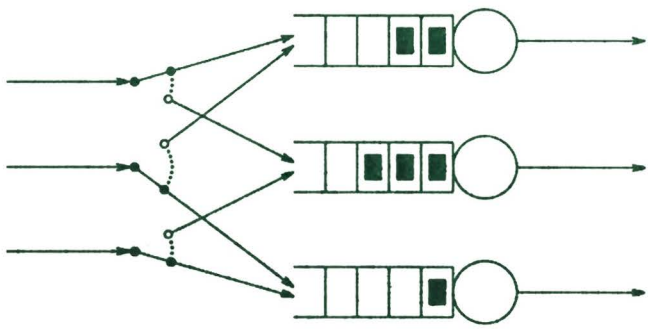
New approaches for multi-dimensional queueing systems

New Approaches for Multi-Dimensional Queueing Systems



Geert-Jan van Houtum

New Approaches for Multi-Dimensional Queueing Systems

New Approaches for Multi-Dimensional Queueing Systems

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de Rector Magnificus, prof. dr. J.H. van Lint, voor een commissie aangewezen door het College van Dekanen in het openbaar te verdedigen op dinsdag 7 februari 1995 om 16.00 uur

door

Geert-Jan Jan Johan Adriaan Nicolaas van Houtum

geboren te Erp

Dit proefschrift is goedgekeurd door
de promotoren
prof.dr. J. Wessels
en
prof.dr. W.H.M. Zijm

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Houtum, Geert-Jan Jan Johan Adriaan Nicolaas van

New approaches for multi-dimensional queueing systems /
Geert-Jan Jan Johan Adriaan Nicolaas van Houtum. -
Eindhoven : Eindhoven University of Technology
With ref. - With summary in Dutch.
ISBN 90-386-0434-3
Subject headings: queueing systems / Markov chains /
steady-state behavior.

Printed and bound by Wibro Dissertatiedrukkerij, Helmond

Contents

Chapter 1

Introduction

1.1. Motivation and objective

In many practical situations, one may have to wait some time before a service request can be fulfilled. This happens for example, when buying bread at a bakery, when passing a traffic crossing point (possibly with traffic lights) or when visiting a dentist. Similarly, products in a factory may have to wait before being processed on machines and computing jobs may have to wait before being handled by the central computer. All these situations have in common that they can be modeled as *queueing systems*.

Having to wait is bad in the first place, but it is even worse if one catches unexpected waiting times, due to a bad planning, for example. When having an appointment somewhere, it is desirable to have some information on the total driving time, including the time spent to waiting, to know what time one should leave home. And, when selling a product to order, one needs information on the production time, including delays caused by machines which are not available at the time they are needed, to know what due date can be promised. Therefore, it is necessary to have methods for obtaining information on waiting times, and other quantities, in queueing systems.

An important class of queueing systems is formed by those systems, for which the behavior can be described by *Markov processes*. For such systems, the information on the relevant performance measures may be obtained from the *equilibrium distribution* of the underlying Markov process. Therefore, much effort has been put in developing techniques for the determination of such equilibrium distributions. In this monograph, we shall focus on techniques for multi-dimensional Markov processes on integer grids and with some *homogeneity* in the transition probabilities/rates. Here, one should think of multi-dimensional state spaces, which are discrete and infinite in each component. Markov processes of this type are also called *multi-dimensional random walks*. They may be useful for modeling queueing systems consisting of two or more servers which all have their own queue, for example.

If the equilibrium distribution of a Markovian queueing system can be determined explicitly by some *analytical method*, then, usually, one also obtains explicit formulae for the relevant performance measures such as the mean waiting time. If the equilibrium distribution cannot be determined explicitly, then *numerical methods* may provide a way to obtain the distribution and the performance measures of interest, and the performance of the queueing system under consideration may be analyzed by means of a numerical study.

Contrary to the case where one has a Markov process or random walk with a state space which is essentially one-dimensional, i.e. which is infinite in at most one direction, for the equilibrium distribution of multi-dimensional random walks only a few *analytical results and*

techniques are available. For so-called *product-form networks* (see Baskett et al. [15]), it has been proved that the equilibrium distribution can be written as a product of powers of fixed factors. These factors may be obtained by substituting a product-form solution in the equilibrium equations and solving the remaining system of non-linear equations. Further two methods based on generating function analysis are available for a class of two-dimensional random walks. They are called the *uniformization technique* (see e.g. Kingman [49] and Flatto and McKean [33]) and the *boundary value method* (see e.g. Cohen and Boxma [23]). The results obtained by these methods are discussed in the next chapter. Here, it suffices to state that, unfortunately, both methods seem not to be extensible to random walks with dimension three and higher. Finally, there is the *compensation approach*, which has been developed for a class of two-dimensional, homogeneous random walks on the integer grid in the positive quadrant of the plane (see Adan [3] and the papers [5, 8, 12, 19]). Apart from the obvious ergodicity requirements, one has to require that no transitions can be made from points in the interior to the North, North-East and East, in order to obtain explicit expressions in the form of infinite series of products of powers of fixed factors. Contrary to the uniformization technique and the boundary value method, the compensation approach is a *direct* method for solving the equilibrium equations. The compensation approach seems to be the most promising method for being extended to a class of random walks with general dimension $N \geq 2$. In the *first part* of this monograph, we shall investigate to what extent this extension is possible.

If one is studying a multi-dimensional queueing system for which it is not possible to derive explicit formulae for the equilibrium distribution of the underlying random walk, then *numerical techniques* may be used to determine the equilibrium distribution and the relevant performance measures within a given accuracy. However, most numerical techniques described in the literature only work for random walks for which the state space is infinite in at most one component (see Stewart [68] and the references therein for an overview of a number of standard numerical techniques; see Neuts [58, 59] for a description of the matrix-geometric approach). The only numerical technique available for random walks with a state space which is infinite in each component, is the *power-series algorithm* (see Hooghiemstra et al. [42] and Blanc [18]). The main idea of the power-series algorithm is that power-series expansions of equilibrium probabilities as a function of the load of a queueing system can be used to solve the equilibrium equations. This technique has successfully been applied to several multi-dimensional queueing problems. It must be noted that for each numerical technique its use will be restricted by the requirements with respect to the computational effort and the memory space. Usually, this means that the equilibrium distribution and the relevant performance measures can be determined within the desired accuracy only for systems with a limited number of servers and a workload which is not too close to the maximum workload.

Instead of solving numerically the exact model within a given accuracy, one can also use *flexible approximation models* which can be solved exactly (or at least within a very high numerical accuracy, if they are solved by some numerical method) and which can approximate the exact model as accurately as desired. Appropriate flexible approximation models are, for example, *truncation models* which can be solved efficiently by a standard numerical technique or the matrix-geometric approach and which depend on some parameter(s) that determine(s) the size of the truncated state space. Approximation models such as truncation models lead to approximations for both the equilibrium distribution and the relevant performance measures of the exact model. Some approximation models can be proved to lead to

bounds for the relevant performance measures of the exact model. Such models are also called *bound models*. By combining a lower bound model and an upper bound model, one may obtain approximations for the relevant performance measures as well as upper bounds for the inaccuracy of these approximations. Defining *flexible lower and upper bound models*, and solving them for varying values of the parameters which determine the quality of the bound models, constitutes an appropriate, alternative way for determining the relevant performance measures within a given accuracy. In the *second part* of this monograph, we will develop the so-called *precedence relation method*, which is an analytical technique that is appropriate for deriving flexible bound models. Usually these flexible bound models will be truncation models which can be solved efficiently by a numerical method. It will be shown that for multi-dimensional queueing systems which cannot be solved by an analytical method and which have a favorable structure for deriving efficient, flexible bound models, an algorithm based on these flexible bound models may be very efficient compared to other algorithms such as the power-series algorithm (i.e. such an algorithm may solve larger systems with less computational effort and less memory space).

The *main objective* of this monograph is the development of methods for the analysis of queueing systems for which the behavior is described by multi-dimensional random walks. Because of its success for two-dimensional random walks, we first extend the compensation approach to random walks with dimension three and higher; and, since, for the three- and higher-dimensional case, the application of the compensation approach will appear to be limited to a relatively small class of problems, after that we will develop the precedence relation method for deriving flexible bound models. In the remainder of this introductory chapter, we will extensively explain the main ideas of both methods by using a special case of a practical queueing problem. The practical queueing problem is described in the next section, and the main ideas of both methods will be shown in the Sections 1.3 and 1.4. An outline of this monograph will be given in the final section of this chapter.

1.2. A practical queueing situation

In this section, we describe a queueing situation stemming from a flexible assembly system consisting of a group of parallel insertion machines, which have to mount vertical components on Printed Circuit Boards. This queueing situation leads to the formulation of a queueing model which we shall call the shortest queue system with a job-dependent parallelism. A special, simple case of this model is represented by the well-known symmetric shortest queue system, which has been studied extensively in the literature and which is known to be a hard problem already.

Let us start the description of the practical queueing situation with explaining how an *insertion machine* operates. An insertion machine mounts vertical components, such as resistors and capacitators, on a Printed Circuit Board (PCB) by the *insertion head*. The components are mounted in a certain sequence, which is prescribed by a Numerical Control program. The insertion head is fed by the *sequencer*, which picks components from tapes and transports them in the right order to the insertion head. Each tape contains only *one* type of components. The tapes are stored in the *component magazine*, which may contain 80 tapes,

say. Each PCB needs, on average, 60 different types of components. If a machine has to mount components on a PCB, then all the components need to be available on that machine. That means that for all those components a tape must be placed in the magazine. So the set of components available on the machine completely determines which types of PCBs can be handled.

In general we have a group of parallel insertion machines which have to process a number of different types of PCBs at the same time. Each insertion machine has its own queue, and the PCBs are transported to the insertion machines by an Automatic Conveyor System. In Figure 1.1, we have depicted a system which consists of three insertion machines and which has to process three different types of PCBs. The machines are basically similar, but due to the fact that they may be loaded with different types of components, the classes of PCB-types that can be handled by the machines may be different. In the situation depicted in Figure 1.1, machine M_1 can handle PCBs of the types A and B, machine M_2 can handle the types A and C, and machine M_3 can handle the types B and C.

In fact, there are two decision problems: the *assignment problem* and the *routing problem*. We first describe the assignment problem, which is the major problem. The assignment problem concerns how the tapes with components have to be divided among the machines. One should try to allocate the tapes with components to the machines such that, for example, the waiting times and/or sojourn times of the PCBs are minimized. There would be no problem if the magazines were big enough to contain all components needed to process all types of PCBs. However, in general they can only contain the components needed for a small subset of the different types of PCBs. So, the *limitations of the magazine capacities* give rise to a *job-dependent parallelism.*

In order to solve the assignment problem, we must be able to evaluate the performance characteristics of a *given assignment* of the components to the machines. These performance characteristics depend on how the second decision problem, i.e. the routing problem, is handled. This problem concerns to which machines the PCBs must be sent upon arrival. For an arriving PCB, we must select one of the machines which can handle that PCB. If for all types the mounting times are roughly the same, then it is reasonable to select the machine with the *shortest queue* (let ties be broken with equal probabilities); this at least (roughly) minimizes the waiting time of the arriving PCB itself, and it may be expected that this also roughly minimizes the average waiting time for all PCBs together, provided that we are in a balanced situation (i.e. a situation in which each server will have to handle the same amount of work on average). Assume that the shortest queue routing is used by the Automatic Conveyer System, and that, once arrived in a queue, the PCBs are served in a First-Come-First-Served (FCFS) manner. Then we have the following problem:

> *Given the shortest queue routing and the FCFS service discipline at each machine, we want to have an efficient method for the determination of the performance characteristics of the flexible assembly system for a given assignment of the components to the machines.*

The main performance characteristics we are interested in, are the waiting times and/or sojourn times for each type of PCBs and for all PCBs together. It is obvious that an efficient method for determining these measures can be exploited for selecting the best possible assignment of the components to the machines.

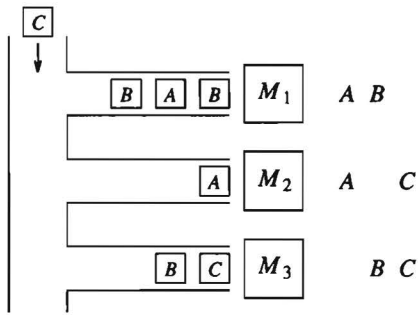


Figure 1.1. A flexible assembly system consisting of three parallel insertion machines, on which three types of PCBs are made.

The assembly of PCBs is often characterized by relatively few job types, large production batches and small processing times (see Zijm [86]). Therefore, a queueing model approach seems natural. The flexible assembly system can be modeled as a job-dependent parallel server system. For example, the system of Figure 1.1 is modeled as a queueing system with three parallel servers, each with its own queue, and three types of jobs; see Figure 1.2, where λ_A, λ_B and λ_C denote the arrival intensities of the three types of jobs (the meaning of the parameter μ is explained later on). The resulting queueing system will be called a *Shortest Queue System with a Job-Dependent Parallelism (SQS-JDP).* The problem is to determine the waiting times and/or sojourn times for each type of jobs and for a job of an arbitrary type.

Apart from the situation described above, queueing systems with job-dependent parallel servers also occur in many other practical situations; for example, in a job shop with a group of identical, parallel machines which are loaded with different sets of tools, in a computer system where each information file is available on a restricted set of a number of parallel disks and requests for information files have to be handled by only one disk, and at a banking office where each clerk is able to carry out a restricted set of tasks. Nevertheless, queueing systems with job-dependent parallel servers have hardly been studied in the literature. To our knowledge, the only contributions are the following ones. Schwartz [63] (see also Roque [61]) studies models with another type of routing than shortest queue routing and a server hierarchy such that the higher the level of the server, the more types of jobs it can handle. Green [37] studies a similar model with two types of jobs and two types of servers: servers which can serve jobs of both types and servers which can serve only jobs of the second type. Places where this situation occurs are, for example, a restaurant with tables that can seat four people and smaller tables for two, and a men's room with toilet stalls and urinals. In this system, there is not a queue at each server (table, toilet stall, urinal), but there is a central queue for the groups of customers. Finally, there is the paper by Adan et al. [7], from which we have adopted the flexible assembly system and the corresponding queueing model as described above.

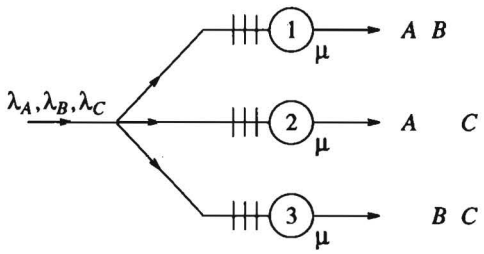


Figure 1.2. The queueing model with job-dependent parallel servers, which corresponds to the flexible assembly system depicted in Figure 1.1.

Schwartz [63], Green [37] and Adan et al. [7] have in common that they make similar assumptions/simplifications. They all assume that:

(i) All jobs arrive according to Poisson streams;

(ii) The service times are exponentially distributed;

(iii) The service times are job-independent.

In Adan et al. [7], it also is assumed that:

(iv) All insertion machines work equally fast.

Even for the simplified models, which are obtained by these assumptions, they all have not been able to derive an analytical solution; in Schwartz [63] and in Adan et al. [7] approximations for the waiting times are given, and Green [37] derives flexible truncation models, which may be solved efficiently by the matrix-geometric approach. This indicates that models with a job-dependent structure are hard to analyze, and that it is sensible to adopt the assumptions (i)-(iv) for the SQS-JDP as described here.

We can now complete the description of our model for the SQS-JDP. Let λ_j denote the intensity of the Poisson arrival process for jobs of type j. The assumptions (ii)-(iv) imply that all service times are exponentially distributed with the same parameter; let this parameter be denoted by μ. The behavior of the queueing model is described by a continuous-time Markov process with states $(m_1, \ldots, m_N)$, where N denotes the number of servers and m_i denotes the number of jobs at server i, including the job in service. The waiting times can easily be obtained from the equilibrium distribution. So, we would like to have a method for the determination of the equilibrium distribution for systems such as the SQS-JDP. If possible, we would like to have an analytical method, otherwise we will have to be content with a numerical method. Up to now, in the literature there is no appropriate, analytical method available.

As already denoted, the main reason for making the assumptions (i)-(iv) is that even the simplified model, which is obtained for the SQS-JDP by these assumptions, will be hard to analyze (see also the last paragraph of this section). For this reason, we better first attempt to find a method with which our simplified model can be solved. If we succeed in finding an appropriate method, then after that we can investigate whether the method itself, or a modified version of it, also works in case some assumptions of our model for the SQS-JDP

are relaxed. If we do not succeed in finding an appropriate method, then we also will not be able to solve a more general model for the SQS-JDP. Further, it must be noted that even the simplified model for the SQS-JDP may already yield a useful contribution to the assignment problem. Our simplified model may lead to insight in how the waiting times depend on the workloads of the different job types and on the job-dependent structure, especially on the number of servers from which a job of a particular type may choose one. Moreover, an efficient method for the determination of the performance characteristics of the simplified model may be useful for finding a good solution for the assignment problem. Here, the real quality of such a solution could be tested afterwards on the basis of a more general queueing model or a simulation model.

A special case of the model for the SQS-JDP is formed by the well-known *Symmetric Shortest Queue System (SSQS)*, which is obtained if there is only one job type, or if each server can handle all job types. The SSQS is described as follows. It consists of N parallel servers, where jobs arrive according to a Poisson process with parameter λ. Each arriving job joins the shortest queue (ties are broken with equal probabilities), and all service times are exponentially distributed with parameter μ. The SSQS has been studied extensively in the literature, and it is known to be a hard problem. This indicates that it is better to focus first on methods for the determination of the equilibrium distribution of the SSQS. We shall use the model for the SSQS as an illustration model at several places in this monograph, and especially in the next two sections, in which we will explain the main ideas of the compensation approach and the precedence relation method for deriving flexible bound models. Only in the last but one chapter of this monograph, i.e. in Chapter 7, we will return to the SQS-JDP.

