switching costs, and a tandem model.

We start this chapter with the analysis of two independent single server queues. The long-run expected average cost and the value function for this system are easy to obtain from the results of Chapter 3, since the two queues are independent. One would also expect that the same holds for the weight function. However, this is not the case; an explicit weight function is difficult to formulate (see Section 3.4 of Spieksma [117]). We show that under a stronger ergodicity condition the weight function has a nice product form.

Next, we discuss a single server priority queue with switching costs. This work is based on earlier work done by Koole and Nain [73] for the discounted cost case and by Groenevelt, Koole, and Nain [55] for the average cost case. In the latter paper an expression for the value function was derived by numerical experimentation, but no formal proof of correctness was given. We derive all solutions to the Poisson equations of this model. Furthermore, we construct a weighted norm such that we can obtain the unique solution that is finite with respect to this norm.

The last model discussed in this chapter concerns a tandem model. The analysis of this model turns out to be very hard, and we are not able to derive an explicit expression for the value function. However, partial results are obtained when no arrivals to the system are allowed, i.e., the analysis is focused on the way the system is depleted. The analysis is hindered by the fact that it is not clear, even when no arrivals are allowed, in what way the system is emptied. In the priority queue it is always the case that the prioritized class of customers leave the system first, then the other customers are served. For the tandem model, however, there is no fixed order in which the queues are emptied; the second queue may become empty several times, while the first is not, when emptying the system.

After discussing the three queueing models, we end this chapter with some concluding remarks.

4.1 Two independent single server queues

Consider a queueing system with two queues each having their own dedicated server (two parallel $M/M/1$ queues). The arrivals to queue i are determined by a Poisson process with parameter λ_i for $i = 1, 2$. The service times at queue i are exponentially distributed with parameter μ_i for $i = 1, 2$. Let state $(x, y) \in \mathbb{N}_0^2$ denote that there are x customers at queue 1, and y customers at queue 2, respectively. The costs occurring in the system are made up of holding costs h_i per unit of time a customer is in queue i for $i = 1, 2$. Let $\rho_i = \lambda_i/\mu_i$ for $i = 1, 2$, and assume that the stability condition $\rho_1 < 1$ and $\rho_2 < 1$ holds. Then the Markov chain satisfies the unichain condition, and the Poisson equations for this system are given by

$$\begin{aligned} g + (\lambda_1 + \lambda_2 + \mu_1 + \mu_2)V(x, y) = {} & h_1 x + h_2 y + \lambda_1 V(x + 1, y) + \lambda_2 V(x, y + 1) + \\ & \mu_1 V\big([x - 1]^+, y\big) + \mu_2 V\big(x, [y - 1]^+\big), \end{aligned}$$

for $x, y \in \mathbb{N}_0$.

Note that the average cost g is composed of g_1 and g_2, which are due to customers in the first and second queue, respectively. Since the two queues are independent, it follows from Theorem 2.11 that the value function of this system is given by the sum of the value functions of the $M/M/1$ queueing system with holding costs h_1 and h_2, respectively. Observe that these value functions can be easily derived from Chapter 3. Thus, the value function $V = V_g + V_{h_1} + V_{h_2}$ is given by

$$\begin{aligned}
V_g(x,y) &= -\frac{x}{\mu_1(1-\rho_1)}\, g_1 + \left[\left(\frac{1}{\rho_1}\right)^x - 1\right] \frac{1}{\mu_1(1-\rho_1)^2}\, g_1 + \\
&\quad -\frac{y}{\mu_2(1-\rho_2)}\, g_2 + \left[\left(\frac{1}{\rho_2}\right)^y - 1\right] \frac{1}{\mu_2(1-\rho_2)^2}\, g_2, \\
V_{h_1}(x,y) &= \frac{x(x+1)\, h_1}{2\,\mu_1(1-\rho_1)} + \frac{\rho_1 x\, h_1}{\mu_1(1-\rho_1)^2} - \left[\left(\frac{1}{\rho_1}\right)^x - 1\right] \frac{\rho_1\, h_1}{\mu_1(1-\rho_1)^3}, \\
V_{h_2}(x,y) &= \frac{y(y+1)\, h_2}{2\,\mu_2(1-\rho_2)} + \frac{\rho_2 y\, h_2}{\mu_2(1-\rho_2)^2} - \left[\left(\frac{1}{\rho_2}\right)^y - 1\right] \frac{\rho_2\, h_2}{\mu_2(1-\rho_2)^3}.
\end{aligned}$$

One could also expect that the weight function has a similar structure, i.e., $w(x,y) = z_1^x\, z_2^y$ with $1 < z_i < z_i^* = \mu_i/\lambda_i$ for $i = 1,2$. However, this turns out to be false due to boundary effects of the transition probabilities. Section 3.4 of Spieksma [117] shows a suitable weight function in such a case is of the form

$$w(x,y) = C \prod_{k=1}^{x} (1+m_k) \prod_{l=1}^{y} (1+n_l),$$

where C is a constant, and the series $\{m_k\}$ and $\{n_l\}$ satisfy

$$\begin{cases}
\{m_k\}_{k\in\mathbb{N}},\ \{n_l\}_{l\in\mathbb{N}} \ \text{non-decreasing}, \\
\sup\limits_{k\in\mathbb{N}} m_k < \infty, \ \text{and} \ \sup\limits_{l\in\mathbb{N}} n_l < \infty, \\
\inf\limits_{k\geq k^*} m_k > 0, \ \text{and} \ \inf\limits_{l\geq l^*} n_l > 0 \ \text{for some} \ k^*, l^* \in \mathbb{N}.
\end{cases}$$

The system with two independent single server queues is a special case of the coupled processor model studied in Section 3.4 of Spieksma [117] with $\nu_i^* = \nu_i$ for $i = 1,2$. However, the expressions are cumbersome and difficult to state explicitly. When the stronger ergodicity condition $\rho_1 + \rho_2 < 1$ is assumed, it is

possible to formulate an explicit weight function. In the sequel we shall carry out the analysis under this assumption.

First, observe that under the stronger ergodicity condition $\rho_i = \lambda_i/\mu_i < 1$ still holds for $i = 1, 2$, and the average cost $g < \infty$. Assume without loss of generality that $\lambda_1 + \lambda_2 + \mu_1 + \mu_2 < 1$ (this can always be obtained through scaling as in Section 3.2). Then the system has the following transition rate matrix.

$$\begin{aligned}
&P_{(x,y)(x+1,y)} = \lambda_1, && P_{(x,y)(x,y+1)} = \lambda_2,\\
&P_{(x,y)(x-1,y)} = \mu_1 \mathbb{1}(x > 0), && P_{(x,y)(x,y-1)} = \mu_2 \mathbb{1}(y > 0),\\
&P_{(x,y)(x,y)} = 1 - P_{(x,y)(x+1,y)} - P_{(x,y)(x-1,y)} - P_{(x,y)(x,y+1)} - P_{(x,y)(x,y-1)}.
\end{aligned}$$

Let $M = \{(0,0)\}$, and assume that $w(x,y) = (1+k_1)^x(1+k_2)^y$ for some constants k_1 and k_2. Now consider

$$\sum_{(x',y') \notin M} \frac{P_{(x,y)(x',y')}\, w(x',y')}{w(x,y)},$$

which is given by

$$\begin{cases}
\lambda_1(1+k_1) + \lambda_2(1+k_2), & (x,y) = (0,0),\\
\lambda_1 k_1 + \lambda_2 k_2 + 1 - \mu_1, & (x,y) = (1,0),\\
\lambda_1 k_1 + \lambda_2 k_2 + 1 - \mu_2, & (x,y) = (0,1),\\
\lambda_1 k_1 + \lambda_2 k_2 - \frac{\mu_1 k_1}{1+k_1} + 1, & x > 1, y = 0,\\
\lambda_1 k_1 + \lambda_2 k_2 - \frac{\mu_2 k_2}{1+k_2} + 1, & x = 0, y > 1,\\
\lambda_1 k_1 + \lambda_2 k_2 - \frac{\mu_1 k_1}{1+k_1} - \frac{\mu_2 k_2}{1+k_2} + 1, & x > 0, y > 0.
\end{cases}$$

We need to choose k_1 and k_2 such that all expressions are strictly less than 1. Observe that if the fourth and fifth expression are less than 1, then all others are also satisfied. Hence, we can restrict our attention to the system

$$\begin{aligned}
f_1(k_1,k_2) &= 1 + \lambda_1 k_1 + \lambda_2 k_2 - \frac{\mu_1 k_1}{1+k_1},\\
f_2(k_1,k_2) &= 1 + \lambda_1 k_1 + \lambda_2 k_2 - \frac{\mu_2 k_2}{1+k_2},
\end{aligned}$$

with the assumptions $\lambda_1 + \lambda_2 + \mu_1 + \mu_2 < 1$ and $\rho_1 + \rho_2 < 1$.

Observe that $f_1(0,0) = f_1\big((\mu_1 - \lambda_1)/\lambda_1, 0\big) = 1$. Thus, the points $(0,0)$ and $\big((\mu_1 - \lambda_1)/\lambda_1, 0\big)$ lie on the curve $f_1(k_1,k_2) = 1$. Furthermore k_2 satisfies $k_2 = \mu_1/\lambda_2 - \mu_1/\big(\lambda_2(1+k_1)\big) - \lambda_1/\lambda_2$. Note that this function has a maximum value at $k_1 = \sqrt{1/\rho_1} - 1$. Hence, this description determines the form of f_1; the curve $f_1(k_1,k_2) = 1$ starts in $(0,0)$, and increases to an extreme point, and then decreases to the k_1-axis again. The curve f_2 has a similar form, but with the role of the k_1-axis interchanged with the k_2-axis.

The curves determine an area of points (k_1,k_2) such that f_1 and f_2 are strictly less than one if the partial derivative to k_1 at $(0,0)$ of the curve $f_1(k_1,k_2) = 1$ is greater than the partial derivative to k_2 of the curve $f_2(k_1,k_2) = 1$ at $(0,0)$. These partial derivatives are given by $(\mu_1 - \lambda_1)/\lambda_2$ and $\lambda_1/(\mu_2 - \lambda_2)$, respectively. Since $\rho_1 + \rho_2 < 1$, we have $\lambda_1\mu_2 + \lambda_2\mu_1 < \mu_1\mu_2$. Adding $\lambda_1\lambda_2$ to both sides gives $\lambda_1\lambda_2 < \mu_1\mu_2 - \lambda_1\mu_2 - \lambda_2\mu_1 + \lambda_1\lambda_2 = (\mu_1 - \lambda_1)(\mu_2 - \lambda_2)$. Hence, the relation $\lambda_1/(\mu_2 - \lambda_2) < (\mu_1 - \lambda_1)/\lambda_2$ holds. Thus, indeed there is an area of pairs (k_1,k_2) such that the Markov chain is w-GR(M), and thus also w-GE. For these points it holds that $(1 + k_i) < 1/\rho_i$ for $i = 1,2$. Observe that any sphere with radius $\varepsilon > 0$ around $(0,0)$ has a non-empty intersection with this area. Hence, the value function cannot contain terms in x and/or y that grow exponentially fast to infinity.

It is clear that the cost function $c \in \mathbb{B}_w(\mathbb{N}_0^2)$ with $c(x,y) = h_1 x + h_2 y$. To meet the final requirement that $V \in \mathbb{B}_w(\mathbb{N}_0^2)$, i.e., $\|V\|_w < \infty$, we need that

$$\frac{1}{\mu_1(1-\rho_1)^2}\, g_1 - \frac{\rho_1}{\mu_1(1-\rho_1)^3}\, h_1 = 0, \quad \frac{1}{\mu_2(1-\rho_2)^2}\, g_2 - \frac{\rho_2}{\mu_2(1-\rho_2)^3}\, h_2 = 0.$$

Finally, we obtain the unique pair (g,V) by solving this equation. The result is as expected

$$g = \frac{\rho_1}{1-\rho_1} h_1 + \frac{\rho_2}{1-\rho_2} h_2.$$

When the obtained average cost is substituted into the solutions to the Poisson equations, one exactly gets the value function of the infinite buffer $M/M/1$ queue (see Section 3.4).

4.2 A priority queue with switching costs

Consider a queueing system with two classes of customers. There is only one server available serving either a class-1 or a class-2 customer with exponentially distributed service times. A class-i customer arrives according to a Poisson process with arrival rate λ_i, and is served with rate μ_i for $i = 1,2$. The system is subject to

holding and switching costs. The cost of holding a class-i customer in the system for one unit of time is h_i for $i = 1, 2$. The cost of switching from serving a class-1 to a class-2 customer (from a class-2 to a class-1 customer) is s_1 (s_2, respectively).

The system follows a priority discipline indicating that class-1 customers have priority over class-2 customers. The priority is also preemptive, i.e., when serving a class-2 customer, the server switches immediately to serve a class-1 customer upon arrival of a class-1 customer. Upon emptying the queue of class-1 customers, the service of class-2 customers, if any present, is resumed from the point where it was interrupted. Due to the exponential service times, this is equivalent to restarting the service for this customer.

Let state (x, y, z) for $x, y \in \mathbb{N}_0$, $z \in \{1, 2\}$ denote that there are x class-1 and y class-2 customers present in the system, with the server serving a class-z customer, if present. Let $\rho_i = \lambda_i/\mu_i$ for $i = 1, 2$ and assume that the stability condition $\rho_1+\rho_2 < 1$ holds. Then the Markov chain is stable and $g < \infty$ holds. Furthermore, the Markov chain satisfies the unichain condition. Let $\lambda = \lambda_1 + \lambda_2$, then the Poisson equations are given by

$$\begin{aligned}
g + (\lambda + \mu_1)V(x, y, 1) &= h_1 x + h_2 y + \lambda_1 V(x + 1, y, 1) + \lambda_2 V(x, y + 1, 1) + \\
&\quad \mu_1 V(x - 1, y, 1), && x > 0, y \geq 0, \\
V(0, y, 1) &= s_1 + V(0, y, 2), && y > 0, \\
V(x, y, 2) &= s_2 + V(x, y, 1), && x > 0, y \geq 0, \\
g + (\lambda + \mu_2)V(0, y, 2) &= h_2 y + \lambda_1 V(1, y, 2) + \lambda_2 V(0, y + 1, 2) + \\
&\quad \mu_2 V(0, y - 1, 2), && y > 0, \\
g + \lambda V(0, 0, z) &= \lambda_1 V(1, 0, z) + \lambda_2 V(0, 1, z), && z = 1, 2.
\end{aligned}$$

This model has been studied by Koole and Nain [73] for the discounted cost case, and by Groenevelt, Koole, and Nain [55] for the average cost case. In the latter paper the value function was derived by means of numerical experimentation and some intuition into the form was mentioned. However, the authors did not derive all solutions to the Poisson equations, and they did not show that their solution is the unique solution to the Poisson equations. We complement their results by providing all solutions to the Poisson equations, and by showing that the norm of Section 4.1 yields the unique solution.

First observe that, given the values of $V(x, y, 1)$, the values of $V(x, y, 2)$ are easily obtained by considering the difference equations: $V(x, y, 2) = V(x, y, 1) + (\lambda_1 s_2 - \lambda_2 s_1)/(\lambda_1 + \lambda_2)\mathbb{1}(x = 0, y = 0) - s_1\mathbb{1}(x = 0, y > 0) + s_2\mathbb{1}(x > 0, y \geq 0)$. Therefore, we only show the expression for $V(x, y, 1)$, as $V(x, y, 2)$ is easily

derived from it. Let $V = V_g + V_{h_1} + V_{h_2} + V_{s_1} + V_{s_2}$ be the decomposition of the value function. Then the previous observation directly prescribes that V_g, V_{h_1}, and V_{h_2} are independent of z. Furthermore, the value function V_{h_1} equals the value function of the single server queue due to the preemptive priority behaviour of the system.

The expressions for the value functions can be derived by setting $V(x,y) = c_1x^2 + c_2x + c_3y^2 + c_4y + c_5xy + c_6$ with constants c_i for $i = 1, \ldots, 6$ to be determined. This quadratic form is to be expected, when one considers the solutions given in Groenevelt, Koole, and Nain [55], and Koole and Nain [73]. Substitution of this equality into the Poisson equations yields a new set of equations that is easier to solve. In order to correct for the boundary $x = 0$, one needs to add the term $c_7[(1/\rho_1)^x - 1]$, which is the homogeneous solution to the equations for $x > 0$ and $y > 0$. Let θ be the unique root $\theta \in (0,1)$ of the equation

$$\lambda_1 x^2 - (\lambda_1 + \lambda_2 + \mu_1)x + \mu_1 = 0.$$

Then the correction term for the boundary $y = 0$ is given by $c_8(1 - \theta^x)$. Observe that for the boundary $x = 0$, we do not have correction terms with powers of y. This has to do with the fact that there are no jumps towards the boundary $y = 0$ from the interior of the state space. Solving for the constants yields the solution to the optimality equations, which is given by

$$\begin{aligned}
V_g(x,y,z) &= -\frac{x\,g}{\mu_1(1-\rho_1)} + \left[\left(\frac{1}{\rho_1}\right)^x - 1\right]\frac{g}{\mu_1(1-\rho_1)^2},\\
V_{h_1}(x,y,z) &= \frac{x(x+1)\,h_1}{2\,\mu_1(1-\rho_1)} + \frac{\rho_1 x\,h_1}{\mu_1(1-\rho_1)^2} - \left[\left(\frac{1}{\rho_1}\right)^x - 1\right]\frac{\rho_1\,h_1}{\mu_1(1-\rho_1)^3},\\
V_{h_2}(x,y,z) &= \frac{\lambda_2 x(x+1)\,h_2}{2\,\mu_1^2(1-\rho_1)(1-\rho_1-\rho_2)} + \frac{\rho_2(\mu_1-\lambda_1+\mu_2\rho_1)x\,h_2}{\mu_1^2(1-\rho_1)^2(1-\rho_1-\rho_2)} +\\
&\quad \frac{y(y+1)\,h_2}{2\,\mu_2(1-\rho_1-\rho_2)} + \frac{xy\,h_2}{\mu_1(1-\rho_1-\rho_2)} -\\
&\quad \left[\left(\frac{1}{\rho_1}\right)^x - 1\right]\frac{\rho_2(\mu_1-\lambda_1+\mu_2\rho_1)\,h_2}{\mu_1^2(1-\rho_1)^3(1-\rho_1-\rho_2)},\\
V_{s_i}(x,y,1) &= \frac{\lambda_1\theta}{\lambda}\,\frac{\rho_1\rho_2 x}{1-\rho_1}\,s_i + \frac{\lambda_1\theta}{\lambda}\,\frac{\lambda_1 y}{\mu_2}\,s_i - \left[\left(\frac{1}{\rho_1}\right)^x - 1\right]\frac{s_i}{\mu_1(1-\rho_1)^2} \times
\end{aligned}$$

$$\left[\lambda_1\left\{\rho_1\left(\frac{\lambda_1\theta}{\lambda}-1\right)+\frac{\lambda_1}{\lambda}(1-\theta)\right\}+\lambda_2\left\{\frac{\lambda_1\theta}{\lambda}\,\frac{\lambda_1}{\mu_2}+\frac{\lambda_1}{\lambda}\right\}\right]+$$

$$\frac{\lambda_1}{\lambda}\,s_i\,\mathbb{1}(y>0)+\frac{\lambda_1}{\lambda}(1-\theta^x)\,s_i\,\mathbb{1}(x>0,y=0),\qquad i=1,2.$$

Observe that, regardless of the position of the server, the system dynamics satisfy equations with the same transition rates. Hence, when the case $z=1$ is studied separately from $z=2$, the transition rates for both cases will be the same. Thus, we can expect that a suitable weighted supremum norm is independent of z.

Due to the stability condition $\rho_1+\rho_2<1$, the average cost $g<\infty$. Furthermore, it makes the Markov chain stable. Assume without loss of generality that $\lambda+\max\{\mu_1,\mu_2\}<1$ (this can always be obtained through scaling as in Section 3.2). Then the system has the following transition rate matrix.

$$\begin{aligned}
&P_{(x,y)(x+1,y)}=\lambda_1, && P_{(x,y)(x,y+1)}=\lambda_2,\\
&P_{(x,y)(x-1,y)}=\mu_1\mathbb{1}(x>0), && P_{(x,y)(x,y-1)}=\mu_2\mathbb{1}(x=0,y>0),\\
&P_{(x,y)(x,y)}=1-P_{(x,y)(x+1,y)}-P_{(x,y)(x-1,y)}-P_{(x,y)(x,y+1)}-P_{(x,y)(x,y-1)}.
\end{aligned}$$

Let $M=\{(0,0)\}$, and assume that $w(x,y)=(1+k_1)^x(1+k_2)^y$ for some k_1 and k_2. Now consider

$$\sum_{(x',y')\notin M}\frac{P_{(x,y)(x',y')}\,w(x',y')}{w(x,y)},$$

which is given by

$$\begin{cases}
\lambda_1(1+k_1)+\lambda_2(1+k_2), & (x,y)=(0,0),\\
\lambda_1k_1+\lambda_2k_2+(1-\mu_1), & (x,y)=(1,0),\\
\lambda_1k_1+\lambda_2k_2+(1-\mu_2), & (x,y)=(0,1),\\
\lambda_1k_1+\lambda_2k_2-\frac{\mu_1k_1}{1+k_1}+1, & x>1,y=0,\\
\lambda_1k_1+\lambda_2k_2-\frac{\mu_2k_2}{1+k_2}+1, & x=0,y>1,\\
\lambda_1k_1+\lambda_2k_2-\frac{\mu_1k_1}{1+k_1}+1, & x>0,y>0.
\end{cases}$$

We need to choose k_1 and k_2 such that all expressions are strictly less than 1. However, this problem is equivalent to the problem studied in Section 4.1. All the assumptions used in that section are also satisfied, since $\lambda_1+\lambda_2+\mu_1+\mu_2<1$

implies $\lambda_1 + \lambda_2 + \max\{\mu_1, \mu_2\} < 1$. Hence, we can use the same weight function. Then, it is clear that the cost function $c \in \mathbb{B}_w(\mathbb{N}_0^2 \cup \{1,2\})$ with $c(x,y) = h_1 x + h_2 y + s_1 \mathbb{1}(x = 0, y > 0, z = 1) + s_2 \mathbb{1}(x > 0, y \geq 0, z = 2)$. To meet the final requirement that $V \in \mathbb{B}_w(\mathbb{N}_0^2 \cup \{1,2\})$, i.e., $\|V\|_w < \infty$, we need that

$$\frac{g}{\mu_1(1-\rho_1)^2} - \frac{\rho_1 h_1}{\mu_1(1-\rho_1)^3} - \frac{\rho_2(\mu_1 - \lambda_1 + \mu_2\rho_1)\, h_2}{\mu_1^2(1-\rho_1)^3(1-\rho_1-\rho_2)} - \frac{s_1+s_2}{\mu_1(1-\rho_1)^2} \times$$

$$\left[\lambda_1 \left\{\rho_1 \left(\frac{\lambda_1\theta}{\lambda} - 1\right) + \frac{\lambda_1}{\lambda}(1-\theta)\right\} + \lambda_2 \left\{\frac{\lambda_1\theta}{\lambda}\,\frac{\lambda_1}{\mu_2} + \frac{\lambda_1}{\lambda}\right\}\right] = 0.$$

Finally, we obtain the unique pair (g, V) by solving this equation. The result is

$$g = \frac{\rho_1}{1-\rho_1}\, h_1 + \frac{\rho_2(\mu_1 - \mu_1\rho_1 + \mu_2\rho_1)}{\mu_1(1-\rho_1)(1-\rho_1-\rho_2)}\, h_2 + (s_1+s_2) \times$$

$$\left[\lambda_1 \left\{\rho_1 \left(\frac{\lambda_1\theta}{\lambda} - 1\right) + \frac{\lambda_1}{\lambda}(1-\theta)\right\} + \lambda_2 \left\{\frac{\lambda_1\theta}{\lambda}\,\frac{\lambda_1}{\mu_2} + \frac{\lambda_1}{\lambda}\right\}\right],$$

as stated in Theorem 3.1 of Groenevelt, Koole, and Nain [55]. The unique value function V is obtained by substituting g into the solutions to the Poisson equations. Doing this shows that V_{h_1} is indeed the value function of the $M/M/1$ queue. A partial explanation for the holding costs of class-2 customers can be found in Groenevelt, Koole, and Nain [55]. For completeness we repeat these arguments as to provide insight into the value function.

Suppose that the system starts with initially no class-1 customers and y class-2 customers. The time until one class-2 customer has left the system is obtained by taking into account that the service of the first class-2 customer that has started service is interrupted each time a class-1 customer joins the system. The service is resumed when all customers (both class-1 and class-2) that have arrived have been served. Thus, this time is the duration of a busy period in an $M/G/1$ queue with arrival rate λ and mean service time $(\rho_1+\rho_2)/\lambda$, and with initial workload equal to the service time of a class-2 customer. Section 4.1 of Jaiswal [68] shows that this duration is given by $1/\big(\mu_2(1-\rho_1-\rho_2)\big)$. The contribution of the holding cost of class-2 customers is therefore given by $h_2 \sum_{i=0}^{y-1}(y-i)/\big(\mu_2(1-\rho_1-\rho_2)\big)$, which is exactly the third term in V_{h_2}.

When both classes of customers are present in the system initially, all initial class-2 customers have to wait for class-1 customers including all arrivals (of class-1 and class-2 customers) before being served. For each class-1 customer this

corresponds to the time needed to empty an $M/G/1$ queue with initially one class-1 customer, arrival rate λ, and mean service time $(\rho_1 + \rho_2)/\lambda$. This time is given by $1/\big(\mu_1(1-\rho_1-\rho_2)\big)$. For each class-1 customer initially present in the system there are holding costs $h_2/\big(\mu_1(1-\rho_1-\rho_2)\big)$ per class-2 customer that is initially present. Hence, this accounts for the fourth term in V_{h_2}.

When the value of g is substituted into the value function the second term in V_{h_2} cancels. What remains is the first term which is due to newly arriving class-2 customers whose service is delayed compared to the initially empty situation because of the presence of class-1 customers. Unfortunately, no simple explanation was found for this term.

4.3 A tandem model

Consider two single server queues in tandem. Customers arrive to the first queue according to a Poisson process with rate λ. After having service at the first server the customer proceeds to the second queue. When the second server has finished serving the customer, the customer leaves the system. The duration of the service in the first (second) queue is exponentially distributed with parameter μ_1 (μ_2). The system is exposed to costs h_i for holding a customer in queue i for one unit of time for $i = 1, 2$. Let $\rho_i = \lambda/\mu_i$ for $i = 1, 2$, and assume that the stability conditions $\rho_1 < 1$ and $\rho_2 < 1$ hold. Then the Markov chain satisfies the unichain condition. The Poisson equations for this system are then given by

$$\begin{aligned}
g + \lambda V(0,0) &= \lambda V(1,0), \\
g + (\lambda+\mu_1)V(x,0) &= h_1 x + \lambda V(x+1,0) + \mu_1 V(x-1,1), && x > 0, \\
g + (\lambda+\mu_2)V(0,y) &= h_2 y + \lambda V(1,y) + \mu_2 V(0,y-1), && y > 0, \\
g + (\lambda+\mu_1+\mu_2)V(x,y) &= h_1 x + h_2 y + \lambda V(x+1,y) + \mu_1 V(x-1,y) + \\
&\quad \mu_2 V(x,y-1), && x > 0, y > 0.
\end{aligned}$$

From the description of the system it is clear that the first queue behaves exactly as an $M/M/1$ queue. Consequently, the value function and the average holding cost for this queue are known (see Section 3.4). It is therefore the average holding cost for the second queue that we are interested in.

It is known that one can treat the second queue as an $M/M/1$ queue with arrival rate λ and service rate μ_2 when one is interested in the long-run average holding costs. Hence, the average cost g for the tandem model is known, and is given by $g = \rho_1 h_1/(1-\rho_1) + \rho_2 h_2/(1-\rho_2)$. However, this perspective does not hold for the derivation of the value function. The value function captures

the transient behaviour of the queueing system and this is not described by the $M/M/1$ system.

In Section 4.1 the value function could be obtained due to complete independence between the queues. The value function for the priority queue in Section 4.2 was obtainable due to the fact that for any state $(x, y) \in \mathbb{N}_0^2$ the most likely way to empty the system was known. First all x class-1 customers are served in an $M/M/1$ manner. Then all y class-2 customers plus the ones that arrived during the service of class-1 customers are served in an $M/G/1$ manner due to service interruptions of class-1 customers. For the tandem model it is not clear in what way the system is emptied, and this accounts for the fact that obtaining the value function for the holding costs in the second queue is more complicated.

In order to gain more insight into the system, set $\lambda = 0$. Thus, the analysis is now focused on the way the system is emptied given that no arrivals to the system occur. Furthermore, it is clear that the average cost $g = 0$, since there are no arrivals. Recall that the weight function was used to determine the value of the average cost g from the set of all solutions expressed in g. Hence, using the fact that $g = 0$, there are no unicity issues to be dealt with. Let for simplicity $h_2 = 1$, then the Poisson equations for this system are given by

$$\begin{aligned}(\mu_1 + \mu_2)V(x, y) = y + \mu_1 V(x, y)\mathbb{1}_{(x=0)} + \mu_1 V(x-1, y+1)\mathbb{1}_{(x>0)} + \\ \mu_2 V\left(x, [y-1]^+\right).\end{aligned} \tag{4.1}$$

Note that an efficient recursive scheme is available to solve the value function for this simplified system. Set $V(0, 0) = 0$, then the solution for $V(0, y)$ is easy to derive due to independence of $V(x, y)$ for $x > 0$. The following lemma illustrates this.

Lemma 4.1: $V(0, y) = y(y + 1)/(2\,\mu_2)$ for $y \in \mathbb{N}_0$.

Proof: Fix $x = 0$. Then Equation (4.1) reduces to $V(0, y) = y/\mu_2 + V(0, y - 1)$ for $y > 0$. The result now follows from Theorem 3.1. □

The result of Lemma 4.1 can be used fruitfully if $V(x, \cdot)$ can be related to $V(x-1, \cdot)$ for $x \in \mathbb{N}$. In that case one can generate $V(x, y)$ iteratively for any value of $x, y \in \mathbb{N}$. Indeed, the following theorem shows that there is a relation between $V(x, \cdot)$ and $V(x - 1, \cdot)$ given as follows.

Theorem 4.2: Fix $x \in \mathbb{N}$. Then $V(x,y)$ for $y \in \mathbb{N}_0$ is given by

$$V(x,y) = \sum_{i=1}^{y} \frac{i}{\mu_1+\mu_2}\left[\frac{\mu_2}{\mu_1+\mu_2}\right]^{y-i} + \frac{\mu_1}{\mu_2}\sum_{i=1}^{y}\left[\frac{\mu_2}{\mu_1+\mu_2}\right]^{y-i+1} V(x-1,i+1)$$

$$+ \left[\frac{\mu_2}{\mu_1+\mu_2}\right]^{y} V(x-1,1). \tag{4.2}$$

Proof: First observe that Equation (4.1) is given by $V(x,0) = V(x-1,1)$ for $x \in \mathbb{N}$. Now, let $x \in \mathbb{N}$ be fixed. The proof is by induction in y. Let $y = 0$, then

$$V(x,0) = V(x-1,1).$$

Suppose that $V(x,n)$ holds for all $n = 0, \ldots, y-1$, then $V(x,y)$ is given by

$$V(x,y) = \frac{y}{\mu_1+\mu_2} + \frac{\mu_1}{\mu_1+\mu_2}V(x-1,y+1) + \frac{\mu_2}{\mu_1+\mu_2}V(x,y-1)$$

$$= \frac{y}{\mu_1+\mu_2} + \frac{\mu_1}{\mu_1+\mu_2}V(x-1,y+1) + \sum_{i=1}^{y-1}\frac{i}{\mu_1+\mu_2}\left[\frac{\mu_2}{\mu_1+\mu_2}\right]^{y-i}$$

$$+ \frac{\mu_1}{\mu_2}\sum_{i=1}^{y-1}\left[\frac{\mu_2}{\mu_1+\mu_2}\right]^{y-i+1} V(x-1,i+1) + \left[\frac{\mu_2}{\mu_1+\mu_2}\right]^{y} V(x-1,1)$$

$$= \sum_{i=1}^{y} \frac{i}{\mu_1+\mu_2}\left[\frac{\mu_2}{\mu_1+\mu_2}\right]^{y-i} + \frac{\mu_1}{\mu_2}\sum_{i=1}^{y}\left[\frac{\mu_2}{\mu_1+\mu_2}\right]^{y-i+1} V(x-1,i+1)$$

$$+ \left[\frac{\mu_2}{\mu_1+\mu_2}\right]^{y} V(x-1,1).$$

The result follows due to the principle of induction. □

Note that the expression in Theorem 4.2 actually expresses the semi-Markovian description of the system dynamics. Recall that $\mu_2/(\mu_1+\mu_2)$ is the probability that a departure from the second queue occurs before a departure from the first queue. Hence, the first term is the total holding cost due to departures in the second queue. The second and the last term account for a departure in the first queue. The last term is the probability of emptying the second queue before a

departure in the first queue occurs. The second term is in fact a sum of geometric probabilities and accounts for a departure from the first queue before emptying the second queue.

Theorem 4.2 is crucial for the analysis since it relates level x to level $x-1$. Together with Lemma 4.1 it determines the structure of the value function $V(x,y)$. Let $\eta = \mu_2/(\mu_1+\mu_2)$, and suppose that $V(x-1,y) = a_0 + a_1 y + a_2 y^2 + \sum_{k=0}^{n} b_k y^k \eta^y$ for arbitrary constants a_0, a_1, a_2 and b_k for $k \in \mathbb{N}_0$. By substitution into Equation (4.2) we derive that $V(x,y)$ is given by

$$\begin{aligned} V(x,y) = &\left\{ \frac{y}{\mu_1} - \frac{\mu_2}{\mu_1^2} + \frac{\mu_2}{\mu_1^2}\,\eta^y \right\} + a_0 + a_1 \left\{ y + \frac{\mu_1-\mu_2}{\mu_1} + \frac{\mu_2}{\mu_1}\,\eta^y \right\} \\ &+ a_2 \left\{ y^2 + \left[\frac{\mu_1-\mu_2}{\mu_1}\right](2y+1) + 2\left[\frac{\mu_2}{\mu_1}\right]^2 + \frac{\mu_2}{\mu_1^2}(\mu_1 - 2\,\mu_2)\,\eta^y \right\} \\ &+ \sum_{k=0}^{n} b_k \left\{ \left[\eta(1-\eta) \sum_{i=1}^{y} (i+1)^k + \eta \right] \cdot \eta^y \right\}. \end{aligned} \tag{4.3}$$

Observe that if $V(x-1,y)$ is a linear combination of functions in the family $\{1, y, y^2, y^k\eta^y; k \in \mathbb{N}_0\}$, then the expression $V(x,y)$ is also composed of functions from the same family. Since $V(0,y)$ is known, it is clear what the structure of $V(x,y)$ is. This is obtained by carefully analyzing the terms at each level, and repeating the argument above. The following theorem adopts this approach.

Theorem 4.3: The value function $V(x,y)$ for $x, y \in \mathbb{N}_0$ is of the following form

$$V(x,y) = \frac{(x+y)(x+y+1)}{2\,\mu_2} - \frac{x(x+1)}{2\,\mu_1} + \sum_{k=0}^{x-1} p(x,k)\, y^k \left[\frac{\mu_2}{\mu_1+\mu_2}\right]^y, \tag{4.4}$$

with $p(x,k)$ some polynomial independent of y.

Proof: First observe that the value function consists only of constants, y, y^2 and $y^k[\mu_2/(\mu_1+\mu_2)]^y$ with $k \in \mathbb{N}_0$. From Equation (4.3) it follows that each y and y^2 generates itself plus some additional terms with y; thus $V(x,y)$ contains $y(y+1)/2\,\mu_2 = V(0,y)$. From repetition of Equation (4.3) starting from 0 to x it follows that the total quantity is given by

$$\frac{y(y+1)}{2\,\mu_2} + \frac{xy}{\mu_1} + 2\left[\frac{\mu_1-\mu_2}{\mu_1}\right]\frac{xy}{2\,\mu_2} = \frac{y(y+1)}{2\,\mu_2} + \frac{xy}{\mu_2}.$$

The constants are partially determined by the coefficient of y, which we just computed. Hence the constants are given by

$$-\frac{x\mu_2}{\mu_1^2}+\left[\frac{\mu_1-\mu_2}{\mu_1}+2\left(\frac{\mu_2}{\mu_1}\right)^2\right]\frac{x}{2\,\mu_2}+\left[\frac{x(x-1)}{2\,\mu_2}+\frac{x}{2\,\mu_2}\right]\frac{\mu_1-\mu_2}{\mu_1}=$$

$$\frac{x(x+1)}{2\,\mu_2}\frac{\mu_1-\mu_2}{\mu_1}.$$

The sum of the two previous quantities leads to

$$\frac{y(y+1)}{2\,\mu_2}+\frac{xy}{\mu_2}+\frac{x(x+1)}{2\,\mu_2}\frac{\mu_1-\mu_2}{\mu_1}=\frac{(x+y)\,(x+y+1)}{2\,\mu_2}-\frac{x(x+1)}{2\,\mu_1}.$$

The last term with the sum in $V(x,y)$ follows from the last line in Expression (4.3) and the fact that $\sum_{i=1}^{y}(i+1)^n$ is a polynomial in y of degree $n+1$. □

Observe that the first two terms in Expression (4.4) can be explained. For that purpose, fix a state $(x,y)\in\mathbb{N}_0^2$. If all customers are in the second queue, then the system becomes a standard $M/M/1$ queue with arrival rate $\lambda=0$ and service rate μ_1. Hence, the value function for this system is given by the first term. However, not all customers are in the second queue, and we have to correct for this situation. The x customers in the first queue arrive to the second queue in an $M/M/1$ fashion with arrival rate $\lambda=0$ and service rate μ_1. This explains the second term. The last term has to do with the fact that one of the queues becomes empty. Unfortunately, we could not find a simple explanation for it.

In order to determine the polynomials $p(x,k)$ in Theorem 4.3, it is necessary to find the coefficients of y^k in $\sum_{i=1}^{y}(i+1)^n$ for $k=1,\ldots,n+1$. This can be done by a change to the polynomial basis $\{y^{(n)}\}_{n\in\mathbb{N}_0}$, with the polynomials $y^{(n)}=\Gamma(y+1)/\Gamma(y+1-n)$. The relation with the standard polynomial basis is given by $y^{(n)}=\sum_{i=1}^{n}S_1(i,n)y^i$, and $y^n=\sum_{i=1}^{n}S_2(i,n)y^{(i)}$, where S_i are Stirling numbers of the i-th kind.

Recall that the Stirling numbers are recursively defined by $S_i(m+1,n)=S_i(m,n-1)+\alpha(i)S_i(m,n)$ for $i=1,2$, with $\alpha(1)=-m$, $\alpha(2)=n$. The initial conditions are given by $S_i(n,n)=1$ and $S_i(m,n)=0$ for $i=1,2$, $n\in\mathbb{N}$, and $m\leq 0, m\geq n+1$ (see page 824 of Goldberg, Newman, and Haynsworth [54] for explicit expressions). The new basis has the nice property that $\sum_{k=0}^{n}k^{(m)}=(n+1)^{(m+1)}/(m+1)$ for $m\neq -1$. The following lemma shows how this can be used to obtain the coefficients we are interested in.

Lemma 4.4: Let $n \in \mathbb{N}$, and fix $k \in \{1, \ldots, n+1\}$. The coefficient of y^k in $\sum_{i=1}^{y}(i+1)^n$ is given by

$$G(n,k) = \sum_{i=k-1}^{n} \sum_{j=k}^{i+1} \frac{S_2(n,i)\, S_1(i+1,j)}{i+1} \binom{j}{k} 2^{j-k},$$

where S_i are Stirling numbers of the i-th kind for $i = 1, 2$.

Proof: Fix $n \in \mathbb{N}$, and $k \in \{1, \ldots, n+1\}$. Then

$$\begin{aligned}
\sum_{i=1}^{y}(i+1)^n = \sum_{i=2}^{y+1} i^n &= \left[\sum_{i=1}^{n} \frac{S_2(n,i)}{i+1} (y+2)^{(i+1)}\right] - 1 \\
&= \left[\sum_{i=1}^{n} \frac{S_2(n,i)}{i+1} \sum_{j=1}^{i+1} S_1(i+1,j)(y+2)^j\right] - 1 \\
&= \left[\sum_{i=1}^{n} \sum_{j=1}^{i+1} \sum_{k=0}^{j} \frac{S_2(n,i)\, S_1(i+1,j)}{i+1} \binom{j}{k} 2^{j-k}\, y^k\right] - 1 \\
&= \left[\sum_{i=1}^{n} \sum_{k=0}^{i+1} \sum_{j=k}^{i+1} \frac{S_2(n,i) S_1(i+1,j)}{i+1} \binom{j}{k} 2^{j-k}\, y^k\right] - 1 \\
&= \sum_{k=1}^{n+1} \sum_{i=k-1}^{n} \sum_{j=k}^{i+1} \frac{S_2(n,i) S_1(i+1,j)}{i+1} \binom{j}{k} 2^{j-k}\, y^k.
\end{aligned}$$

The first line consists of a transformation to a space of polynomials with a particular basis. The polynomials in this basis are defined by $y^{(n)} = \Gamma(y+1)/\Gamma(y+1-n)$. The second line constitutes the transformation back to the standard basis. The third line is due to the binomial formula. Finally, the result follows from isolating the coefficients in the last line. □

Observe that Lemma 4.4 tell us that $\sum_{i=1}^{y}(i+1)^n = \sum_{k=1}^{n+1} G(n,k) y^k$. When substituted into the last term of Expression (4.3), it is possible to isolate terms $y^k \eta^y$. Hence, we can formulate a recursive scheme for computation of the polynomials $p(x,k)$ in Expression (4.4) as follows.

Theorem 4.5: The polynomials $p(x,k)$ for $x \in \mathbb{N}_0$ and $k \in \{0, \ldots, x-1\}$ are determined by

$$p(x,0) = \left[\frac{\mu_2}{\mu_1+\mu_2}\right] \sum_{i=0}^{x-2} p(x-1,i) + \frac{x}{\mu_1},$$

$$p(x,k) = \left[\frac{\mu_1\mu_2}{(\mu_1+\mu_2)^2}\right] \sum_{i=k-1}^{x-2} G(i,k)\, p(x-1,i), \qquad k = 1, \ldots, x-1.$$

Proof: Fix x and consider $p(x,0)$. Note that we are looking at the coefficients of $[\mu_2/(\mu_1+\mu_2)]^y$ in $V(x,y)$. Therefore, the sum is explained by the last term in Expression (4.3). However, the terms y and y^2 also contribute to $p(x,0)$. By Theorem 4.3 we know that at $V(x-1,y)$ the terms $y(y+1)/2\,\mu_2 + (x-1)y/\mu_2$ are present. Therefore, their contribution to $V(x,y)$ is given by

$$\frac{1}{2\,\mu_2}\,\frac{\mu_2}{\mu_1^2}\,(\mu_1 - 2\,\mu_2) + \frac{2x-1}{2\,\mu_2}\,\frac{\mu_2}{\mu_1} + \frac{\mu_2}{\mu_1^2} = \frac{x}{\mu_1}.$$

The equation for $p(x,k)$ directly follows by the last term in Expression (4.3) and Lemma 4.4. □

Theorem 4.5 prescribes a recursive equation for obtaining the polynomials $p(x,k)$. Numerically they are easy to obtain. However, it is difficult to derive a closed-form expression for them analytically due to the dependence on the terms $G(n,k)$. Combining Theorem 4.3 and Theorem 4.5 gives us the following final result.

Theorem 4.6: The value function $V(x,y)$ for $x,y \in \mathbb{N}_0$ is given by

$$V(x,y) = \frac{(x+y)(x+y+1)}{2\,\mu_2} - \frac{x(x+1)}{2\,\mu_1} + \sum_{k=0}^{x-1} p(x,k)\, y^k \left[\frac{\mu_2}{\mu_1+\mu_2}\right]^y,$$

where the polynomials $p(x,k)$ are determined by

$$p(x,0) = \left[\frac{\mu_2}{\mu_1+\mu_2}\right] \sum_{i=0}^{x-2} p(x-1,i) + \frac{x}{\mu_1},$$

$$p(x,k) = \left[\frac{\mu_1\mu_2}{(\mu_1+\mu_2)^2}\right] \sum_{i=k-1}^{x-2} G(i,k) p(x-1,i), \qquad k = 1, \ldots, x-1,$$

with

$$G(n,k) = \sum_{i=k-1}^{n} \sum_{j=k}^{i+1} \frac{S_2(n,i)\, S_1(i+1,j)}{i+1} \binom{j}{k} 2^{j-k},$$

where S_i are Stirling numbers of the i-th kind for $i = 1, 2$.

4.4 Concluding remarks

In this chapter we have seen that the analysis of multi-dimensional models is substantially harder than the approach for one-dimensional models. Whereas in the latter case recursive equations are easy to formulate in known quantities, recursive equations for multi-dimensional models are usually expressed in one of the (unknown) boundaries. Therefore, it is difficult to formulate general results for multi-dimensional models, since the behaviour on the boundaries plays an important role (see, e.g., Bousqet-Mélou and Petkovšek [29]). Moreover, seemingly minor changes in the system dynamics can result in completely different solutions with different complexities.

The obtained value functions of the queueing systems studied in Chapters 3 and 4 all have in common that they are quadratic with perhaps a term that grows exponentially fast to zero for large states. Since the exponential terms in the models were caused by the boundaries, one might wonder if the value function grows at most quadratically for large states in all queueing systems with exponential arrival and service times. Although the answer to this question is open, it seems plausible for the tandem queue when y grows large; see Equation (4.4). Some intuition for this might be that since there is enough work in the second queue, the queue behaves as a standard $M/M/1$ queue. Thus, the transition of a customer leaving the first queue and joining the second queue does not have a substantial contribution to the value function.

Although there are no standard methods and techniques to solve the Poisson equations for the multi-dimensional models analytically, there is a different approach to this problem. Insight into the queueing model under study already can give the structure of the value function (e.g., linear or quadratic). Koole and Nain [73] were able to explain all the terms appearing in the value function for the discounted cost case, mostly in terms of busy periods of single server queues. This understanding was in fact used to find the exponential term in Groenevelt, Koole, and Nain [55] for the average cost case. Moreover, this insight was helpful for the generalization to multiple queues as well (see Koole and Nain [74]). Hence, explanations for the value functions are very important as this might provide a means to guess (the structure of) value functions for other queueing models.

Chapter 5

Partial Information Models

In this chapter we study Markov decision problems in which the decision maker does not have full access to all information. In spite of this, the decision maker desires to design decision rules in order to optimize his criterion function. The design of decision rules cannot ignore the feature of incompleteness, e.g., by replacing some unknown quantities with their expected values, when some actions in the system provide additional information. In the latter case, the decision maker is faced with a control problem and an estimation problem simultaneously.

We start this chapter with a description of partially observed Markov decision problems. In this class of problems the decision maker cannot directly observe the state of the system. Instead, probabilistic information about the state the process occupies is available through observations of a stochastically related process. We show that by using a Bayesian approach to estimate the state of the system, the problem can be transformed into a standard Markov decision problem with complete information. Based on Koole [71], we show how equivalence between the partially observed Markov decision problem and the Markov decision problem with complete information can be established. The results are very general and allow for different information structures, e.g., the case where the decision rules depend on the complete history of observations, the last observation, or no observations at all.

Next, we study Markov decision problems with a transition law that is not completely known. It is assumed that the transition law depends on a parameter that is unknown to the decision maker. We show that this problem is a special case of the partially observed Markov decision problem. Due to the special structure of the problem additional results are available. Based on van Hee [58], we show that the sequence of estimates converges to the true value of the unknown parameter. Furthermore, by studying statistical sampling theory, we show that sufficient statistics can be used to recast the state space to lower dimensional state

spaces.

In this chapter it is convenient to have terminology and notation that makes it possible to discuss both discrete and absolute continuous distributions simultaneously. Accordingly, we use the term generalized probability density function, abbreviated by g.p.d.f., to denote a function f that is either a probability mass function or a probability density function. Let X be the random variable with g.p.d.f. f, then it follows from measure theory that for some suitable probability measure μ (e.g., a Borel measure or a counting measure), the expectation can be denoted with $\mathbb{E}X = \int_{\mathbb{R}} x\, f(x)\, \mu(\mathrm{d}x)$ for both the discrete and continuous case. We will use this notation in the sequel. However, it should be emphasized that the values of any g.p.d.f. can be arbitrarily changed on any set of points that has probability zero. Hence, properties of g.p.d.f.'s will be understood to apply to versions of these g.p.d.f.'s that have those properties.

5.1 Partial observation Markov decision problems

A partially observed Markov decision process is determined by the tuple

$$(\mathcal{X}, \mathcal{Y}, \{\mathcal{A}_y \mid y \in \mathcal{Y}\}, p, q, c, \Psi_0).$$

In this setting, the state space $\mathcal{X}$, the transition law p, and the cost function c have the usual interpretation. In contrast to the previous chapters, we shall allow the state space $\mathcal{X}$ to be a Borel space, i.e., a set isomorphic to a measurable subset of a Polish space (a complete separable metric space). It is not our intention to present the model in full generality; when discussing applications we shall restrict our attention to denumerable state spaces. The main purpose for allowing Borel state spaces is to include the model discussed in Section 5.2 as a special case of the partially observed Markov decision process. Therefore, we shall not develop the mathematical apparatus to guarantee existence of optimal policies in the case of Borel state spaces.

In the partially observed Markov decision process the decision maker cannot directly observe the state of system, i.e., the decision maker cannot observe the realization of $\{X_t\}_{t\in\mathbb{N}_0}$. Instead, the decision maker observes the stochastic process $\{Y_t\}_{t\in\mathbb{N}_0}$, which takes values in the set $\mathcal{Y}$. The set $\mathcal{Y}$ is called the observation space, and we assume that it is denumerable. The random variable Y_t is associated with the random variable X_t and provides information about the true value of X_t. The probabilistic relation between X_t and Y_t is known to the decision maker, and is given by the probability function q, where $q(y \mid x)$ for $y \in \mathcal{Y}$ and $x \in \mathcal{X}$ is the probability of observing y in state x.

Suppose that the system is in state $x \in \mathcal{X}$, and that the decision maker observes $y \in \mathcal{Y}$ with probability $q(y \mid x)$. Due to the unobservability of the state of the process, the feasible action set in this case may not depend on the state x, since otherwise uncertainty would also reside in the feasible action set itself. The feasible action set is allowed to depend on the observation y under the assumption that an action $a \in \mathcal{A}_y$ is applicable to all states $x' \in \mathcal{X}$ for which $q(y \mid x') > 0$. This assumption guarantees that whenever $y \in \mathcal{Y}$ is observed, any action $a \in \mathcal{A}_y$ can be applied to the system independent of the underlying state of the process. Finally, we assume that the decision maker is given a prior probability distribution Ψ_0 (with generalized probability density function ψ_0 with respect to a measure μ) on the initial states.

The partially observed Markov decision process generalizes the standard Markov decision process; it coincides when $\mathcal{Y} = \mathcal{X}$, with $q(y \mid x) = 1$ if $y = x$, and zero otherwise. Due to the fact that we allow Borel state spaces, additional measurability conditions on p, q, and c have to be imposed. In this context, we shall assume that p is a generalized probability density function (with respect to the measure μ) of some stochastic kernel (see Appendix C in Hernández-Lerma and Lasserre [60]). Define the set of feasible observation-action pairs by $\mathcal{K} = \{(y, a) \mid y \in \mathcal{Y}, a \in \mathcal{A}_y\}$. Let $\mathcal{W}$ be the set of probability measures on the measurable space $(\mathcal{X}, \sigma(\mathcal{X}))$, where $\sigma(\mathcal{X})$ is the Borel σ-algebra of the state space $\mathcal{X}$; observe that $\mathcal{W}$ is a Borel space (see Appendix 5 of Dynkin and Yushkevich [45]). The set of histories in the partially observed Markov decision process is given by $\mathcal{H}_t = \mathcal{W} \times \mathcal{K}^t \times \mathcal{Y}$ for $t \in \mathbb{N}_0$. A generic element $h_t = (\Psi_0, y_0, a_0, y_1, \ldots, a_{t-1}, y_t) \in \mathcal{H}_t$ depends on the initial distribution. Therefore, the decision rules may only depend on the observed values and the initial distribution, but not on the unobserved states. In the sequel we shall use the notation developed in Section 2.1 also for partially observed decision models without explicitly mentioning the dependency on the initial distribution.

In partially observed Markov decision models the decision maker is faced with two problems simultaneously: the control problem as in standard Markov decision processes, and an identification problem for the unobserved states. Each time the decision maker takes an action, the transition to a new state implicitly provides new information about the underlying state the process occupies. The decision maker can use this information to his advantage for the next decision to be made. The key idea in partially observed Markov decision models is to revise the initial distribution at every transition, so that it takes into account new information through the observations. The revised distribution is called the posterior distribution, and is used to identify the unobserved state and to control

the system at that decision epoch. The dual effect of control and identification is also called adaptive control.

Suppose that at epoch $t \in \mathbb{N}_0$ we have constructed a probability distribution Ψ_t based on the history $h_t \in \mathcal{H}_t$. Furthermore, assume that the decision maker chooses action $a_t \in \mathcal{A}_{y_t}$ after which the decision maker observes $y_{t+1} \in \mathcal{Y}$. Then the Bayesian approach to revise the probability distribution is according to Bayes' rule, i.e., for $h_{t+1} = (h_t, a_t, y_{t+1})$ we have $\psi_{t+1}(h_{t+1})(x_{t+1}) =$

$$\frac{\int_{\mathcal{X}} q(y_{t+1} \,|\, x_{t+1})\, p(x_{t+1} \,|\, x_t, a_t)\, \psi_t(h_t)(x_t)\, \mu(\mathrm{d}x_t)}{\int_{\mathcal{X}} \int_{\mathcal{X}} q(y_{t+1} \,|\, x_{t+1})\, p(x_{t+1} \,|\, x_t, a_t)\, \psi_t(h_t)(x_t)\, \mu(\mathrm{d}x_t)\, \mu(\mathrm{d}x_{t+1})}. \tag{5.1}$$

The distribution Ψ_{t+1} is called the posterior distribution at epoch t. Starting with the prior distribution $\Psi_0 \in \mathcal{W}$ and repeatedly revising the probability distribution results in a sequence of probability distributions $\{\Psi_t\}_{t \in \mathbb{N}_0}$. At each epoch $t \in \mathbb{N}_0$ we are interested in the random variable $X_t \,|\, H_t = h_t$; the estimate of the unobserved state at epoch t. The estimation procedure has to revise the prior distribution consistently, in the sense that the probability distribution of $X_t \,|\, H_t = h_t$ is given by Ψ_t for all $t \in \mathbb{N}_0$. The following theorem shows that this is indeed the case.