1.3. The compensation approach

In the literature, explicit results for the equilibrium distribution of the SSQS have only been obtained for the case with $N=2$ servers. Methods applied to the two-dimensional SSQS are the uniformization technique (see Kingman [49] and Flatto and McKean [33]), the boundary value method (see Cohen and Boxma [23]), and the compensation approach (see Adan et al. [8]). The most explicit results have been obtained by the *compensation approach*, which is an *analytical method* leading to explicit formulae for all equilibrium probabilities. In this section, we shall describe both the way in which the compensation approach works for the two-dimensional SSQS and the general main idea behind this approach. Precedingly, an extensive numerical experiment is performed to explain the origin of the compensation approach as developed for the two-dimensional SSQS.

This section consists of four parts and is organized as follows. In the first part, we describe a Markov model for the two-dimensional SSQS. Next, the extensive numerical experiment is performed in order to gain some insight in the structure of the corresponding equilibrium distribution. Subsequently, it is shown that by exploiting the conjectured structure stemming from the numerical experiment, explicit expressions for the equilibrium distribution can be derived. Thereafter, in the last part of this section, it is shown how these explicit expressions may be derived in a similar, but *purely analytical* way by using the so-called compensation approach. In this last part, we also formulate the general main idea behind the

compensation approach and we discuss the other two-dimensional problems for which the equilibrium distribution can be determined by the compensation approach.

The Markov model for the SSQS with $N=2$ servers

Consider the SSQS with $N=2$ servers, which is depicted in Figure 1.3. The behavior of this system may be described by a continuous-time Markov process with states (m',n'), where m' and n' denote the queue lengths (including the jobs in service) at the servers 1 and 2. Because of the shortest queue routing, for this state description the state space is divided into two, similar, homogeneous regions (i.e. two similar regions with uniform transition rates) by the diagonal. A more attractive Markov process, of which the state space consists of *one* homogeneous region, is obtained by choosing the following alternative state description, which exploits the symmetry. Let the system be described by a continuous-time Markov process with states (m,n), where m denotes the length of the shortest queue and n denotes the difference between both queue lengths. The transition rates for this Markov process are depicted in Figure 1.4.

The system may be shown to be ergodic if and only if the workload $\rho=\lambda/(2\mu)$ is smaller than 1. This condition is assumed to be satisfied, and consequently we may characterize the equilibrium distribution $\{p_{m,n}\}$ as the unique normalized solution of the equilibrium equations:

$$(\lambda+2\mu)p_{m,n} = \lambda p_{m-1,n+1} + \mu p_{m,n+1} + \mu p_{m+1,n-1} \quad \text{if } m\geq 1,\ n\geq 2, \tag{1.1}$$

$$(\lambda+2\mu)p_{m,1} = \lambda p_{m-1,2} + \mu p_{m,2} + \lambda p_{m,0} + 2\mu p_{m+1,0} \quad \text{if } m\geq 1,\ n=1, \tag{1.2}$$

$$(\lambda+2\mu)p_{m,0} = \lambda p_{m-1,1} + \mu p_{m,1} \quad \text{if } m\geq 1,\ n=0, \tag{1.3}$$

$$(\lambda+\mu)p_{0,n} = \mu p_{0,n+1} + \mu p_{1,n-1} \quad \text{if } m=0,\ n\geq 2, \tag{1.4}$$

$$(\lambda+\mu)p_{0,1} = \mu p_{0,2} + \lambda p_{0,0} + 2\mu p_{1,0} \quad \text{if } m=0,\ n=1, \tag{1.5}$$

$$\lambda p_{0,0} = \mu p_{0,1} \quad \text{if } m=0,\ n=0. \tag{1.6}$$

Equation (1.1) is called the equilibrium equation for the *interior*, (1.2) and (1.3) are called the equilibrium equations for the *horizontal boundary*, and (1.4) is the equilibrium equation for the *vertical boundary*. Note that the horizontal boundary consists of two layers, since the rates for the outgoing transitions to the north for the points on the horizontal axis differ from the corresponding rates for the states (m,n) with $m,n\geq 1$. Dividing all equations by μ shows that the equilibrium distribution $\{p_{m,n}\}$ only depends on the workload ρ.

It is easily verified that the equilibrium equations do not have a simple product-form solution. So, we must look for a solution with a more complicated structure. In order to gain some insight in the structure of the equilibrium distribution, we shall perform a numerical experiment. The numerical experiment itself will lead to a conjecture with respect to the form of the equilibrium distribution.

Numerical experiment

We shall gain insight in the structure of the equilibrium distribution $\{p_{m,n}\}$ for the SSQS with $N=2$ servers by studying a numerically determined equilibrium distribution, which can be obtained by the method of successive substitutions after truncation of the state space. An appropriate truncated Markov process is obtained by truncating all states (m,n) with

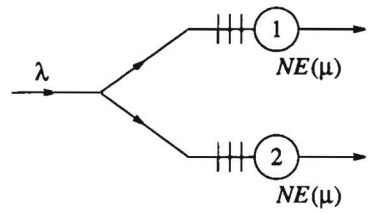


Figure 1.3. The symmetric shortest queue system (SSQS) with $N=2$ servers.

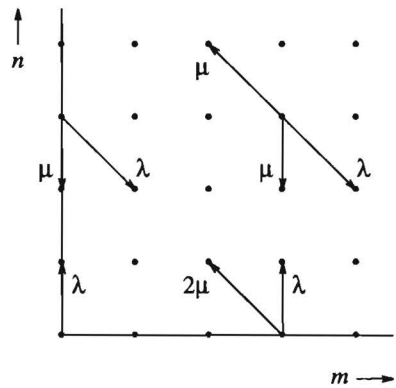


Figure 1.4. The transition rate diagram for the SSQS with $N=2$ servers.

$m+n > T$, where T is a fixed positive integer, and cutting off the transition from $(T, 0)$ to $(T, 1)$, which is caused by an arrival of a new job when there are T jobs present at each server. In fact, this truncated Markov process describes the behavior of the SSQS with finite buffers of size $T-1$. The equilibrium probabilities of the truncated process will serve as quite accurate approximations for the equilibrium probabilities $p_{m,n}$ of the original Markov process, at least for the states (m,n) near the origin, if the truncation level T is taken sufficiently large.

Below, we shall study the equilibrium distribution $\{p_{m,n}\}$ for the case $\rho=0.6$. We shall restrict ourselves to the probabilities $p_{m,n}$ for the states in the region $0 \le m,n \le 15$. Accurate approximations for these probabilities are obtained from the equilibrium distribution of the Markov process truncated at level $T=60$, which we have computed with a relative accuracy of $0.5 \cdot 10^{-14}$. The level $T=60$ seemed to be sufficiently large to exclude the impact of the truncation on the probabilities in the region $0 \le m,n \le 15$. The probabilities for the states (m,n) with $m,n \le 8$ are depicted in Table 1.1. As expected, a large part of the probability mass appears to be concentrated around the origin. Further, there appears to be more mass concentrated around the horizontal axis than around the vertical axis. This undoubtedly is due to the shortest queue routing, which causes a strong drift to the states corresponding to situations with equal queue lengths (i.e. to the states on the horizontal axis).

n ↑	m →								
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	0.0011	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0080	0.0023	0.0008	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000
2	0.0544	0.0166	0.0058	0.0021	0.0007	0.0003	0.0001	0.0000	0.0000
1	0.2761	0.1117	0.0410	0.0148	0.0053	0.0019	0.0007	0.0002	0.0001
0	0.2301	0.1384	0.0547	0.0200	0.0072	0.0026	0.0009	0.0003	0.0001
	0	1	2	3	4	5	6	7	8

Table 1.1. The equilibrium probabilities $p_{m,n}$ for the symmetric shortest queue system with two servers and workload $\rho = 0.6$.

n ↑	m →					
5	0.286	0.343	0.357	0.359	0.360	0.360
4	0.286	0.343	0.357	0.359	0.360	0.360
3	0.289	0.344	0.357	0.359	0.360	0.360
2	0.304	0.347	0.358	0.360	0.360	0.360
1	0.405	0.367	0.361	0.360	0.360	0.360
0	0.602	0.395	0.366	0.361	0.360	0.360
	0	1	2	3	4	5

n ↑	m →					
5	0.139	0.138	0.138	0.138	0.138	0.138
4	0.139	0.139	0.138	0.138	0.138	0.138
3	0.140	0.139	0.139	0.138	0.138	0.138
2	0.148	0.140	0.139	0.139	0.138	0.138
1	0.197	0.148	0.140	0.139	0.139	0.138
0	1.200	0.807	0.750	0.741	0.739	0.739
	0	1	2	3	4	5

Table 1.2. The ratios $p_{m+1,n}/p_{m,n}$ (on the left-hand side) and the ratios $p_{m,n+1}/p_{m,n}$ (on the right-hand side) of the equilibrium probabilities $p_{m,n}$ for the symmetric shortest queue system with two servers and workload $\rho = 0.6$.

Although the equilibrium distribution $\{p_{m,n}\}$ is not a product-form solution, it may still behave as a product-form solution for the states (m,n) corresponding to large queue lengths. It is known that for several complex (non-product-form) queueing systems the equilibrium distribution has such a behavior (which usually results in geometric/exponential tails in the distributions of queue lengths and waiting times; see e.g. Takahashi [69]). A product-form (geometric) behavior may be established by considering ratios of equilibrium probabilities. For the equilibrium distribution $\{p_{m,n}\}$, which we have numerically determined for the two-dimensional SSQS with workload $\rho = 0.6$, we have computed the ratios $p_{m+1,n}/p_{m,n}$ and $p_{m,n+1}/p_{m,n}$ of the equilibrium probabilities for two neighboring states in the m- and n-direction, respectively; see Table 1.2, where these ratios are depicted for all $m,n \leq 5$. As we see, the same values are obtained for all states (m,n) with sufficiently large m and n, which means that $\{p_{m,n}\}$ behaves as a product-form solution $\alpha^m \beta^n$, with $\alpha \approx 0.360$ and $\beta \approx 0.138$, for sufficiently large m and n (note that we would have found the same values for all m and n, in case the equilibrium distribution $\{p_{m,n}\}$ would have been equal to a product-form solution). We define the product-form solution

$$p^{(1)}_{m,n} = c_0 \alpha_0^m \beta_0^n \quad \text{for all } m,n \geq 0,$$

such that it accurately describes the behavior for the states around some point far from the origin, say around point (14,14); we take

$$\alpha_0 = \frac{p_{15,14}}{p_{14,14}} = 0.3600\,, \quad \beta_0 = \frac{p_{14,15}}{p_{14,14}} = 0.1385\,, \quad c_0 = \frac{p_{14,14}}{\alpha_0^{14}\beta_0^{14}} = 2.2974\,.$$

Then, studying the values for the difference $p_{m,n} - p^{(1)}_{m,n}$, see Table 1.3, shows that $\{p^{(1)}_{m,n}\}$ may serve as a rather good first-order approximation for $\{p_{m,n}\}$, at least for all states (m,n) with $n \geq 1$. The probabilities $p_{m,n}$ for the states on the horizontal axis appear to have a deviant behavior, which probably is due to the fact that for these states the rates for the transitions to the north differ from the corresponding rates for the states above them, and therefore we will neglect them during the remainder of our numerical experiment.

Just like for the equilibrium distribution $\{p_{m,n}\}$, we can also compute ratios for the values found for the difference $p_{m,n} - p^{(1)}_{m,n}$. In Table 1.4, we have depicted the ratios $(p_{m+1,n} - p^{(1)}_{m+1,n})/(p_{m,n} - p^{(1)}_{m,n})$ for two neighboring states in the m-direction (on the left-hand side) and the ratios $(p_{m,n+1} - p^{(1)}_{m,n+1})/(p_{m,n} - p^{(1)}_{m,n})$ for two neighboring states in the n-direction (on the right-hand side). We again observe that the same values occur for sufficiently large m and n, which means that also the second-order behavior of $\{p_{m,n}\}$ is described by a product form $\alpha^m\beta^n$. Since approximately the same value as before is found for the ratios in the n-direction, it seems that the factor β must be taken equal to β_0. We define

$$p^{(2)}_{m,n} = c_0\alpha_0^m\beta_0^n + c_1\alpha_1^m\beta_0^n \quad \text{for all } m,n \geq 0,$$

and choose α_1 and c_1 such that $\{p^{(2)}_{m,n}\}$ accurately describes the equilibrium behavior around the point (6,6), i.e. we choose

$$\alpha_1 = \frac{p_{7,6} - p^{(1)}_{7,6}}{p_{6,6} - p^{(1)}_{6,6}} = 0.0639\,, \quad c_1 = \frac{p_{6,6} - p^{(1)}_{6,6}}{\alpha_1^6\beta_0^6} = 0.7732\,.$$

Studying the values for the difference $p_{m,n} - p^{(2)}_{m,n}$, see Table 1.5, shows that the solution $\{p^{(2)}_{m,n}\}$ (slightly) improves the first-order approximation $\{p^{(1)}_{m,n}\}$.

The solution $\{p^{(2)}_{m,n}\}$ is called the second-order approximation and may be further improved to a third-order approximation $\{p^{(3)}_{m,n}\}$. Let us compute the ratios in the m- and n-direction for $p_{m,n} - p^{(2)}_{m,n}$; see Table 1.6. Since $\{p^{(2)}_{m,n}\}$ has been defined such that it accurately describes the behavior of $\{p_{m,n}\}$ around the point (6,6), the values obtained for $p_{m,n} - p^{(2)}_{m,n}$ for the states close to (6,6) are very small and cannot be used to determine the third-order behavior of $\{p_{m,n}\}$. The ratios for the points which are sufficiently close to the origin suggest to define

$$p^{(3)}_{m,n} = c_0\alpha_0^m\beta_0^n + c_1\alpha_1^m\beta_0^n + c_2\alpha_1^m\beta_1^n \quad \text{for all } m,n \geq 0,$$

and let β_1 and c_2 be chosen such that $\{p^{(3)}_{m,n}\}$ accurately describes the equilibrium behavior around the point (2,2):

$$\beta_1 = \frac{p_{2,3} - p^{(2)}_{2,3}}{p_{2,2} - p^{(2)}_{2,2}} = 0.0233\,, \quad c_2 = \frac{p_{2,2} - p^{(2)}_{2,2}}{\alpha_1^2\beta_1^2} = -6.5469\,.$$

The results presented in Table 1.7 show that $\{p^{(3)}_{m,n}\}$ very closely approximates $\{p_{m,n}\}$ for all states (m,n) with $n \geq 1$. Mainly because of the numerical inaccuracy with which we have computed the equilibrium probabilities $p_{m,n}$ for the states (m,n) with $m,n \leq 15$, the third-order

n \ m	0	1	2	3	4	5	6	7	8
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0019	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0104	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	-0.0420	-0.0028	-0.0002	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
0	-2.0674	-0.6886	-0.2430	-0.0872	-0.0314	-0.0113	-0.0041	-0.0015	-0.0005

Table 1.3. The values for the difference $p_{m,n}-p^{(1)}_{m,n}$ between the equilibrium probability $p_{m,n}$ and its first order approximation $p^{(1)}_{m,n}$.

n \ m	0	1	2	3	4	5
5	0.064	0.064	0.064	0.064	0.064	0.064
4	0.064	0.064	0.064	0.064	0.064	0.064
3	0.065	0.064	0.064	0.064	0.064	0.064
2	0.069	0.065	0.064	0.064	0.064	0.064
1	0.067	0.064	0.064	0.064	0.064	0.064
0	0.333	0.353	0.359	0.360	0.360	0.360

n \ m	0	1	2	3	4	5
5	0.139	0.139	0.139	0.139	0.139	0.139
4	0.140	0.139	0.139	0.139	0.139	0.139
3	0.145	0.144	0.143	0.143	0.143	0.143
2	0.187	0.177	0.175	0.175	0.175	0.175
1	-0.247	-0.252	-0.253	-0.253	-0.253	-0.253
0	0.020	0.004	0.001	0.000	0.000	0.000

Table 1.4. The ratios in the m-direction (on the left-hand side) and the n-direction (on the right-hand side) for $p_{m,n}-p^{(1)}_{m,n}$.

behavior of $\{p_{m,n}\}$ is the highest-order behavior which we can determine numerically, and therefore the numerical experiment must be ended here.