Theorem 5.1: Let $B \in \sigma(\mathcal{X})$ and $\pi = \{\varphi_t\}_{t \in \mathbb{N}_0} \in \Pi_R$. Then,

$$\mathbb{P}^{\pi}_{\Psi_0}(X_t \in B \,|\, H_t = h_t) = \int_B \psi_t(h_t)(x)\, \mu(\mathrm{d}x), \quad t \in \mathbb{N}_0.$$

Proof: The proof is by induction in t. By definition the statement holds for $t = 0$. Now suppose that the statement holds for $t \in \mathbb{N}$, then

$$\begin{aligned}
\mathbb{P}^{\pi}_{\Psi_0}(X_{t+1} \in B \,|\, H_{t+1} = h_{t+1}) &= \frac{\mathbb{P}^{\pi}_{\Psi_0}\big(\{X_{t+1} \in B\} \cap \{H_{t+1} = h_{t+1}\}\big)}{\mathbb{P}^{\pi}_{\Psi_0}(H_{t+1} = h_{t+1})} \\
&= \frac{\int_B \int_{\mathcal{X}} \varphi(a_t \,|\, h_t)\, q(y_{t+1} \,|\, x)\, p(x \,|\, z, a_t)\, \psi_t(h_t)(z)\, \mu(\mathrm{d}z)\, \mu(\mathrm{d}x)}{\int_{\mathcal{X}} \int_{\mathcal{X}} \varphi(a_t \,|\, h_t)\, q(y_{t+1} \,|\, x)\, p(x \,|\, z, a_t)\, \psi_t(h_t)(z)\, \mu(\mathrm{d}z)\, \mu(\mathrm{d}x)} \\
&= \frac{\int_B \int_{\mathcal{X}} q(y_{t+1} \,|\, x)\, p(x \,|\, z, a_t)\, \psi_t(h_t)(z)\, \mu(\mathrm{d}z)\, \mu(\mathrm{d}x)}{\int_{\mathcal{X}} \int_{\mathcal{X}} q(y_{t+1} \,|\, x)\, p(x \,|\, z, a_t)\, \psi_t(h_t)(z)\, \mu(\mathrm{d}z)\, \mu(\mathrm{d}x)} \\
&= \int_B \psi_{t+1}(h_{t+1})(x)\, \mu(\mathrm{d}x).
\end{aligned}$$

The second equality follows from the induction hypothesis and the definition of $\mathbb{P}^{\pi}_{\Psi_0}$, see Equations (2.2) and (2.3). The third equality is valid since the decision rules may not depend on the unobserved states. □

The partially observed Markov decision process can be reformulated as a Markov decision process $(\overline{\mathcal{X}}, \{\overline{\mathcal{A}}_{\overline{x}} \mid \overline{x} \in \overline{\mathcal{X}}\}, \overline{p}, \overline{c})$ with complete information such that the costs incurred under equivalent policies in both processes are the same. The key idea in this transformation is to augment all information available to the decision maker to the new state space. Thus, apart from the observation space $\mathcal{Y}$, it should also include the space $\mathcal{W}$. The purpose of this construction is to store a probability distribution at every decision epoch that reflects the likely values of the state the process occupies. The first systematic proof of this technique for denumerable state spaces was given by Hinderer [62]. Rhenius [97] extended this result to Borel state spaces. We shall illustrate the technique for the partially observed Markov decision process.

Define the state space by $\overline{\mathcal{X}} = \mathcal{Y} \times \mathcal{W}$. For each $\overline{x} = (y, \Psi) \in \overline{\mathcal{X}}$ define the feasible action set by $\overline{\mathcal{A}}_{\overline{x}} = \mathcal{A}_y$. Note that Equation (5.1) cannot be used directly in the construction, as this would result in a non-Markovian process due to the dependency on the complete history. However, by construction, the state space has sufficient information to compute the posterior distribution. Hence, by virtue of the augmented state space the dependency on the history in Equation (5.1) can be dropped by defining the mapping $S : \mathcal{Y} \times \mathcal{W} \times \mathcal{A} \times \mathcal{Y} \to \mathcal{W}$ by

$$S(y, \Psi, a, y')(x) = \frac{\int_{\mathcal{X}} q(y' \mid x)\, p(x \mid z, a)\, \psi(z)\, \mu(\mathrm{d}z)}{\int_{\mathcal{X}} \int_{\mathcal{X}} q(y' \mid x)\, p(x \mid z, a)\, \psi(z)\, \mu(\mathrm{d}z)\, \mu(\mathrm{d}x)}, \tag{5.2}$$

for $(y, \Psi) \in \overline{\mathcal{X}}$, $a \in \mathcal{A}_y$, $y' \in \mathcal{Y}$, and $x \in \mathcal{X}$. This mapping is equivalent to the expression in Equation (5.1), and thus provides consistent estimates for the unobserved states. The transition probabilities of the Markov decision process are expressed in the mapping S by

$$\begin{aligned} \overline{p}(y', B \mid y, \Psi, a) = \int_{\mathcal{X}} \int_{\mathcal{X}} q(y' \mid x)\, p(x \mid z, a)\, \psi(z)\, \mu(\mathrm{d}z)\, \mu(\mathrm{d}x) \cdot \\ \mathbb{1}_B\big(S(y, \Psi, a, y')\big), \end{aligned} \tag{5.3}$$

for $(y, \Psi) \in \overline{\mathcal{X}}$, $a \in \mathcal{A}_y$, $y' \in \mathcal{Y}$, and $B \in \sigma(\mathcal{W})$, i.e., B is an element of the Borel σ-algebra of $\mathcal{W}$. The description of the Markov decision process is completed by

defining the cost function $\overline{c} : \mathcal{Y} \times \mathcal{W} \times \mathcal{A} \to \mathbb{R}$ by

$$\overline{c}(y, \Psi, a) = \int_{\mathcal{X}} c(x, a)\, \psi(x)\, \mu(\mathrm{d}x), \tag{5.4}$$

for $(y, \Psi) \in \overline{\mathcal{X}}$ and $a \in \mathcal{A}_y$. By Lemma 2 in Appendix 5 of Dynkin and Yushkevich [45] it follows that Expressions (5.3) and (5.4) are well-defined in the sense that they are measurable.

It is possible to study partially observed Markov decision processes with different information structures as well, e.g., the decision maker only recalls the last n observations for $n \in \mathbb{N}_0$. In these cases it is still possible to augment the state space with the available knowledge about the uncertainty in the states. Hence, the transformation method for the reduction to a standard Markov decision process can be done in a similar manner. In all cases, one needs to show that the optimal policies in both formulations yield the same cost in order to complete the reduction to the Markov decision process with complete information. For this purpose, define the equivalence relation $\sim$ on $\overline{\Pi}_R$ as follows: $\overline{\pi} \sim \overline{\pi}'$ if the histories $\overline{H}_t(\overline{\pi})$ and $\overline{H}_t(\overline{\pi}')$ induced by $\overline{\pi}$ and $\overline{\pi}'$, respectively, are equal in distribution for all $t \in \mathbb{N}_0$. This equivalence relation permits us to formulate a general theorem to prove equivalence of both models.

Theorem 5.2 (Thm. 2.1 in [71]): Consider a partially observed Markov decision problem, determined by the tuple $(\mathcal{X}, \mathcal{Y}, \{\mathcal{A}_y \mid y \in \mathcal{Y}\}, p, q, c, \Psi_0)$, and a Markov decision problem, determined by the tuple $(\overline{\mathcal{X}}, \{\overline{\mathcal{A}}_{\overline{x}} \mid \overline{x} \in \overline{\mathcal{X}}\}, \overline{p}, \overline{c})$. Assume that there exists a function $\Gamma : \Pi_R \to \overline{\Pi}_R$. If

- in every equivalence class there is a $\overline{\pi} \in \overline{\Pi}_R$ (which is called the representative of its class) such that there exists a $\pi \in \Pi_R$ with $\overline{\pi} = \Gamma(\pi)$;
- $\mathbb{E}^{\pi}_{\Psi_0}\, c(X_t, A_t) = \mathbb{E}^{\Gamma(\pi)}_{\Psi_0}\, \overline{c}(\overline{X}_t, \overline{A}_t)$ for all $\pi \in \Pi_R$ and $t \in \mathbb{N}_0$;

then every $\pi \in \Pi_R$ with $\Gamma(\pi)$ optimal in the second problem is also optimal in the original problem.

Theorem 5.2 permits us to study various information structures for the partially observed Markov decision problem. Note that one can restrict attention to the set of deterministic stationary policies, when it is known that the problem has an optimal deterministic stationary policy. The idea behind Theorem 5.2 is the following. Suppose that there exists an optimal policy $\overline{\pi}' \in \overline{\Pi}_R$ for the Markov decision problem. By the first condition there exists a policy $\overline{\pi} \in \overline{\Pi}_R$ in the

equivalence class of $\overline{\pi}'$, and a policy $\pi \in \Pi_R$ such that $\overline{\pi} = \Gamma(\pi)$. Note that the policy $\overline{\pi}$ is also optimal, since it generates the same history as $\overline{\pi}'$, by definition of the equivalence class, and thus also the same costs. Finally, the second condition establishes the optimality of π in the partially observed problem.

At epoch $t \in \mathbb{N}_0$ the decision rules in the partially observed process are based on information of the form $(\Psi_0, y_0, a_0, y_1, \ldots, a_{t-1}, y_t)$. This information is sufficient to derive information of the form $(y_0, \Psi_0, a_0, y_1, \Psi_1, \ldots, a_{t-1}, y_t, \Psi_t)$ in the equivalent Markov decision process. The reverse is also true by dropping the arguments $\Psi_1, \Psi_2, \ldots, \Psi_t$. Hence, there is a bijective relation between the set of policies Π_R in the partially observed process and the set of policies $\overline{\Pi}_R$ in the Markov decision model. Following Theorem 5.2, the reduction to the Markov decision model is completed by showing that equivalent policies in both formulations yield the same cost. This result for the partially observed Markov decision process is well-known (see, e.g., Section 7.2 in Araposthatis et al. [9], Chapter 3 in Bertsekas [20], and Chapter 6 in Kumar and Varaiya [79]).

Theorem 5.3: Let $\Psi_0 \in \mathcal{W}$, $\pi \in \Pi_R$, and let $\overline{\pi} \in \overline{\Pi}_R$ be the equivalent policy of π in the Markov decision process. Then,

$$\mathbb{E}^{\pi}_{\Psi_0} c(X_t, A_t) = \mathbb{E}^{\overline{\pi}}_{\Psi_0} \overline{c}(\overline{X}_t, \overline{A}_t), \quad t \in \mathbb{N}_0.$$

5.2 Bayesian decision problems

Bayesian decision models are completely determined by the tuple

$$(\mathcal{X}, \{\mathcal{A}_x \mid x \in \mathcal{X}\}, p^{\theta}, c^{\theta}, \Theta, F_0).$$

In this setting, the denumerable state space $\mathcal{X}$ and the feasible action set $\mathcal{A}_x$ for $x \in \mathcal{X}$ are defined as usual. The transition law and the cost function depend on a parameter that is unknown to the decision maker. Instead, he knows that the transition law and the cost function are given by p^{θ} and c^{θ}, respectively, when $\theta \in \Theta$ is the true value of the parameter. The set Θ is called the parameter space, and represents prior knowledge of the decision maker about the possible values that the unknown parameter can assume. In addition to the parameter space, the decision maker is also given a probability distribution F_0 defined on Θ that summarizes available background knowledge about the likely parameter values, prior to the first decision epoch.

To avoid measure theoretical problems, we assume that the parameter space Θ is a Borel space. We denote with $\mathcal{F}$ the σ-algebra of Borel subsets of Θ. The

set of all probability measures on the measurable space $(\Theta, \mathcal{F})$ is denoted with $\mathcal{W}$; this is also a Borel space (see Appendix 5 of Dynkin and Yushkevich [45]). As mentioned before, the decision maker is given a probability distribution $F_0 \in \mathcal{W}$, called the prior distribution, defined on Θ. We denote the generalized probability density function of F_0 by f_0 with corresponding probability measure μ.

The Bayesian decision model is a generalization of the Markov decision process described in Section 2.1; it coincides when the parameter space is a singleton. The problem can be formulated as a partially observed Markov decision process determined by $\big(\mathcal{X} \times \Theta, \mathcal{X}, \{\mathcal{A}_x \mid x \in \mathcal{X}\}, p, q, c, (x_0, F_0)\big)$, where $p(x', \theta' \mid x, \theta, a) = \mathbb{1}_{\{\theta'=\theta\}}\, p^\theta(x' \mid x, a)$, $q(y \mid x, \theta) = \mathbb{1}_{\{y=x\}}$, and $c(x, \theta, a) = c^\theta(x, a)$. Indeed, by specifying that the parameter is part of the state space, its true value becomes unknown. The transition probabilities reflect that we deal with a fixed parameter that does not change over time. Due to the definition of q, the process $\{X_t\}_{t \in \mathbb{N}_0}$ in the Bayesian decision model is observed in the partially observed process. This also implies that the feasible action sets may depend on the values in $\mathcal{X}$, and that the initial state x_0 is known. It follows that we can use the results of Section 5.1 to reformulate the Bayesian decision model to a Markov decision process with complete information.

Let the random variable Y, with values in the parameter space Θ, denote the value of the unknown parameter. At each epoch $t \in \mathbb{N}_0$ we are interested in the random variable $Y_t = \{Y \mid H_t = h_t\}$; the estimate of the unknown parameter at epoch t. From Theorem 5.1 we know that Equation (5.1) gives a consistent estimate for the unknown parameter. In order to derive a Markovian description, the transformation method in Section 5.1 prescribes that the new state space should be defined as $\mathcal{X} \times \sigma(\mathcal{X} \times \Theta)$, with $\sigma(\mathcal{X} \times \Theta)$ the Borel σ-algebra of $\mathcal{X} \times \Theta$. Suppose that at some epoch the decision maker has constructed a probability distribution F for the unknown parameter while he observed $x \in \mathcal{X}$. Furthermore, assume that the decision maker chooses action $a \in \mathcal{A}_x$ after which he observes $x' \in \mathcal{X}$. Then, by definition of q, Equation (5.2) becomes

$$S\big(x, (x, F), a, x'\big)(x', \theta) = \frac{p^\theta(x' \mid x, a)\, f(\theta)}{\int_\Theta p^\theta(x' \mid x, a)\, f(\theta)\, \mu(\mathrm{d}\theta)},$$

for $\theta \in \Theta$. Note that the notation can be simplified by replacing $\mathcal{X} \times \sigma(\mathcal{X} \times \Theta)$ with $\mathcal{X} \times \mathcal{W}$, since the underlying state in the Bayesian decision model is known with certainty through the observations. Hence, the formulation of the Markov decision process with complete information is as follows.

Define the state space by $\overline{\mathcal{X}} = \mathcal{X} \times \mathcal{W}$. For each $\overline{x} = (x, F) \in \overline{\mathcal{X}}$ define the feasible action set by $\overline{\mathcal{A}}_{\overline{x}} = \mathcal{A}_x$. Defining the mapping $S : \mathcal{X} \times \mathcal{W} \times \mathcal{A} \times \mathcal{X} \to \mathcal{W}$

by

$$S(x, F, a, x')(\theta) = \frac{p^\theta(x' \mid x, a)\, f(\theta)}{\int_\Theta p^\theta(x' \mid x, a)\, f(\theta)\, \mu(\mathrm{d}\theta)},$$

for $(x, F) \in \overline{\mathcal{X}}$, $a \in \mathcal{A}_x$, $x' \in \mathcal{X}$, and $\theta \in \Theta$. This mapping is equivalent to the expression in Equation (5.1), and thus provides consistent estimates for the unknown parameter. The transition probabilities of the Markov decision process are expressed in the mapping S by

$$\overline{p}(x', B \mid x, F, a) = \int_\Theta p^\theta(x' \mid x, a)\, f(\theta)\, \mu(\mathrm{d}\theta)\, \mathbb{1}_B\big(S(x, F, a, x')\big), \tag{5.5}$$

for $(x, F) \in \overline{\mathcal{X}}$, $a \in \mathcal{A}_x$, $x' \in \mathcal{X}$, and $B \in \sigma(\mathcal{W})$, i.e., B is an element of the σ-algebra of $\mathcal{W}$. The description of the Markov decision process is completed by defining the cost function $\overline{c} : \mathcal{X} \times \mathcal{W} \times \mathcal{A} \to \mathbb{R}$ by

$$\overline{c}(x, F, a) = \int_\Theta c^\theta(x, a)\, f(\theta)\, \mu(\mathrm{d}\theta), \tag{5.6}$$

for $(x, F) \in \overline{\mathcal{X}}$ and $a \in \mathcal{A}_x$. By Lemma 2 in Appendix 5 of Dynkin and Yushkevich [45] it follows that Expressions (5.5) and (5.6) are well-defined in the sense that they are measurable.

As mentioned before, a transition in the model results in additional information about the unknown parameter through observations of the state. Hence, the decision maker learns about the unknown parameter during the course of the process. This learning process is different from the one in general partially observed Markov decision processes where the unknown state continually changes. In Bayesian decision models the decision maker accumulates information about a fixed unknown parameter. Therefore, one expects the sequence $\{F_t\}_{t \in \mathbb{N}_0}$ to converge to a probability distribution that is degenerate at the true value of the unknown parameter. This assertion is true under some recurrence conditions. Loosely speaking, the recurrence condition guarantees that enough information is obtained about the value of the unknown parameter. The following theorem makes this statement more precise.

Theorem 5.4 (Thm. 2.4 in [58]): Let $F_0 \in \mathcal{W}$ be a given prior distribution, and let w be a bounded real-valued measurable function defined on Θ. For all $\pi \in \Pi_R$ such that the induced Markov chain given $Y = \theta$ is irreducible for all $\theta \in \Theta$, we have

$$\lim_{n \to \infty} \int_\Theta w(\theta) f_n(H_n)(\theta)\, \mu(\mathrm{d}\theta) = w(Y).$$

The result of Theorem 5.4 holds in particular for all bounded continuous functions w. Hence, f_n converges weakly to the distribution that is degenerate in Y, i.e., $\int_B f_n(H_n)(\theta)\,\mathrm{d}\theta$ converges almost surely to $\mathbb{1}_B(Y)$ for all $B \in \mathcal{F}$. This result also holds for a prior distribution $F_0' \in \mathcal{W}$ for which $F_0'(\theta) \neq 0$ whenever $F_0(\theta) \neq 0$ for $\theta \in \Theta$. For any such choice of the prior distribution it follows that the sequence of posterior distributions converges to the distribution that is degenerate in Y.

Note that at epoch $t \in \mathbb{N}_0$ the decision rules in the Bayesian decision model are based on information of the form $(F_0, x_0, a_0, x_1, \ldots, a_{t-1}, x_t)$. This information is sufficient to derive information of the form $(x_0, F_0, a_0, x_1, F_1, \ldots, a_{t-1}, x_t, F_t)$ in the equivalent Markov decision process. The reverse is also true by dropping the arguments $F_1, F_2, \ldots, F_t$. Hence, there is a bijective relation between the set of policies Π_R in the Bayesian decision model and the set of policies $\overline{\Pi}_R$ in the Markov decision model. In order to complete the reduction of the Bayesian decision problem with incomplete information to the equivalent Markov decision problem with complete information, we need to show that equivalent policies in both formulations yield the same cost (see Theorem 5.2). From Theorem 5.3 it follows that this statement holds.

Theorem 5.5: Let $x_0 \in \mathcal{X}$, $\pi \in \Pi_R$, and let $\overline{\pi} \in \overline{\Pi}_R$ be the equivalent policy of π in the Markov decision process. Then,

$$\mathbb{E}_{x_0}^{\pi} c^{Y_t}(X_t, A_t) = \mathbb{E}_{x_0}^{\overline{\pi}} \overline{c}(\overline{X}_t, \overline{A}_t), \quad t \in \mathbb{N}_0.$$

5.3 Computational aspects

So far, we have not mentioned a criterion function for the partial information model. By Theorem 5.3 the described reduction to a Markov decision process is valid under both the discounted and the average cost criterion. The state space of the equivalent Markov decision process is augmented with the space of probability measures $\mathcal{W}$. The purpose of this construction was to keep all necessary information in the state space instead of deriving the same information from the history of the process. The drawback of this technique, however, is that the state space is not denumerable anymore, even if the original state space is finite. Therefore, the obtained Markov decision process does not fit within the framework of Chapter 2, so that the results and algorithms derived in Chapter 2 do not apply.

The theory of Markov decision processes can be generalized to Borel state spaces (see, e.g., Dynkin and Yushkevich [45], Hinderer [62], and Hernández-

Lerma and Lasserre [60, 61]). The limitation of these models is that the derivation of optimal policies require substantial amount of computational effort. Sondik [115, 116], and Smallwood and Sondik [114] were the first to address and partially resolve computational difficulties associated with partial information models. However, these results are of limited practical use, since the complexity of the algorithms is very large.

Papadimitriou and Tsitsiklis [95] show that the complexity is most likely a generic property of the partial information model rather than a defect of the formulation as a Markov decision process. They study partial information models with a finite horizon, and show that the memory cells consumed by operations necessary for the computation of optimal policies is not polynomially bounded in the length of the horizon. They argue that, most likely, it is not possible to have an exact efficient implementation, involving polynomial time on-line computations and polynomial memory requirements, for the computation of optimal policies. This result even holds when an arbitrary amount of pre-processing is allowed.

Observe that Theorem 5.2 is not limited to the construction with the augmented state space only. Suppose that one has an alternative description for the state space, such that the information about the states are preserved, then Theorem 5.2 can be used to show equivalence of the original model with the alternative model. This transformation is worth doing when the state space of the alternative model has a lower dimension. In Chapter 6 we shall illustrate this approach for a class of routing problems. We show that the alternative description allows us to obtain both structural and computational results.

In Bayesian decision models it turns out that under the average cost criterion the model seems to be trivial when there is a policy such that the recurrence condition described in Theorem 5.4 holds. Indeed, due to the fact that the costs incurred in any finite number of epochs is not of influence on the average cost, the decision maker may learn about the unknown parameter until he has enough information. Afterwards, he solves the problem as a non-adaptive decision problem. This procedure results in a nearly optimal control policy. The optimal policy gathers information in the non-adaptive problem as well, but so rarely in time that its contribution to the average cost vanishes. The mathematical formalization of this idea can be found in van Hee [59]. In the case of the discounted cost criterion, there is a possible tension between the desire to obtain low immediate costs and to acquire information for parameter estimation in the model. In particular, the information and immediate costs cannot be separated and analyzed individually as in the average cost case. Therefore, the derivation of optimal policies is considerably harder in this case.

The computational complexity in the transformed Markov decision process is mainly caused by storing a probability distribution at each epoch. Assume that the distribution is fully determined by a vector of fixed dimension. If this property holds for the posterior distribution as well, then the probability distribution in the state space can be replaced by the vector instead. In that case, the state space becomes denumerable, so that the Markov decision process falls within the framework of Chapter 2. In the next section we shall illustrate this approach for the Bayesian decision model. These results will be applied to multi-armed bandit problems in Chapter 7.

5.4 Conjugate family of distributions

When a Bayesian decision model is transformed to a Markov decision process the dimensionality of the state space becomes too large to allow computations. In this section we shall construct a class of probability distributions parameterized by a vector of low dimension with the feature that the class is closed under the adaptation by Equation (5.1). This allows us to greatly reduce the dimensionality of the state space; instead of keeping record of a probability distribution, only the parameters of it need to be stored. A class of distributions that possess this property is called a conjugate family of distributions. We shall discuss the construction of such a family in the setting of statistical sampling theory.

Let Y be an unknown parameter that is an element of some parameter set Θ. Let $x = (x_1, \ldots, x_n)$ be a sample of size $n \in \mathbb{N}$ from a probability distribution defined on the set $\mathcal{X}$. We assume that the g.p.d.f. of this distribution depends on the unknown parameter Y, and denote it with p^θ when $Y = \theta$ for $\theta \in \Theta$. It will be understood that the sample x will be used for making inferences and decisions relating to the unknown parameter Y. In this context, a procedure T that can take as input $x \in \mathcal{X}^n$ for all $n \in \mathbb{N}$ is called a statistic. The statistic is said to be of fixed dimension when there exists a finite-dimensional set $\mathcal{T}$ such that $T(x) \in \mathcal{T}$ with $x \in \mathcal{X}^n$ for all $n \in \mathbb{N}$.

A statistic T is called sufficient if for any prior distribution of Y, its posterior distribution depends on a sample $x \in \mathcal{X}^n$ only through the statistic $T(x)$. More formally, let f_0 be a prior g.p.d.f. of Y, and denote with $f_n(\cdot \mid x)$ the posterior g.p.d.f. based on a sample $x \in \mathcal{X}^n$. Then, a statistic T is called sufficient if $f_n(\cdot \mid x) = f_n(\cdot \mid x')$ for any two samples $x, x' \in \mathcal{X}^n$ with $n \in \mathbb{N}$ such that $T(x) = T(x')$. This definition implies that in order to be able to compute the posterior distribution of Y from any prior distribution, it suffices to know the value of the statistic T instead of the sample x, which may be a vector of high

dimension. An efficient way to determine if a statistic is sufficient, is given by the following factorization criterion, known as the Fisher-Neyman criterion, from statistical analysis.

Theorem 5.6 (Fisher-Neyman criterion): A statistic T is sufficient if and only if there are functions g^θ and h_n such that for all $\theta \in \Theta$ and $x \in \mathcal{X}^n$ with $n \in \mathbb{N}$

$$p^\theta(x) = g^\theta\big(T(x)\big)\, h_n(x). \tag{5.7}$$

If there exists a sufficient statistic T of fixed dimension, then one can construct a conjugate family of distributions for Y, under the assumption that the joint probability distribution p^θ is the product of its marginal distributions (see Chapter 9 in DeGroot [36]). Before we show the construction, we need some additional terminology. Let p^θ and q^θ be functions that depend on $\theta \in \Theta$. Then p^θ is said to be proportional to q^θ, denoted by $p^\theta \propto q^\theta$, if p^θ is equal to q^θ multiplied by a factor, which does not depend on θ.

Theorem 5.7 (Ch. 9 in [36]): Let T be a sufficient statistic of fixed dimension taking values in some set $\mathcal{T}$. Suppose that the joint probability distribution p^θ is the product of its marginal distributions, and that the function g^θ in Equation (5.7) satisfies $\int_\Theta g^\theta(t)\, \mu(\mathrm{d}\theta) < \infty$ for all $t \in \mathcal{T}$. Then there exists a conjugate family of distributions for Y.

Proof: Let $x = (x_1, \ldots, x_n)$ be a sample of size $n \in \mathbb{N}$, and fix $\theta \in \Theta$. From Equation (5.7) it follows that

$$p^\theta(x) \propto g^\theta\big(T(x)\big). \tag{5.8}$$

Let $T(x) = t \in \mathcal{T}$, then by assumption $c = \int_\Theta g^\theta(t)\, \mu(\mathrm{d}\theta) < \infty$. Therefore, g^θ as a function of θ, can be seen as a probability density factored by c. Consequently, there exists a generalized probability density function $h(\,\cdot \,|\, t, n)$ such that

$$g^\theta(t) \propto h(\theta \,|\, t, n). \tag{5.9}$$

Now consider the family $\mathcal{G}$ of generalized probability density functions $h(\,\cdot \,|\, t, n)$ for all samples $x \in \mathcal{X}^n$ of all sizes $n \in \mathbb{N}$ with corresponding values of $T(x) \in \mathcal{T}$. It follows from Equations (5.8) and (5.9) that $p^\theta(x)$ is proportional to one of the generalized probability density functions in this family.

Suppose that $h(\,\cdot \,|\, s, m)$ and $h(\,\cdot \,|\, t, n)$ are two g.p.d.f.'s that belong to the family $\mathcal{G}$. By construction, there are values $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ such that

$T(x_1, \ldots, x_m) = s$ and $T(y_1, \ldots, y_n) = t$. The combination of the two samples forms a sample of size $m + n$, and by assumption it satisfies

$$p^\theta(x_1, \ldots, x_m, y_1, \ldots, y_n) = p^\theta(x_1, \ldots, x_m)\, p^\theta(y_1, \ldots, y_n). \tag{5.10}$$

Let $T(x_1, \ldots, x_m, y_1, \ldots, y_n) = u \in \mathcal{T}$, then from Equations (5.8) to (5.10) it follows that

$$h(\theta \,|\, u, m + n) \propto h(\theta \,|\, s, m)\, h(\theta \,|\, t, n).$$

Thus, we found a family $\mathcal{G}$ of generalized probability density functions that is closed under multiplication. Therefore, if the prior distribution f_0 is a member of this family, then by Bayes' rule we have $f_n(\theta \,|\, x_1, \ldots, x_n) \propto p^\theta(x_1, \ldots, x_n) f_0(\theta)$ for $\theta \in \Theta$ and $x \in \mathcal{X}^n$ with $n \in \mathbb{N}$. Hence, the posterior distribution is also a member of the family $\mathcal{G}$. □

The proof of Theorem 5.7 is constructive and provides an easy method for constructing conjugate families of distributions. In order to find such a family we need to determine a family of g.p.d.f.'s of θ, such that the joint probability distribution $p^\theta(x_1, \ldots, x_n)$ regarded as function of θ is proportional to a g.p.d.f. in the family. Furthermore, one has to check that this family is closed under multiplication as well. We shall illustrate this for samples from a Bernoulli distribution with an unknown parameter.

Theorem 5.8: Let $x = (x_1, \ldots, x_n)$ be a sample of size $n \in \mathbb{N}$ from a Bernoulli distribution with an unknown parameter. Suppose that the prior distribution for the unknown parameter is given by a Beta distribution with parameters $\alpha > 0$ and $\beta > 0$. Then the posterior distribution is given by a Beta distribution with parameters $\alpha + \sum_{i=1}^n x_i$ and $\beta + n - \sum_{i=1}^n x_i$.

Proof: Let $y = \sum_{i=1}^n x_i$, and fix $\theta \in [0, 1]$. Recall that a Beta distribution with parameters $\alpha > 0$ and $\beta > 0$ has a probability density function $f_{(\alpha,\beta)}$ given by

$$f_{(\alpha,\beta)}(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, \theta^{\alpha-1}\, (1 - \theta)^{\beta-1}, \tag{5.11}$$

for $\theta \in [0, 1]$, where $\Gamma(z) = \int_0^\infty u^{z-1} e^{-u}\, \mathrm{d}u$ for $z > 0$. It follows that the probability $p^\theta(x_1, \ldots, x_n) \propto \theta^y\, (1 - \theta)^{n-y}$ is proportional to a probability density in the family of Beta distributions, since $f_{(\alpha,\beta)}(\theta) \propto \theta^{\alpha-1}\, (1 - \theta)^{\beta-1}$. The posterior distribution also belongs to the family of Beta distributions, because $f_n(\theta) \propto \theta^{\alpha+y-1}\, (1 - \theta)^{\beta+n-y-1}$. Thus, the posterior distribution is a Beta distribution with parameters $\alpha + \sum_{i=1}^n x_i$ and $\beta + n - \sum_{i=1}^n x_i$. □

Distribution	Prior	Statistic T	Posterior
Bernoulli(θ)	Beta(α, β)	$\sum_{i=1}^{n} x_i$	Beta($\alpha + T$, $\beta + n - T$)
Exponential(θ)	Gamma(α, β)	$\sum_{i=1}^{n} x_i$	Gamma($\alpha + n$, $\beta + T$)
Neg. Bin(r, θ)	Beta(α, β)	$\sum_{i=1}^{n} x_i$	Beta($\alpha + rn$, $\beta + T$)
Normal(θ, r)	Normal(μ, τ)	$\frac{1}{n}\sum_{i=1}^{n} x_i$	Normal$\left(\frac{\tau\mu+nrT}{\tau+nr}, \tau + nr\right)$
Normal(μ, θ)	Gamma(α, β)	$\sum_{i=1}^{n}(x_i - \mu)^2$	Gamma($\alpha + n/2$, $\beta + T/2$)
Poisson(θ)	Gamma(α, β)	$\sum_{i=1}^{n} x_i$	Gamma($\alpha + T$, $\beta + n$)
Uniform(0, θ)	Pareto(R, α)	$\max_i\{x_i\}$	Pareto($\max\{T, R\}, \alpha + n$)

Table 5.1: Conjugate families of distributions.

Conjugate families of distributions can be derived for various other distributions with an unknown parameter as well. Table 5.1 summarizes results for the Beta, exponential, negative binomial, normal, Poisson, and the uniform distribution, where the parameter θ denotes the unknown parameter. The parameters of the normal distribution in this table are given by the mean μ and the precision r, i.e., $r = 1/\sigma^2$ where σ^2 is the variance. The definitions of the probability distributions and the proofs for the conjugate families can be found in Chapter 4 and Chapter 9, respectively, of DeGroot [36].

Chapter 6

Open-loop Routing Problems

Consider the problem of routing arriving customers to several parallel servers with no buffers. The servers work independently of each other and have a general service distribution, independent of the arrival process. The customers arrive according to a stationary arrival process, and have to be routed to one of the servers in order to receive service. We assume that the decision maker does not have information on the state of the servers, and a customer routed to a busy server preempts the customer being served. The objective in this setting is to minimize the average number of lost customers in the system due to preemption.

Koole [72] studied this problem under the assumption that the interarrival times are i.i.d. and the service times are exponentially distributed. He showed that there exists a periodic optimal policy that routes customers to every server infinitely often. Similar results have been obtained in a dual model of server assignment in Coffman, Liu, and Weber [34] with a different cost structure. The results in both problems heavily depend on the Markovian structure of the arrivals. An alternative approach based on the multimodularity of the cost functions has been presented in Altman, Gaujal, Hordijk, and Koole [7]. This approach uses the theory for optimization of multimodular functions in order to obtain structural results for the optimal policy without the Markovian assumptions.

In this chapter we extend the work of Coffman, Liu, and Weber [34], and Koole [72] to stationary arrival processes, general service distributions, and general convex cost functions. Based on Altman, Bhulai, Gaujal, and Hordijk [3, 4], we show that this problem is a special case of a partially observed Markov decision problem for which it is possible to obtain structural results. Computational techniques are also obtained by showing that the cost function in this problem is multimodular. These results are then used to obtain explicit expressions for the cost function for arrival processes that can model bursty and regular traffic. Finally, an application to robot scheduling for web search engines is discussed.

6.1 Problem formulation

Consider a system to which customers arrive at times $\{T_n\}_{n\in\mathbb{N}}$. We use the convention that $T_1 = 0$, and we assume that the process $\{\tau_n\}_{n\in\mathbb{N}}$ of interarrival times $\tau_n = T_{n+1} - T_n$ is stationary, thus $(\tau_m, \ldots, \tau_{m+h})$ and $(\tau_n, \ldots, \tau_{n+h})$ are equal in distribution for all $m, n, h \in \mathbb{N}_0$.

Assume that the system has M parallel servers with no waiting room, and independent service times, independent of the arrival process. Upon arrival of a customer the decision maker must route this customer to one of the M servers in the system. The service time has a general service distribution G_m when routed to server $m \in \{1, \ldots, M\}$. If there is still a customer present at the server, where the customer is routed to, then the customer in service is lost (we call this the preemptive discipline). In the special case of an exponential distribution, one can consider instead the non-preemptive discipline in which the arriving customer is lost; the results in the sequel will still hold.

We assume that the decision maker, who wishes to minimize the average number of lost customers in the system, has no information on the state of the servers (i.e., busy or idle). The only information that the decision maker possesses is its own previous actions, i.e., to which server previous customers were routed. We assume that this information does not include the actual time that has elapsed since a customer was last routed to that server. This assumption enables us to study the embedded Markov chain of the continuous-time process, i.e., we can consider the discrete-time process embedded at the epochs when customers arrive (see Section 2.5). The mathematical formulation of this problem is given by the following partially observed Markov decision process.

Let $\overline{\mathcal{X}} = \mathbb{N}_0^M$ denote the state space. For state $\overline{x} = (\overline{x}_1, \ldots, \overline{x}_M) \in \overline{\mathcal{X}}$ component $\overline{x}_m = 0$ for $m \in \{1, \ldots, M\}$ indicates that there is no customer at server m before the next decision. If $\overline{x}_m > 0$, then there is a customer present at server m and $\overline{x}_m$ is the total number of arrivals that occurred at other servers since this customer arrived. We assume that the system is initially empty, thus the initial state is given by $(0, \ldots, 0)$. Let $\overline{\mathcal{A}} = \{1, \ldots, M\}$ denote the action set, where action $\overline{a} \in \overline{\mathcal{A}}$ corresponds to sending the customer to server $\overline{a}$. Let S_m be a random variable with distribution G_m, then define $f_m(n)$ by

$$f_m(n) = \mathbb{P}(S_m \geq \tau_1 + \cdots + \tau_n), \quad n \in \mathbb{N}. \tag{6.1}$$

Note that $f_m(n)$ represents the probability that a customer did not leave server m during n arrivals. The actual value of the transition probability $p(x' \mid x, a)$ does not matter: Theorem 6.1 shows that only the transition probabilities for

the state components are relevant. The transition probabilities p_m for the state components are given by

$$\overline{p}_m(\overline{x}'_m \,|\, \overline{x}_m, \overline{a}) = \begin{cases} 1, & \text{if } \overline{x}_m = 0, \overline{a} \neq m \text{ and } \overline{x}'_m = 0 \\ f_m(1), & \text{if } \overline{x}_m = 0, \overline{a} = m \text{ and } \overline{x}'_m = 1 \\ 1 - f_m(1), & \text{if } \overline{x}_m = 0, \overline{a} = m \text{ and } \overline{x}'_m = 0 \\ f_m(n+1), & \text{if } \overline{x}_m = n \text{ and } \overline{x}'_m = n+1 \text{ for } n \in \mathbb{N} \\ 1 - f_m(n+1), & \text{if } \overline{x}_m = n \text{ and } \overline{x}'_m = 0 \text{ for } n \in \mathbb{N} \\ 0, & \text{otherwise.} \end{cases}$$

The direct costs are given by

$$\overline{c}(\overline{x}, \overline{a}) = \begin{cases} 0, & \overline{x}_{\overline{a}} = 0 \\ 1, & \overline{x}_{\overline{a}} > 0. \end{cases}$$

In this setting, the history at epoch $t \in \mathbb{N}$ of the process is defined by $\mathcal{H}_t = \mathcal{A}^{t-1}$. Thus, the decision maker is not allowed to design decision rules based on the occupancy information of the server. The only information available for decision making are the previously taken actions.

Koole [72] studied this problem under the assumption that the interarrival times are i.i.d. and the service times are exponential. In that case, the transition probabilities only depend on the presence of a customer at the server, whereas in our model we have to record the time a customer spends at the server. For the special case of symmetrical servers and arrival processes, the optimality of the round-robin policy was established. For the case of 2 servers, it was shown that a periodic policy whose period has the form $(1, 2, 2, 2, \ldots, 2)$ is optimal, where 2 means sending a customer to the faster server. Similar results have been obtained in a dual model of server assignment in Coffman, Liu, and Weber [34] with a somewhat different cost. The results in Coffman, Liu, and Weber [34] and Koole [72] heavily depend on the Markovian structure of the arrivals.

An alternative approach based on the multimodularity and Schur convexity of the cost functions has been presented in Altman, Gaujal, Hordijk, and Koole [7]. This approach uses the theory for optimization of multimodular functions (see, e.g., Altman, Gaujal, and Hordijk [5, 6], and Hajek [57]) in order to obtain structural results for the optimal policy without the Markovian assumptions made in Coffman, Liu, and Weber [34], and Koole [72]. The objective was to minimize losses or maximize the throughput.

In the next sections we shall obtain new insight on the structure of optimal policies under general stationary arrival processes. We show that customers should be routed to every server infinitely often, and that there exists an periodic optimal policy. We then present some further properties for the case of two servers. As a first application, we consider an optimal routing problem in which losses are to be minimized under the Poisson arrival process. In addition to the structure of the optimal policy for this problem, we provide explicit expressions for the costs for Markov modulated Poisson process and Markovian arrival process. Finally, we apply our results to the problem of robot scheduling for web search engines. Let us first start with the formulation of the Markov decision problem with complete information.

Let $\mathcal{X} = \big(\mathbb{N} \cup \{\infty\}\big)^M$ be the state space. The m-th component x_m of $x \in \mathcal{X}$ denotes the number of arrivals that occurred since the last arrival to server m. We assume that the initial state in this case is $(\infty, \ldots, \infty)$. Let $\mathcal{A} = \overline{\mathcal{A}}$ be the corresponding action space. The transition probabilities are now given by

$$p(x, a, x') = \begin{cases} 1, & \text{if } x'_a = 1 \text{ and } x'_m = x_m + 1 \text{ for all } m \neq a \\ 0, & \text{otherwise.} \end{cases}$$

Let the direct cost $c(x, a) = f_a(x_a)$, where $f_a(\infty)$ equals zero. Note that the policies in Π_R are allowed to depend on X_t, since these random variables do not refer to the state of the servers. Moreover, by Theorem 2.8 we know that there exists an optimal deterministic stationary policy. Hence, we can restrict our attention to $\overline{\Pi}_{DS}$ and Π_{DS} in Theorem 5.2. The equivalence of the previously defined models is now given by the following theorem.

Theorem 6.1: The models $(\overline{\mathcal{X}}, \overline{\mathcal{A}}, \overline{p}, \overline{c})$ and $(\mathcal{X}, \mathcal{A}, p, c)$ are equivalent in the sense that policies in the two models can be transformed into each other leading to the same performances. Furthermore, optimal policies exist for both models and the expected average cost of the optimal policies are equal.

Proof: Consider an arbitrary policy $\overline{\pi} \in \overline{\Pi}_{DS}$. Now, define the string of actions $(\overline{a}_1, \overline{a}_2, \ldots)$ by $\overline{a}_1 = \overline{\pi}_1$, $\overline{a}_2 = \overline{\pi}_2(\overline{a}_1)$, and so forth. Define $\Gamma : \overline{\Pi}_{DS} \to \Pi_{DS}$ such that $\pi = \Gamma(\overline{\pi})$ is the constant policy given by $\pi_t(h_t) = \overline{a}_t$ for all $h_t \in \mathcal{H}_t$.

From the definition of x_t it follows that the m-th component of x_t is equal to $(x_t)_m = t - \max_s\{\overline{a}_s = m\}$ if such an s exists, and infinity otherwise. It follows that $\mathbb{P}\big((\overline{X}_t)_m = n\big)$ is non-zero only if $\overline{a}_{t-n} = m$. Therefore, $\mathbb{P}\big((\overline{X}_t)_m = n\big) = f_m\big((x_t)_m\big)$ for $n = (x_t)_m$, and zero otherwise. Note that the expected cost only depends on the marginal probabilities. From the structure of the cost function and

the transition probabilities it follows that $\mathbb{E}\,\overline{c}(\overline{X}_t, \overline{a}) = f_{\overline{a}}\big((x_t)_{\overline{a}}\big) = \mathbb{E}\,c(X_t, \overline{a})$. From this fact it follows that the average cost obtained under the policies $\overline{\pi}$ and π in the model with partial and full information, respectively, is the same.

Let $\pi \in \Pi_{DS}$ be an arbitrary policy. Note that there is only one sample path $h_t \in \mathcal{H}_t$ that occurs with non-zero probability, since the initial state is known and the transition probabilities are zero or one. Define the equivalence relation $\sim$ on Π_{DS} such that $\pi \sim \pi'$ if $\pi_t(h_t) = \pi'_t(h_t)$ for all t. Now it directly follows that policies in the same class have the same cost. Furthermore, it follows that the constant policy π'_t for each $t \in \mathbb{N}_0$ is a representative element in the class, which is the image of a policy in $\overline{\Pi}_{DS}$.

At this point we know that if there exists an optimal policy for the full observation problem, then there is also an optimal policy for the partially observed problem with the same value. In order to prove the existence of an optimal policy for our expected average cost problem, we first introduce the discounted cost

$$V^\alpha(\pi, x) = \mathbb{E}_x^\pi \sum_{t=0}^{\infty} \alpha^t\, c(X_t, A_t), \quad \pi \in \Pi_{DS},\ x \in \mathcal{X}.$$

The optimal discounted cost is defined as $V^\alpha(x) = \min\{V^\alpha(\pi, x) \mid \pi \in \Pi_{DS}\}$. A sufficient condition for the existence of optimal stationary policy for the expected average cost is that $\big|V^\alpha(x) - V^\alpha(z)\big|$ is uniformly bounded in α and x for some z, where $V^\alpha(x)$ represents the minimal α-discounted costs (see Theorem 2.2, Chapter 5 in Ross [101]).

Note that using the same policy for different initial states can only lead to a difference in cost when a customer is sent to a server for the first time. Since $0 \le \alpha^t\, c(x, a) \le 1$, it follows that the difference in cost cannot be larger than M. Now let π_z denote the α-optimal policy for initial state z, and $V^\alpha(\pi_z, x)$ the value for the assignment done for initial state x using π_z. Then

$$V^\alpha(x) - V^\alpha(z) \le V^\alpha(\pi_z, x) - V^\alpha(z) = V^\alpha(\pi_z, x) - V^\alpha(\pi_z, z) \le M.$$

The proof is completed by repeating the argument for $V^\alpha(z) - V^\alpha(x)$ with π_x. □

Consider the following situation as an illustration of the class of routing problems just described. Suppose that the decision maker wishes to minimize the expected number of lost customers (i.e., the number of preempted customers) per unit time. The cost function $f_m(n)$ for $m \in \{1, \ldots, M\}$ will typically be a decreasing function in $n \in \mathbb{N}$, because a longer time interval between an assignment to server m results in a smaller probability that a customer, that was previously assigned

there, is still in service. Assume that the arrival process is a Poisson process with rate λ and that services at server $i \in \{1, \ldots, M\}$ are exponentially distributed with rate μ_i independent of the other servers. Let S_i be a random variable which is exponentially distributed with rate μ_i. Then $f_m(n) = \mathbb{P}(S_m \geq \tau_1 + \cdots + \tau_n) = \big[\lambda/(\lambda + \mu_m)\big]^n$.

In the previous example we observed that $f_m(n)$ was a decreasing function in $n \in \mathbb{N}$ for all $m \in \{1, \ldots, M\}$. In Section 6.5 we want to study an application for which $f_m(n)$ should be increasing in n. Therefore, we extend our framework to general convex functions $f_m(n)$, that are not necessarily decreasing. For that purpose, we shall disregard the partial information model in the sequel, and consider the T-horizon Markov decision process, determined by $(\mathcal{X}, \mathcal{A}, p, c_t)$, to be given. Here, the sets $\mathcal{X}, \mathcal{A}$, and p are defined as in the Markov decision process with complete information. The immediate costs are given by $c_t(x, a) = f_a(x_a)$ for $t < T$. The terminal costs are given by

$$c_T(x) = \sum_{a \in \mathcal{A}} f_a(x_a), \quad x \in \mathcal{X}. \tag{6.2}$$

Defining these terminal costs will be essential for the mathematical results as will be illustrated later. It has also a natural physical interpretation as will be illustrated in Section 6.5. Similarly to Equation (2.6), the expected average cost criterion function $g(\pi, x)$ is defined by

$$g(\pi, x) = \lim_{T \to \infty} \mathbb{E}_x^\pi \frac{1}{T} \sum_{t=0}^{T} c_t(X_t, A_t), \quad \pi \in \Pi_{DS},\ x \in \mathcal{X},$$

where $x = (x_1, \ldots, x_M) \in \mathcal{X}$ is the initial state. We shall also write $g(\pi)$ when the initial state is not of importance.

6.2 Structural results

In this section we shall investigate structural properties of the optimal policy. For that purpose, we need some general assumptions on the cost function in order to cover different cost structures. Therefore, assume that all the f_m are convex. Moreover, one of Conditions (6.3)–(6.5) defined below holds for all $m \in \{1, \ldots, M\}$ simultaneously.

$$\lim_{x \to \infty} \big(f_m(x+1) - f_m(x)\big) = \infty, \tag{6.3}$$

$$f_m \text{ are strictly convex and } \lim_{x \to \infty} \big(f_m(x) - a_m x\big) = C, \tag{6.4}$$

$$f_m(x) = a_m x + C, \tag{6.5}$$

where strictly convex means that $f_m(x+2) - f_m(x+1) > f_m(x+1) - f_m(x)$ for all $x \in \mathbb{N}$, and where $a_m \geq 0$ and C are constants.

Note that Condition (6.5) is not included in Condition (6.4): it violates its first part. Condition (6.3) covers the case where f_m grows more than linearly, whereas Conditions (6.4) and (6.5) cover the case where f_m grows asymptotically linearly. These conditions are complementary to Condition (6.3) since any one of them implies that $\lim_{x\to\infty} \big(f_m(x+1) - f_m(x)\big) < \infty$. In Conditions (6.3) and (6.4), f_m is strictly convex.

Theorem 6.2: Assume that one of Conditions (6.3)–(6.5) holds, then

(i) There exists an optimal policy that uses every server infinitely many times, thus $\sup\{j \mid \pi_j = m\} = \infty$ for $m \in \{1, \ldots, M\}$.

(ii) There exists an periodic optimal policy.

Before proving the theorem we illustrate the necessity of some parts of Conditions (6.3)–(6.5). We present cost functions for which the above theorem does not hold.

Example 6.3: Consider the case of two servers with the costs

$$f_i(x) = a_i\, x + b_i\, \exp(-d_i\, x) + c_i, \quad i = 1, 2,$$

where $c_1 < c_2$, and where $b_i \geq 0$ and $d_i > 0$ are some constants (as follows from the next remark, the sign of a_i is not important). For a sufficiently small value of b_1, the policy that always routes to server 1 is the only optimal policy for any finite horizon T.

Indeed, assume first $b_i = 0$ for $i = 1, 2$ and let π be any policy that routes at its n-th step to server 2. By changing the action at this step into an assignment to server 1 we gain $c_2 - c_1$. By continuity of the cost in the b_i's, we also gain a positive amount using this modified policy if $b_i \neq 0$, provided that the b_i's are sufficiently small. Hence, for b_i sufficiently small, a policy cannot be optimal if it does not always route to server 1.

When using the average cost criterion, the cost is not affected anymore by any changes in the actions provided that the frequency of such changes converges to zero. Hence, for the average cost, there may be other policies that are optimal, but still, any policy for which the fraction of customers routed to server 2 does not converge to 0 cannot be optimal. We conclude that we cannot relax Conditions (6.4) or (6.5) and replace C by C_m.

Remark 6.4: Suppose that the cost $f_i(x)$ contains a linear term $a_i x$ for $i = 1, 2$. Then the total accumulated cost that corresponds to the term $a_i x$ over any horizon of length T is $a_i(T + x_i)$, where $x = (x_1, x_2)$ is the initial state. This part does not depend on the policy. If we use a policy π and then modify it by changing an assignment at epoch $t < T$ from server i to server $j \neq i$, then the linear part of the cost at epoch t under the modified policy decreases by $a_i\, x_i(t) - a_j\, x_j(t)$, but it then increases by the same amount at epoch $t + 1$. Thus, the accumulated linear cost is independent of the policy. Note that this argument is valid due to the definition of the cost at epoch T in Equation (6.2).

Example 6.5: Consider the case of two servers with the costs $f_1(x) = a_1 x$ and $f_2(x) = \exp(-d_2 x)$. For any finite horizon T and for $d_2 > 0$, the only optimal policy is the one that always routes to server 1. Note that the average cost until epoch T of the policy that always routes to server 1 is

$$\frac{T f_2(1) + f_1(T)}{T} = e^{-d_2} + a_1.$$

The average cost of the policy that always routes to server 2 is

$$\frac{T f_1(1) + f_2(T)}{T} = a_1 + \frac{e^{-d_2 T}}{T}.$$

Again, for the average cost there are other optimal policies but they have to satisfy the following: the fraction of customers routed to queue 2 by epoch T should converge to zero as $T \to \infty$. This illustrates the necessity of the first part in Condition (6.4). For $d_2 < 0$, the only optimal policy is the one that always routes to server 2.

Next we present an example to illustrate the importance of the terminal cost.

Example 6.6: Assume that there are two servers, and that the costs are given by $f_1(x) = x^2$ and $f_2(x) = 2\,x^2$. Assume that the terminal costs $c_T(x)$ were zero, then the policy that always routes to server 1 is optimal.

Proof of Theorem 6.2 : First suppose that the cost function satisfies Condition (6.5). Interchanging assignments between any two servers for any finite horizon does not result in changes in cost for that horizon, due to the linearity of the cost function and the terminal cost. Hence, any periodic policy that routes customers to all servers is optimal.

We consider next Conditions (6.3) and (6.4). Instead of describing a policy using a sequence π, we use an equivalent description using time distances between customers routed to each server. More precisely, given an initial state $x \in \mathcal{X}$, define the j-th instant at which a customer is routed to server m by

$$\eta^m(0) = -x_m, \quad \text{and} \quad \eta^m(j) = \min\left\{i \,\middle|\, [\eta^m(j-1)]^+ < i \le T \text{ and } \pi_i = m\right\},$$

for $j \in \mathbb{N}$, where $[x]^+ = \max\{x, 0\}$. We use the convention that the minimum of an empty set is taken to be infinity. Define the distance sequence δ^m by $\delta^m = (\delta^m_1, \delta^m_2, \ldots)$, with $\delta^m_j = \eta^m(j) - \eta^m(j-1)$ for $j \in \mathbb{N}$. For simplicity we do not include the T in the notation.