The main conclusion, which may be drawn from the numerical experiment, is the following one. The equilibrium distribution $\{p_{m,n}\}$ for the SSQS is not equal to a simple product-form solution, but *it seems to consist of a linear combination of product-form solutions*, at least for all states (m,n) with $n \geq 1$. More specifically, the equilibrium distribution $\{p_{m,n}\}$ for the states (m,n) with $n \geq 1$ seems to be equal to a linear combination

$$c_0\alpha_0^m\beta_0^n + c_1\alpha_1^m\beta_0^n + c_2\alpha_1^m\beta_1^n + c_3\alpha_2^m\beta_1^n + c_4\alpha_2^m\beta_2^n + \cdots \tag{1.7}$$

of product forms $\alpha_0^m\beta_0^n$, $\alpha_1^m\beta_0^n$, $\alpha_1^m\beta_1^n$, $\cdots$ (note that alternately a new α-factor and a new β-factor is taken), presumably with real-valued, positive and monotonously strictly decreasing product factors α_i and β_i (which must be smaller than 1, since the linear combination given by (1.7) must lead to a normalized solution of the equilibrium equations), and with real-valued coefficients c_i. It is noted that the linear combination given by (1.7) consists of an infinite number of terms; the number of *nonnull* terms may be finite, of course (this is the case, if there is an index k such that $c_i = 0$ for all $i \geq k$, for example). Now, the question is whether the equilibrium distribution indeed has the form as denoted by (1.7), and, if so, then

n \ m	0	1	2	3	4	5	6	7	8
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	-0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	0.0000	0.0000	-0.0000
5	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
4	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
3	-0.0001	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
2	-0.0044	-0.0002	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
1	-0.1491	-0.0097	-0.0006	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
0	-2.8405	-0.7380	-0.2462	-0.0874	-0.0314	-0.0113	-0.0041	-0.0015	-0.0005

Table 1.5. The values for the difference $p_{m,n}-p^{(2)}_{m,n}$ between the equilibrium probability $p_{m,n}$ and its second order approximation $p^{(2)}_{m,n}$.

n \ m	0	1	2	3	4	5
5	0.049	0.059	0.062	0.062	0.062	0.062
4	0.051	0.061	0.063	0.064	0.064	0.064
3	0.051	0.061	0.063	0.064	0.064	0.064
2	0.053	0.062	0.064	0.064	0.064	0.064
1	0.065	0.064	0.064	0.064	0.064	0.064
0	0.260	0.334	0.355	0.359	0.360	0.360

n \ m	0	1	2	3	4	5
5	0.022	0.018	0.015	0.011	0.007	0.003
4	0.023	0.023	0.022	0.021	0.021	0.020
3	0.023	0.023	0.023	0.023	0.023	0.023
2	0.024	0.023	0.023	0.023	0.023	0.023
1	0.030	0.024	0.024	0.023	0.023	0.023
0	0.052	0.013	0.003	0.000	0.000	0.000

Table 1.6. The ratios in the m-direction (on the left-hand side) and the n-direction (on the right-hand side) for $p_{m,n}-p^{(2)}_{m,n}$.

n \ m	0	1	2	3	4	5	6	7	8
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	-0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	-0.0000	-0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	-0.0000	-0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	-0.0000	-0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	-0.0009	-0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.0036	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0	3.7063	-0.3197	-0.2195	-0.0857	-0.0313	-0.0113	-0.0041	-0.0015	-0.0005

Table 1.7. The values for the difference $p_{m,n}-p^{(3)}_{m,n}$ between the equilibrium probability $p_{m,n}$ and its third order approximation $p^{(3)}_{m,n}$.

we would like to know how the product factors α_i and β_i and the coefficients c_i have to be chosen. The answers are given in the next part of this section.

Explicit expressions for the equilibrium distribution

Suppose that the equilibrium distribution $\{p_{m,n}\}$ for the states (m,n) with $n \geq 1$ is equal to a linear combination of product forms as denoted by (1.7), where the α_i, β_i and c_i have the presumed properties. Then the linear combination of (1.7) has to satisfy all equilibrium equations (1.1)-(1.6). We shall show that this observation leads to explicit expressions for the product factors α_i and β_i and the coefficients c_i. Besides, it will be shown that if the α_i, β_i and c_i are defined by these explicit expressions, then the equilibrium distribution $\{p_{m,n}\}$ indeed is equal to the linear combination given by (1.7) for the states (m,n) with $n \geq 1$ and it is equal to a similar expression for the states $(m, 0)$.

Let us first exploit the fact that the linear combination given in (1.7) has to satisfy the equilibrium equation (1.1) for the interior for all $m \geq 1$ and $n \geq 2$. For large m and n, the linear combination is dominated by the first product form $\alpha_0^m \beta_0^n$, which has a larger α-factor and/or a larger β-factor than all other product forms. Therefore, $\alpha_0^m \beta_0^n$ itself will be a solution of the equilibrium equation (1.1) for the interior. But, then, because of the linearity of this equilibrium equation, also the linear combination without the first term $c_0 \alpha_0^m \beta_0^n$ will be a solution of this equilibrium equation. For this remaining part of the linear combination the same reasoning can be given as for the full linear combination, and similarly for the linear combination without the terms $c_0 \alpha_0^m \beta_0^n$ and $c_1 \alpha_1^m \beta_0^n$, and so on. This leads to the conjecture that all product forms of the linear combination given by (1.7) must be solutions of the equilibrium equation (1.1) for the interior.

Substituting a product form $\alpha^m \beta^n$ into (1.1), and dividing by $\mu \alpha^{m-1} \beta^{n-1}$, shows that a product form $\alpha^m \beta^n$ is a (nonnull) solution of the equilibrium equation (1.1) for the interior if and only if the pair (α,β) is a solution of the equation

$$2(\rho+1)\alpha\beta = 2\rho\beta^2 + \alpha\beta^2 + \alpha^2 . \tag{1.8}$$

By substituting the pairs (α_0,β_0), (α_1,β_0) and (α_1,β_1), which we found during the numerical experiment for the case $\rho = 0.6$, into equation (1.8), it is easily verified that for that case the product forms $\alpha_0^m \beta_0^n$, $\alpha_1^m \beta_0^n$ and $\alpha_1^m \beta_1^n$ indeed satisfy the equilibrium equation (1.1) for the interior, which supports our conjecture.

Explicit formulae for the factors α_i and β_i for all $i \geq 1$ are now obtained by noting that equation (1.8) is a *quadratic equation* in α for fixed β, and vice versa. This means that equation (1.8) always has two (possibly complex) solutions α for fixed β and two (possibly complex) solutions β for fixed α. Therefore, the factor α_1 of the second product form $\alpha_1^m \beta_0^n$ of the linear combination given by (1.7) must be the companion solution to α_0 of the quadratic equation (1.8) for fixed $\beta = \beta_0$, the factor β_1 of the third product form $\alpha_1^m \beta_1^n$ must be the companion solution to β_0 of the quadratic equation (1.8) for fixed $\alpha = \alpha_1$, and so on. This leads to the recursive formulae

$$\alpha_{i+1} = 2\rho\beta_i^2 \cdot \frac{1}{\alpha_i} , \quad \beta_{i+1} = \frac{\alpha_{i+1}^2}{2\rho + \alpha_{i+1}} \cdot \frac{1}{\beta_i} , \quad i \geq 0, \tag{1.9}$$

which are obtained from the formula for the product of the two roots of a quadratic equation. With these formulae we can easily generate the factors α_i and β_i for all $i \geq 1$, once the starting

factors α_0 and β_0 are known. In the left part of Figure 1.5 we have depicted the real-valued solutions of equation (1.8) for the case $\rho = 0.6$, and in the right part of Figure 1.5 we have visualized the generation of the factors α_i and β_i; here, the factors α_0 and β_0 have been taken equal to the values which we found during the numerical experiment.

Next, exploiting the fact that the linear combination of (1.7) has to satisfy the equilibrium equation (1.4) for the vertical boundary for all $n \geq 2$, leads to an explicit expression for the coefficients c_i. For the states (m,n) around the vertical axis and large n, the linear combination of (1.7) is dominated by the first two product forms, since they have the largest β-factor. Therefore the sum $c_0\alpha_0^m\beta_0^n + c_1\alpha_1^m\beta_0^n$ of the first two terms will be a solution of the equilibrium equation (1.4) for the vertical boundary. But, then also the remaining part of the linear combination will be a solution of this equilibrium equation, by which we find that also the sum $c_2\alpha_1^m\beta_1^n + c_3\alpha_2^m\beta_1^n$ will be a solution of this equilibrium equation; and so on. This leads to the conjecture that all pairs of product forms with the same β-factor must be solutions of the equilibrium equation (1.4) for the vertical boundary. By using elementary algebra, it may be shown that a sum $c_{2k}\alpha_k^m\beta_k^n + c_{2k+1}\alpha_{k+1}^m\beta_k^n$ satisfies equation (1.4), if the coefficients c_{2k} and c_{2k+1} satisfy the equation

$$c_{2k+1} = \frac{\beta_k - \alpha_{k+1}}{\alpha_k - \beta_k}\, c_{2k}\,, \quad k \geq 0. \tag{1.10}$$

This expression may be used as a definition for the coefficients c_i for all odd i. It is easily verified that the coefficients c_0 and c_1 which we found for the case $\rho = 0.6$ satisfy equation (1.10).

Subsequently, we exploit the fact that the linear combination of (1.7) has to satisfy the equilibrium equations (1.2) and (1.3) for the horizontal boundary for all $m \geq 1$. This will lead to a definition for the coefficients c_i for all even $i \geq 2$, and also to explicit expressions for the product factors α_0 and β_0. Note that, to let the linear combination of (1.7) satisfy the equations (1.2) and (1.3), we also have to give a definition for the equilibrium probabilities for the states $(m, 0)$ with $m \geq 1$, since the linear combination of (1.7) is assumed to describe the equilibrium behavior only for the states (m,n) with $n \geq 1$, and in the equations (1.2) and (1.3) also the equilibrium probabilities for the states $(m, 0)$ with $m \geq 1$ are present.

For the states (m,n) around the horizontal axis and large m, the linear combination of (1.7) is dominated by the first term $c_0\alpha_0^m\beta_0^n$, which has the largest α-factor. Therefore this term itself, accompanied by some solution for the states $(m, 0)$ with $m \geq 1$, must be a solution of the equilibrium equations (1.2) and (1.3). Since these equations have to be satisfied for all $m \geq 1$, it seems sensible to take a solution of the form $c_0'\alpha_0^m$ for the states $(m, 0)$ with $m \geq 1$. This choice is supported by values for the ratios $p_{m,1}/p_{m,0}$ which we found for the case $\rho = 0.6$; see Table 1.2. By replacing the coefficient c_0' by $f_0 c_0$, we obtain the solution

$$\begin{cases} c_0\alpha_0^m\beta_0^n & \text{if } m \geq 0,\ n \geq 1; \\ f_0 c_0\alpha_0^m & \text{if } m \geq 1,\ n = 0, \end{cases} \tag{1.11}$$

which should satisfy the equilibrium equations (1.2) and (1.3). By substituting this solution into (1.2) and (1.3) and using that $c_0\alpha_0^m\beta_0^n$ also is a solution of the equilibrium equation (1.1) for the interior, we find that α_0, β_0 and f_0 must be equal to

$$\alpha_0 = \rho^2\,, \quad \beta_0 = \frac{\rho^2}{\rho + 2}\,, \quad f_0 = \frac{\alpha_0}{2(\rho + \alpha_0)}\,. \tag{1.12}$$

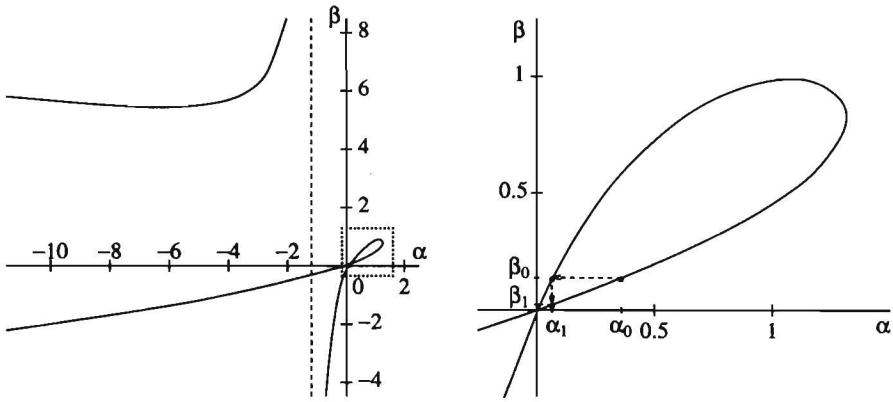


Figure 1.5. The real-valued solutions (α,β) of the quadratic equation (1.8) for the case $\rho=0.6$. The part outlined by the dotted line in the figure on the left hand, has been blown up and is depicted in the figure on the right hand, where the bold dots denote the solutions needed for the equilibrium distribution $\{p_{m,n}\}$.

Note that for the case $\rho=0.6$, we obtain $\alpha_0=0.3600$ and $\beta_0=0.1385$, which are precisely the values which we found during the numerical experiment.

Since the first term of the linear combination given by (1.7) is a solution of the equilibrium equations (1.2) and (1.3) for the horizontal boundary, also the remaining part has to satisfy these equations. This leads to the property that $c_1\alpha_1^m\beta_0^n+c_2\alpha_1^m\beta_1^n$, accompanied by some solution for the states $(m, 0)$ with $m\geq 1$, must be a solution of the equations (1.2) and (1.3); and, similarly for all other sums of pairs of product forms with the same α-factor. For all $k\geq 0$, we let the sums $c_{2k+1}\alpha_{k+1}^m\beta_k^n+c_{2k+2}\alpha_{k+1}^m\beta_{k+1}^n$ be accompanied by solutions $f_{k+1}(c_{2k+1}+c_{2k+2})\alpha_{k+1}^m$ for the states $(m, 0)$ with $m\geq 1$. By using elementary algebra, it may be shown that the solution

$$\begin{cases} c_{2k+1}\alpha_{k+1}^m\beta_k^n+c_{2k+2}\alpha_{k+1}^m\beta_{k+1}^n & \text{if } m\geq 0,\ n\geq 1; \\ f_{k+1}(c_{2k+1}+c_{2k+2})\alpha_{k+1}^m & \text{if } m\geq 1,\ n=0, \end{cases} \tag{1.13}$$

satisfies the equations (1.2) and (1.3), if the coefficients c_{2k+1}, c_{2k+2} and f_{k+1} satisfy the equations

$$c_{2k+2}=-\frac{(\rho+\alpha_{k+1})/\beta_{k+1}-(\rho+1)}{(\rho+\alpha_{k+1})/\beta_k-(\rho+1)}\,c_{2k+1}\ ,\quad f_{k+1}=\frac{\alpha_{k+1}}{2(\rho+\alpha_{k+1})}\ ,\quad k\geq 0. \tag{1.14}$$

The first one of these equations is used as a definition for the coefficients c_i for all even $i\geq 2$.

Let us recapitulate what we found up to now. The linear combination of (1.7) for the states (m,n) with $n\geq 1$ is accompanied by the linear combination

$$f_0c_0\alpha_0^m+f_1(c_1+c_2)\alpha_1^m+f_2(c_3+c_4)\alpha_2^m+\cdots \tag{1.15}$$

for the states $(m, 0)$ with $m\geq 1$ (see (1.11) and (1.13)). Further, explicit definitions have been given for all product factors α_i and β_i and for all coefficients c_i and f_i, except for c_0 (see

(1.9), (1.10), (1.12) and (1.14)). The formulae for the coefficients c_i may be slightly simplified by replacing, for all $k \geq 0$, the c_{2k} by $a_k b_k$ and the c_{2k+1} by $a_{k+1} b_k$. We then obtain the formulae

$$a_{k+1} = \frac{\beta_k - \alpha_{k+1}}{\alpha_k - \beta_k} a_k \ , \quad b_{k+1} = -\frac{(\rho + \alpha_{k+1})/\beta_{k+1} - (\rho+1)}{(\rho + \alpha_{k+1})/\beta_k - (\rho+1)} b_k \ , \quad k \geq 0, \tag{1.16}$$

instead of the formulae for the coefficients c_i as given in (1.10) and (1.14), and the formal solution $\{x_{m,n}\}$ defined by

$$x_{m,n} = \begin{cases} \overbrace{a_0 b_0 \alpha_0^m \beta_0^n}^{\text{V}} + \overbrace{a_1 b_0 \alpha_1^m \beta_0^n + a_1 b_1 \alpha_1^m \beta_1^n}^{\text{V}} + a_2 b_1 \alpha_2^m \beta_1^n + \cdots & \text{if } m \geq 0,\ n \geq 1; \\ \underbrace{f_0 a_0 b_0 \alpha_0^m}_{\text{H}} + \underbrace{f_1 a_1 (b_0 + b_1) \alpha_1^m}_{\text{H}} + \cdots & \text{if } m \geq 1,\ n = 0; \end{cases}$$

is obtained instead of the linear expressions given by (1.7) and (1.15). Let the coefficients a_0 and b_0 be defined by $a_0 = b_0 = 1$. Then the terms of $\{x_{m,n}\}$ satisfy the following properties:

(i) all terms $a_i b_i \alpha_i^m \beta_i^n$ and $a_{i+1} b_i \alpha_{i+1}^m \beta_i^n$, $i \geq 0$, individually satisfy the equilibrium equation (1.1) for the interior;

(ii) all pairs $a_i b_i \alpha_i^m \beta_i^n + a_{i+1} b_i \alpha_{i+1}^m \beta_i^n$, $i \geq 0$, of terms with the same β-factor satisfy the equilibrium equation (1.4) for the vertical boundary (V);

(iii) for all $i \geq 0$, the pairs $a_{i+1} b_i \alpha_{i+1}^m \beta_i^n + a_{i+1} b_{i+1} \alpha_{i+1}^m \beta_{i+1}^n$ of terms with the same α-factor, which are accompanied by the terms $f_{i+1} a_{i+1} (b_i + b_{i+1}) \alpha_{i+1}^m$ for the states $(m, 0)$ with $m \geq 1$, satisfy the equilibrium equations (1.2) and (1.3) for the horizontal boundary (H), while the first term $a_0 b_0 \alpha_0^m \beta_0^n$, which is accompanied by the term $f_0 a_0 b_0 \alpha_0^m$ for the states (m,n) with $n = 0$, individually satisfies these equations.

So, we find that $\{x_{m,n}\}$ is an unnormalized solution of the equilibrium equations (1.1)-(1.4), at least in principle; it remains to show that $\{x_{m,n}\}$ is well-defined, i.e. that $\{x_{m,n}\}$ converges (absolutely) for all m and n, $(m,n) \neq (0,0)$).