Let $\pi \in \Pi_{DS}$ be an arbitrary policy, and let m be an arbitrary server. Assume that the sequence $\delta = \delta^m$ for this server satisfies $\limsup_{T\to\infty}\{\delta_j \,|\, j \in \mathbb{N}\} = \infty$. We shall construct a policy $\pi' \in \Pi_{DS}$ with distance sequence $\delta^{\pi'}$, such that $\limsup_{T\to\infty}\{\delta^{\pi'}_j \,|\, j \in \mathbb{N}\}$ is finite and $g(\pi') \le g(\pi)$. Assume first that f satisfies Condition (6.3). Choose n_0 such that for all $n > n_0$

$$\min_{1\le k\le 2M+1} \big(f_m(n+k) - f_m(k) - f_m(n)\big) > \max_{1\le k<2M+1} \big(f_l(2M+1) - f_l(k) - f_l(2M+1-k)\big),$$

for all $l \in \{1, \ldots, M\}$. Since the supremum of the distance sequence is infinite, there is a $j \in \mathbb{N}$ (and T) such that $\delta_j > n+2M+1$. Consider the $2M+1$ consecutive assignments starting at $\eta^m(j-1)$. Since there are M servers, it follows that there is at least one server to which a customer is assigned three times during that period, say m'. Denote the distance (or interarrival) times of assignments to m' by δ^1 and δ^2. Replace the second assignment to m' in this sequence by an assignment to server m. Denote the new distance (or interarrival) times to m by δ'_j and δ''_j (if $\eta(j) = T$, then the distance δ''_j is not a real interarrival time). Consider the cost for a horizon of length l, where l is an arbitrary integer larger than $\eta(j-1) + n_0 + 2M + 1$. Then,

$$\big[f_m(\delta_j) + f_{m'}(\delta^1) + f_{m'}(\delta^2)\big] - \big[f_m(\delta'_j) + f_m(\delta''_j) + f_{m'}(\delta^1+\delta^2)\big] =$$

$$\big[f_m(\delta_j) - f_m(\delta'_j) - f_m(\delta''_j)\big] - \big[f_{m'}(\delta^1+\delta^2) - f_{m'}(\delta^1) - f_{m'}(\delta^2)\big] > 0,$$

where the last inequality follows from the choice of n_0.

Now consider Condition (6.4). Since by assumption the supremum of the distance sequence $\delta_l = \delta^m_l$ for $l = 1, 2, \ldots$ is infinite, there is a $j \in \mathbb{N}$ (and T) such

that $\delta_j > 2n+2M+1$ for some $n \in \mathbb{N}$. Let $p = \min\{f_m(k)+a_m-f_m(k+1) \mid m = 1, \ldots, M,\ k = 1, \ldots, M\}$. Note that p is positive, since Condition (6.4) implies that $\big(f_l(x) - a_l x - C\big)$ is positive and strictly decreasing for all $l \in \mathbb{N}$. Choose n such that $2q = 2\big(f_m(n) - a_m n - C\big) < p$. Note that this is possible, since $f_m(n) - a_m n - C$ goes to zero as n goes to infinity. Consider the $2M+1$ consecutive assignments starting n units after $\eta(j-1)$. There is at least one server to which a customer is assigned three times, say m'. Replace the second assignment to m' in this sequence by an assignment to server m.

Define the function $g_i(k) = f_i(k) - a_i k - C$ for all $i \in \{1, \ldots, M\}$, and consider the cost for a horizon of length l, where $l \in \mathbb{N}$ is an arbitrary integer larger than $\eta(j-1)+2n+2M+1$. The decrease in cost due to the interchange is

$$
\begin{aligned}
&\big[f_m(\delta_j) + f_{m'}(\delta^1) + f_{m'}(\delta^2)\big] - \big[f_m(\delta'_j) + f_m(\delta''_j) + f_{m'}(\delta^1+\delta^2)\big] = \\
&\big[g_m(\delta_j) + g_{m'}(\delta^1) + g_{m'}(\delta^2)\big] - \big[g_m(\delta'_j) + g_m(\delta''_j) + g_{m'}(\delta^1+\delta^2)\big] > \\
&\big[g_m(\delta_j) + g_{m'}(\delta^1) + g_{m'}(\delta^2)\big] - \big[2g_m(n) + g_{m'}(\delta^1+1)\big] = \\
&g_m(\delta_j) + g_{m'}(\delta^2) + \big[g_{m'}(\delta^1) - g_{m'}(\delta^1+1)\big] - \big[2g_m(n)\big] > \\
&g_m(\delta_j) + g_{m'}(\delta^2) + p - 2q > 0,
\end{aligned}
$$

where δ^1, δ^2, δ'_j, and δ''_j are defined as before. The first inequality follows from the fact that $n < \delta'_j$, $n < \delta''_j$, $\delta^1+1 \le \delta^1+\delta^2$, and $f_m(x) - a_m x - C$ is decreasing. The second inequality follows from the definition of p. Since $f_m - a_m - C$ is positive it follows by construction of n that the last inequality holds.

Repeating the interchange argument for every $j \in \mathbb{N}$ for which $\delta_j > 2n + 2M + 1$ when dealing with Condition (6.4), or for which $\delta_j > n + 2M + 1$ when dealing with Condition (6.3), provides us a policy π' such that $g(\pi') \le g(\pi)$ and $\sup\{\delta_j^{\pi'} \mid j \in \mathbb{N}\} < 2n + 2M + 1$. By repeating this procedure for every server, we get an optimal policy that visits a finite number of states. By Theorem 2.8 we know that the optimal policy can be chosen stationary. It follows that $\pi_n(h_n) = \pi_0(x_n)$. Since the state transitions are deterministic it follows that the optimal policy is periodic. □

6.3 Multimodularity

Let $\mathbb{Z}$ be the set of integers. Let $e_i \in \mathbb{Z}^n$ for $i = 1, \ldots, n$ denote the vector having all entries zero except for a 1 in the i-th entry. Let d_i be given by $d_i = e_{i-1} - e_i$ for $i = 2, \ldots, n$. The base of vectors for multimodularity is defined as the collection

$\mathcal{B} = \{b_0 = -e_1, b_1 = d_2, \ldots, b_{n-1} = d_n, b_n = e_n\}$. Following Hajek [57] a function g defined on $\mathbb{Z}^n$ is called multimodular if for all $x \in \mathbb{Z}^n$ and b_i, $b_j \in \mathcal{B}$ with $i \neq j$

$$g(x + b_i) + g(x + b_j) \geq g(x) + g(x + b_i + b_j). \tag{6.6}$$

We define the notion of an atom in order to study some properties of multimodularity. In $\mathbb{R}^n$ the convex hull of $n + 1$ affine independent points in $\mathbb{Z}^n$ forms a simplex. This simplex, defined on $x_0, \ldots, x_n$, is called an atom if for some permutation σ of $(0, \ldots, n)$ we have

$$\begin{array}{lcccl} x_1 & = & x_0 & + & b_{\sigma(0)}, \\ x_2 & = & x_1 & + & b_{\sigma(1)}, \\ \vdots & & \vdots & & \vdots \\ x_n & = & x_{n-1} & + & b_{\sigma(n-1)}, \\ x_0 & = & x_n & + & b_{\sigma(n)}. \end{array}$$

This atom is referred to as $A(x_0, \sigma)$, and the points $x_0, \ldots, x_n$ are called the extreme points of this atom. Each unit cube is partitioned into $n!$ atoms and all atoms together tile $\mathbb{R}^n$. We can use this to extend the function $g : \mathbb{Z}^n \to \mathbb{R}$ to the function $\overline{g} : \mathbb{R}^n \to \mathbb{R}$ as follows. If $x \in \mathbb{Z}^n$, then define the function by $\overline{g}(x) = g(x)$; thus $\overline{g}$ agrees with g on $\mathbb{Z}^n$. If x is not an extreme point of an atom, then x is contained in some atom A. The value of $\overline{g}$ on such a point $x \in \mathbb{R}^n$ is obtained as the corresponding linear interpolation of the values of g on the extreme points of A. The following theorem shows the relation between multimodularity and convexity.

Theorem 6.7 ([57], [6], Theorem 2.1 in [14]): A function g is multimodular if and only if the function $\overline{g}$ is convex.

Let $g : \mathbb{Z}^n \to \mathbb{R}$ be an objective function in a mathematical model that has to be minimized. In general this is a hard problem to solve. However, if the function g is multimodular, then we can use a local search algorithm, which converges to the globally optimal solution. In Koole and Van der Sluis [76] the local search algorithm has been proven for the search space $\mathbb{Z}^n$. The following theorem shows that it also holds for any convex subset S, which is a union of a set of atoms; thus in particular for $S = \mathbb{N}_0^n$, which is essential for our application. Multimodularity of g and convexity of $\overline{g}$ still hold when restricted to S. Multimodularity on this convex subset means that Expression (6.6) must hold for all $x \in S$ and b_i, $b_j \in \mathcal{B}$ with $i \neq j$, such that $x + b_i$, $x + b_j$, and $x + b_i + b_j \in S$.

Theorem 6.8: Let g be a multimodular function on S, a convex subset which is a union of atoms. A point $x \in S$ is a global minimum of g if $g(x) \leq g(y)$ for all $y \neq x$ that are an extreme point of $A(x, \sigma)$ for some σ.

Proof: Let $x \in S$ be a fixed point. Suppose that there is a $z \in \mathbb{R}^n$ such that $x+z \in S$ and $\overline{g}(x+z) < \overline{g}(x) = g(x)$. We show that there is an atom $A(x, \sigma)$ in S with an extreme point y such that $g(y) < g(x)$.

Since any n vectors of $\mathcal{B}$ form a basis of $\mathbb{R}^n$, we can write z as $z = \sum_{i=0}^{n} \beta_i \, b_i$. Furthermore, this can be done such that $\beta_i \geq 0$, since any $\beta_i \, b_i$ with $\beta_i < 0$ can be replaced by $-\beta_i \sum_{j=0, j\neq i}^{n} b_j$.

Now, reorder the elements of $\mathcal{B}$ as $(b'_0, \ldots, b'_n)$ and the elements $(\beta_0, \ldots, \beta_n)$ as $(\beta'_0, \ldots, \beta'_n)$, such that $\beta'_0 \geq \cdots \geq \beta'_n \geq 0$ and $z = \sum_{i=0}^{n} \beta'_i \, b'_i$. This notation is equivalent to the notation $z = \beta''_0 \, b'_0 + \beta''_1 \, (b'_0 + b'_1) + \cdots + \beta''_n \, (b'_0 + \cdots + b'_n)$ with all $\beta''_i \geq 0$ and $b'_0 + \cdots + b'_n = 0$. Fix a non-zero point $z' = \alpha \, z$ with $\alpha < 1$ such that $\alpha \, (\beta''_0 + \cdots + \beta''_{n-1}) \leq 1$. The set S is convex with x, $x + z \in S$, hence $x + z' \in S$. Since by Theorem 6.7 $\overline{g}$ is convex and $\overline{g}(x+z) < g(x)$, it follows that $\overline{g}(x + z') < g(x)$.

Let σ be the permutation induced by $(\beta''_0, \ldots, \beta''_n)$. Now consider the atom $A(x, \sigma)$, then by construction $x + z' \in A(x, \sigma)$. Since x is an extreme point and $\overline{g}$ is linear, there must be another extreme point, say y, such that $g(y) < g(x)$. □

This theorem shows that to check whether $x \in S$ is a globally optimal solution, it is sufficient to consider all extreme points of all atoms with x as an extreme point. When a particular extreme point of such an atom has a lower value than x, then we repeat the same algorithm with that point. Repeating this procedure is guaranteed to lead to the globally optimal solution.

Every multimodular function is also integer convex (see Theorem 2.2 in Altman, Gaujal, and Hordijk [6]). One could wonder if local search also works with integer convexity instead of multimodularity, which is a stronger property. The next counter-example in $(\{0, 1, 2\})^2$ shows that this is not true, and that multimodularity is indeed the property needed for using local search. Define the function g such that

$$\begin{array}{lll} g(0,2) = -1, & g(1,2) = 2, & g(2,2) = 5, \\ g(0,1) = 2, & g(1,1) = 1, & g(2,1) = 0, \\ g(0,0) = 5, & g(1,0) = 4, & g(2,0) = 3. \end{array}$$

One can easily check that g is an integer convex function, but not multimodular since $g\big((1,2)+b_0\big)+g\big((1,2)+b_2\big) = 0 < 4 = g\big((1,2)\big)+g\big((1,2)+b_0+b_2\big)$. Starting

the local search algorithm at coordinate $(2,1)$ shows that all neighbours have values that are greater than 0. However, the global minimum is $g\big((0,2)\big) = -1$.

Since all the extreme points of all atoms with x as an extreme point can be written as $x + \sum_{i=0}^{n} \alpha_i\, b_i$ with $\alpha_i \in \{0,1\}$ and $b_i \in \mathcal{B}$, the neighborhood of a point x consists of $2^{n+1} - 2$ points (we subtract 2 due to the fact that when $\alpha_i = 0$ or when $\alpha_i = 1$ for all $i \in \{1, \ldots, n\}$, then the points coincide). Although the complexity of the local search algorithm is big for large $n \in \mathbb{N}$, the algorithm is worthwhile studying. First of all, comparing to minimizing a convex function on the lattice, this algorithm gives a reasonable improvement. Secondly, the algorithm can serve as a basis for heuristic methods.

In the previous section we showed that the optimal policy for all stationary arrival processes is periodic. In this section we shall show that the value function is multimodular when restricted to periodic policies. This can then be used to get a much more detailed structure of optimal policies, in particular, the regularity of optimal policies in the case of two servers (see Altman, Gaujal, and Hordijk [6]). Moreover, the local search algorithm can be applied to find an optimal policy.

We note that some results on multimodularity have been obtained already in Altman, Gaujal, Hordijk, and Koole [7] for each individual server with respect to the routing sequence to that server. This allows to obtain some structure for the optimal policies as in Altman, Gaujal, and Hordijk [5]. The novelty of the results here is that, for the case of two servers to which we restrict, multimodularity is obtained directly for the global cost of all servers rather than for each one separately.

The notation of the distance sequence can be beneficially used to approach the decision problem. After the first assignment to server $m \in \{1, \ldots, M\}$, the distance sequence δ^m for server m is periodic, say with period $d(m)$. Therefore, in future discussions we will write $\pi = (\pi_1, \ldots, \pi_n)$ for the periodic assignment sequence with period $n \in \mathbb{N}$, and with a slight abuse of notation we denote the periodic distance sequence for server m by $\delta^m = (\delta^m_1, \ldots, \delta^m_{d(m)})$.

The periodicity reduces the cost function in complexity. Since we use the expected average cost function, we only have to consider the costs incurred during one period. It would be interesting to establish multimodular properties for any $M \in \mathbb{N}$. Unfortunately, it is not clear how even to define multimodularity for $M > 2$. Therefore, we restrict attention to $M = 2$ in the sequel. The expected average cost for periodic policies $\pi \in \Pi_{DS}$ is given by

$$g(\pi) = \sum_{m=1}^{M} g_m(\pi) = \frac{1}{n} \sum_{m=1}^{M} \sum_{j=1}^{d(m)} f_m(\delta^m_j). \tag{6.7}$$

It is tempting to formulate that $g_m(\pi)$ is multimodular in π for $m = 1, 2$. Note that this is not necessarily true, since an operation $v \in \mathcal{B}$ applied to π leads to different changes in the distance sequences for the different servers.

We shall thus use an alternative description for the average cost through the period of server 1. Define g'_m by

$$g'_m(\pi) = \frac{1}{n} \sum_{j=1}^{d(1)} f_m(\delta_j^1). \tag{6.8}$$

We note that the function $g'_m(\pi)$ only looks at the distance sequence assigned to the first server with respect to π using cost function f_m. By the symmetry between the assignments to the two servers, $g(\pi)$ can now be expressed as $g(\pi) = g'_1(\pi) + g'_2(\vec{3} - \pi)$, with $\vec{3}$ the vector whose components are all 3. Note that $\delta_j^1 = \delta_j^1(\pi)$ is a function of π, and we have

$$\delta_j^1(\vec{3} - \pi) = \delta_j^2(\pi).$$

We first prove that $g'_m(\pi)$ is multimodular in π. Then, multimodularity of $g(\pi)$ follows as the sum of two multimodular functions.

Lemma 6.9: Assume that f_m are convex. Let $\pi \in \Pi_{DS}$ be a fixed periodic policy with period $n \in \mathbb{N}$. Let $g'_m(\pi)$ be defined as in Equation (6.8), then $g'_m(\pi)$ is a multimodular function in π.

Proof: Since π is a periodic sequence, the distance sequence $\delta = \delta^1$ is also a periodic function, say with period $p = d(1)$. Now, define the function h_j for $j = 1, \ldots, p$ by $h_j(\pi) = f_m(\delta_j)$. The function h_j represents the cost of the $(j+1)$-st assignment to server m by looking at the j-th interarrival time. We will first check the conditions for multimodularity for $\mathcal{V} = \{b_1, \ldots, b_{n-1}\}$.

Let $v, w \in \mathcal{V}$ with $v \neq w$. If none of these elements changes the length of the j-th interarrival time, then $h_j(\pi) = h_j(\pi + v) = h_j(\pi + w) = h_j(\pi + v + w)$. Suppose that only one of the elements changes the length of the interarrival time, say v, then $h_j(\pi + v) = h_j(\pi + v + w)$ and $h_j(\pi) = h_j(\pi + w)$. In both cases the function $h_j(\pi)$ satisfies the conditions for multimodularity.

Now suppose that v increases and w decreases the length of the j-th interarrival time by one. Then $\delta_j^1(\pi + v) - \delta_j^1(\pi) = \delta_j^1(\pi + v + w) - \delta_j^1(\pi + w)$. Since h_j is a convex function, it follows that $h_j(\pi+w)-h_j(\pi+v+w) \geq h_j(\pi)-h_j(\pi+v)$. Now the multimodularity condition in Equation (6.6) directly follows by rearranging

the terms. Since $g'_m(\pi)$ is a sum of $h_j(\pi)$ it follows that $g'_m(\pi)$ is multimodular for $\mathcal{V}$.

Now consider the elements b_0 and b_n and note that the application of b_0 and b_n to π splits an interarrival period and merges two interarrival periods, respectively. Therefore,

$$n\, g'_m(\pi + b_0) = n\, g'_m(\pi) - f_m(\delta_1) - f_m(\delta_p) + f_m(\delta_1 + \delta_p),$$

$$n\, g'_m(\pi + b_n) = n\, g'_m(\pi) - f_m(\delta_p) + f_m(\delta_p - 1) + f_m(1),$$

$$n\, g'_m(\pi + b_0 + b_n) = n\, g'_m(\pi) - f_m(\delta_1) - f_m(\delta_p) + f_m(\delta_1 + 1) + f_m(\delta_p - 1).$$

Now $n\big[g'_m(\pi+b_0)+g'_m(\pi+b_n)-g'_m(\pi)-g'_m(\pi+b_0+b_n)\big] = \big[f_m(\delta_1+\delta_p)+f_m(1)\big] - \big[f_m(\delta_1+1)+f_m(\delta_p)\big]$. Let $k = \delta_1+\delta_p+1$. Since the function $f_m(x)+f_m(y)$ with $x+y=k$ is a symmetric and convex function, it follows from Proposition C2 of Chapter 3 in Marshall and Olkin [87], that $f_m(x) + f_m(y)$ is also Schur-convex. Since $(\delta_1 + 1, \delta_p) \prec (\delta_1 + \delta_p, 1)$, the quantity above is non-negative.

In the case that we use $w = b_0$ and $v \in \mathcal{V}$ such that v does not alter δ_1, then it follows that $g'_m(\pi + v + w) = g'_m(\pi + v) + g'_m(\pi + w) - g'_m(\pi)$. The same holds for $w = b_n$ and $v \in \mathcal{V}$ such that v does not alter δ_p. Suppose that v does alter δ_1, then we have $n\big[g'_m(\pi + b_0) + g'_m(\pi + v) - g'_m(\pi) - g'_m(\pi + b_0 + v)\big] = \big[f_m(\delta_1+\delta_p)+f_m(\delta_1-1)\big] - \big[f_m(\delta_1+\delta_p-1)+f_m(\delta_1)\big]$. When v alters δ_p we have $n\big[g'_m(\pi + b_n) + g'_m(\pi + v) - g'_m(\pi) - g'_m(\pi + b_n + v)\big] = \big[f_m(\delta_p + 1) + f_m(l)\big] - \big[f_m(\delta_p) + f_m(l + 1)\big]$ for some $l < \delta_p$. By applying the same argument as in the case of b_0 and b_n, we derive multimodularity of $g'_m(\pi)$ for the base $\mathcal{B}$. □

Now we will prove that $g(\pi)$, which is given by $g(\pi) = g'_1(\pi) + g'_2(\vec{3} - \pi)$ is multimodular. The proof is based on the fact that if a function is multimodular with respect to a base $\mathcal{B}$, then it is also multimodular with respect to $-\mathcal{B}$.

Theorem 6.10: Let g'_1 and g'_2 be multimodular functions. Then the function $g(\pi)$ given by $g(\pi) = c_1\, g'_1(\pi) + c_2\, g'_2(\vec{3} - \pi)$ for positive constants c_1 and c_2 is multimodular in π.

Proof: Let $v, w \in \mathcal{B}$, such that $v \neq w$. Then

$$\begin{aligned} &g(\pi+v) + g(\pi + w) \\ &\quad = c_1\, g'_1(\pi + v) + c_2\, g'_2(\vec{3} - \pi - v) + c_1\, g'_1(\pi + w) + c_2\, g'_2(\vec{3} - \pi - w) \\ &\quad = c_1\big[g'_1(\pi + v) + g'_1(\pi + w)\big] + c_2\big[g'_2(\vec{3} - \pi - v) + g'_2(\vec{3} - \pi - w)\big] \end{aligned}$$

$$\geq c_1\big[g_1'(\pi) + g_1'(\pi + v + w)\big] + c_2\big[g_2'(\vec{3} - \pi) + g_2'(\vec{3} - \pi - v - w)\big]$$
$$= c_1\, g_1'(\pi) + c_2\, g_2'(\vec{3} - \pi) + c_1\, g_1'(\pi + v + w) + c_2\, g_2'(\vec{3} - \pi - v - w)$$
$$= g(\pi) + g(\pi + v + w).$$

The inequality in the fourth line holds, since g_1' is multimodular with respect to $\mathcal{B}$, and g_2' is multimodular with respect to $-\mathcal{B}$. □

6.4 Examples of arrival processes

In today's information and communication systems the traffic pattern may be quite complex, as they may represent a variety of data, such as customer phone calls, compressed video frames, and other electronic information. Modern communication systems are designed to accommodate such a heterogeneous input. Therefore, the arrival process used in mathematical models is of crucial importance to the engineering and performance analysis of these systems. In this section we elaborate on the setting of the example at the end of Section 6.1 with different arrival processes. We derive explicit formulae for the cost function of the corresponding arrival processes.

Poisson Process

Assume that the decision maker wishes to minimize the number of lost customers (i.e., the number of preempted customers) per unit time; note that this is equivalent to maximizing the throughput of the system. Furthermore, let the services at server $i = 1, \ldots, M$ be exponentially distributed with rate μ_i, independent of the other servers. Since we know that there exists a periodic optimal policy (see Theorem 6.2), we can write the cost induced by using a period policy $\pi \in \Pi_{DS}$ by

$$g(\pi) = \frac{1}{n} \sum_{m=1}^{M} \sum_{j=1}^{d(m)} \left[\frac{\lambda}{\lambda + \mu_m}\right]^{\delta_j^m},$$

in case the arrival process is a Poisson process with parameter λ. In Koole [72] it was shown that the optimal policy has a period of the form $(1, 2, \ldots, 2)$, where 2 is the faster server. In Altman, Bhulai, Gaujal, and Hordijk [3] this result was generalized to general stationary arrival processes. Hence, suppose that $\lambda = 1$ and $\mu_1 = 1$, then the cost function can be parameterized by the period n and the server speed $\mu_2 \geq \mu_1$. This yields

$$g(n, \mu_2) = \frac{1}{n}\left(\frac{1}{2}\right)^n + \frac{1}{n}\left(\frac{1}{1 + \mu_2}\right)^2 + \frac{n-2}{n}\left(\frac{1}{1 + \mu_2}\right).$$

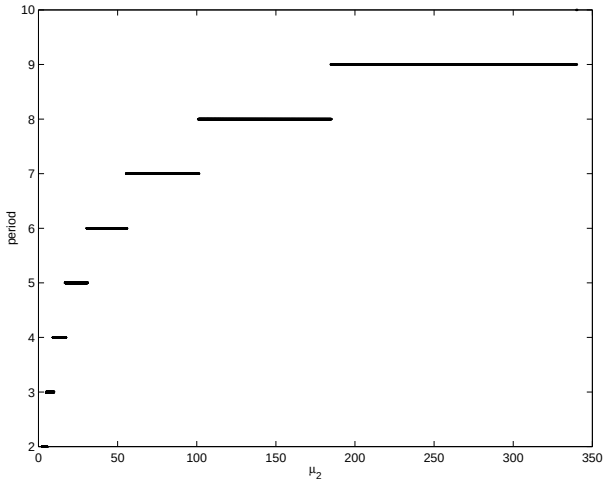

Figure 6.1: Relationship between n and μ_2.

By solving the equations $g(n, \mu_2) = g(n+1, \mu_2)$ for $n \geq 2$, we can compute the server rates μ_2 for which the optimal policy changes period. For example: the optimal policy changes from $(1, 2)$ to $(1, 2, 2)$ when $\mu_2 \geq 1 + \sqrt{2}$. The results of the computation are shown in Figure 6.1.

Markov Modulated Poisson Process

An important class of models for arrival processes is given by Markov modulated models. The key idea here is to use an explicit notion of states of an auxiliary Markov process into the description of the arrival process. The Markov process evolves as time passes and its current state modulates the probability law of the arrival process. The utility of such arrival processes is that they can capture bursty inputs.

The Markov Modulated Poisson Process (MMPP) is the most commonly used Markov modulated model. It is constructed by varying the arrival rate of a Poisson process according to an underlying continuous time Markov process, which is independent of the arrival process. Therefore, let $\{X_n \mid n \geq 0\}$ be a continuous time irreducible Markov process taking values in the set $\mathcal{X} = \{1, \ldots, k\}$ for some fixed $k \in \mathbb{N}$. When the Markov process is in state $i \in \mathcal{X}$, arrivals occur according to a Poisson process with rate λ_i. Let p_{ij} denote the transition probability to go from state $i \in \mathcal{X}$ to state $j \in \mathcal{X}$, and let Q be the infinitesimal generator of the Markov process. Let $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$ be the matrix with the arrival rates

on the diagonal, and let $\lambda = (\lambda_1, \ldots, \lambda_k)$ be the vector of arrival rates. With this notation, we can use the matrix analytic approach, described in Section 5.3 of Neuts [91], to derive a formula for $f_m(n)$.

Theorem 6.11: The sequence $\{(X_n, \tau_n) \,|\, n \geq 0\}$ is a Markov renewal sequence with transition probability matrix $F(t)$ given by

$$F(t) = \int_0^t e^{(Q-\Lambda)u}\,\mathrm{d}u\;\Lambda = \left[I - e^{(Q-\Lambda)t}\right](\Lambda - Q)^{-1}\Lambda.$$

The interpretation of the matrix $F(t)$ is as follows. The elements $F_{ij}(t)$ are given by the conditional probabilities $\mathbb{P}(X_{n+1} = j, \tau_{n+1} \leq t \,|\, X_n = i)$ for $n \geq 1$. Since $F(t)$ is a transition probability matrix, it follows that $F(\infty)$ given by $(\Lambda - Q)^{-1}\Lambda$ is a stochastic matrix.

The MMPP is fully parameterized by specifying the initial probability vector q, the infinitesimal generator Q of the Markov process, and the vector λ of arrival rates. Let the row vector s be the steady state vector of the Markov process. Then s satisfies the equations $sQ = 0$ and $se = 1$, where $e = (1, \ldots, 1)$. Define the row vector $q = s\Lambda/s\lambda$, then q is the stationary vector of $F(\infty)$ and makes the MMPP interval stationary (see Fisher and Meier-Hellstern [49]). This is intuitively clear since the stationary vector of $F(\infty)$ means that we obtain the MMPP started at an arbitrary arrival epoch.

In order to find an explicit expression for the cost function, we compute the Laplace-Stieltjes transform $f^*(\mu)$ of the matrix F. Since F is a matrix, we use matrix operations in order to derive $f^*(\mu)$, which will also be a matrix. The interpretation of the elements $f^*_{ij}(\mu)$ are given by $\mathbb{E}\left[e^{-\mu\tau_{n+1}}\,\mathbb{1}_{\{X_{n+1}=j\}} \,|\, X_n = i\right]$ for $n \geq 1$. Let I denote the identity matrix, then $f^*(\mu)$ is given by

$$f^*(\mu) = \int_0^\infty e^{-\mu It} F(\mathrm{d}t) = \int_0^\infty e^{-\mu It} e^{(Q-\Lambda)t}\,(\Lambda - Q)(\Lambda - Q)^{-1}\Lambda\,\mathrm{d}t$$

$$= \int_0^\infty e^{-(\mu I - Q + \Lambda)t}\,\mathrm{d}t\,\Lambda = (\mu I - Q + \Lambda)^{-1}\Lambda.$$

Now we can compute $f_m(n) = \mathbb{P}(S_m \geq \tau_1 + \cdots + \tau_n) = \mathbb{E}\exp[-\mu\sum_{k=1}^n \tau_k]$. The next lemma shows that this is simply given by the product of $f^*(\mu)$ with itself. Note that we do not need the assumption of independence of the interarrival times to derive this result.

Lemma 6.12: Let $f^*(\mu)$ be the Laplace-Stieltjes transform of F, with F a transition probability matrix of a stationary arrival process. Then,

$$\mathbb{E}\,\exp\left(-\mu\sum_{k=1}^{n}\tau_k\right) = q\left[f^*(\mu)\right]^n e.$$

Proof: Define a matrix Q_n with entries $Q_n(i,j)$ given by

$$Q_n(i,j) = \mathbb{E}\left[\exp\left(-\mu\sum_{k=1}^{n}\tau_k\right)\mathbb{1}_{\{X_n=j\}}\,\middle|\,X_0=i\right].$$

Note that Q_1 is given by $f^*(\mu)$. By using the stationarity of the arrival process it follows that $Q_n(i,j)$ is recursively defined by

$$\begin{aligned}
Q_n(i,j) &= \sum_{l=1}^{m} Q_{n-1}(i,l)\;\mathbb{E}\big[\exp(-\mu\tau_n)\,\mathbb{1}_{\{X_n=j\}}\,|\,X_{n-1}=l\,\big]\\
&= \sum_{l=1}^{m} Q_{n-1}(i,l)\;\mathbb{E}\big[\exp(-\mu\tau_1)\,\mathbb{1}_{\{X_1=j\}}\,|\,X_0=l\,\big]\\
&= \sum_{l=1}^{m} Q_{n-1}(i,l)\cdot Q_1(l,j).
\end{aligned}$$

Note that the last line exactly denotes the matrix product, thus $Q_n = Q_{n-1}\cdot Q_1$. By induction it follows that $Q_n = (Q_1)^n$. Thus, it follows that

$$\mathbb{E}\,\exp\left(-\mu\sum_{k=1}^{n}\tau_k\right) = \sum_{i=1}^{m}\sum_{j=1}^{m}\mathbb{P}(X_0=i)\,Q_n(i,j) = q\left[f^*(\mu)\right]^n e.$$

The last equation holds since the row vector q is the initial state of the Markov process and summing over all j is the same as right multiplying by e. □

Hence, $g(\pi)$ is given by

$$g(\pi) = \frac{1}{n}\sum_{m=1}^{M}\sum_{j=1}^{d(m)} q\big[(\mu_m I - Q + \Lambda)^{-1}\Lambda\big]^{\delta_j^m} e.$$

Note that although in case of two servers we know the structure of the optimal policy, it is not intuitively clear that it is optimal in the case of the MMPP. The

following argument will clarify this statement. Suppose that one has an MMPP with two states. Choose the rates λ_1 and λ_2 of the Poisson processes such that the policies would have period 2 and 3, respectively, if the MMPP is not allowed to change state. One could expect that if the transition probabilities to go to another state are very small, the optimal policy should be a mixture of both policies. But this is not the case.

Markovian Arrival Process

The Markovian arrival process model (MAP) is a broad subclass of models for arrival processes. It has the special property that every marked point process is the weak limit of a sequence of Markovian arrival processes (see Asmussen and Koole [10]). In practice, this means that very general point processes can be approximated by appropriate MAP's. The utility of the MAP follows from the fact that it is a versatile, yet tractable, family of models, which captures both bursty and regular inputs. The MAP can be described as follows.

Let $\{X_n \,|\, n \geq 0\}$ be a continuous time irreducible Markov process taking values in the set $\mathcal{X} = \{1, \ldots, k\}$ for some fixed $k \in \mathbb{N}$. Assume that the Markov process is in state $i \in \mathcal{X}$. The sojourn time in this state is exponentially distributed with parameter γ_i. After this time has elapsed, there are two transition possibilities. Either the Markov process moves to state $j \in \mathcal{X}$ with probability p_{ij} with generating an arrival, or the process moves to state $j \neq i$ with probability q_{ij} without generating an arrival.

This definition also gives rise to a natural description of the model in terms of matrix algebra. Define the matrix C with elements $C_{ij} = \gamma_i\, q_{ij}$ for $i \neq j$. Set the elements C_{ii} equal to $-\gamma_i$. Define the matrix D with elements $D_{ij} = \gamma_i\, p_{ij}$. The interpretation of these matrices is given as follows. The elementary probability that there is no arrival in an infinitesimal interval of length $\mathrm{d}t$ when the Markov process moves from state i to state j is given by $C_{ij}\,\mathrm{d}t$. A similar interpretation holds for D, but in this case it represents the elementary probability that an arrival occurs. The infinitesimal generator of the Markov process is then given by $C + D$. Note that the MMPP can be derived by choosing $C = Q - \Lambda$ and $D = \Lambda$.

In order to derive an explicit expression for the cost function, we use the same approach as in the case of the MMPP. The transition probability matrix $F(t)$ of the Markov renewal process $\{(X_n, \tau_n) \,|\, n \geq 0\}$ is of the form (see Lucantoni, Meier-Hellstern, and Neuts [83])

$$F(t) = \int_0^t e^{Cu}\,\mathrm{d}u\, D = \left[I - e^{Ct}\right]\left(-C^{-1}D\right).$$

Again, the elements of the matrix $F(t)$ have the interpretation that $F_{ij}(t)$ is given by $\mathbb{P}(X_{n+1} = j, \tau_{n+1} \le t \,|\, X_n = i)$ for $n \ge 1$. It also follows that $F(\infty)$ defined by $-C^{-1}D$ is a stochastic matrix. Let the row vector s be the steady state vector of the Markov process. Then s satisfies the equations $s(C+D) = 0$ and $se = 1$. Define the vector row vector $q = sD/sDe$, then q is the stationary vector of $F(\infty)$. This fact can be easily seen upon noting that $sD = s(C+D-C) = s(C+D) - sC = -sC$. With this observation it follows that $q\,F(\infty) = (sDe)^{-1}\,s\,CC^{-1}D = q$. The MAP defined by q, C and D has stationary interarrival times.

The Laplace-Stieltjes transform $f^*(\mu)$ of the matrix F is given by

$$f^*(\mu) = \int_0^\infty e^{-\mu It} F(\mathrm{d}t) = \int_0^\infty e^{-\mu It} e^{Ct} \big(-C\big)\big(-C^{-1}\big) D \,\mathrm{d}t$$

$$= \int_0^\infty e^{-(\mu I - C)t} \,\mathrm{d}t\, D = (\mu I - C)^{-1} D.$$

The interpretation of f^* is given by the elements $f^*_{ij}(\mu)$, which represent the expectation $\mathbb{E}\big[e^{\mu\tau_{n+1}}\,\mathbb{1}_{\{X_{n+1}=j\}} \,|\, X_n = i\big]$. By Lemma 6.12 we know that $f_m(n)$ is given by the product of f^*. Therefore the cost function, when using the MAP as arrival process, is given by

$$g(\pi) = \frac{1}{n} \sum_{m=1}^{M} \sum_{j=1}^{d(m)} q\big[(\mu_m I - C)^{-1} D\big]^{\delta_j^m} e.$$

6.5 Robot scheduling for web search engines

Altman, Gaujal, Hordijk, and Koole [7] specified a routing problem in which the expected average weighted loss rate was to be minimized (or equivalently, the average weighted throughput or average weighted number of customers at service was to be maximized). This gave rise to an immediate cost of the form

$$c(x, a) = \mathbb{E}\,\exp\left[-\mu_a \sum_{i=1}^{x_a} \tau_i\right].$$

Due to stationarity of the interarrival times, this cost function satisfies Condition (6.4) with $a_m = C = 0$. Evidently, we assume that not all τ_i are zero, which then implies the strict convexity of f_m. Indeed, denote

$$y = \exp\left[-\mu_a \sum_{k=2}^{m} \tau_k\right].$$

Let x be a state such that $x_a = m > 0$ for a particular action a. Since the interarrival times form a stationary sequence,

$$\begin{aligned} c(x,a) &= \mathbb{E}\, y e^{-\mu_a \tau_{m+1}} = \mathbb{E}\, y e^{-\mu_a \tau_1}, \\ c(x+e_a,a) &= \mathbb{E}\, y e^{-\mu_a [\tau_{m+1}+\tau_{m+2}]} = \mathbb{E}\, y e^{-\mu_a [\tau_1+\tau_{m+1}]}, \\ c(x-e_a,a) &= \mathbb{E}\, y. \end{aligned}$$

Since the function $r(x) = y e^{-\mu_a x}$ is convex in x, it follows that $r(\tau_1 + z) - r(z)$ is increasing in z, so that

$$r(\tau_1 + \tau_{m+1}) - r(\tau_{m+1}) \geq r(\tau_1) - r(0).$$

By taking expectations, this implies the convexity. Consequently, the results of the previous sections apply. In this section we present another application studied in Coffman, Liu, and Weber [34] under assumptions that are more restrictive than ours.

The World Wide Web offers search engines, such as Altavista, Google, Lycos, Infoseek, and Yahoo, that serve as a database that allow to search information on the web. These search engines often use robots that periodically traverse a part of the web structure in order to keep the database up-to-date.

We consider a problem where we assume that there is a fixed number of M web-pages. The contents of page i is modified at time instants that follow a Poisson process with parameter μ_i. The time a page is considered up-to-date by the search engine is the time since the last visit by the robot until the next time instant of modification; at this point the web-page is considered out-of-date until the next time it is visited by the robot. The times between updates by the robot are given by a sequence τ_n. In Coffman, Liu, and Weber [34] these times are assumed to be i.i.d., but in our framework we may allow them to form a general stationary sequence.

Let r_i denote the obsolescence rate of page i, i.e., the fraction of time that page i is out-of-date. Then the problem is to find an optimal visiting schedule such that the sum of the obsolescence rates r_i weighted by specified constants c_i is minimized. A reasonable choice for the weights c_i would be the customer page-access frequency, because the total cost then represents the customer total error rate. The case where the customer access frequency $c_i = k\,\mu_i$ is proportional to the page-change rate μ_i is reasonable under this interpretation, since the greater the interest for a particular page is, the more likely the frequency of page modification is.

We now show that this problem is equivalent to the problem studied in Section 6.1. Indeed, the robot can be considered as the decision maker in the previous problem. An assignment of the n-th customer to server i in the original problem corresponds with sending the robot to page i and requires τ_n time in order to complete the update. The lengths of the periods that page i is up-to-date correspond to the service times of customers before server i in the original problem. Page i is considered to be out-of-date when server i is idle.

Let S_i be an exponential random variable with parameter μ_i. The cost which one incurs when sending a customer to server a should reflect the expected obsolescence time. Some straightforward computations yield

$$c(x,a) = k\mu_a \,\mathbb{E}\left[\sum_{i=1}^{x_a} \tau_i - S_a\right]^+ = k\mu_a \,\mathbb{E}\left[\sum_{i=1}^{x_a} \tau_i - S_a\right] + k\mu_a \,\mathbb{E}\left[S_a - \sum_{i=1}^{x_a} \tau_i\right]^+$$

$$= k\mu_a x_a \,\mathbb{E}\tau_1 + k\left[\mathbb{E}\exp\left(-\mu_a \sum_{i=1}^{x_a} \tau_i\right) - 1\right].$$

This cost function clearly satisfies Condition (6.4). Hence, the theorems from the previous sections can indeed be applied to this scheduling problem.

The assumption that the weights c_i are proportional to the page-change rates μ_i is essential in this problem. The cost function for general c_i is given by

$$c(x,a) = c_a x_a \,\mathbb{E}\tau_1 + \frac{c_a}{\mu_a}\left[\mathbb{E}\exp\left(-\mu_a \sum_{i=1}^{x_a} \tau_i\right) - 1\right].$$

When the c_i are not proportional to the page-change rates, then the cost function is of the type mentioned in Example 6.3. Therefore, if for some i, c_i/μ_i is sufficiently large in comparison to the others, then it becomes optimal never to update page i. This is an undesirable situation, and it shows that the problem is not always well posed when the costs are not proportional to the page-change rates μ_i.

This problem of robot scheduling for web search engines illustrates the importance of including the terminal cost c_T in terms of modeling. Indeed, for any finite horizon T, if we wish that the cost represents indeed the obsolescence time of a page, we have to make sure that if this page is never updated (or at least it stops to be updated after some time), this will be reflected in the cost. It is easy to see that the terminal cost defined in Section 6.1 indeed takes care of that.

Chapter 7

Multi-Armed Bandit Problems

In this chapter we study multi-armed bandit problems. This class of problems is motivated by the design of clinical trials (see Bellman [15]). In a general setting the problem can be formulated as follows. Suppose that a decision maker periodically has to select a project to work on from a given set of projects. The work conducted on a particular project results in either a success or a failure with a fixed unknown probability. The decision maker receives a reward of one unit in case of a success, and zero otherwise. The problem in this setting is to choose a sequence of projects such that the long-term discounted rewards are maximized. Note that the problem indeed models clinical trials when the projects are seen as different treatments, and a reward is seen as a healed patient.

Multi-armed bandit problems are perhaps the simplest problems in the important class of Bayesian decision problems. It captures very well the possible tension between obtaining high direct rewards and acquiring information for better decision making in the future. The amount of information one has about the projects plays an important role in the selection of a project. Gittins and Wang [53] have investigated the relationship between the importance of acquiring additional information and the amount of information that is already available. Berry and Kertz [19] have studied the worth of perfect information for multi-armed bandits. However, both methods for quantifying the value of information are cumbersome in practice, since the computational complexity is too large.

Based on Bhulai and Koole [24], we shall discuss the relationship between acquiring additional information about the unknown parameters and the immediate costs. This is done by considering two extreme situations when a fixed number of information samples have been acquired: the situation where the decision maker stops learning, and the situation where the decision maker obtains full information. We show that the difference in rewards between this lower and upper bound goes to zero as the number of information samples grows large.

7.1 Problem formulation

In this section we apply the results of Sections 5.2 and 5.4 to multi-armed bandit problems. The multi-armed bandit problem is a classical problem in Bayesian adaptive control introduced by Bellman [15]. It is perhaps the simplest problem in the important class of Bayesian adaptive control problems that captures very well the tension between acquiring information and incurring low costs. The importance of this model follows from its many applications in the areas of stochastic scheduling, searching, planning of research, and design of experiments (see, e.g., Berry and Fristedt [18], Gittins [52], and Kumar [77]). A version of the multi-armed bandit problem, stated in terms of rewards instead of costs, is defined as follows.

Suppose that there are M independent projects to work on. At every epoch $t = 1, 2, \ldots$ one of the projects must be selected. The work conducted on a particular project can result in either a success or a failure with a fixed unknown probability. When the conducted work is successful the project yields a reward of one unit, and zero otherwise. The resulting sequence of successes and failures forms a Bernoulli process with an unknown parameter. The problem in this setting is to choose a project at each epoch such that the discounted rewards are maximized.

There are two reasons for selecting a particular project to work on. The first reason is to obtain a high reward. The second is to acquire information which can be used to determine whether subsequent selections of the same project are profitable in the future. The possible contradiction between these two reasons for choosing a project makes the problem difficult and mathematically interesting. The amount of information one has about the projects plays an important role. It helps to answer the question whether one should take a less rewarding but more informative action over one that is more rewarding but less informative.

Gittins and Wang [53] have investigated the relationship between the importance of acquiring additional information and the amount of information that is already available. They have quantified a learning component in terms of dynamic allocation indices and have shown that this component is a decreasing function in the number of selections.

Berry and Kertz [19] have studied the worth of perfect information for multi-armed bandits. They have defined an information comparison region in order to compare the reward of the decision maker with the reward of the decision maker who has perfect information. Relations between these comparisons and the concept of regret in the minimax approach to bandit processes were established.

The methods for quantifying the value of information studied in Berry and

Kertz [19], and Gittins and Wang [53] are cumbersome in practice, since the computational complexity is too large. In the next section we adopt a direct approach. We consider two extreme situations, which occur when a project has been selected N times: the situation where the decision maker stops learning, and the situation where the decision maker acquires full information about the project. We express the difference in rewards between this lower and upper bound in N, and show that it goes to zero as N grows large. Let us first start with the mathematical formulation of the problem.

Since the M projects are independent of each other, the unknown parameter takes values in the set $\Theta = [0,1]^M$. From Theorems 5.4 and 5.8 it follows that we can take the family of Beta distributions for each project independently as a conjugate family. From Theorem 5.8 it also follows that the parameters of the posterior distributions count the number of successes and failures for each project observed during the course of the process. Therefore, let the state space $\mathcal{X} = (\mathbb{N}_0 \times \mathbb{N}_0)^M$ represent the observed number of successes and failures for all projects. Thus, state $x = (r_1, s_1, \ldots, r_M, s_M) \in \mathcal{X}$ denotes that r_i successes and s_i failures have been observed for project $i = 1, \ldots, M$. Suppose that the initial state is $x_0 = (0, \ldots, 0) \in \mathcal{X}$; this corresponds to the situation where no prior knowledge about any of the projects is available. In this case, when project i is in state (r_i, s_i), the corresponding parameters of its Beta distribution are given by $r_i + 1$ and $s_i + 1$. Hence, the probability density function of the Beta distribution, given by Equation (5.11), can be written as

$$f_{(r_i+1,s_i+1)}(z) = \frac{(r_i + s_i + 1)!}{r_i!\, s_i!}\, z^{r_i}\, (1-z)^{s_i},$$

for $r_i, s_i \in \mathbb{N}_0$ and $z \in [0,1]$.

Let $\mathcal{A}_x = \mathcal{A} = \{1, \ldots, M\}$ for $x \in \mathcal{X}$ denote the action space, where action $a \in \mathcal{A}$ represents selecting project a to work on. The transition probabilities defined by Equation (5.5) are given by

$$p\big(x' \,\big|\, x, a\big) = \begin{cases} \frac{r_a+1}{r_a+s_a+2}, & \text{for } x' = x + e_{2a-1} \\ \frac{s_a+1}{r_a+s_a+2}, & \text{for } x' = x + e_{2a} \\ 0, & \text{otherwise,} \end{cases}$$

where e_i is the $2M$-dimensional unit vector with all entries zero except for the i-th, which is one. Given state $x \in \mathcal{X}$ and action $a \in \mathcal{A}$ the expected direct reward

is determined by $F_{(r_a+1,s_a+1)}$. From Equation (5.6) it follows that the expected direct reward is given by

$$r(x,a) = \tfrac{r_a+1}{r_a+s_a+2}.$$

The tuple $(\mathcal{X}, \{\mathcal{A}_x \mid x \in \mathcal{X}\}, p, r)$ determines the Markov decision process for the multi-armed bandit problem. The description of the Markov decision problem is now completed by describing the criterion function. For that purpose, let $\alpha \in (0,1)$ be the discount factor, and $\pi \in \Pi_R$ be a fixed policy. Then the discounted reward criterion function $V^\pi(x_0)$ is defined by

$$V^\pi(x_0) = \mathbb{E}^\pi_{x_0} \sum_{t=0}^{\infty} \alpha^t \, r(X_t, A_t).$$

The Markov decision problem is to find a policy $\pi^* \in \Pi_R$ such that $V(x_0) = V^{\pi^*}(x_0) = \sup\{V^\pi(x_0) \mid \pi \in \Pi_R\}$. Since the rewards are bounded by 1, it follows by Theorem 2.2 that there exists an optimal deterministic stationary policy. Moreover, V^{π^*} is the unique solution to the optimality equations.

7.2 The value of information

The Markov decision problem of the previous section satisfies the following optimality equation

$$V(x) = \max_{i=1,\ldots,M} \left\{ \tfrac{r_i+1}{r_i+s_i+2} \Big[1 + \alpha V(x + e_{2i-1}) \Big] + \tfrac{s_i+1}{r_i+s_i+2} \, \alpha V(x + e_{2i}) \right\},$$

where $V(x)$ denotes the optimal discounted reward starting from state $x \in \mathcal{X}$ satisfying $x = (r_1, s_1, \ldots, r_i, s_i, \ldots, r_M, s_M)$. For ease of notation we denote the expression between the brackets as $T_i\, V(x)$. Even though the dynamic programming equation can be explicitly written down, it is difficult to obtain closed form solutions or computational results because of the large state space.

The optimality equation shows that the decision maker not only receives a direct reward in selecting a project, but also gains information that can lead to better decisions in future. When action a is chosen, project a either has a posterior distribution $F_{(r_a+2,s_a+1)}$ or $F_{(r_a+1,s_a+2)}$, depending on whether a success or failure is observed. Since a random variable with such a distribution has lower variance than a random variable with probability distribution $F_{(r_a+1,s_a+1)}$, the decision maker is better informed. A formal proof of this statement is given in Lemma 7.1.

One could argue that when a particular project has been selected N times, where N can be large, enough information about the project has been obtained.

Therefore, basing future decisions only on this information for this project should not result in a great difference in the obtained discounted reward. In that case, the decision maker does not need to keep record of changes in the state for this project anymore. This means that the decision maker stops learning about the unknown parameter of this particular project.

To compare this situation with the original situation, we formulate the new situation again as a Markov decision problem. Since the decision maker stops learning about a project when it has been selected N times, the corresponding state is frozen. This results in the finite state space $\underline{\mathcal{X}} = \{(r,s) \in \mathbb{N}_0 \times \mathbb{N}_0 \,|\, r+s \leq N\}^M$. Take $\underline{\mathcal{A}} = \mathcal{A}$, and change the transition probabilities as follows.

$$\underline{p}\big(x' \,\big|\, x, a\big) = \begin{cases} \frac{r_a+1}{r_a+s_a+2}, & \text{for } r_a + s_a < N \text{ and } x' = x + e_{2a-1} \\ \frac{s_a+1}{r_a+s_a+2}, & \text{for } r_a + s_a < N \text{ and } x' = x + e_{2a} \\ 1, & \text{for } r_a + s_a = N \text{ and } x' = x \\ 0, & \text{otherwise.} \end{cases}$$

Finally, define $\underline{r} = r$, then $(\underline{\mathcal{X}}, \underline{\mathcal{A}}, \underline{p}, \underline{r})$ defines the Markov decision process in case the decision maker stops learning about project $i = 1, \ldots, M$ when the information for this project has been accumulated by N samples. The appropriate modification to the optimality equation becomes

$$\underline{V}(x) = \max_{i=1,\ldots,M} \Big\{ \mathbb{1}_{\{r_i+s_i=N\}} \Big[\tfrac{r_i+1}{r_i+s_i+2} + \alpha \underline{V}(x) \Big] + \mathbb{1}_{\{r_i+s_i<N\}} \, T_i \, \underline{V}(x) \Big\}.$$

Note that in a situation where it is optimal to select an action that does not change state, that action will remain optimal. Therefore, $\underline{V}(x)$ can be rewritten as follows

$$\underline{V}(x) = \max_{i=1,\ldots,M} \big\{ \mathbb{1}_{\{r_i+s_i=N\}} \, \tfrac{1}{1-\alpha} \, \tfrac{r_i+1}{r_i+s_i+2} + \mathbb{1}_{\{r_i+s_i<N\}} \, T_i \, \underline{V}(x) \big\}.$$

Intuitively, it is clear that if the decision maker does not use new information for future selections anymore, then the discounted reward will be less than $V(x)$, where this information is taken into account. The next lemma formalizes this statement.

Lemma 7.1: $\underline{V}(x) \leq V(x)$ for all $x \in \underline{\mathcal{X}}$.

Proof: Define the operator T for functions $W : \mathcal{X} \to \mathbb{R}$ as

$$T\,W(x) = \max_{k=1,\ldots,M} \Big\{ \tfrac{r_k+1}{r_k+s_k+2} \Big[1 + \alpha W(x + e_{2k-1}) \Big] + \tfrac{s_k+1}{r_k+s_k+2} \, \alpha W(x + e_{2k}) \Big\}.$$

Define $V_0(x) = 0$ for all $x \in \mathcal{X}$ and $V_n(x) = T\,V_{n-1}(x)$. From Section 2.3 we know that the operator T is a contraction mapping on the Banach space of all bounded real valued functions on $\mathcal{X}$ endowed with the supremum norm. Therefore, $V(x)$ is the unique solution to $T\,W(x) = W(x)$ and $V(x) = \lim_{n\to\infty} V_n(x)$ for arbitrary V_0.