By exploiting the explicit formulae for the product factors α_i and β_i, it may be proved that they are real-valued numbers satisfying the monotonicity result

$$1 > \alpha_0 > \beta_0 > \alpha_1 > \beta_1 > \cdots > 0$$

(see also Figure 1.5), and that both the α_i and the β_i decrease exponentially fast to 0 as $i \to \infty$. From these properties, it immediately follows that the coefficients a_i, b_i and f_i are well-defined, real-valued, nonnull numbers and that $\{x_{m,n}\}$ consists of infinitely many (nonnull) terms. Further, it may be shown that for all m and n, $(m,n) \neq (0,0)$, the terms $a_i b_i \alpha_i^m \beta_i^n$, $a_{i+1} b_i \alpha_{i+1}^m \beta_i^n$ and $f_{i+1} a_{i+1} (b_i + b_{i+1}) \alpha_{i+1}^m$, in absolute value, decrease exponentially fast to 0, as $i \to \infty$, which implies that all series occurring in the expressions for $x_{m,n}$ are absolutely convergent for all m and n, $(m,n) \neq (0,0)$. So, all variables $x_{m,n}$ are well-defined and $\{x_{m,n}\}$ indeed is a solution of the equilibrium equations (1.1)-(1.4).

The solution $\{x_{m,n}\}$ is completed by defining $x_{0,0}$ such that the equilibrium equation (1.5) for the state $(0,1)$ is satisfied. Then, $\{x_{m,n}\}$ satisfies all equilibrium equations. The equilibrium equation (1.6) for the state $(0,0)$ is also satisfied, since it is well-known that the

system of equilibrium equations of a Markov process is dependent. Next, it may be shown that $\Sigma_{m,n\geq 0} |x_{m,n}| < \infty$. As a result, the equilibrium distribution $\{p_{m,n}\}$ may be obtained by simply normalizing the solution $\{x_{m,n}\}$:

$$p_{m,n} = C^{-1} x_{m,n} \quad \text{for all } m,n \geq 0, \tag{1.17}$$

where C is the normalizing constant and may be proved to be equal to (see [8])

$$C = \frac{\rho(2+\rho)}{4(1-\rho^2)(2-\rho)}.$$

The main result stated in (1.17) shows that the equilibrium distribution $\{p_{m,n}\}$ consists of a linear combination of an infinite number of product-form solutions, and it confirms the conjecture which we obtained from the numerical experiment, i.e. it confirms that for the states (m,n) with $n \geq 1$ the equilibrium behavior indeed is described by a linear combination of the form denoted by (1.7). Note that, for the case $\rho = 0.6$, this main result states that

$$\begin{aligned} p_{m,n} &= C^{-1} (a_0 b_0 \alpha_0^m \beta_0^n + a_1 b_0 \alpha_1^m \beta_0^n + a_1 b_1 \alpha_1^m \beta_1^n + a_2 b_1 \alpha_2^m \beta_1^n + \cdots) \\ &= 2.2974\,(0.3600)^m (0.1385)^n + 0.7732\,(0.0639)^m (0.1385)^n \\ &\quad - 6.4977\,(0.0639)^m (0.0233)^n - 2.0998\,(0.0102)^m (0.0233)^n + \cdots, \end{aligned}$$

for all states (m,n) with $n \geq 1$, which (almost completely) corresponds to the values for the product factors and coefficients which we found during the numerical experiment (for the coefficient of the third product form we found $c_1 = -6.5469$ instead of -6.4977, which seems to be due to numerical inaccuracy).

The main result stated in (1.17) and the formulae for the solution $\{x_{m,n}\}$, the product factors α_i and β_i, and the coefficients a_i, b_i and f_i provide us of explicit expressions for the equilibrium distribution $\{p_{m,n}\}$. We have obtained these expressions by exploiting the conjectured structure stemming from the numerical experiment. In the next and last part of this section, it is explained how these expressions may be derived in a similar, but *purely analytical* way by using the so-called *compensation approach*, which for the two-dimensional SSQS has been developed in Adan et al. [8].

The compensation approach

Consider the solution $\{x_{m,n}\}$, which is equal to the equilibrium distribution $\{p_{m,n}\}$ for the SSQS with $N = 2$ servers, up to a normalizing constant. The first term $a_0 b_0 \alpha_0^m \beta_0^n$ of $\{x_{m,n}\}$, which is accompanied by the term $f_0 a_0 b_0 \alpha_0^m$ for the states $(m, 0)$ with $m \geq 1$, is the dominating term, which satisfies the equilibrium equations (1.1)-(1.3) for the interior and the horizontal boundary. This first term already quite accurately describes the equilibrium behavior, provided that it is normalized by the normalizing constant C; see Table 1.3, in which we have depicted the difference between the equilibrium distribution $\{p_{m,n}\}$ and the first-order approximation $\{p^{(1)}_{m,n}\}$ consisting of only the first term. It is easily verified that the first term violates the equilibrium equation (1.4) for the vertical boundary, and therefore the second term $a_1 b_0 \alpha_1^m \beta_0^n$, which satisfies equation (1.1) for the interior and, together with the first term, also equation (1.4) for the vertical boundary, may be seen as a *compensation term*, which is added to compensate the error of the first term on the vertical boundary. However, the solution consisting of the first two terms of $\{x_{m,n}\}$ violates the equations (1.2) and (1.3) for the horizontal boundary, i.e. the second term has introduced a *new* error on this horizontal

boundary. Fortunately, this new error is smaller than the old one; see Table 1.5, in which the quality of the second-order approximation $\{p^{(2)}_{m,n}\}$ consisting of the first two terms of $\{x_{m,n}\}$ has been depicted. Next, in order to compensate the error of the second term, the third term $a_1 b_1 \alpha_1^m \beta_1^n$, together with the term $f_1 a_1 (b_0 + b_1)\alpha_1^m$ for the states $(m, 0)$ with $m \geq 1$, is added. The third term, which has been defined such that it satisfies equation (1.1) for the interior and, together with the first two terms, also the equations (1.2) and (1.3) for the horizontal boundary, introduces a new, but smaller error at the vertical boundary (see also Table 1.7). This error is compensated by the forth term $a_2 b_1 \alpha_2^m \beta_1^n$ similarly to the compensation of the error of the first term on the vertical boundary by the second term; and so on.

The reasoning in the previous paragraph shows that the solution $\{x_{m,n}\}$ may be obtained in an alternative way by starting with an initial term (the first term of $\{x_{m,n}\}$), which satisfies the equilibrium equations (1.1)-(1.3) for the interior and the horizontal boundary, but violates the equilibrium equation (1.4) for the vertical boundary, and subsequently, step by step, adding compensation terms such that each compensation term compensates the error of the previous term on alternately the vertical boundary and the horizontal boundary. This precisely describes how for the two-dimensional SSQS, a solution of the equilibrium equations is generated by the *compensation approach*. The compensation approach is based on the following, simple *main idea*, which easily can be applied to several other problems:

1. *Characterize a set P of product-form solutions which satisfy the equilibrium equation for the interior;*
2. *Try to construct a linear combination of product-form solutions of this set P, such that also the equilibrium equations for the boundaries are satisfied.*

In general, the application of the compensation approach leads to the generation of one or more *formal* solutions of the equilibrium equations for the interior and the boundaries. After that, the formal solutions must be shown to be absolutely convergent, and subsequently these solutions must be linearly combined and normalized in order to obtain the equilibrium distribution. For the two-dimensional SSQS, the application of the compensation approach leads to one formal solution, the solution $\{x_{m,n}\}$, which is just sufficient for obtaining the equilibrium distribution (see the explicit result stated in (1.17)).

The SSQS with $N = 2$ servers has been the first problem which has been solved by the compensation approach, i.e. for which the compensation approach has been developed (see Adan et al. [8]). In the meantime, the compensation approach has been applied to several other two-dimensional problems. The success for the two-dimensional SSQS inspired Adan et al. [12] to apply the compensation approach to the class of homogeneous, nearest-neighboring random walks on the lattice of the first quadrant; see Figure 1.6. They established that the compensation approach is successful for a problem of this class if and only if the transition probabilities/rates $q_{i,j}$ for the interior satisfy the condition that

$$q_{0,1} = q_{1,0} = q_{1,1} = 0. \tag{1.18}$$

This condition states that no transitions to the North, East and North-East are allowed from points in the interior of the state space, and stems from convergence requirements for the generated formal solutions. Problems belonging to the class considered in [12] and satisfying condition (1.18), are besides the two-dimensional SSQS, a multiprogramming queues system (see [5]), and the 2×2 buffered switch (see [19]). A problem belonging to the class considered in [12], but violating condition (1.18), is the symmetric SQS-JDP with $N = 2$ servers,

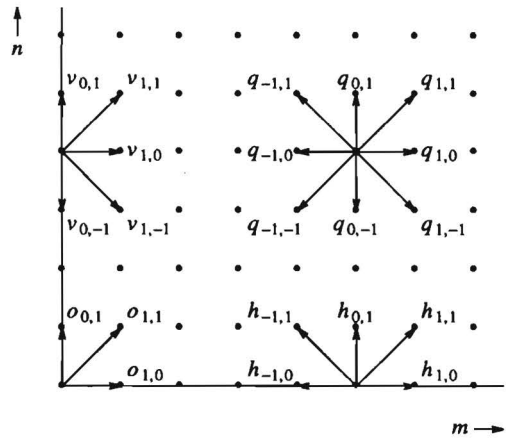


Figure 1.6. The transition rates for a two-dimensional, homogeneous, nearest-neighboring random walk on the lattice of the first quadrant; for all states the transitions to themselves have been left away.

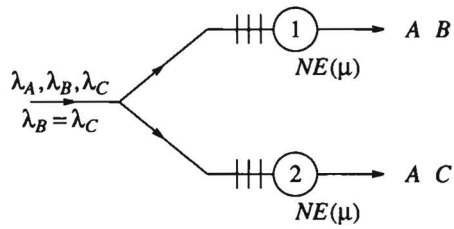


Figure 1.7. A SQS-JDP with $N=2$ servers and 3 types of jobs. The system is symmetric since the arrival intensity λ_B for the jobs which can be served only by server 1 is equal to the arrival intensity λ_C for the jobs which can be served only by server 2.

as depicted in Figure 1.7. Just like the SSQS, this system can be modeled as a Markov process with states (m,n), where m denotes the length of the shortest queue and n denotes the difference between both queue lengths. For this process, the positive transition rates for the interior are

$$q_{1,-1} = \lambda_A + \lambda_B \,, \quad q_{0,1} = \lambda_B \,, \quad q_{-1,1} = q_{0,-1} = \mu \,,$$

which shows that there is a positive rate $q_{0,1}$ for the transitions to the North from interior points (it is noted that a similar Markov process is obtained for the model studied by Hassin and Haviv [40]). Apart from the problems belonging to the class of two-dimensional random walks studied in [12], and satisfying (1.18), the compensation approach has been proved to work for the asymmetric shortest queue system with 2 servers working at different speeds (see [9]), and for the system consisting of two parallel Erlang servers, where jobs arrive

according to a Poisson stream and join the queue where they expect to have the shortest delay (see [6]). All two-dimensional problems for which the compensation approach works, have in common that for each interior point (m,n) only transitions are possible to points (m',n') with $|m'|+|n'| \leq |m|+|n|$; for points (m,n) in the interior of the first quadrant this means that only transitions are possible to points $(m+i,n+j)$ with $i+j \leq 0$ (cf. (1.18)).

In the first part of this monograph, we will apply the compensation approach to a class of N-dimensional homogeneous, nearest-neighboring random walks (see also [78, 80]). Among others, we will generalize the condition (1.18), under which the compensation approach works. The analysis will point out that the compensation approach does not work for the SSQS with $N \geq 3$ servers, which supports the conjecture stated in [77]. In that paper, on the basis of a numerical study of ratios of equilibrium probabilities, we formulated the conjecture that the equilibrium distribution of the SSQS with $N=3$ cannot be expressed as a linear combination of product-form solutions, but must have a more complicated structure. The same will hold for the SQS-JDP, of which the SSQS is a special case. Since also no other analytical methods are available for the SSQS and the SQS-JDP with $N \geq 3$ servers (the other two methods mentioned at the beginning of this section, the uniformization technique and the boundary value method, seem to be not extensible to random walks with dimension $N \geq 3$ at all), these conclusions constitute a *justification for the use of numerical or numerically oriented methods* for these queueing systems.

1.4. The precedence relation method for deriving flexible bound models

Since no analytical methods are available for the determination of the equilibrium distribution of the N-dimensional SSQS with $N \geq 3$, in the literature many numerical studies have appeared. However, most studies deal with approximation models, which can be solved by a standard numerical technique and which lead to approximations for the equilibrium distribution and/or the relevant performance measures, or they deal with simple approximations for the relevant performance measures (of which subsequently the quality is tested). Only a few studies may be characterized as exact numerical studies, i.e. as studies which lead to numerical procedures with which the equilibrium distribution and/or the relevant performance measures can be computed as accurately as desired.

The exact studies by Blanc [17], Lui et al. [54] (see also [53]), and Adan et al. [2] have led to the most successful numerical procedures. In Blanc [17], the power-series algorithm has been applied to the SSQS, with which all equilibrium probabilities as well as all performance measures may be computed (as accurately as desired). In the papers by Lui and Muntz [54] and Adan et al. [2], the analysis has mainly been focused on the determination of the *mean waiting time*, which may be considered to be the most interesting performance measure for the SSQS, and for that purpose, in both papers a flexible lower bound model and a flexible upper bound model are derived; by combining these two models, approximations for the mean waiting time as well as upper bounds for the corresponding inaccuracies are obtained. For the computation of the mean waiting time for the SSQS, the numerical procedure based on the flexible bound models as derived in [2], seems to be the most powerful procedure; up to now, the largest systems (viz. systems with up to $N=50$ servers and workloads up to 0.95) have been solved with this procedure.

The flexible bound models derived in [2] are truncation models of which the state space and the structure of the transitions are such that they can be solved efficiently by the matrix-geometric approach (as described by Neuts [58]). To prove that one model (a threshold jockeying model) leads to lower bounds for the mean waiting time, and the other model (a threshold blocking model) to upper bounds, an *analytical* method based on Markov cost/reward theory has been used. This analytical method is similar to the technique used in the papers by Van der Wal [72], Van Dijk and Van der Wal [76], and Van Dijk and Lamond [75], and we call this method the *precedence relation method.*

In this section, we shall explain the main idea of the precedence relation method, as it is used for comparing truncation models to their original model. From this main idea, it immediately follows that, apart from proving that an approximation or truncation model is a bound model, the precedence relation method is also appropriate for *deriving* bound models, and especially for deriving *flexible* bound models. The main idea will be explained on the basis of the SSQS with $N=2$ servers. For this model, we shall describe the two flexible truncation models as presented in [2], and subsequently it will be shown how the precedence relation method may be used to prove that these truncation models lead to bounds for the mean waiting time. Finally, some numerical results will be presented to show how tight these bounds are.

Consider the Markov process for the SSQS with $N=2$ servers as described at the beginning of the previous section. To obtain approximations for the equilibrium probabilities $p_{m,n}$ for the states close to the origin, we defined a simple truncated Markov process with a finite state space of the form $\{ (m,n) \mid m+n \leq T \}$, where T is a fixed, positive integer. We used this simple truncation model to determine quite accurately the probabilities $p_{m,n}$ depicted in Table 1.1 for the case $\rho=0.6$. The results depicted in this table show that most of the probability mass is concentrated around the states on the horizontal axis. An explanation for this property follows from the shortest queue routing discipline, which causes a strong drift to the states corresponding to situations with equal queue lengths, i.e. to the states on the horizontal axis. This observation suggests that it might be better to define a truncated Markov process with a state space of the form $\{ (m,n) \mid n \leq T \}$, since a set of this form better matches the states where most of the probability mass is present.

Suppose that we take a truncated state space $M' = \{ (m,n) \mid n \leq T \}$, where T is a fixed, positive integer. This means that we do not allow the difference between both queue lengths to exceed the threshold value T. Therefore we must modify the transitions from the states (m,T), $m \geq 1$, inside the state space M' to the states $(m-1,T+1)$ outside M'. These transitions are due to a service completion at the shortest queue when the difference between both queue lengths has already reached its maximum allowed value T, and occur with rate μ. We consider the following two modifications:

* A very natural modification is that we just cut off each transition from a state (m,T), $m \geq 1$, to a state $(m-1,T+1)$, i.e. the transition is redirected to the state (m,T) itself. The physical interpretation is that, if a service completion at the shortest queue causes a too large difference between the longest and the shortest queue, then the job in service at the shortest queue stays in the system and has to be served again. Since we have exponentially distributed service times (which satisfy the memoryless property), this is equivalent to saying that the server at the shortest queue is blocked, if the length of the

shortest queue is T less than the length of the longest queue. The model corresponding to the truncated Markov process which we obtain in this case, is called the *Threshold Blocking (TB) model.*

* Another modification is that each transition from a state (m,T), $m \geq 1$, to the state $(m-1,T+1)$ is redirected to the state $(m,T-1)$, in which we have one job more in the shortest queue and one job less in the longest queue (compared to the state $(m-1,T+1)$). The physical interpretation for this redirection is that if a service completion at the shortest queue occurs in a situation that the length of the shortest queue is T less than the length of the longest queue, then at the same time we let a job of the longest queue jockey to the shortest queue. The truncation model which we obtain in this case is called the *Threshold Jockeying (TJ) model.*

For both truncation models, the modifications are depicted in Figure 1.8. Note that both the TJ and TB model can approximate the original SSQS as accurately as desired, since the size of the truncated state space M' depends on the threshold parameter T and M' is equal to the state space of the original SSQS in case $T=\infty$; for that reason, they are called *flexible* truncation models.

The reason why it makes sense to consider the TJ model and the TB model, is that for both models the equilibrium distribution, and the relevant performance measures, can be determined by using the matrix-geometric approach. For both models, one may even obtain very efficient numerical procedures by exploiting the special structure of the transitions (i.e. by partitioning the state space into levels l consisting of all states (m,n) for which the length $m+n$ of the longest queue is equal to l, by which each level has only one state from which a transition to a higher level can be made, and the rate matrix R has only one row with nonnull elements). It must be noted that, besides in [2], the TJ model has also been studied in several other papers (see, for example, Haight [38], Gertsbakh [35], and Adan et al. [10, 13]; actually, Gertsbakh [35] uses the TJ model as an approximation model for the two-dimensional SSQS).

We shall use the TJ and TB model to determine the mean *normalized* waiting time W for the original SSQS. The normalized waiting time is defined as the ratio of the waiting time and the mean of a service time ($= 1/\mu$), and only depends on the workload ρ (and on the number of servers N in the case with general $N \geq 2$). By using Little's formula, it follows that

$$W = \frac{L_w}{2\rho}, \tag{1.19}$$

where L_w is the mean of the total number of waiting jobs in the system. The two bound models are expected to give bounds for L_w, and therefore also for W. Since in the TJ model some jobs are allowed to jockey from the longest queue to the shortest queue, which leads to more balance in the system and probably less frequently to the 'bad' situations in which one server is idle while there are still waiting jobs at the other queue (in fact, these situations cause the difference between the SSQS and the $M|M|2$ queue), the mean $L_w^{TJ}(T)$ of the total number of waiting jobs in the TJ model is expected to produce a lower bound for L_w. The TB model is obtained by introducing a type of service blocking, and therefore the mean $L_w^{TB}(T)$ of the total number of waiting jobs in the TB model is expected to produce an upper bound for L_w. If $L_w^{TJ}(T)$ and $L_w^{TB}(T)$ produce a lower bound and an upper bound for L_w, then, by equation (1.19), the variables $W_{TJ}(T)$ and $W_{TB}(T)$ defined by $W_{TJ}(T) = L_w^{TJ}(T)/(2\rho)$ and $W_{TB}(T) = L_w^{TB}(T)/(2\rho)$, produce a lower bound and an upper bound for W; and, vice versa (note that, by these definitions, the variables $W_{TJ}(T)$ and $W_{TB}(T)$ are precisely equal to the

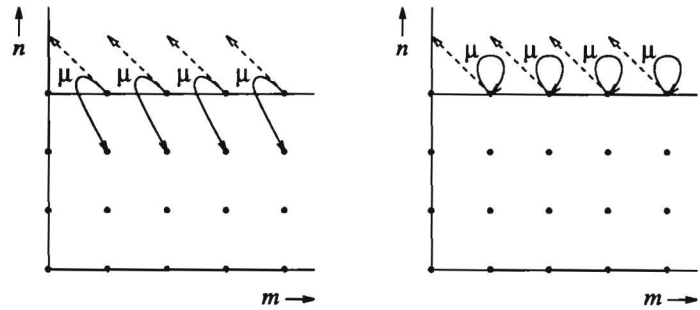


Figure 1.8. The two truncation models, both with threshold parameter $T=3$, for the SSQS with $N=2$ servers. The dashed arrows denote transitions of the original model to states outside the truncated state space and the corresponding uninterrupted arrows show how these transitions are modified.

mean normalized waiting times in the TJ and TB model itself). Except that the two bound models are expected to produce bounds for L_w and W, it is also expected that the larger T, the more accurate the bounds will be. So, our *conjecture* is that

$$0 \leq L_w^{TJ}(1) \leq \cdots \leq L_w^{TJ}(T) \leq L_w^{TJ}(T+1) \leq \cdots \leq L_w \leq \cdots \leq L_w^{TB}(T+1) \leq L_w^{TB}(T) \leq \cdots \leq L_w^{TB}(1), \tag{1.20}$$

which, by (1.19) and the definitions for $W_{TJ}(T)$ and $W_{TB}(T)$, is equivalent to the conjecture that

$$0 \leq W_{TJ}(1) \leq \cdots \leq W_{TJ}(T) \leq W_{TJ}(T+1) \leq \cdots \leq W \leq \cdots \leq W_{TB}(T+1) \leq W_{TB}(T) \leq \cdots \leq W_{TB}(1). \tag{1.21}$$

It is noted that, since both truncation models are identical to the SSQS in case $T=\infty$, for both truncation models the produced bounds will tend to L_w and W as $T \to \infty$. Together with the conjectures stated in (1.20) and (1.21), this implies that

$$L_w^{TJ}(T) \uparrow L_w \quad \text{and} \quad L_w^{TB}(T) \downarrow L_w, \quad \text{as } T \to \infty, \tag{1.22}$$

$$W_{TJ}(T) \uparrow W \quad \text{and} \quad W_{TB}(T) \downarrow W, \quad \text{as } T \to \infty. \tag{1.23}$$

The latter property for the bounds $W_{TJ}(T)$ and $W_{TB}(T)$ shows that the mean normalized waiting time W may be determined within an arbitrary, desired accuracy by solving both bound models for increasing values of T. Further, it is noted that since the behavior of the $M|M|2$ queue may be described by a Markov process that is equivalent to the Markov process for the TJ model with $T=1$, the means for the total number of waiting jobs and the normalized waiting time for the $M|M|2$ queue are the same as for the TJ model with $T=1$; and, thus, the conjectures stated in (1.20) and (1.21) also imply the intuitively obvious result that the means for the total number of waiting jobs and the normalized waiting time for the $M|M|2$ queue are smaller than or equal to the corresponding quantities L_w and W for the SSQS.

Let us now discuss the so-called *precedence relation method.* This method in principle is an *analytical* method for the comparison between an original Markov cost model on a state space M and a second model Markov cost model on a state space M' which is a subset of M ($M'=M$ is also allowed). The *main idea* of the precedence relation method is that this comparison may be based on so-called *precedence pairs* of states of the original model, which satisfy a certain *precedence relation* which denotes that the first state of a pair is more attractive with respect to certain costs than the second state. The precedence pairs can be derived in a first, preliminary, step. After that, in a second step, they can be exploited to compare the costs (or the relevant performance measures) in the second model to the corresponding costs (or the relevant performance measures) in the original model (if the second model is a truncation model, then this second step appears to result in an extremely simple step; see step 2 as described in the next paragraph). The precedence relation method will be described in detail in Chapter 5 of this monograph. It will be developed mainly for the comparison of performance measures for truncation models to the corresponding quantities for their original model. Here, in this section, we shall only globally show how the precedence relation method may be used to prove the monotonicity results as stated in (1.20) and (1.21).

The precedence relation method, as it is used for the comparison of performance measures for truncation models to the corresponding performance measures for the original model, consists of the following two steps:

1. *The derivation of* precedence pairs *for the original model (i.e. proving for pairs of states of the original model that they satisfy the given* precedence relation*);*
2. *Establishing, for each truncation model, whether each transition which originally ended in a state n outside the truncated state space M', has been redirected to a state n' inside M' which is a* more/less attractive *state than the state n according to the precedence pairs derived in step 1.*

Usually, most of the effort has to be devoted to the first step. Further, it is noted that which precedence pairs can be derived depends on, among others, the performance measures which are considered. Below, both the first and the second step are further explained on the basis of the comparison of the TJ and TB model to the original SSQS.

The first step of the precedence relation method consists of the derivation of precedence pairs for the original model, which satisfy a certain, given precedence relation. Let us consider the mean L_w of the total number of waiting jobs in the original SSQS. Then the corresponding precedence relation is satisfied by the following pairs of states. It can be shown that a state (m,n) is a more attractive state than its neighbor $(m,n+1)$, which represents a situation with one extra job in the longest queue. This means that the prospects with respect to the total number of waiting jobs are better when being in the state (m,n) than when being in the state $(m,n+1)$. Further, it can be shown that a state $(m,n+1)$ is more attractive than a state $(m+1,n)$, which represents a situation with one extra job in the shortest queue, and that a state $(m+1,n)$ is more attractive than a state $(m,n+2)$, which represents a situation with the same total number of jobs in the system, but with *less balance* between the queue lengths.

In the second step, for a truncation model which is expected to be a lower bound model (upper bound model), it must be established whether the transitions which originally were ending in states outside the truncated state space, have been redirected to more (less) attractive states. If so, then for the truncation model, independently of the starting state, the future expectations are better (worse) than for the original model, and therefore lower bounds (upper

bounds) for the performance measure(s) under consideration are obtained. In the TJ model, the transitions starting in the states (m,T), $m \geq 1$, and originally ending in the states $(m-1,T+1)$, have been redirected to the states $(m,T-1)$. According to the precedence pairs obtained in the first step, a state $(m,T-1)$ is more attractive than a state $(m-1,T+1)$, and therefore the TJ model leads to a lower bound for the mean number L_w of waiting jobs in the system. In the TB model, the transitions from the states (m,T), $m \geq 1$, to the states $(m-1,T+1)$, have been redirected to the less attractive states (m,T), by which the TB model leads to an upper bound for L_w. So, this indicates that

$$L_w^{TJ}(T) \leq L_w \quad \text{and} \quad L_w^{TB}(T) \geq L_w \quad \text{for all } T \geq 1. \tag{1.24}$$

By considering the TJ model with threshold parameter T as a truncation model of the TJ model with threshold parameter $T+1$, and similarly for the TB model, it may be shown that

$$L_w^{TJ}(T) \leq L_w^{TJ}(T+1) \quad \text{and} \quad L_w^{TB}(T) \geq L_w^{TB}(T+1) \quad \text{for all } T \geq 1. \tag{1.25}$$

Together with (1.24), (1.25) implies the monotonicity results as stated in (1.20), and therefore also the monotonicity results as stated in (1.21) for the mean normalized waiting time.

Due to the simplicity of the second step of the precedence relation method as described above for *proving* that a truncation model is a bound model, the precedence relation method is also very appropriate for *deriving* (or *constructing*) bound models, and especially for deriving *flexible* bound models. The *precedence relation method for deriving flexible bound models* consists of the following two steps (note that step 2 is a constructive step in this case):

1. *The derivation of precedence pairs for the original model;*
2. *The definition of flexible lower and upper bound models: to obtain a flexible lower (upper) bound model, first a flexible truncated state space M' must be defined, and next each transition from a state m inside M' to a state n outside M' must be redirected to a state n' inside M' which, according to the precedence pairs derived in step 1, is more (less) attractive than the state n in which the transition originally ended.*

In this way, once the precedence pairs have been derived, a whole set of flexible bound models can be obtained.

In Chapter 5, the precedence relation method is described in detail (see also [79, 81]). In that chapter, the SSQS with $N=2$ servers will serve as an illustration model, and the precedence relation method will be used to derive *six* flexible bound models, among which the TJ model and the TB model (and also the truncation models as presented by Conolly [24] and Rao and Posner [60]). In Chapter 5, we shall also see that the precedence relation method is appropriate for deriving several other results than those given by (1.24) and (1.25).

After having indicated that the TJ model and the TB model indeed give bounds for the means of the total number of waiting jobs and the normalized waiting time, we would like to know whether they also give *accurate* bounds. We shall investigate this for the mean normalized waiting time W by computing the bounds $W_{TJ}(T)$ and $W_{TB}(T)$ for varying values of ρ and T and comparing them to the values which are found for W. As noted earlier in this section, both bound models can be solved very efficiently by the matrix-geometric approach (see also Chapter 6).

The values for W itself can be determined by solving the model for the SSQS by the compensation approach, or by solving the bound models for large values of T. In Table 1.8,

the accuracy of the bounds $W_{TJ}(T)$ and $W_{TB}(T)$ is depicted for varying values of ρ and T. Note that, due to the blocking, which leads to a destruction of capacity, the TB model may be not ergodic for large values of ρ and/or small values of T, in which case $W_{TB}(T)=\infty$ and a value ∞ is obtained for the difference $W_{TB}(T)-W$. In Figure 1.9, the values for the bounds and W itself are depicted graphically. In this figure, a logarithmic axis has been taken for ρ in order to blow up the relevant region near $\rho=1$.

The numerical results illustrate that the bounds are tight for already small values of the threshold parameter T, which seems to be due to the appropriate choice for the truncated state space M' for both bound models (by which only few redirections occur). Contrary to the bounds obtained from the TB model, the tightness of the bounds obtained from the TJ model appears to decrease only slowly as the workload ρ increases to its maximum value, by which for large workloads ρ the bounds obtained from the TJ model are considerably tighter than the bounds obtained from the TB model; this indicates that for large workloads ρ the jockeying of jobs in the TJ model has a much smaller impact than the destruction of capacity in the TB model. Further, the numerical results show that the larger ρ, the larger T must be taken to obtain bounds of which the difference with respect to W is smaller than a given, required absolute or relative accuracy.

Due to the tightness of the bounds $W_{TJ}(T)$ and $W_{TB}(T)$, an efficient numerical procedure for the determination of the mean normalized waiting time W within a given absolute or relative accuracy, is obtained by computing the lower bound $W_{TJ}(T)$ and the upper bound $W_{TB}(T)$ for increasing values of T. Here, for each fixed value of T, W may be approximated by the mean $(W_{TJ}(T)+W_{TB}(T))/2$, and upper bounds for the absolute and relative inaccuracy of this approximation are given by $(W_{TB}(T)-W_{TJ}(T))/2$ and $(W_{TB}(T)-W_{TJ}(T))/(2W_{TJ}(T))$, respectively. The computation process may be stopped as soon as the desired accuracy is reached for some T.

We end this section with some remarks on the precedence relation method, as used for deriving flexible bound models. In principle, this method can be applied to any Markovian (queueing) system, but it depends on the structure of a particular problem whether the precedence relation method can lead to flexible bound models which are appropriate for the determination of the relevant performance measure(s). We are satisfied, if we can determine the relevant performance measure(s) of the original system *in an analytical way* (this may be possible in case the bound models can be solved analytically) *or by an efficient numerical procedure* (for this it is required that the bound models can be solved efficiently by a standard numerical technique, for example, and that they closely approximate the original model for relatively small values of the parameters which determine the flexible sizes of the truncated state spaces). In this section, for the SSQS with $N=2$ servers, we have been able to derive appropriate, flexible bound models by exploiting the property that, due to the shortest queue routing, most of the probability mass is concentrated around the horizontal axis. These flexible bound models were named the TJ model and the TB model, and they have appeared to be appropriate for being used in an efficient numerical procedure for the determination of the mean normalized waiting time. In the Chapters 6 and 7 of this monograph, we will apply the precedence relation method in order to obtain appropriate, flexible bound models for the determination of the waiting times in the SSQS and the SQS-JDP, both with $N\geq 2$ servers. In Chapter 6, we shall see that the TJ model and the TB model are also appropriate for the N-

ρ	$W_{TJ}(T)-W$				W	$W_{TB}(T)-W$			
	$T=1$	$T=3$	$T=5$	$T=10$		$T=10$	$T=5$	$T=3$	$T=1$
0.1	-0.008	-0.000	-0.000	-0.000	0.018	+0.000	+0.000	+0.000	+0.003
0.2	-0.024	-0.000	-0.000	-0.000	0.066	+0.000	+0.000	+0.000	+0.026
0.3	-0.045	-0.000	-0.000	-0.000	0.144	+0.000	+0.000	+0.000	+0.100
0.4	-0.068	-0.001	-0.000	-0.000	0.259	+0.000	+0.000	+0.001	+0.296
0.5	-0.093	-0.003	-0.000	-0.000	0.426	+0.000	+0.000	+0.004	+0.824
0.6	-0.119	-0.006	-0.000	-0.000	0.682	+0.000	+0.000	+0.017	+2.731
0.7	-0.147	-0.012	-0.001	-0.000	1.108	+0.000	+0.002	+0.058	+67.899
0.8	-0.178	-0.022	-0.002	-0.000	1.956	+0.000	+0.012	+0.227	$+\infty$
0.9	-0.212	-0.037	-0.005	-0.000	4.475	+0.000	+0.096	+1.500	$+\infty$
0.95	-0.230	-0.046	-0.007	-0.000	9.487	+0.002	+0.504	+10.345	$+\infty$
0.98	-0.242	-0.053	-0.009	-0.000	24.494	+0.012	+3.828	$+\infty$	$+\infty$
0.99	-0.246	-0.055	-0.010	-0.000	49.497	+0.053	+18.737	$+\infty$	$+\infty$

Table 1.8. The values for the mean normalized waiting time W for the SSQS with $N=2$ servers and for the differences between the bounds $W_{TJ}(T)$ and $W_{TB}(T)$ and W itself.