We have to prove that a state with more information is more rewarding, therefore it suffices to prove that

$$V(x) \leq \tfrac{r_i+1}{r_i+s_i+2}\,V(x+e_{2i-1}) + \tfrac{s_i+1}{r_i+s_i+2}\,V(x+e_{2i}),$$

for all $i = 1, \ldots, M$. We prove this relation by induction on n for the functions V_n. Clearly, this relation holds for V_0. Fix i and suppose that the relation holds for $n \in \mathbb{N}$. Assume that the maximizing action for $V_{n+1}(x)$ is a. Since $V_{n+1}(x) = T\,V_n(x)$ we derive

$$\begin{aligned} V_{n+1}(x) &= \max_{k=1,\ldots,M} \left\{ \tfrac{r_k+1}{r_k+s_k+2}\Big[1 + \alpha V_n(x+e_{2k-1})\Big] + \tfrac{s_k+1}{r_k+s_k+2}\,\alpha V_n(x+e_{2k}) \right\} \\ &= \tfrac{r_a+1}{r_a+s_a+2}\Big[1 + \alpha V_n(x+e_{2a-1})\Big] + \tfrac{s_a+1}{r_a+s_a+2}\,\alpha V_n(x+e_{2a}). \end{aligned}$$

First suppose that $i \neq a$, then by applying the induction hypothesis we derive that the latter expression is less or equal than

$$\begin{aligned} &\tfrac{r_a+1}{r_a+s_a+2} + \alpha \tfrac{r_a+1}{r_a+s_a+2}\Big[\tfrac{r_i+1}{r_i+s_i+2} V_n(x+e_{2a-1}+e_{2i-1}) + \tfrac{s_i+1}{r_i+s_i+2} V_n(x+e_{2a-1}+e_{2i})\Big] + \\ &\qquad \alpha\,\tfrac{s_a+1}{r_a+s_a+2}\Big[\tfrac{r_i+1}{r_i+s_i+2} V_n(x+e_{2a}+e_{2i-1}) + \tfrac{s_i+1}{r_i+s_i+2} V_n(x+e_{2a}+e_{2i})\Big]. \end{aligned}$$

By rearranging terms we find that this expression is equal to

$$\begin{aligned} &\tfrac{r_i+1}{r_i+s_i+2}\Big[\tfrac{r_a+1}{r_a+s_a+2} + \alpha\,\tfrac{r_a+1}{r_a+s_a+2} V_n(x+e_{2i-1}+e_{2a-1}) + \alpha\,\tfrac{s_a+1}{r_a+s_a+2} V_n(x+e_{2i-1}+e_{2a})\Big] + \\ &\tfrac{s_i+1}{r_i+s_i+2}\Big[\tfrac{r_a+1}{r_a+s_a+2} + \alpha\,\tfrac{r_a+1}{r_a+s_a+2} V_n(x+e_{2i}+e_{2a-1}) + \alpha\,\tfrac{s_a+1}{r_a+s_a+2} V_n(x+e_{2i}+e_{2a})\Big]. \end{aligned}$$

By definition of T_a, the latter expression is equal to

$$\begin{aligned} &\tfrac{r_i+1}{r_i+s_i+2}\,T_a V_n(x+e_{2i-1}) + \tfrac{s_i+1}{r_i+s_i+2}\,T_a V_n(x+e_{2i}) \\ &\qquad \leq \max_{k=1,\ldots,M} \left\{ \tfrac{r_i+1}{r_i+s_i+2}\,T_k V_n(x+e_{2i-1}) + \tfrac{s_i+1}{r_i+s_i+2}\,T_k V_n(x+e_{2i}) \right\}. \end{aligned}$$

The argument for $i = a$ is completely analogous. Hence, we have

$$\begin{aligned} V_{n+1}(x) &= \max_{k=1,\ldots,M} \left\{ \tfrac{r_k+1}{r_k+s_k+2} \Big[1 + \alpha V_n(x + e_{2k-1}) \Big] + \tfrac{s_k+1}{r_k+s_k+2}\, \alpha V_n(x + e_{2k}) \right\} \\ &\leq \max_{k=1,\ldots,M} \left\{ \tfrac{r_i+1}{r_i+s_i+2}\, T_k V_n(x + e_{2i-1}) + \tfrac{s_i+1}{r_i+s_i+2}\, T_k V_n(x + e_{2i}) \right\} \\ &\leq \tfrac{r_i+1}{r_i+s_i+2} \max_{k=1,\ldots,M} \big\{ T_k V_n(x+e_{2i-1}) \big\} + \tfrac{s_i+1}{r_i+s_i+2} \max_{k=1,\ldots,M} \big\{ T_k V_n(x+e_{2i}) \big\} \\ &= \tfrac{r_i+1}{r_i+s_i+2} V_{n+1}(x + e_{2i-1}) + \tfrac{s_i+1}{r_i+s_i+2}\, V_{n+1}(x + e_{2i}). \end{aligned}$$

The proof is concluded by taking the limit in n. □

At this point we know that $\underline{V}(x) \leq V(x)$, and our main interest is to compare $\underline{V}(x)$ with $V(x)$, i.e., to give an upper bound to $V(x) - \underline{V}(x)$. However, it is not straightforward to carry out this computation. Therefore, we define a Markov decision process, which has a discounted reward higher than $V(x)$, such that it will facilitate the comparison.

Suppose that when a particular project has been selected N times, the decision maker obtains full information about the unknown parameter of that project. Hence, the decision maker does not need to learn anything about the unknown parameter and can base his future decisions on the realization of the unknown parameter.

The state space of this process is equal to $\underline{\mathcal{X}}$ when none of the projects have been selected N times. When a project has been selected N times, the true value of its unknown parameter becomes known. The state space of this process is complicated to represent precisely. For ease of notation, we define the state space slightly bigger by $\overline{\mathcal{X}} = \big\{ \overline{x} \mid \overline{x} = (x, z) \in \underline{\mathcal{X}} \times [0,1]^M \big\}$. In this case the state $\overline{x}$ consists of x with augmented extra information z. The value of z_i represents the realization of the unknown parameter of project i. Note that this information can only be used after a project has been selected N times.

Let $\overline{V}(\overline{x})$ denote the optimal discounted reward starting in state $\overline{x} \in \overline{\mathcal{X}}$. Note that the set of states which V and $\overline{V}$ share is given by $\underline{\mathcal{X}}$. For states $x \in \underline{\mathcal{X}}$ the extra information z is not available and can be disregarded. Therefore, we will write $\overline{V}(x)$ instead of $\overline{V}\big((x, z)\big)$ for all $x \in \underline{\mathcal{X}}$ and $z \in [0,1]$. Intuitively, it is clear that if the decision maker has full information, then the discounted reward will be greater than $V(x)$ for all $x \in \underline{\mathcal{X}}$. The following lemma justifies this intuition.

Lemma 7.2: $V(x) \leq \overline{V}(x)$ for all $x \in \underline{\mathcal{X}}$.

Proof: Let $x \in \underline{\mathcal{X}}$ and fix $i \in \mathcal{A}$. Define $\overline{V}_n^i(x)$ inductively by

$$\overline{V}_n^i(x) = \tfrac{r_i+1}{r_i+s_i+2}\,\overline{V}_{n-1}^i(x+e_{2i-1}) + \tfrac{s_i+1}{r_i+s_i+2}\,\overline{V}_{n-1}^i(x+e_{2i}),$$

with $\overline{V}_0^i(x) = V(x)$. Note that $\overline{V}_n^i(x)$ represents the situation where the decision maker has already acquired more information about project i. The decision maker looks n steps ahead and, consequently, knows more about project i. In Lemma 7.1 it was proven that $V(x) \leq \frac{r_i+1}{r_i+s_i+2}V(x+e_{2i-1}) + \frac{s_i+1}{r_i+s_i+2}\,V(x+e_{2i})$ for all $i \in \mathcal{A}$. By repeating this argument we derive $0 \leq \overline{V}_0^i(x) \leq \overline{V}_1^i(x) \leq \cdots$. Now define $\overline{V}_n(x)$ as follows

$$\overline{V}_n(x) = \max_{i=1,\ldots,M} \Big\{ \mathbb{1}_{\{r_i+s_i=N\}} \overline{V}_n^i(x) + \mathbb{1}_{\{r_i+s_i<N\}}\, T_i\, V(x) \Big\}.$$

It follows that $V(x) \leq \overline{V}_n(x)$ for all $n \in \mathbb{N}$. Note that $\overline{V}(x) = \lim_{n\to\infty} \overline{V}_n(x)$ is the case where the decision maker has full information. By the Monotone Convergence Theorem it finally follows that $V(x) \leq \overline{V}(x)$. □

Comparing $\overline{V}(x)$ with $\underline{V}(x)$ for a given $x \in \underline{\mathcal{X}}$ is still not easy, since both processes differ in the amount of information when a project has already been selected N times. In case of $\overline{V}(x)$ the decision maker knows $\overline{\mathcal{X}}$, which includes extra information z. However, in case of $\underline{V}(x)$ the decision maker has less information, because he only knows $\underline{\mathcal{X}}$. Therefore, for a given a policy $\overline{\pi} \in \overline{\Pi}$, with decision rules based on z, one cannot compare $\overline{V}^{\overline{\pi}}(x) - \underline{V}^{\overline{\pi}}(x)$, since the latter term is not well defined.

The value of the realization z is determined by the probability distribution of the unknown parameter. If we adjust this probability distribution, then we will be able to carry out the comparison. Let $\overline{\overline{F}}_{(r+1,s+1)}$ denote the probability distribution with positive probability mass concentrated on only two points as follows. Let $\beta = (r+1)/(r+s+2)$, and choose $0 < \delta < 1/(N+2)$, then $\beta+\delta < 1$. Define the probability mass function $\overline{\overline{f}}$ by

$$\overline{\overline{f}}_{(r+1,s+1)}(z) = \begin{cases} \int_0^{\beta+\delta} f_{(r+1,s+1)}(u)\,\mathrm{d}u & \text{for } z = \beta+\delta \\ \int_{\beta+\delta}^1 f_{(r+1,s+1)}(u)\,\mathrm{d}u & \text{for } z = 1. \end{cases}$$

The Markov decision process, determined by $(\overline{\overline{\mathcal{X}}}, \overline{\overline{\mathcal{A}}}, \overline{\overline{p}}, \overline{\overline{r}})$, can now be defined as follows. Let $\overline{\overline{\mathcal{X}}} = \overline{\mathcal{X}}$. Define the action space $\overline{\overline{\mathcal{A}}} = \mathcal{A}$, and the transition

probabilities by $\overline{\overline{p}}(\overline{\overline{x}}' \mid \overline{\overline{x}}, a) =$

$$\begin{cases} \frac{r_a+1}{r_a+s_a+2}, & \text{for } r_a+s_a < N-1 \text{ and } x' = x + e_{2a-1} \\ \frac{s_a+1}{r_a+s_a+2}, & \text{for } r_a+s_a < N-1 \text{ and } x' = x + e_{2a} \\ \frac{r_a+1}{r_a+s_a+2}\,\overline{\overline{f}}_{(r_a+2,s_a+1)}(u) & \text{for } r_a+s_a = N-1, x' = x + e_{2a-1} \text{ and } z_a' = u \\ \frac{s_a+1}{r_a+s_a+2}\,\overline{\overline{f}}_{(r_a+1,s_a+2)}(u) & \text{for } r_a+s_a = N-1, x' = x + e_{2a} \text{ and } z_a' = u \\ 1, & \text{for } r_a+s_a = N \text{ and } \overline{x}' = \overline{x} \\ 0, & \text{otherwise,} \end{cases}$$

where $u \in [0,1]$. Finally, define the direct reward by

$$\overline{\overline{r}}\big((x,z),a\big) = \begin{cases} \frac{r_a+1}{r_a+s_a+2}, & \text{for } r_a+s_a < N \\ z_a, & \text{for } r_a+s_a = N. \end{cases}$$

Note that the difference between $\overline{V}$ and $\overline{\overline{V}}$ is reflected in the transition probabilities. In the former case, one had to deal with a continuous probability distribution. In the latter case, the probability distribution is discrete and is concentrated on two specific points only. If Z_1 and Z_2 are two random variables with probability distribution $F_{(r,s)}$ and $\overline{\overline{F}}_{(r,s)}$, respectively, then Z_2 is stochastically larger than Z_1; i.e., $\mathbb{P}(Z_1 > z) \leq \mathbb{P}(Z_2 > z)$ for all $z \in [0,1]$. From Proposition 8.1.2 in Ross [100] we know that in this case $\mathbb{E}[h(Z_1)] \leq \mathbb{E}[h(Z_2)]$ for all increasing functions h. Therefore, we have the following corollary.

Corollary 7.3: $0 \leq \underline{V}(x) \leq V(x) \leq \overline{V}(x) \leq \overline{\overline{V}}(x)$ for all $x \in \underline{\mathcal{X}}$.

The process with discounted reward $\overline{\overline{V}}(x)$ is constructed in such a way, that the information structure when a project has been selected N times is nearly the same as in $\underline{V}(x)$. The decision maker either observes $\beta + \delta$ or 1 as the realization of the unknown parameter. In the first case, the decision maker has the same information as in $\underline{V}(x)$, namely the expectation. In the other case, we know that the decision maker is going to select that project continuously in future, since 1 is the highest possible reward. This fact enables us to prove the main theorem.

Theorem 7.4: $0 \leq V(x) - \underline{V}(x) \leq \max_{i \in \mathcal{A}} \frac{\alpha^{N-(r_i+s_i)}}{1-\alpha} \left[\delta + \frac{l(x)}{\delta^2\,(N+3)}\right]$ for all $x \in \underline{\mathcal{X}}$ and $\delta < \frac{1}{N+2}$, where $l(x) = \sum_{i=1}^{M} \mathbb{1}_{\{r_i+s_i<N\}}$.

Proof: Because of Lemma 7.1, the difference $V(x) - \underline{V}(x)$ is non-negative. By Corollary 7.3 we know that $V(x) - \underline{V}(x) \leq \overline{\overline{V}}(x) - \underline{V}(x)$. Therefore, it suffices to prove the bound for the latter term. We adopt the same approach as in the proof of Lemma 7.1. Define the operator $\underline{T}$ for functions $W : \underline{\mathcal{X}} \to \mathbb{R}$ as

$$\underline{T}\,W(x) = \max_{i=1,\,\ldots,M} \left\{ \mathbb{1}_{\{r_i+s_i=N\}} \left[\tfrac{1}{1-\alpha} \, \tfrac{r_i+1}{r_i+s_i+2} \right] + \mathbb{1}_{\{r_i+s_i<N\}} \, T_i \, W(x) \right\}.$$

Define $\underline{V}_0(x) = 0$ for all $x \in \underline{\mathcal{X}}$, and $\underline{V}_n(x) = \underline{T}\,\underline{V}_{n-1}(x)$ for $n \in \mathbb{N}$. By Section 2.3 we know that the operator $\underline{T}$ is a contraction mapping on the Banach space of all bounded real valued functions on $\underline{\mathcal{X}}$ endowed with the supremum norm. Therefore, $\underline{V}(x)$ is the unique solution to $\underline{T}\,W(x) = W(x)$, and $\underline{V}(x) = \lim_{n\to\infty} \underline{V}_n(x)$ for arbitrary $\underline{V}_0$. Similarly, the operator $\overline{\overline{T}}$ is defined for functions $W : \overline{\overline{\mathcal{X}}} \to \mathbb{R}$ as

$$\begin{aligned}
\overline{\overline{T}}\,W(x) = \max_{i=1,\,\ldots,M} \Bigg\{ & \mathbb{1}_{\{r_i+s_i=N\}} \tfrac{1}{1-\alpha} \left(\tfrac{r_i+1}{r_i+s_i+2} + \delta \right) + \\
& \mathbb{1}_{\{r_i+s_i=N-1\}} \Bigg[\tfrac{r_i+1}{r_i+s_i+2} + \alpha \tfrac{r_i+1}{r_i+s_i+2} \left(q_{(r_i+1,s_i)} W(x+e_{2i-1}) + \left(1 - q_{(r_i+1,s_i)}\right) \tfrac{1}{1-\alpha} \right) + \\
& \alpha \tfrac{s_i+1}{r_i+s_i+2} \left(q_{(r_i,s_i+1)} W(x+e_{2i}) + \left(1 - q_{(r_i,s_i+1)}\right) \tfrac{1}{1-\alpha} \right) \Bigg] + \mathbb{1}_{\{r_i+s_i<N-1\}} \, T_i \, W(x) \Bigg\},
\end{aligned}$$

where $q_{(r_i,s_i)} = \int_0^{\beta+\delta} f_{(r_i+1,s_i+1)}(z)\,\mathrm{d}z$ with $\beta = (r_i+1)/(r_i+s_i+2)$. Note that this equation represents the situation where the decision maker receives the realization of the unknown parameter (under the modified probability distribution) after selecting a project $N-1$ times. When it is optimal to select this project again after this moment, then it will be optimal to select it continuously thereafter, since the state does not change.

Now we prove the statement by induction. Let $x \in \underline{\mathcal{X}}$, then clearly the statement holds for $\overline{\overline{V}}_0(x) - \underline{V}_0(x) = 0$. Suppose that the statement holds for $n \in \mathbb{N}$. Assume without loss of generality that the first m projects have reached level N, the second m' projects level $N-1$, and suppose that the remaining $M-m-m'$ projects have not reached level $N-1$ yet for arbitrary $m \in \{0,\,\ldots,M\}$ and $m' \in \{0,\,\ldots,M-m\}$. Now assume that it is optimal to choose one of the first m projects. Then, for $i = 1,\,\ldots,m$ the difference is less than

$$\left[\tfrac{1}{1-\alpha} \left(\tfrac{r_i+1}{r_i+s_i+2} + \delta \right) \right] - \left[\tfrac{1}{1-\alpha} \, \tfrac{r_i+1}{r_i+s_i+2} \right] = \tfrac{\delta\,\alpha^{N-(r_i+s_i)}}{1-\alpha} \leq \max_{k\in\mathcal{A}} \tfrac{\alpha^{N-(r_k+s_k)}}{1-\alpha} \left[\delta + \tfrac{l(x)}{\delta^2\,(N+3)} \right].$$

Next, consider the second m' projects. First note that for a random variable U

with a Beta distribution with parameters $(r_i + 1, s_i + 1)$ the following holds.

$$\begin{aligned} 1 - q_{(r_i,s_i)} &= \mathbb{P}\big(U > \beta + \delta\big) \leq \mathbb{P}\big(\{U < \beta - \delta\} \cup \{U > \beta + \delta\}\big) \\ &= \mathbb{P}\big(\{U - \beta < -\delta\} \cup \{U - \beta > \delta\}\big) = \mathbb{P}\big(|U - \beta| > \delta\big) \\ &\leq \tfrac{\mathbb{V}\mathrm{ar}\, U}{\delta^2} \leq \tfrac{1}{\delta^2\,(r_i+s_i+3)}. \end{aligned}$$

The last inequality follows by Chebyshev's Inequality. The difference for $j = m + 1, \ldots, m + m'$ is given by

$$\begin{aligned} &\Big[\tfrac{r_j+1}{r_j+s_j+2} + \alpha\,\tfrac{r_j+1}{r_j+s_j+2}\Big(q_{(r_j+1,s_j)}\overline{\overline{V}}_n(x + e_{2j-1}) + \big(1 - q_{(r_j+1,s_j)}\big)\tfrac{1}{1-\alpha}\Big) + \\ &\alpha\,\tfrac{s_j+1}{r_j+s_j+2}\Big(q_{(r_j,s_j+1)}\overline{\overline{V}}_n(x + e_{2j}) + \big(1 - q_{(r_j,s_j+1)}\big)\tfrac{1}{1-\alpha}\Big)\Big] - T_j\,\underline{V}_n(x) \\ \leq\ &\alpha\,\tfrac{r_j+1}{r_j+s_j+2}\Big(q_{(r_j+1,s_j)}\overline{\overline{V}}_n(x + e_{2j-1}) + \big(1 - q_{(r_j+1,s_j)}\big)\tfrac{1}{1-\alpha} - \underline{V}_n(x + e_{2j-1})\Big) + \\ &\alpha\,\tfrac{s_j+1}{r_j+s_j+2}\Big(q_{(r_j,s_j+1)}\overline{\overline{V}}_n(x + e_{2j}) + \big(1 - q_{(r_j,s_j+1)}\big)\tfrac{1}{1-\alpha} - \underline{V}_n(x + e_{2j})\Big) \\ \leq\ &\alpha\,\tfrac{r_j+1}{r_j+s_j+2}\Big(\Big[\overline{\overline{V}}_n(x + e_{2j-1}) - \underline{V}_n(x + e_{2j-1})\Big] + \tfrac{1}{\delta^2\,(N+3)}\,\tfrac{1}{1-\alpha}\Big) + \\ &\alpha\,\tfrac{s_j+1}{r_j+s_j+2}\Big(\Big[\overline{\overline{V}}_n(x + e_{2j}) - \underline{V}_n(x + e_{2j})\Big] + \tfrac{1}{\delta^2\,(N+3)}\,\tfrac{1}{1-\alpha}\Big). \end{aligned}$$

Note that $l(x + e_{2j-1}) = l(x + e_{2j}) = l(x) - 1$. By applying the induction hypothesis we derive that the last expression is less than

$$\begin{aligned} &\tfrac{\alpha}{1-\alpha}\,\Big(\max\Big\{\alpha^{N-(r_i+s_i)}; i \in \mathcal{A} \setminus \{j\},\, \alpha^{N-(r_j+s_j+1)}\Big\}\Big[\delta + \tfrac{l(x)-1}{\delta^2\,(N+3)}\Big] + \tfrac{1}{\delta^2\,(N+3)}\Big) \\ \leq\ &\max_{i\in\mathcal{A}}\,\tfrac{\alpha^{N-(r_i+s_i)}}{1-\alpha}\,\Big[\delta + \tfrac{l(x)}{\delta^2\,(N+3)}\Big]. \end{aligned}$$

Finally, consider the last $M - m - m'$ projects. Note that for $k = m + m', \ldots, M$ we have $l(x) = l(x + e_{2k-1}) = l(x + e_{2k})$. Therefore, $T_k\,\big[\overline{\overline{V}}_n(x) - \underline{V}_n(x)\big]$ is given

by

$$\begin{aligned}
&\tfrac{r_k+1}{r_k+s_k+2}\,\alpha\Big(\overline{\overline{V}}_n(x+e_{2k-1}) - \underline{V}_n(x+e_{2k-1})\Big) + \tfrac{s_k+1}{r_k+s_k+2}\,\alpha\Big(\overline{\overline{V}}_n(x+e_{2k}) - \underline{V}_n(x+e_{2k})\Big) \\
&\leq \tfrac{\alpha}{1-\alpha}\,\max\Big\{\alpha^{N-(r_i+s_i)}; i\in\mathcal{A}\setminus\{j\},\, \alpha^{N-(r_j+s_j+1)}\Big\}\Big[\delta + \tfrac{1}{\delta^2\,(N+3)}\Big] \\
&\leq \max_{i\in\mathcal{A}} \tfrac{\alpha^{N-(r_i+s_i)}}{1-\alpha}\,\Big[\delta + \tfrac{1}{\delta^2\,(N+3)}\Big].
\end{aligned}$$

It follows that $\overline{\overline{V}}_{n+1}(x) - \underline{V}_{n+1}(x)$ satisfies the statement of the theorem. The proof is concluded by taking the limit in n. □

The bounds in the previous theorem still contain $\delta > 0$. Since δ was arbitrarily chosen, we can minimize the bound for fixed N with respect to δ. This results in a bound independent of δ, and the result is stated in the following theorem.

Theorem 7.5: For all $x \in \underline{\mathcal{X}}$, we have

$$0 \leq V(x) - \underline{V}(x) \leq \max_{i\in\mathcal{A}} \tfrac{\alpha^{N-(r_i+s_i)}}{1-\alpha}\,\tfrac{\sqrt[3]{2}+\sqrt[3]{\frac{1}{4}}}{\sqrt[3]{N+3}}\,\sqrt[3]{l(x)}.$$

Proof: Let N be fixed, and define $g(\delta) = \max_{i\in\mathcal{A}} \tfrac{\alpha^{N-(r_i+s_i)}}{1-\alpha}\,\Big[\delta + \tfrac{l(x)}{\delta^2\,(N+3)}\Big]$. Then,

$$\hat{\delta} = \sqrt[3]{\tfrac{2\,l(x)}{N+3}} \quad \text{solves} \quad \frac{\mathrm{d}\,g(\delta)}{\mathrm{d}\,\delta} = \max_{i\in\mathcal{A}} \tfrac{\alpha^{N-(r_i+s_i)}}{1-\alpha}\,\Big[1 - \tfrac{2\,l(x)}{\delta^3\,(N+3)}\Big] = 0.$$

Hence $\hat{\delta}$ minimizes g. Although Theorem 7.4 is formulated for $\delta < 1/(N+2)$, the theorem holds for general $\delta > 0$. When $\overline{\overline{f}}$ is defined to be degenerate at 1 when $\beta + \delta \geq 1$, one can easily check that Theorem 7.4 still holds. The proof is concluded by substituting $\hat{\delta}$ in g, since $0 \leq V(x) - \underline{V}(x) \leq g(\hat{\delta})$. □

Observe that the bound in Theorem 7.5 has the property that the difference goes to zero as N grows large. However, this is not due to discounting, since

$$V(x) - \underline{V}(x) \leq \max_{i\in\mathcal{A}} \tfrac{\alpha^{N-(r_i+s_i)}}{1-\alpha}\,\tfrac{\sqrt[3]{2}+\sqrt[3]{\frac{1}{4}}}{\sqrt[3]{N+3}}\,\sqrt[3]{l(s)} \leq \tfrac{1}{1-\alpha}\,\tfrac{\sqrt[3]{2}+\sqrt[3]{\frac{1}{4}}}{\sqrt[3]{N+3}}\,\sqrt[3]{M}.$$

The latter bound, which is less tight, also goes to zero as N grows large even without the discount factor. Moreover, observe that this bound also holds for any state. One can even show that the bound for the reward at any epoch t is $\big[(\sqrt[3]{2} + \sqrt[3]{\frac{1}{4}})\sqrt[3]{M}\,\big]/\sqrt[3]{N+3}$.

7.3 Numerical results

In this section we illustrate Theorem 7.5 derived in Section 7.2. We show that in practice the state space can indeed be chosen finite in order to be close to the optimal solution.

Let the initial state $x_0 = (0, \ldots, 0)$ be given. Suppose that the decision maker wants to obtain a solution that differs less than $\varepsilon = 10^{-3}$ from the optimal solution. We call such a solution an ε-optimal solution. Note that the initial state x_0 represents the situation where the decision maker does not have any information about the unknown parameters of the projects.

By using the bounds derived in Theorem 7.5 one can determine the value of N for which the decision maker can stop learning about the unknown parameter of a particular project. However, since the total reward will increase for α close to one, the value of N will grow large. Therefore, it is better to look at the relative difference $\frac{V(x)-\underline{V}(x)}{V(x)}$. This leads to the following table when M and α are varied.

α	$M = 2$	$M = 3$
0.7	21	21
0.8	32	33
0.9	66	66

In practice it suffices to take smaller values of N. Figures 7.1 and 7.2 depict two situations with two and three projects, respectively. The three lines from top to down reflect the cases with $\alpha = 0.90$, $\alpha = 0.80$, and $\alpha = 0.70$, respectively, for $\underline{V}(x_0)$. The dashed lines represent the bounds on the error of the discounted reward obtained by using $\underline{V}(x_0)$ instead of $V(x_0)$; thus the dashed lines represent the upper bound $V(x_0) + \big[\overline{V}(x_0) - \underline{V}(x)\big]$.

One can see that the discounted reward converges very fast already for small values of N. It turns out that for $\alpha = 0.90$ one can take $N = 28$ instead of $N = 108$ in order to derive an ε-optimal discounted reward. The following table summarizes the values of N which can be taken instead of the larger values which one derives from Theorem 7.5.

α	$M = 2$	$M = 3$
0.7	9	10
0.8	13	14
0.9	28	33

These values of N are small enough to make the problem computationally tractable, and to derive an ε-optimal solution.