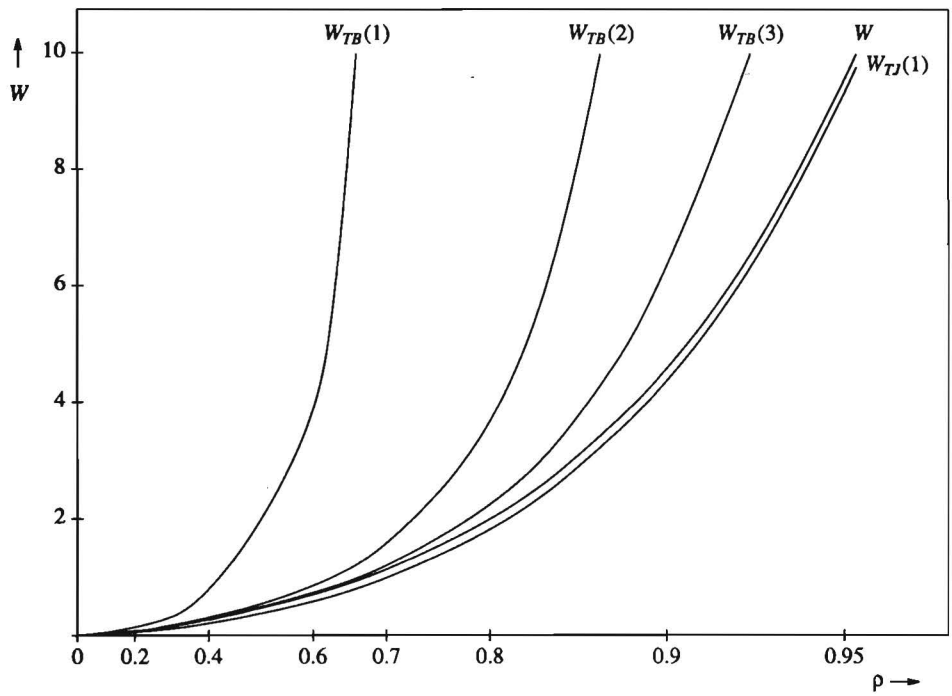


Figure 1.9. The mean normalized waiting time W and the bounds $W_{TJ}(T)$ and $W_{TB}(T)$ for the SSQS with $N=2$ servers.

dimensional SSQS (see also [2]), and in Chapter 7, similar bound models are shown to be appropriate for the SQS-JDP with N servers. We finally remark that in [1] appropriate, flexible bound models have been derived for the symmetric *longest* queue system.

1.5. Outline

This monograph is devoted to two methods for the analysis of multi-dimensional Markov processes/random walks, which are used for describing the behavior of queueing systems, for example. The main ideas of both methods have been described in this introductory chapter.

In the first part of this monograph, consisting of the Chapters 2-4, we shall extend the *compensation approach*, as developed by Adan et al. [12] for the class of 2-dimensional, homogeneous, nearest-neighboring random walks, to the corresponding class of N-dimensional random walks. An additional property, called the projection property, is introduced to avoid complex notations and to simplify the analysis. The analysis leads to two main results. First of all, it is shown that the compensation approach works for a random walk of the considered class if and only if from states $(m_1, \ldots, m_N)$ in the interior no transitions can be made into directions which for some pair of components m_i and m_j enlarge the distance to the origin, which generalizes the condition derived in [12] for the case $N=2$ (see also condition (1.18)). Secondly, it is shown that the equilibrium distribution of a random walk satisfying this condition can be expressed as an alternating sum of infinitely many, pure product-form distributions, which are obtained from $(N-1)$-fold trees of product-form solutions of the equilibrium equation for the interior. The results are proved by induction with respect to the dimension N. The case $N=2$ is treated in Chapter 2. The step from $N=2$ to $N=3$, which contains all elements of the general step, is described in Chapter 3. In Chapter 4, an extensive analysis of the structure of the solution obtained for the equilibrium distribution will be presented; this structure analysis will lead to the development of efficient numerical procedures for the computation of the equilibrium distribution and related quantities.

The second part of this monograph, consisting of the Chapters 5-7, is devoted to the so-called *precedence relation method*. This method is an analytical method, which is based on Markov cost/reward theory, and which is appropriate for comparing the costs of two Markov cost models, of which the state space of one model is a subset of the state space of the other model. We shall mainly focus on how the precedence relation method may be used for deriving flexible truncation models which lead to lower and upper bounds for the relevant performance measure(s) of the original model. Such models are called *flexible lower and upper bound models*, and, provided that they can be solved, they may be used to determine the relevant performance measure(s) of the original model (which itself may be a model that cannot be solved). The precedence relation method is developed in Chapter 5, and in the Chapters 6 and 7 the method is applied to the well-known Symmetric Shortest Queue System (SSQS), consisting of $N \geq 2$ parallel servers, and to a generalization of it, called the Shortest Queue System with a Job-Dependent Parallelism (SQS-JDP). For both systems, we derive flexible bound models which lead to efficient numerical procedures for the determination of the mean waiting times within a given accuracy.

Finally, in Chapter 8, we draw the main conclusions and we discuss some extensions which might be interesting for future research.

Chapter 2

The Compensation Approach for a Class of Two-Dimensional Random Walks

2.1. Introduction

For several queueing systems, the behavior may be described by Markov processes/random walks with a discrete, possibly multi-dimensional, state space. Therefore, much effort has been put in investigating the equilibrium distribution of such random walks. Explicit expressions for the equilibrium distribution have been derived for many one-dimensional systems. However, for the multi-dimensional case, only a few results are available. For *product-form networks*, which are modeled by random walks with a dimension that is at least equal to the number of queues in the network, it has been shown that the equilibrium distribution can be written as a product of powers of fixed factors (see Baskett et al. [15]). Further, explicit expressions have been derived for some multi-dimensional random walks with a state space being infinite in only one dimension (see [4], and also Bertsimas [16], for a treatment of the $E_k | E_r | c$ queue; see [13] for a treatment of the shortest queue system with N servers and threshold jockeying). Finally, explicit expressions, in the form of infinite series of products of powers of fixed factors, are available for a number of two-dimensional random walks with a state space being infinite in both dimensions (for example, as we have shown in Chapter 1, for the symmetric shortest queue system, for which explicit expressions may be constructed step by step with the help of the compensation approach). This type of random walks is considered in this chapter.

The *body of this chapter* consists of the application of the compensation approach to a subclass of the class of two-dimensional random walks studied by Adan et al. [12] (see also [3]), and serves as a preparation for the analysis in Chapter 3, where a class of two- and higher-dimensional random walks, with a state space being infinite in each dimension, is studied.

For two-dimensional random walks such as the one describing the behavior of the symmetric shortest queue system (see Figure 1.4), there essentially exist *three methods* to obtain explicit expressions for the equilibrium distribution. The most recent method is the compensation approach, developed by Adan et al. [12] for the class of homogeneous, nearest-neighboring random walks which satisfy the *additional condition* that no transitions from interior points are possible to the North, East and North-East (cf. (1.18)). This method may be characterized as a direct approach for solving the equilibrium equations without resorting

to generating functions. The two other methods are indirect, complex-variable methods, which focus on explicit expressions for the generating function of the equilibrium distribution. Let us shortly discuss the main results for all three methods.

The oldest of the complex-variable approaches is the *uniformization technique*, which has been developed by Kingman [49] and Flatto and McKean [33] for the symmetric shortest queue system with two servers. For the generating function $f(x,y)$ of the equilibrium distribution of the lengths of the two queues, they derive a functional equation with as unknown functions $f(x,y)$ on one side and the generating functions $f(x,0)$ and $f(0,y)$ for the equilibrium probabilities on the axes on the other side of the equation. The functions $f(x,0)$ and $f(0,y)$ and, hence, $f(x,y)$ are shown to be meromorphic, and explicit formulae are derived for the poles and their residues after having introduced a uniformizing variable. By decomposing the meromorphic function $f(x,y)$ into partial fractions it follows that the equilibrium probabilities may be written as infinite linear combinations of product forms. The same technique has been used by Hofri [41] for a multiprogramming queueing model (see also [11]) and by Jaffe [47] for the 2×2 clocked buffered switch. All three cases for which the uniformization technique has been worked out, have the property that there are no transitions from interior points to the North, East and North-East. In all three cases the generating function is meromorphic and partial fraction decomposition of this function yields expressions for the equilibrium probabilities in the form of infinite linear combinations of product forms, although it appears to be difficult to give explicit formulae for the coefficients of the linear combinations.

The uniformization technique has also been employed by Flatto and Hahn [32] to analyze the fork and join model with two servers. They show that the generating functions $f(x,0)$ and $f(0,y)$ can be extended to multiple-valued algebraic functions. However, partial fraction decomposition is not available for multiple-valued functions, hence it is no longer possible to derive exact formulae for the equilibrium probabilities via this decomposition. Recently, Wright [82] analyzed a generalization of the fork and join model by using the uniformization technique. In his analysis he encounters the same difficulties as Flatto and Hahn [32], i.e. multiple-valued functions. Until recently, there has been no general result for two-dimensional random walks based on a derivation via the uniformization technique, although it seems possible to derive a general result for cases which satisfy the additional condition that no transitions from interior points are possible to the North, East and North-East. Nevertheless, the results of Kingman [49] and Flatto and McKean [33] for the symmetric shortest queue problem inspired the compensation procedure (see [8] and [12]), which indeed gives explicit formulae for the coefficients of the linear combination for all cases satisfying the additional condition.

Extension of the uniformization technique to random walks with more than two dimensions has never succeeded. One reason for this failure might be that the additional condition has never appeared as essential in these investigations, whereas the shortest queue problem with N servers ($N>2$) appears not to satisfy the generalization of the extra condition to higher dimensions (this generalization is derived in the next chapter; see Condition 3.1). Recently, inspired by [12], Cohen [22] has shown that a technique, which is actually a direct generalization of the uniformization technique, may be used for the same class of problems as the class to which the compensation approach is applicable (which has to do with the resemblance between both methods; see Section 2.6). Therefore we can conclude that for a two-

dimensional, irreducible, positive recurrent, homogeneous, nearest-neighboring random walk the compensation approach and the uniformization technique are only usable, in the sense that they give explicit expressions for the equilibrium probabilities in the form of series of products of powers, if there are no transitions from interior points to the North, East and North-East. But, if this condition is satisfied, then these two methods are very suitable. Further, we believe that the compensation approach is preferable to the uniformization technique, since it leads to more explicit results (explicit formulae for *all* equilibrium probabilities, for example) and it avoids complex analysis.

A more recent indirect method for solving the functional equation for the generating function of the equilibrium distribution, is the *boundary value method.* This method aims at reducing the functional equation to a standard problem of the theory of boundary value problems and integral equations for complex functions and has established itself as a powerful method for a large class of two-dimensional random walks in the first quadrant; see Cohen and Boxma [23]. Queueing problems solved by the boundary value method are the symmetric shortest queue model, the M/G/2 queue, a polling model with two queues and 1-limited service (see [23] for all these examples), the coupled processor model (see [23], the work of Fayolle and Iasnogorodski [29,30,45] and also Konheim et al. [51]), the longest queue model with nonpreemptive priority (see Cohen [20]; the longest queue model with preemptive priority has been treated by Zheng and Zipkin [85], who solve the equilibrium equations iteratively, and by Flatto [31], who explicitly solves the functional equation for the generating function), the fork and join model (see De Klein [25]), and the 2×2 clocked buffered switch (see Jaffe [46]). For a review of the boundary value method for two-dimensional problems, see Cohen [21]. Some examples mentioned show already that the boundary value method is not restricted to random walks without transitions from interior points to the North, East and North-East. It seems to be the only really general method for two-dimensional random walks on the integer grid in the positive quadrant. However, the compensation method gives more complete results in the cases in which it works. Concerning extensions of the boundary value method to higher dimensions, the review paper [21] states that it should be possible in principle, but the mathematical as well as the numerical analysis becomes very intricate.

Let us finally discuss the main results obtained by the *compensation approach.* In Adan et al. [12], the compensation approach has been applied to the class of two-dimensional, homogeneous, nearest-neighboring random walks on the integer grid in the first quadrant of the plane; see Figure 1.6. This method starts with characterizing the product forms which satisfy the equilibrium equations in the interior points ($m \geq 2$, $n \geq 2$). Subsequently, it is attempted to construct an infinite series of such solutions which also satisfies the boundary equations. The construction starts by taking a product form which satisfies the interior equations as well as the equations for one of the boundaries. It is then corrected by adding a product form which not only satisfies the interior equations, but also makes the sum satisfy the equations on the other boundary. Then a new correction term is added to make the solution again satisfy the equations on the first boundary, etc. Requirements for this method to work are:

1. In each step it should be possible to find a new correction term which satisfies the needs;
2. The resulting series should converge.

In [12], it appeared that these requirements are fulfilled if and only if the random walk is irreducible, positive recurrent and satisfies the additional condition that no transitions from interior points can be made to the North, East and North-East. The latter condition stems from convergence requirements for the infinite linear combinations of product forms constructed by the compensation approach, and certainly limits the applicability of this method. However, if this condition is satisfied, then the compensation approach is very powerful. Application of this method shows that the equilibrium distribution consists of a linear combination of at most four series of product-form solutions and explicit formulae are produced for all coefficients and factors. It is noted that, if the additional condition is not satisfied, then the equilibrium distribution may be expected to have a more complicated structure than a linear combination of product-form distributions (this conjecture follows from a study of the ratios of numerically determined equilibrium probabilities for various instances for which the additional condition is not satisfied).

There are a number of well-known queueing problems present in the class of two-dimensional, homogeneous, nearest-neighboring random walks, as studied in [12]. For these problems, the additional condition stemming from the convergence requirements is satisfied by the symmetric shortest queue problem, Hofri's multiprogramming queues model and the 2×2 clocked buffered switch (see [5, 8, 19]), while this condition is violated by the coupled processor model, the longest queue model, and the fork and join model (and, as we observed in Chapter 1, also by the two-dimensional, symmetric shortest queue system with a job-dependent parallelism; see Figure 1.7). Except for the problems belonging to the class studied in [12] and satisfying the additional condition, the compensation approach has appeared to work also for some other two-dimensional problems; see [6, 9]. From the analysis in [6], it follows that the restriction to nearest-neighbor transitions is not essential, however, it simplifies the arguments considerably. Particularly, to find good starting solutions becomes much more complex in the other case with not only nearest-neighbor transitions. From the analysis in [9], it follows that the compensation approach can also be used for random walks on integer grids of a more complex form. Contrary to the complex-variable methods, extension of the compensation approach to higher-dimensional random walks appears to be possible, as will be shown in the next chapter. But, the main question remains: under which condition does it work?

The *main objective* of this chapter is to derive, by using the compensation approach, explicit expressions for the equilibrium distribution for a relevant subclass of the class of random walks studied in [12]. This derivation and the explicit expressions for the equilibrium distribution serve as a preparation for the analysis presented in the next chapter. In that chapter, the so-called *projection property* will be introduced to avoid complex notations and to simplify the analysis, and therefore this property will also be introduced here. Moreover, from the beginning, we shall restrict ourselves to the class of random walks satisfying the condition under which the compensation approach works. The assumed properties are satisfied by the 2×2 switch, for which the compensation approach has been worked out in detail by Boxma and Van Houtum [19]. In fact, this chapter largely coincides with that paper.

The organization of this chapter is as follows. In Section 2.2, we present in detail the class of random walks for which we shall describe the compensation approach, and we derive

the equilibrium equations. The equilibrium equations are solved in the Sections 2.3 and 2.4, by using the compensation approach; in its application several simplifications arise which are due to the projection property. In Section 2.3, we show that for the present model the compensation approach generates two alternating series of pure, two-dimensional product-form (geometric) distributions, and, in Section 2.4, we prove that the equilibrium distribution $\{p_{m,n}\}$ is obtained by simply taking the sum of these two series:

$$p_{m,n} = \sum_{i=0}^{\infty}(1-\beta_i)\beta_i^n\,[(1-\alpha_i)\alpha_i^m - (1-\alpha_{i+1})\alpha_{i+1}^m] + \sum_{i=0}^{\infty}(1-\hat{\alpha}_i)\hat{\alpha}_i^m\,[(1-\hat{\beta}_i)\hat{\beta}_i^n - (1-\hat{\beta}_{i+1})\hat{\beta}_{i+1}^n]\,, \quad m \geq 0,\ n \geq 0,\ m+n \geq 1. \tag{2.1}$$

The α_i, β_i, $\hat{\alpha}_i$ and $\hat{\beta}_i$ are specified in Section 2.3 and the convergence of the above sums for all $m \geq 0$, $n \geq 0$ with $m+n \geq 1$ is discussed in Section 2.4. Due to divergence of the series of product forms for $m=n=0$, formula (2.1) does not hold for $p_{0,0}$. The main results obtained from the application of the compensation approach, are summarized in the *Main Theorem* at the end of Section 2.4. In the Sections 2.5 and 2.6, some additional results are derived. In Section 2.5, we derive error bounds for the computation of the series which constitute the equilibrium distribution, and numerical results are presented for the 2×2 buffered switch, which belongs to the class studied in this chapter. In Section 2.6, on the basis of the results which are obtained for the symmetric 2×2 buffered switch, the compensation approach is compared to the two complex-variable methods, which have been used by Jaffe [46,47] for this problem. Finally, Section 2.7 is devoted to the conclusions.

2.2. The class of two-dimensional random walks

In this section, we describe a class of two-dimensional random walks, for which we shall show in detail how the compensation approach works. Three examples of queueing systems are presented to illustrate for which kind of problems the assumed properties are satisfied. Finally, the equilibrium equations are formulated, which will be used by the compensation approach to determine the equilibrium distribution.

For the class of two-dimensional random walks considered in this chapter, we shall assume some properties which are rather natural for random walks stemming from queueing systems. Since several queueing systems can be modeled as discrete-time or continuous-time Markov processes/random walks on the lattice of the first quadrant, the state space M is assumed to be equal to

$$M = \{\,(m,n) \mid m,n \in \mathbb{N}_0\,\},$$

where $\mathbb{N}_0$ is the set of nonnegative integers. Further, some reasonable assumptions may be made with respect to the transitions and the corresponding probabilities/rates.

The components m and n of the states (m,n) of a random walk describing the behavior of a queueing system, usually represent quantities such as queue lengths, which in most cases leads to a certain *homogeneity* in the transition probabilities/rates. This means that the state space may be partitioned into a finite number of subsets consisting of states with the same

outgoing transition probabilities/rates. We shall assume that the same transition probabilities/rates occur for all states in the interior, i.e. for all states (m,n) with $m,n \geq 1$, and similarly for all states $(m, 0)$, $m \geq 1$, on the horizontal boundary and for all states $(0,n)$, $n \geq 1$, on the vertical boundary. In case the components m and n represent queue lengths, this assumption means that the same transition probabilities/rates occur for all states corresponding to situations with the same set of empty queues.

The assumed homogeneity in the transition probabilities/rates is *essential* for having a chance that the equilibrium distribution can be determined by using the compensation approach. The other two assumptions which we make with respect to the transitions, and the corresponding probabilities/rates, are mainly made to simplify the analysis. These assumptions are also satisfied by several random walks stemming from queueing systems. They read as follows. We assume that only transitions to *nearest neighbors* occur, i.e. that for each state $(m,n) \in M$ it is only possible to jump to states $(m',n') \in M$ with $|m'-m| \leq 1$ and $|n'-n| \leq 1$, and further we assume that the so-called *projection property* is satisfied. This projection property may be formulated only for random walks with homogeneity in the transition probabilities/rates and is explained below.

The first two assumptions on the transitions imply that we have a homogeneous, nearest-neighboring random walk, as depicted in Figure 1.6. Let the transition probabilities/rates for the interior, the horizontal boundary, the vertical boundary and the origin be denoted by the variables q_{t_1,t_2}, h_{t_1,t_2}, v_{t_1,t_2} and o_{t_1,t_2}, respectively. Then the projection property means the following. For the horizontal boundary, the transition probabilities/rates $h_{t,1}$ are the same as the probabilities/rates $q_{t,1}$ and the probabilities/rates $h_{t,0}$ are equal to the sums of $q_{t,0}$ and $q_{t,-1}$. One might say that for the horizontal boundary, the set of transitions, accompanied by the corresponding probabilities/rates, is obtained by pushing the set of transitions for the interior against this horizontal boundary. Similarly, the set of transitions for the vertical boundary is obtained by pushing the set of transitions for the interior against the vertical boundary. For the origin the impact of the projection property is a little bit more complex. The set of transitions for the origin is obtained by pushing the set of transitions for the horizontal boundary against the vertical boundary, or by pushing the set of transitions for the vertical boundary against the horizontal boundary. We say that for both the horizontal boundary and the vertical boundary, the set of transitions is a kind of *projection* of the set of transitions for the interior; and, similarly the set of transitions for the origin is the projection of the set of transitions for the horizontal boundary as well as the set of transitions for the vertical boundary. In Figure 2.1, we have depicted the random walk that is obtained in the end. Note that for a random walk with the projection property all transition probabilities/rates are uniquely determined by the transition probabilities/rates for the interior.

A formal description of the three assumptions on the transitions is given in Assumption 2.1. After that, three examples of queueing systems are presented. All three systems satisfy the Assumptions 2.1(i) and 2.1(ii), and two of them also satisfy Assumption 2.1(iii).

Assumption 2.1.

(i) For all states only transitions to nearest neighbors occur, i.e. for all states $(m,n) \in M$, only transitions to states $(m',n') \in M$ with $|m'-m| \leq 1$ and $|n'-n| \leq 1$ occur;

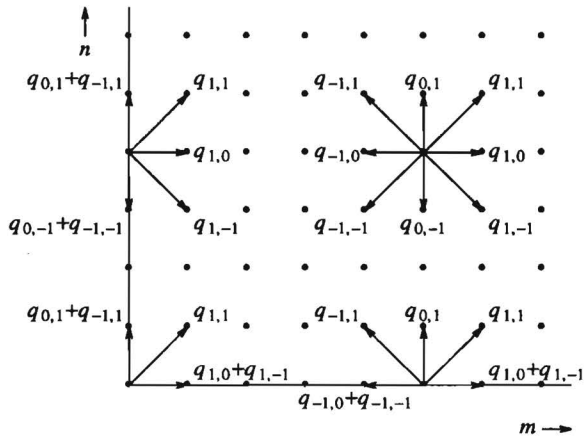


Figure 2.1. The transition probabilities/rates for a two-dimensional, homogeneous, nearest-neighboring random walk with the projection property; for all states the transitions to themselves have been left away.

(ii) Homogeneity: for the interior, the horizontal boundary as well as the vertical boundary, all states have the same outgoing transition probabilities/rates; the probabilities/rates for the transitions from an interior state $(m,n) \in M$, $m,n \geq 1$, *to the states* $(m+t_1, n+t_2)$, $t_1, t_2 \in \{-1,0,1\}$, *are denoted by* q_{t_1,t_2}, *and similarly the transition probabilities/rates for the horizontal boundary and the vertical boundary are denoted by* h_{t_1,t_2} *and* v_{t_1,t_2}, *respectively; let the transition probabilities/rates for the origin be denoted by* o_{t_1,t_2};

(iii) Projection property: the transition probabilities/rates q_{t_1,t_2}, h_{t_1,t_2}, v_{t_1,t_2} *and* o_{t_1,t_2} *satisfy the following equations:*

$$h_{t,1} = q_{t,1} \quad \text{and} \quad h_{t,0} = q_{t,0} + q_{t,-1} \quad \text{for all } t \in \{-1,0,1\},$$

$$v_{1,t} = q_{1,t} \quad \text{and} \quad v_{0,t} = q_{0,t} + q_{-1,t} \quad \text{for all } t \in \{-1,0,1\},$$

$$o_{1,1} = q_{1,1}, \quad o_{0,1} = q_{0,1} + q_{-1,1}, \quad o_{1,0} = q_{1,0} + q_{1,-1} \quad \text{and}$$

$$o_{0,0} = q_{0,0} + q_{-1,0} + q_{0,-1} + q_{-1,-1}.$$

Example 2.1: The symmetric shortest queue system

This system has been described at the end of Section 1.2 and the beginning of Section 1.3. The transition rates have been depicted in Figure 1.4, and it is easily seen that the Assumptions 2.1(i) and 2.1(ii) are satisfied. The projection property appears to be violated; however, the projection property is not essential for the application of the compensation approach and violation of the projection property only leads to more complex formulae for the equilibrium distribution (compare the expressions obtained in Section 1.3 (see (1.17) and the form of the solution $\{x_{m,n}\}$) to formula (2.1)).

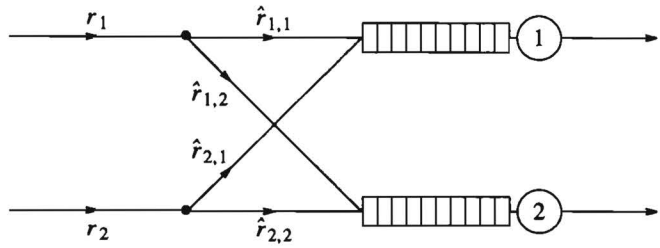


Figure 2.2. The 2×2 switch consisting of 2 parallel servers, where 2 types of jobs arrive.

Example 2.2: The 2×2 switch

The 2×2 clocked buffered switch is a discrete-time queueing system with 2 types of arriving jobs and 2 parallel servers; see Figure 2.2. Jobs of type j, $j=1,2$, arrive according to a Bernoulli stream with rate r_j, $0<r_j\le 1$, i.e. every time unit (clock cycle) the number of arriving jobs of type j is one with probability r_j and zero with probability $1-r_j$. Upon arrival a job of type j joins the queue at server i, $i=1,2$, with probability $\hat{r}_{j,i}$, $\hat{r}_{j,i}>0$, where $\hat{r}_{j,1}+\hat{r}_{j,2}=1$ for $j=1,2$. As a result, every time unit the number of arriving jobs of type j at server i is one with probability $r_{j,i}=r_j\hat{r}_{j,i}$ and zero with probability $1-r_{j,i}$. Jobs are assumed to arrive at the beginning of a time unit and they are immediately candidates for service. Each server serves exactly one job per time unit, if one present. Since we want to have an ergodic system, it is assumed that $r_{1,i}+r_{2,i}<1$ for $i=1,2$. The behavior of the 2×2 switch is described by a discrete-time Markov process with states (m,n), where m and n denote the numbers of waiting jobs at server 1 and 2 at the beginning of a time unit (just before the arrival instant). The only positive transition probabilities for the states in the interior are

$$q_{1,-1}=r_{1,1}r_{2,1}\,,\quad q_{0,0}=r_{1,1}r_{2,2}+r_{1,2}r_{2,1}\,,\quad q_{-1,1}=r_{1,2}r_{2,2}\,,$$
$$q_{-1,0}=(1-r_1)r_{2,2}+(1-r_2)r_{1,2}\,,\quad q_{0,-1}=(1-r_1)r_{2,1}+(1-r_2)r_{1,1}\,,$$
$$q_{-1,-1}=(1-r_1)(1-r_2)\,.$$

It is easily verified that the 2×2 switch satisfies all three properties stated in Assumption 2.1.

Example 2.3: The fork and join model

This system consists of 2 parallel servers, where customers arrive according to a Poisson stream with intensity λ, $\lambda>0$. Each customer brings along 2 subjobs, one subjob for each server, and may leave the system if and only if both subjobs have been served. Each server uses a FCFS service discipline and for server i the service times are assumed to be exponentially distributed with mean $1/\mu_i$, $\mu_i>\lambda$ (which implies the ergodicity of the system). This system may be described by a continuous-time Markov chain with states (m,n), where m and n denote the numbers of unfinished subjobs at the servers 1 and 2. The positive transition rates for the interior points are

$$q_{1,1}=\lambda\,,\quad q_{-1,0}=\mu_1\,,\quad q_{0,-1}=\mu_2\,.$$

Also the fork and join model satisfies all three properties stated in Assumption 2.1.

In the paper by Adan et al. [12], the compensation approach has been applied to the class of random walks which satisfy the Assumptions 2.1(i) and 2.1(ii). They established that the equilibrium distribution can be obtained by means of the compensation approach, *if and only if* there are no transitions from the states in the interior to the North, East and North-East. Here, for simplicity, it is assumed from the beginning that this condition is satisfied.

Assumption 2.2.

$$q_{0,1} = q_{1,0} = q_{1,1} = 0. \tag{2.2}$$

Note that this assumption/condition is satisfied by the symmetric shortest queue system and the 2×2 switch. Assumption 2.2 is violated by the fork and join model (this could be expected, since the asymptotic formula derived by Flatto and Hahn (see [32], Theorem 7.1) for the equilibrium probability $p_{m,n}$ as m is fixed and $n \to \infty$ involves a factor $n^{3/2}$ (if $\mu_1 < \mu_2$), which suggests that its equilibrium distribution cannot be expressed as a linear combination of product forms $\alpha^m \beta^n$).

Apart from satisfying the Assumptions 2.1 and 2.2, the random walks which we consider are assumed to be *irreducible* and *positive recurrent* (= ergodic). Due to the projection property, we can give a simple, necessary and sufficient condition for the irreducibility and the positive recurrence. This simple condition is derived by analyzing the two one-dimensional random walks with the aggregate states associated with the marginal distributions; see the next two paragraphs.

Let $\{p_{m,n}\}$ be the equilibrium distribution of the (full) random walk and let its marginal distributions for the m-component and n-component be denoted by $\{p_m^{(1)}\}$ and $\{p_n^{(2)}\}$:

$$p_m^{(1)} = \sum_{n=0}^{\infty} p_{m,n} \quad \text{for all } m \geq 0, \qquad p_n^{(2)} = \sum_{m=0}^{\infty} p_{m,n} \quad \text{for all } n \geq 0.$$

Due to the projection property, for all states (m,n) with $m \geq 1$ the total probability/rate for transitions to states $(m+t,n')$ equals $q_t^{(1)} = q_{t,-1} + q_{t,0} + q_{t,1}$, where t is fixed and $t \in \{-1,0,1\}$, and for all states $(0,n)$ the total probability/rate for transitions to states (t,n') equals $q_t^{(1)}$ for $t = 1$ and $q_0^{(1)} + q_{-1}^{(1)}$ for $t = 0$; similarly for the n-direction. This shows that the distributions $\{p_k^{(i)}\}$ may be characterized as the equilibrium distributions of one-dimensional, homogeneous, nearest-neighboring random walks with the projection property; here, the transition probabilities/rates for the interior are given by the variables $q_t^{(i)}$ defined by

$$q_t^{(1)} = q_{t,-1} + q_{t,0} + q_{t,1}, \quad q_t^{(2)} = q_{-1,t} + q_{0,t} + q_{1,t} \quad \text{for } t \in \{-1,0,1\}.$$

It can be shown that the full random walk is positive recurrent if and only if both component random walks are positive recurrent, i.e. if and only if the component random walks have negative drifts (for a proof, see Malyshev [55], who discusses ergodicity conditions of two-dimensional random walks with bounded jumps). So, we obtain the following necessary and sufficient condition for the positive recurrence:

$$q_{-1}^{(i)} > q_1^{(i)} \quad \text{for all } i = 1,2. \tag{2.3}$$

If $q_1^{(i)} = 0$ for some $i = 1,2$, then all states (m,n) with $m \geq 1$ (if $i = 1$) or all states (m,n) with $n \geq 1$ (if $i = 2$) are transient, and we can restrict ourselves to a one-dimensional problem. To exclude this special case, we must require that $q_1^{(i)} > 0$ for all i. Together with (2.3), this

leads to the condition that

$$q_{-1}^{(i)} > q_{1}^{(i)} > 0 \quad \text{for all } i=1,2. \tag{2.4}$$

Given Assumption 2.2, condition (2.4) is necessary and sufficient for having an irreducible and positive recurrent random walk, since then $q_{1,-1} > 0$ and $q_{-1,1} > 0$ (without Assumption 2.2, (2.4) does not guarantee the irreducibility; in that case, one might have the situation with $q_{-1,-1} > q_{1,1} > 0$ and $q_{t_1,t_2} = 0$ for all other directions (t_1,t_2), for example).

Since the random walks which we consider are assumed to be irreducible and positive recurrent, condition (2.4) must be satisfied, and we obtain the following geometric distributions for the one-dimensional marginal distributions $\{p_k^{(i)}\}$:

$$p_k^{(i)} = \left[1 - \frac{q_1^{(i)}}{q_{-1}^{(i)}}\right]\left[\frac{q_1^{(i)}}{q_{-1}^{(i)}}\right]^k, \quad k \geq 0, \ i=1,2. \tag{2.5}$$

In the next sections, we apply the compensation approach to the class of two-dimensional random walks which are irreducible and positive recurrent and which satisfy the Assumptions 2.1 and 2.2, i.e. to the class of two-dimensional, irreducible, positive recurrent, homogeneous, nearest-neighboring random walks which satisfy the projection property and the property that from interior points no transitions are possible to the North, East and North-East. Note that the 2×2 switch belongs to this class. The equilibrium distribution $\{p_{m,n}\}$ for a random walk of this class is characterized as the unique normalized solution of the equilibrium equations, which are given below.

The equilibrium equations, and also some other formulae in the remainder of this chapter, slightly simplify in case we have the property that for each state the total probability/rate of outgoing transitions adds up to 1, i.e. in case

$$q_{-1,1} + q_{0,0} + q_{1,-1} + q_{-1,0} + q_{0,-1} + q_{-1,-1} = 1. \tag{2.6}$$

Therefore, from now on we assume that we have a *discrete-time* random walk, since then the property stated in (2.6) is satisfied by definition. This assumption also implies that we may use the term transition probabilities instead of transition probabilities/rates. For a continuous-time random walk, the property stated in (2.6) may be satisfied by rescaling time, after which the same equilibrium equations and, hence, also the same equilibrium distribution are obtained as in the discrete-time case. So, the analysis in the remainder of this chapter also applies to the continuous-time case; this also follows from the property that, by using the uniformization technique as described in e.g. Tijms [70] (not to be confused with the uniformization technique developed by Kingman [49] and Flatto and McKean [33]), a continuous-time random walk/Markov process can be transformed to an equivalent discrete-time random walk/Markov process. By (2.6), we obtain the following equilibrium equations:

$$\begin{aligned} p_{m,n} = {} & q_{1,-1}\, p_{m-1,n+1} + q_{0,0}\, p_{m,n} + q_{-1,1}\, p_{m+1,n-1} \\ & + q_{0,-1}\, p_{m,n+1} + q_{-1,0}\, p_{m+1,n} + q_{-1,-1}\, p_{m+1,n+1} \qquad \text{if } m \geq 1,\ n \geq 1, \end{aligned} \tag{2.7}$$

$$\begin{aligned} p_{m,0} = {} & q_{1,-1}\, p_{m-1,0} + q_{1,-1}\, p_{m-1,1} + (q_{0,0} + q_{0,-1})\, p_{m,0} \\ & + q_{0,-1}\, p_{m,1} + (q_{-1,0} + q_{-1,-1})\, p_{m+1,0} + q_{-1,-1}\, p_{m+1,1} \qquad \text{if } m \geq 1,\ n = 0, \end{aligned} \tag{2.8}$$

$$p_{0,n} = q_{-1,1}p_{0,n-1} + (q_{0,0}+q_{-1,0})p_{0,n} + q_{-1,1}p_{1,n-1} + (q_{0,-1}+q_{-1,-1})p_{0,n+1} + q_{-1,0}p_{1,n} + q_{-1,-1}p_{1,n+1} \quad \text{if } m=0,\ n\geq 1, \tag{2.9}$$

$$p_{0,0} = (q_{0,0}+q_{-1,0}+q_{0,-1}+q_{-1,-1})p_{0,0} + (q_{0,-1}+q_{-1,-1})p_{0,1} + (q_{-1,0}+q_{-1,-1})p_{1,0} + q_{-1,-1}p_{1,1} \quad \text{if } m=0,\ n=0. \tag{2.10}$$

Equation (2.7) is called the equilibrium equation for the interior, (2.8) is called the equilibrium equation for the horizontal boundary, (2.9) is the equation for the vertical boundary, and (2.10) is the equation for the origin.