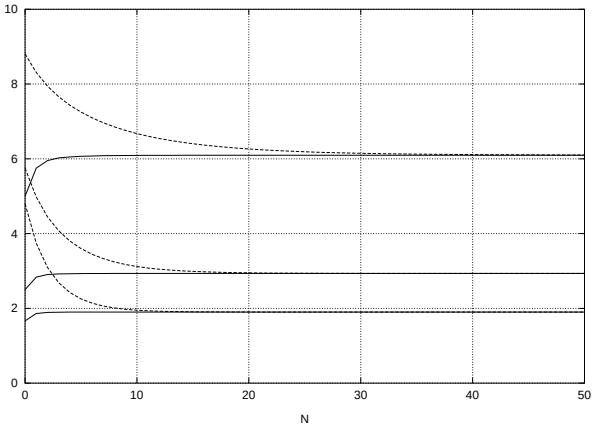

Figure 7.1: Comparison of bounds for $M = 2$.

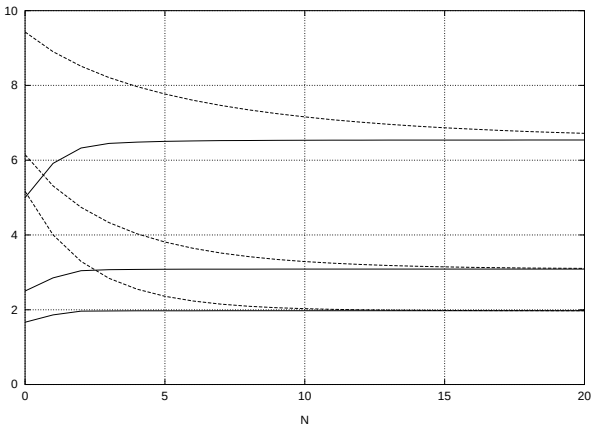

Figure 7.2: Comparison of bounds for $M = 3$.

Notation

a_t	The action chosen at epoch t
A_t	Random variable denoting the action chosen at epoch t
$\mathcal{A}$	The action space; $\mathcal{A} = \cup_{x \in \mathcal{X}} \mathcal{A}_x$
$\mathcal{A}_x$	The set of feasible actions when the system is in state $x \in \mathcal{X}$
$\mathbb{B}(\mathcal{X})$	The Banach space of bounded real-valued functions on $\mathcal{X}$
$\mathbb{B}_w(\mathcal{X})$	The Banach space of w-bounded real-valued functions on $\mathcal{X}$
c	The cost function in Markov (decision) processes
$\mathbb{E}^{\pi}_{x_0}$	The expectation with respect to $\mathbb{P}^{\pi}_{x_0}$
$F^{(n)}_{ij}$	The probability $({}_M P^{n-1} \cdot P)_{ij}$ with $M = \{j\}$
F_{ij}	The probability of reaching state j from state i; $F_{ij} = \sum_{n \in \mathbb{N}} F^{(n)}_{ij}$
F_t	The distribution of the unknown parameter at epoch t
$\mathcal{F}$	The σ-algebra of Borel subsets of a given space
$g(\pi, x)$	The expected average cost under policy π when starting in state x
$g^*(x)$	The optimal long-run expected average cost; $g^*(x) = \inf_{\pi \in \Pi_R} g(\pi, x)$
h_t	The history up to epoch t
H_t	Random variable denoting the history up to epoch t
$\mathcal{H}_t$	The set of admissible histories up to epoch t
$\mathcal{K}$	The set of feasible state-action pairs
M	A finite subset of $\mathcal{X}$
$\mathbb{N}$	The set of positive numbers $\{1, 2, 3, \ldots\}$
$\mathbb{N}_0$	The set of non-negative numbers $\{0, 1, 2, \ldots\}$
p	The transition law
P	The matrix of transition probabilities of a Markov chain
P^*	The stationary probability matrix; $P^*_{ij} = \lim_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n-1} P^t_{ij}$
${}_M P$	The taboo transition matrix; ${}_M P_{ij} = \mathbb{1}_{\{j \notin M\}} P_{ij}$
$\mathbb{P}^{\pi}_{x_0}$	Probability measure under policy π with respect to the initial state x_0

$\mathbb{R}$	The set of real numbers
$V(x)$	The average cost value function corresponding to g^*
$V(\pi, x)$	The α-discount value function under policy π when starting in state x
$V^*(x)$	The optimal α-discount value function; $V^*(x) = \inf_{\pi \in \Pi_R} V(\pi, x)$
w	A weight function, i.e., a function $w : \mathcal{X} \to [1, \infty)$
$\mathcal{W}$	A set of probability measures on a given space
x_t	The state of the system at epoch t
X_t	Random variable denoting the state of the system at epoch t
$\mathcal{X}$	The state space of a Markov (decision) process
$\mathcal{Y}$	The observation space
$\mathbb{Z}$	The set of integers

$\mathbb{1}$	The indicator function
α	The discount factor
Δ	The backward difference operator; $\Delta V(x) = V(x) - V(x-1)$
κ	The number of closed classes in a Markov chain
ν	Mean duration between state transitions
ω	A sample path from Ω
Ω	The sample space
φ_t	Decision rule for epoch t
π	A control policy (a sequence of decision rules); $\pi = \{\varphi_t\}_{t \in \mathbb{N}_0}$
Π_K	The set of all policies having property $K \in \{R, RM, RS, D, DM, DS\}$
$\propto$	Proportionality; $p^\theta(x) \propto q^\theta(x)$ if $\exists\, f(x) : p^\theta(x) = f(x)\, q^\theta(x)$
Ψ_t	The distribution of the state the process occupies at epoch t
$\sigma(\mathcal{W})$	The Borel σ-algebra of $\mathcal{W}$
Θ	The parameter space

Bibliography

[1] I.J.B.F. Adan. *A Compensation Approach for Queueing Problems*. CWI, Amsterdam, 1994. CWI Tract 104.

[2] E. Altman. Applications of Markov decision processes in communication networks: A survey. In E.A. Feinberg and A. Shwartz, editors, *Handbook of Markov Decision Processes*. Kluwer, 2002.

[3] E. Altman, S. Bhulai, B. Gaujal, and A. Hordijk. Optimal routing problems and multimodularity. Technical Report RR–3727, INRIA, 1999.

[4] E. Altman, S. Bhulai, B. Gaujal, and A. Hordijk. Open-loop routing to M parallel servers with no buffers. *Journal of Applied Probability*, 37:668–684, 2000.

[5] E. Altman, B. Gaujal, and A. Hordijk. Balanced sequences and optimal routing. Technical Report RR–3180, INRIA, 1997.

[6] E. Altman, B. Gaujal, and A. Hordijk. Multimodularity, convexity, and optimization properties. Technical Report TW–97–07, Leiden University, 1997.

[7] E. Altman, B. Gaujal, A. Hordijk, and G.M. Koole. Optimal admission, routing and service assignment control: The case of single buffer queues. In *IEEE Conference on Decision and Control*, December 1998.

[8] M. Aoki. *Optimization of Stochastic Systems*. Academic Press, 1967.

[9] A. Arapostathis, V.S. Borkar, E. Fernández-Gaucherand, M.K. Ghosh, and S.I. Marcus. Discrete-time controlled Markov processes with average cost criterion: A survey. *SIAM Journal of Control and Optimization*, 31:282–344, 1993.

[10] S. Asmussen and G.M. Koole. Marked point processes as limits of Markovian arrival streams. *Journal of Applied Probability*, 30:365–372, 1993.

[11] K.S. Azoury. Bayes solution to dynamic inventory models under unknown demand distribution. *Management Science*, 31:1150–1160, 1985.

[12] K.S. Azoury and B.L. Miller. A comparison of the optimal ordering levels of Bayesian and non-Bayesian inventory models. *Management Science*, 30:993–1003, 1984.

[13] A.G. Barto, R.S. Sutton, and C.W. Anderson. Neuron-like elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:835–846, 1983.

[14] M. Bartroli and S. Stidham, Jr. Towards a unified theory of structure of optimal policies for control of network of queues. Technical report, University of North Carolina, 1987.

[15] R. Bellman. A problem in the sequential design of experiments. *Sankhya*, 16:221–229, 1956.

[16] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[17] R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[18] A. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, 1985.

[19] D.A. Berry and R.P. Kertz. Worth of perfect information in Bernoulli bandits. *Advances in Applied Probability*, 23:1–23, 1991.

[20] D.P. Bertsekas. *Dynamic Programming*. Prentice-Hall, 1987.

[21] D.P. Bertsekas. *Dynamic Programming and Optimal Control, Volume One*. Athena Scientific, 1995.

[22] D.P. Bertsekas. *Dynamic Programming and Optimal Control, Volume Two*. Athena Scientific, 1995.

[23] D.P. Bertsekas and J.N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[24] S. Bhulai and G.M. Koole. On the value of learning for Bernoulli bandits with unknown parameters. *IEEE Transactions on Automatic Control*, 45:2135–2140, 2000.

[25] S. Bhulai and G.M. Koole. Scheduling time-constrained jobs in the presence of background traffic. In *Proceedings of the 39th IEEE Conference on Decision and Control*, pages 1421–1426, 2000.

[26] S. Bhulai and G.M. Koole. On the structure of value functions for threshold policies in queueing models. Technical Report 2001-4, Vrije Universiteit Amsterdam, 2001 (submitted).

[27] S. Bhulai and F.M. Spieksma. On the uniqueness of solutions to the Poisson equations for average cost Markov chains with unbounded cost functions. Technical Report 2001-7, Vrije Universiteit Amsterdam, 2001 (submitted).

[28] D. Blackwell. Discrete dynamic programming. *Annals of Mathematical Statistics*, 33:719–726, 1962.

[29] M. Bousquet-Mélou and M. Petkovšek. Linear recurrences with constant coefficients: The multivariate case. *Discrete Mathematics*, 225:51–75, 2000.

[30] O.J. Boxma and J.W. Cohen. *Boundary Value Problems in Queueing System Analysis*. North-Holland, Amsterdam, 1983.

[31] J. Carlstrom and E. Nordstrom. Control of self-similar ATM call traffic by reinforcement learning. In *Proceedings of the International Workshop on Applications of Neural Networks to Telecommunication*, pages 54–62, 1997.

[32] R-R. Chen and S. Meyn. Value iteration and optimization of multiclass queueing networks. *Queueing Systems*, 32:65–97, 1999.

[33] K.L. Chung. *Markov Chains with Stationary Transition Probabilities*. Springer-Verlag, 1967.

[34] E.G. Coffman, Jr., Z. Liu, and R.R. Weber. Optimal robot scheduling for web search engines. *Journal of Scheduling*, 1:15–29, 1998.

[35] M.B. Combé and O.J. Boxma. Optimization of static traffic allocation policies. *Theoretical Computer Science*, 125:17–43, 1994.

[36] M.H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, 1970.

[37] R. Dekker. *Denumerable Markov Decision Chains: Optimal Policies for Small Interest Rates*. PhD thesis, Leiden University, 1985.

[38] R. Dekker and A. Hordijk. Average, sensitive, and Blackwell optimality in denumerable state Markov decision chains with unbounded rewards. *Mathematics of Operations Research*, 13:395–421, 1989.

[39] R. Dekker and A. Hordijk. Recurrence conditions for average and Blackwell optimality in denumerable state Markov decision chains. *Mathematics of Operations Research*, 17:271–290, 1992.

[40] R. Dekker, A. Hordijk, and F.M. Spieksma. On the relation between recurrence and ergodicity properties in denumerable Markov decision chains. *Mathematics of Operations Research*, 19:1–21, 1994.

[41] C. Derman. Denumerable state Markov decision processes: Average cost criterion. *Annals of Mathematical Statistics*, 37:1545–1553, 1966.

[42] C. Derman and A.F. Veinott, Jr. A solution to a countable system of equations arising in Markovian decision processes. *Annals of Mathematical Statistics*, 38:582–584, 1967.

[43] A. Drake. *Observation of a Markov Process Through a Noisy Channel*. PhD thesis, Massachusetts Institute of Technology, 1962.

[44] E.B. Dynkin. Controlled random sequences. *Theory of Probability*, 10:1–14, 1965.

[45] E.B. Dynkin and A.A. Yushkevich. *Controlled Markov Processes*. Springer-Verlag, 1979.

[46] A. Federgruen, A. Hordijk, and H.C. Tijms. Recurrent conditions in denumerable state Markov decision processes. In M.L. Puterman, editor, *Dynamic Programming and its Applications*. Academic Press, 1978.

[47] A. Federgruen, A. Hordijk, and H.C. Tijms. Denumerable state semi-Markov decision processes with unbounded costs. *Stochastic Processes and Applications*, 9:223–235, 1979.

[48] A. Federgruen, P.J. Schweitzer, and H.C. Tijms. Contraction mappings underlying undiscounted Markov decision problems. *Journal of Mathematical Analysis and Application*, 65:711–730, 1978.

[49] W. Fisher and K.S. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18:149–171, 1992.

[50] L.S. Freedman, D.J. Spiegelharter, and M.K.B. Parmar. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society*, 157:357–416, 1994.

[51] F.D. Gakhov. *Boundary Value Problems*. Dover Publications, 1990.

[52] J.C. Gittins. *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, 1989.

[53] J.C. Gittins and Y. Wang. The learning component of dynamic allocation indices. *Annals of Statistics*, 20:1625–1636, 1992.

[54] K. Goldberg, M. Newman, and E. Haynsworth. Combinatorial analysis. In M. Abramowitz and I.A. Stegun, editors, *Handbook of Mathematical Functions*. Dover Publications, 1972.

[55] R. Groenevelt, G.M. Koole, and P. Nain. On the bias vector of a two-class preemptive priority queue. *ZOR - Mathematical Methods of Operations Research*, to appear, 2001.

[56] D. Gross and C.M. Harris. *Fundamentals of Queueing Theory*. John Wiley & Sons, 1985.

[57] B. Hajek. Extremal splitting of point processes. *Mathematics of Operations Research*, 10:543–556, 1985.

[58] K.M. van Hee. *Bayesian Control of Markov Chains*. PhD thesis, Technical University of Eindhoven, 1978.

[59] K.M. van Hee. Markov decision processes with unknown transition law: The average return case. In R. Hartley, L.C. Thomas, and D.J. White, editors, *Recent Developments in Markov Decision Processes*. Academic Press, 1980.

[60] O. Hernández-Lerma and J.B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer-Verlag, 1996.

[61] O. Hernández-Lerma and J.B. Lasserre. *Further Topics on Discrete-Time Markov Control Processes*. Springer-Verlag, 1999.

[62] K. Hinderer. *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameters*, volume 33 of *Lecture Notes Operations Research*. Springer-Verlag, 1970.

[63] A. Hordijk. *Dynamic Programming and Markov Potential Theory*. Mathematisch Centrum, Amsterdam, 1974. Math. Centre Tract 51.

[64] A. Hordijk and F.M. Spieksma. On ergodicity and recurrence properties of a Markov chain with an application to an open Jackson network. *Advances in Applied Probability*, 24:343–376, 1992.

[65] R. Howard. *Dynamic Programming and Markov Decision Processes*. MIT Press, 1960.

[66] E. Hyytiä and J. Virtamo. Dynamic routing and wavelength assignment using first policy iteration. Technical Report COST 257, Helsinki University of Technology, 2000.

[67] D.L. Iglehart. The dynamic inventory problem with unknown demand distribution. *Management Science*, 10:429–440, 1964.

[68] N.K. Jaiswal. *Priority Queues*. Academic Press, 1968.

[69] S. Karlin. Dynamic inventory policy with varying stochastic demands. *Management Science*, 6:231–258, 1960.

[70] F.P. Kelly. Multi-armed bandits with discount factor near one: The Bernoulli case. *Annals of Statistics*, 9:987–1001, 1981.

[71] G.M. Koole. A transformation method for stochastic control problems with partial observations. *Systems and Control Letters*, 35:301–308, 1998.

[72] G.M. Koole. On the static assignment to parallel servers. *IEEE Transactions on Automatic Control*, 44:1588–1592, 1999.

[73] G.M. Koole and P. Nain. On the value function of a priority queue with an application to a controlled polling model. *Queueing Systems*, 34:199–214, 2000.

[74] G.M. Koole and P. Nain. An explicit solution for the value function of a priority queue. Technical Report 2001-5, Vrije Universiteit Amsterdam, 2001 (submitted).

[75] G.M. Koole and F. Spieksma. On deviation matrices for birth-death processes. *Probability in the Engineering and Informational Sciences*, 15:239–258, 2001.

[76] G.M. Koole and E. van der Sluis. An optimal local search procedure for manpower scheduling in call centers. Technical report, Vrije Universiteit Amsterdam, 1998.

[77] P.R. Kumar. A survey of some results in stochastic adaptive control. *SIAM Journal of Control and Optimization*, 23:329–380, 1985.

[78] P.R. Kumar and T.I. Seidman. On the optimal solution of the one-armed bandit adaptive control problem. *IEEE Transactions on Automatic Control*, 26:1176–1184, 1981.

[79] P.R. Kumar and P. Varaiya. *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Prentice-Hall, 1986.

[80] S.A. Lippman. Semi-Markov decision processes with unbounded rewards. *Management Science*, 19:717–731, 1973.

[81] S.A. Lippman. Applying a new device in the optimization of exponential queueing systems. *Operations Research*, 23:687–710, 1975.

[82] S.A. Lippman. On dynamic programming with unbounded rewards. *Management Science*, 21:1225–1233, 1975.

[83] D.M. Lucantoni, K.S. Meier-Hellstern, and M.F. Neuts. A single-server queue with server vacations and a class of non-renewal arrival processes. *Advances in Applied Probability*, 22:675–705, 1990.

[84] A. Maitra. *Dynamic Programming for Countable State Systems*. PhD thesis, University of California, 1964.

[85] A. Manne. Linear programming and sequential decisions. *Management Science*, 6:259–267, 1960.

[86] P. Marbach, O. Mihatsch, and J.N. Tsitsiklis. Call admission control and routing in integrated service networks using neuro-dynamic programming. *IEEE Journal on Selected Areas in Communications*, 18:197–208, 2000.

[87] A.W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and its Applications*. Academic Press, 1979.

[88] S. Meyn. The policy improvement algorithm for Markov decision processes with general state space. *IEEE Transactions on Automatic Control*, 42:191–196, 1997.

[89] R.E. Mickens. *Difference Equations: Theory and Applications*. Chapman & Hall, 1990.

[90] G.E. Monahan. A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, 28:1–16, 1982.

[91] M.F. Neuts. *Structured Stochastic Matrices of M/G/1 Type and their Applications*. Marcel Dekker, 1989.

[92] E. Nordstrom and J. Carlstrom. A reinforcement learning scheme for adaptive link allocation in ATM networks. In *Proceedings of the International Workshop on Applications of Neural Networks to Telecommunication*, pages 88–95, 1995.

[93] J.M. Norman. *Heuristic Procedures in Dynamic Programming*. Manchester University Press, 1972.

[94] T.J. Ott and K.R. Krishnan. Separable routing: A scheme for state-dependent routing of circuit switched telephone traffic. *Annals of Operations Research*, 35:43–68, 1992.

[95] C.H. Papadimitriou and J.N. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12:441–450, 1987.

[96] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

[97] D. Rhenius. Incomplete information in Markovian decision models. *Annals of Statistics*, 2:1327–1334, 1974.

[98] U. Rieder. Bayesian dynamic programming. *Advances in Applied Probability*, 7:330–348, 1975.

[99] Z. Rosberg and D. Towsley. Customer routing to parallel servers with different rates. *IEEE Transactions on Automatic Control*, 30:1140–1143, 1985.

[100] S.M. Ross. *Stochastic Processes*. John Wiley & Sons, 1983.

[101] S.M. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, 1993.

[102] H. Rummukainen and J. Virtamo. Polynomial cost approximations in Markov decision theory based least cost routing. *IEEE/ACM Transactions on Networking*, to appear.

[103] S.A.E. Sassen, H.C. Tijms, and R.D. Nobel. A heuristic rule for routing customers to parallel servers. *Statistica Neerlandica*, 51:107–121, 1997.

[104] Y. Sawaragi and T. Yoshikawa. Discrete-time Markovian decision processes with incomplete state observation. *Annals of Mathematical Statistics*, 41:78–86, 1970.

[105] H.E. Scarf. Bayes solutions of the statistical inventory problem. *Annals of Mathematical Statistics*, 30:490–508, 1959.

[106] H.E. Scarf. Some remarks on Bayes solutions to the inventory problem. *Naval Research Logistics Quarterly*, 7:591–596, 1960.

[107] L. Sennott. A new condition for the existence of optimal stationary policies in average cost Markov decision processes. *Operations Research Letters*, 5:17–23, 1986.

[108] L. Sennott. A new condition for the existence of optimum stationary policies in average cost MDPs – unbounded cost case. In *Proceedings of the 25th IEEE Conference on Decision and Control*, pages 1719–1721, 1986.

[109] L. Sennott. Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs. *Operations Research*, 37:626–633, 1989.

[110] L. Sennott. Average cost semi-Markov decision processes and the control of queueing systems. *Probability in the Engineering and Informational Sciences*, 3:247–272, 1989.

[111] L. Sennott. *Stochastic Dynamic Programming and the Control of Queueing Systems*. John Wiley & Sons, 1999.

[112] R.F. Serfozo. An equivalence between continuous and discrete time Markov decision processes. *Operations Research*, 27:616–620, 1979.

[113] A.N. Shiryaev. On the theory of decision functions and control of an observation process with incomplete data. In *Transactions of the Third Prague Conference on Information Theory*, pages 657–682, 1964.

[114] R. Smallwood and E.J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21:1071–1088, 1973.

[115] E.J. Sondik. *The Optimal Control of Partially Observable Markov Processes*. PhD thesis, Stanford University, 1971.

[116] E.J. Sondik. The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Operations Research*, 26:282–304, 1978.

[117] F.M. Spieksma. *Geometrically Ergodic Markov Chains and the Optimal Control of Queues*. PhD thesis, Leiden University, 1990.

[118] S. Stidham, Jr. and R.R. Weber. A survey of Markov decision models for control of networks of queues. *Queueing Systems*, 13:291–314, 1993.

[119] L.W.G. Strijbosch and J.J.A. Moors. Inventory control: The impact of unknown demand distribution. Technical Report FEW 770, Tilburg University, 1999.

[120] L.C. Thomas. Connectedness conditions for denumerable state Markov decision processes. In R. Hartley, L.C. Thomas, and D.F. White, editors, *Recent Developments in Markov Decision Processes*. Academic Press, 1980.

[121] H.C. Tijms. *Stochastic Models: An Algorithmic Approach*. John Wiley & Sons, 1994.

[122] H. Tong and T.X. Brown. Adaptive call admission control under quality of service constraints: A reinforcement learning solution. *IEEE Journal on Selected Areas in Communications*, 18:209–221, 2000.

[123] D. Towsley, R.H. Hwang, and J.F. Kurose. MDP routing for multi-rate loss networks. *Computer Networks and ISDN*, 34:241–261, 2000.

[124] A. Wald. *Sequential Analysis*. John Wiley, 1947.

[125] A. Wald. *Statistical Decision Functions*. John Wiley, 1950.

[126] C.J.C.H. Watkins. *Learning from delayed rewards*. PhD thesis, Cambridge University, 1989.

[127] D.J. White. Dynamic programming of Markov chains and the method of successive approximations. *Journal of Mathematical Analysis and Application*, 6:373–376, 1963.

[128] C.C. White III. A survey of solution techniques for the partially observed Markov decision process. *Annals of Operations Research*, 32:215–230, 1991.

Samenvatting

Markov beslissingsprocessen: het sturen van hoog-dimensionale systemen
Veel bedrijfsproblemen kenmerken zich door een regelmatig terugkerend beslissingsvraagstuk. Het herhalende karakter van het vraagstuk maakt de besluitvorming er niet noodzakelijkerwijs makkelijker op. Dit komt doordat beslissingen consequenties hebben op zowel korte als lange termijn. Het korte termijn-effect bestaat voornamelijk uit de kosten die gemoeid zijn met het uitvoeren van de beslissing tot aan het volgende beslismoment. Het effect op de lange termijn uit zich in potentieel andere omstandigheden op het volgende beslismoment; een beslissing, die op dit moment lage kosten met zich meebrengt, kan immers leiden tot ongunstige situaties waarin alleen maar beslissingen die met hoge kosten gepaard gaan, genomen kunnen worden. Om de kosten te minimaliseren kunnen beslissingen dus niet in isolatie bestudeerd worden, maar moeten de relaties met toekomstige omstandigheden ook in overweging genomen worden.

Een Markov beslissingsproces (Markov decision process) is een wiskundig formalisme om het bovengeschetste probleem te beschrijven en te bestuderen. De term 'Markov' geeft aan dat het gedrag van het systeem in de toekomst slechts afhankelijk is van de huidige toestand en niet van het verleden. Deze eigenschap lijkt de toepasbaarheid te beperken, maar veel praktisch relevante problemen voldoen aan deze eigenschap. Hierbij valt onder andere te denken aan het bepalen van de optimale bestelgrootte in het voorraadbeheer, het maken van werkroosters zodanig dat er voldoende capaciteit aanwezig is om aan een bepaalde servicegraad te voldoen, en het routeren in communicatienetwerken zodanig dat wachttijden minimaal gehouden worden. Het direct oplossen van dergelijke problemen met behulp van Markov beslissingstheorie is niet zonder meer mogelijk, omdat de ruimte van oplossingen hoog-dimensionaal is. Dat houdt in dat het berekenen van optimale beslisregels binnen een redelijke tijd niet haalbaar is.

In dit proefschrift richten we ons op het ontwikkelen van algoritmen voor

het bepalen van (bijna) optimale beslisregels in hoog-dimensionale systemen. In Hoofdstuk 1 starten we met een literatuuroverzicht van het onderwerp van onderzoek. Vervolgens bespreken we in Hoofdstuk 2 de formele theorie van Markov beslissingsprocessen. Hierbij wordt extra nadruk gelegd op de rol van de waardefunctie binnen deze theorie. De waardefunctie speelt een centrale rol in de één-staps strategieverbetering en de neuro-dynamische programmering; dit zijn twee technieken om optimale beslisregels te benaderen. In Hoofdstuk 3 wordt een systematische analyse van de waardefunctie voor geboorte-sterfte processen gegeven. De bekende wachtrijsystemen, zoals de multi-server, de single server en de infinite server queue, zijn speciale gevallen hiervan. In Hoofdstuk 4 wordt gekeken naar meer ingewikkelde wachtrijsystemen, zoals de priority queue en de tandem queue.

In Hoofdstuk 5 bestuderen we Markov beslissingsprocessen waarbij de beslisser geconfronteerd wordt met extra onzekerheid, naast de onzekerheid over de ontwikkeling van de toekomst. Deze onzekerheden kunnen onder andere ontstaan doordat de beslisser de toestand van het systeem niet geheel kan observeren, of doordat de overgangen tussen de verschillende toestanden van het systeem niet geheel gespecificeerd kunnen worden. We laten zien dat het laatste een speciaal geval is van het eerste, en dat het probleem omgeschreven kan worden als een standaard Markov beslissingsproces. Echter, de herformulering zorgt wederom voor een hoog-dimensionaal systeem. Hoewel het aantal toestandconfiguraties heel erg groot is, kan er slechts een beperkt aantal daarvan aangenomen worden. Hierdoor is het mogelijk het systeem in dimensionaliteit te reduceren en te analyseren. Deze technieken worden in Hoofdstuk 6 en 7 geïllustreerd aan de hand van routeringsproblemen en multi-armed bandits.

Index