2.3. The compensation approach

Application of the compensation approach to a random walk of the class described in the previous section, leads to the generation of two formal solutions of the equilibrium equations. In this section, we show how these solutions are obtained. In the next section, it is shown that these solutions are well-defined and that they lead to the equilibrium distribution $\{p_{m,n}\}$.

When applying the compensation approach, according to its main idea (see Section 1.3), we have to start with characterizing the set of product-form solutions which satisfy the equilibrium equation (2.7) for the interior. Substitution of the product form $\alpha^m\beta^n$ into equation (2.7), and dividing by common terms, leads to the following lemma.

Lemma 2.1

The product form $\alpha^m\beta^n$ is a solution of the equilibrium equation (2.7) for the interior if and only if (α,β) satisfies

$$\alpha\beta = q_{1,-1}\beta^2 + q_{0,0}\alpha\beta + q_{-1,1}\alpha^2 + q_{0,-1}\alpha\beta^2 + q_{-1,0}\alpha^2\beta + q_{-1,-1}\alpha^2\beta^2 . \tag{2.11}$$

Equation (2.11) is a quadratic equation in α for fixed β, and vice versa. The set P of appropriate product-form solutions of (2.7) is defined by

$$P = \{\, (\alpha,\beta)\in \mathbb{C}^2 \mid (\alpha,\beta) \text{ satisfies (2.11)},\ \alpha,\beta\neq 0 \text{ and } |\alpha|,|\beta| < 1 \,\}.$$

Product forms with α or β equal to 0 are excluded, since they only lead to non-relevant solutions; and, product forms with α or β larger than or equal to 1 in absolute value are excluded, since later on it must be possible to normalize the ultimate solution of product forms.

Due to the linearity of the equilibrium equations, each linear combination of appropriate product-form solutions of the equilibrium equation (2.7) for the interior, also satisfies this equation (2.7). This property is exploited in the second step of the main idea of the compensation approach. This second step prescribes to construct a linear combination of product-form solutions of P, such that also the equilibrium equations (2.8) and (2.9) for the boundaries are satisfied. For the class under consideration, we obtain two series of product-form solutions, i.e. two linear combinations of countably many product-form solutions. These series are called *formal solutions*. Here, the adjective *formal* is used, since we do not pay attention to the convergence of the series during the construction process. A formal solution consists of an *initial term*, which satisfies the equilibrium equations for the interior and one of

the two boundaries, and a countable number of *compensation terms.* Each compensation term corrects the error made by the previous term at one of the two boundaries.

The number of formal solutions needed for the equilibrium distribution is equal to the number of different initial terms that can be found. For the general case discussed in [12] one finds at least one and at most four initial terms. In our case, due to the projection property, we have *exactly two* initial terms: one for the horizontal boundary and one for the vertical boundary. Contrary to the general case, we have explicit formulae for the initial terms.

Lemma 2.2.

(i) There exists exactly one solution $(\alpha,\beta)\in P$ *which also satisfies the equilibrium equation (2.8) for the horizontal boundary. The factors* α *and* β *of this solution are equal to*

$$\alpha = \frac{q_1^{(1)}}{q_{-1}^{(1)}}, \quad \beta = \frac{q_{-1,1}\alpha^2}{q_{1,-1}+q_{0,-1}\alpha+q_{-1,-1}\alpha^2}. \tag{2.12}$$

(ii) There exists exactly one solution $(\alpha,\beta)\in P$ *which also satisfies the equilibrium equation (2.9) for the vertical boundary. The factors* α *and* β *of this solution are equal to*

$$\beta = \frac{q_1^{(2)}}{q_{-1}^{(2)}}, \quad \alpha = \frac{q_{1,-1}\beta^2}{q_{-1,1}+q_{-1,0}\beta+q_{-1,-1}\beta^2}. \tag{2.13}$$

Proof.

We only prove part (i). Part (ii) can be proved along the same lines. Let $\alpha^m\beta^n$, $0<|\alpha|<1$ and $0<|\beta|<1$, be a solution of (2.7) and (2.8). Substitution of $\alpha^m\beta^n$ in (2.7) and (2.8) gives the quadratic equation stated in (2.11) and

$$\alpha = q_{1,-1}+q_{1,-1}\beta+(q_{0,0}+q_{0,-1})\alpha+q_{0,-1}\alpha\beta+(q_{-1,0}+q_{-1,-1})\alpha^2+q_{-1,-1}\alpha^2\beta. \tag{2.14}$$

Multiplying both sides of (2.14) by β and subtracting from both sides of (2.11) leads to

$$0 = -q_{1,-1}\beta-q_{0,-1}\alpha\beta+q_{-1,1}\alpha^2-q_{-1,-1}\alpha^2\beta,$$

which shows that β has to be taken as presented by (2.12). To find α, we first rearrange the terms of (2.11):

$$(q_{1,-1}+q_{0,-1}\alpha+q_{-1,-1}\alpha^2)\beta^2-(\alpha-q_{0,0}\alpha-q_{-1,0}\alpha^2)\beta+q_{-1,1}\alpha^2 = 0.$$

Now, dividing by β, and substituting the formula for β as stated in (2.12), leads to

$$q_{-1,1}\alpha^2-(\alpha-q_{0,0}\alpha-q_{-1,0}\alpha^2)+(q_{1,-1}+q_{0,-1}\alpha+q_{-1,-1}\alpha^2) = 0.$$

Finally, rearranging terms and using (2.6) leads to

$$(q_{-1,1}+q_{-1,0}+q_{-1,-1})\alpha^2-(q_{1,-1}+q_{-1,1}+q_{-1,0}+q_{-1,-1})\alpha+q_{1,-1} = 0.$$

This quadratic equation has two real-valued solutions, viz. $\alpha=1$ and α as given by (2.12). Here $\alpha=1$ is not feasible, but, by (2.4), the other solution is. By (2.4), also the related solution for β is feasible, which completes the proof. □

As we shall see in Section 2.4, for large m the product form $\alpha^m\beta^n$ with α and β as given by (2.12) will be the dominating term (largest α-factor) of the equilibrium distribution

$\{p_{m,n}\}$. So this product form describes the behavior of $\{p_{m,n}\}$ for large m, which explains that this α is equal to the parameter of the marginal distribution $\{p_m^{(1)}\}$. In the same way it is explained that the factor β given by (2.13) is equal to the parameter of the marginal distribution $\{p_n^{(2)}\}$.

For both initial terms we get a formal solution. Let us first consider the formal solution $\{x_{m,n}\}$ with initial term $a_0\alpha_0^m\beta_0^n$, where α_0 and β_0 are defined by (2.12) and a_0 is a nonnull constant. Instead of only giving explicit formulae for $\{x_{m,n}\}$, we prefer to start with showing how $\{x_{m,n}\}$ is constructed step by step. This also enables us to make clear why we get such simple formal solutions for our class of random walks.

The initial term $a_0\alpha_0^m\beta_0^n$ satisfies the equilibrium equations (2.7) and (2.8) for the interior of the state space and the horizontal boundary. However, it violates equation (2.9) for the vertical boundary. To obtain a solution which satisfies the equations (2.7)-(2.9), we add step by step product-form solutions $\alpha^m\beta^n$ of equation (2.7) for the interior, i.e. product forms $\alpha^m\beta^n$ for which (α,β) satisfies equation (2.11) (in the next section, it will be checked whether these solutions have product factors which are not equal to 0 and which are smaller than 1 in absolute value, i.e. whether they are in the set P). In the first compensation step, to the initial term, a compensation term $a_1\alpha^m\beta^n$ is added to compensate the error of the initial term at the vertical boundary. This error is compensated by chosing the coefficient a_1 and the product factors α and β such that (α,β) satisfies (2.11) and the new solution $a_0\alpha_0^m\beta_0^n + a_1\alpha^m\beta^n$ satisfies (2.9). Since the compensation term generates a new error at the horizontal boundary, after this step more compensation terms have to be added. To show the details of the construction of a compensation term, we give an extensive description of the first compensation step in the next paragraph. All other compensation terms are constructed in the same way.

In the first compensation step, we have to define a_1, α and β such that (α,β) is a solution of (2.11) and the linear combination $a_0\alpha_0^m\beta_0^n + a_1\alpha^m\beta^n$ satisfies the equilibrium equation (2.9) for the vertical boundary. Substitution of the linear combination into (2.9) gives the condition

$$a_0 K(\alpha_0,\beta_0)\beta_0^{n-1} + a_1 K(\alpha,\beta)\beta^{n-1} = 0 \quad \text{for all } n \geq 1, \tag{2.15}$$

where

$$K(\alpha,\beta) = q_{-1,1} - (1 - q_{0,0} - q_{-1,0})\beta + (q_{0,-1} + q_{-1,-1})\beta^2 + (q_{-1,1} + q_{-1,0}\beta + q_{-1,-1}\beta^2)\alpha .$$

Because $a_0\alpha_0^m\beta_0^n$ violates (2.9), $K(\alpha_0,\beta_0) \neq 0$ and hence condition (2.15) forces us to take $\beta = \beta_0$. Next, we use equation (2.11) for the choice for α. Substitution of $\beta = \beta_0$ into (2.11) and rearrangement of the terms of (2.11) gives the following quadratic equation for α:

$$(q_{-1,1} + q_{-1,0}\beta_0 + q_{-1,-1}\beta_0^2)\alpha^2 - (\beta_0 - q_{0,0}\beta_0 - q_{0,-1}\beta_0^2)\alpha + q_{1,-1}\beta_0^2 = 0 . \tag{2.16}$$

Of course, α_0 is one root of this quadratic equation. Let α_1 be the other root. Since we would have no compensation if we would take $\alpha = \alpha_0$, we have to take $\alpha = \alpha_1$. For the computation of α_1, one can use one of the formulae

$$\alpha_0\alpha_1 = \frac{q_{1,-1}\beta_0^2}{q_{-1,1} + q_{-1,0}\beta_0 + q_{-1,-1}\beta_0^2} , \tag{2.17}$$

$$\alpha_0 + \alpha_1 = \frac{(1 - q_{0,0} - q_{0,-1}\beta_0)\beta_0}{q_{-1,1} + q_{-1,0}\beta_0 + q_{-1,-1}\beta_0^2}, \tag{2.18}$$

which are the formulae for the product and the sum of the roots of the quadratic equation (2.16). Finally, the first compensation step is completed by defining the factor a_1 such that the linear combination $a_0\alpha_0^m\beta_0^n + a_1\alpha_1^m\beta_0^n$ satisfies equation (2.9). By (2.15), we find the expression

$$a_1 = -\frac{K(\alpha_0, \beta_0)}{K(\alpha_1, \beta_0)} a_0 .$$

This expression can be simplified considerably. By substituting (2.18) in the formulae for $K(\alpha_0, \beta_0)$ and $K(\alpha_1, \beta_0)$, we find the expressions

$$K(\alpha_0, \beta_0) = (1 - \alpha_1)(q_{-1,1} + q_{-1,0}\beta_0 + q_{-1,-1}\beta_0^2),$$

$$K(\alpha_1, \beta_0) = (1 - \alpha_0)(q_{-1,1} + q_{-1,0}\beta_0 + q_{-1,-1}\beta_0^2),$$

which are due to the projection property (it is easily verified that this factorization is not obtained for a homogeneous, nearest-neighboring random walk without the projection property). As a result, the expression for a_1 simplifies to

$$a_1 = -\frac{1 - \alpha_1}{1 - \alpha_0} a_0 . \tag{2.19}$$

The solution $a_0\alpha_0^m\beta_0^n + a_1\alpha_1^m\beta_0^n$, which we have after the first compensation step, satisfies the equilibrium equations (2.7) and (2.9) for the interior and the vertical boundary. However, the compensation term $a_1\alpha_1^m\beta_0^n$ has generated a new error (a smaller one, as will be shown in Section 2.4) at the horizontal boundary. To compensate for this error, we again have to add a compensation term, and so on; here, a compensation step on the horizontal boundary is symmetric to the compensation step on the vertical boundary. Ultimately, we obtain the following formal solution, where b_0 is a second nonnull constant:

$$x_{m,n} = \overbrace{\underbrace{a_0 b_0 \alpha_0^m \beta_0^n}_{\text{V}} + \underbrace{a_1 b_0 \alpha_1^m \beta_0^n}_{\text{V}}}^{\text{H}} + \overbrace{\underbrace{a_1 b_1 \alpha_1^m \beta_1^n}_{\text{V}} + \underbrace{a_2 b_1 \alpha_2^m \beta_1^n}_{\text{V}}}^{\text{H}} + \ldots$$

The construction is such that each term in this series satisfies (2.7), each sum of terms with the same α-factor satisfies (2.8) and each sum of terms with the same β-factor satisfies (2.9). As a consequence, $\{x_{m,n}\}$ is a formal solution of the equilibrium equations (2.7)-(2.9).

The formulae for $\{x_{m,n}\}$ are as follows. By taking pairs of product forms in two different ways, we get two expressions:

$$x_{m,n} = a_0 b_0 \alpha_0^m \beta_0^n + \sum_{i=0}^{\infty} a_{i+1}\alpha_{i+1}^m (b_i \beta_i^n + b_{i+1}\beta_{i+1}^n) \tag{2.20}$$

$$= \sum_{i=0}^{\infty} b_i \beta_i^n (a_i \alpha_i^m + a_{i+1}\alpha_{i+1}^m), \qquad m \geq 0,\ n \geq 0. \tag{2.21}$$

The factors α_0 and β_0 are given by (2.12). The other α- and β-factors are obtained by using the formula for the product of the roots of (2.11) (compare (2.17)):

$$\alpha_{i+1} = \frac{q_{1,-1}\beta_i^2}{q_{-1,1}+q_{-1,0}\beta_i+q_{-1,-1}\beta_i^2}\cdot\frac{1}{\alpha_i}\ , \quad i\geq 0, \tag{2.22}$$

$$\beta_{i+1} = \frac{q_{-1,1}\alpha_{i+1}^2}{q_{1,-1}+q_{0,-1}\alpha_{i+1}+q_{-1,-1}\alpha_{i+1}^2}\cdot\frac{1}{\beta_i}\ , \quad i\geq 0. \tag{2.23}$$

The coefficients a_0 and b_0 are only required to be nonnull constants. Formulae for the other coefficients are derived in the same way as formula (2.19) for a_1:

$$a_{i+1} = -\frac{1-\alpha_{i+1}}{1-\alpha_i}\,a_i\ , \quad i\geq 0, \tag{2.24}$$

$$b_{i+1} = -\frac{1-\beta_{i+1}}{1-\beta_i}\,b_i\ , \quad i\geq 0. \tag{2.25}$$

The simple form of the recursive formulae (2.24) and (2.25) for the coefficients a_i and b_i (which is due to the projection property, cf. the derivation of formula (2.19) for a_1), leads to an elegant expression for $\{x_{m,n}\}$. Define $a_0 := 1-\alpha_0$ and $b_0 := 1-\beta_0$, then the recursive formulae (2.24) and (2.25) for a_i and b_i are easily rewritten to

$$a_i = (-1)^i(1-\alpha_i)\ , \quad i\geq 0,$$

$$b_i = (-1)^i(1-\beta_i)\ , \quad i\geq 0.$$

Substitution of these formulae in (2.20) and (2.21) yields

$$x_{m,n} = (1-\alpha_0)\alpha_0^m(1-\beta_0)\beta_0^n - \sum_{i=0}^{\infty}(1-\alpha_{i+1})\alpha_{i+1}^m[(1-\beta_i)\beta_i^n-(1-\beta_{i+1})\beta_{i+1}^n] \tag{2.26}$$

$$= \sum_{i=0}^{\infty}(1-\beta_i)\beta_i^n[(1-\alpha_i)\alpha_i^m-(1-\alpha_{i+1})\alpha_{i+1}^m]\ , \quad m\geq 0,\ n\geq 0, \tag{2.27}$$

which show that $\{x_{m,n}\}$ is an alternating sum of pure, two-dimensional product-form solutions.

For the other formal solution $\{\hat{x}_{m,n}\}$ generated by the initial term with product factors defined by (2.13), we get similar expressions as for $\{x_{m,n}\}$:

$$\hat{x}_{m,n} = (1-\hat{\alpha}_0)\hat{\alpha}_0^m(1-\hat{\beta}_0)\hat{\beta}_0^n - \sum_{i=0}^{\infty}(1-\hat{\beta}_{i+1})\hat{\beta}_{i+1}^n[(1-\hat{\alpha}_i)\hat{\alpha}_i^m-(1-\hat{\alpha}_{i+1})\hat{\alpha}_{i+1}^m] \tag{2.28}$$

$$= \sum_{i=0}^{\infty}(1-\hat{\alpha}_i)\hat{\alpha}_i^m[(1-\hat{\beta}_i)\hat{\beta}_i^n-(1-\hat{\beta}_{i+1})\hat{\beta}_{i+1}^n]\ , \quad m\geq 0,\ n\geq 0. \tag{2.29}$$

Here, the factors $\hat{\beta}_0$ and $\hat{\alpha}_0$ are defined by (2.13). The other factors are defined such that each product form satisfies quadratic equation (2.11):

$$\hat{\beta}_{i+1} = \frac{q_{-1,1}\hat{\alpha}_i^2}{q_{1,-1}+q_{0,-1}\hat{\alpha}_i+q_{-1,-1}\hat{\alpha}_i^2}\cdot\frac{1}{\hat{\beta}_i}\ , \quad i\geq 0, \tag{2.30}$$

$$\hat{\alpha}_{i+1} = \frac{q_{1,-1}\hat{\beta}_{i+1}^2}{q_{-1,1}+q_{-1,0}\hat{\beta}_{i+1}+q_{-1,-1}\hat{\beta}_{i+1}^2}\cdot\frac{1}{\hat{\alpha}_i}\ , \quad i\geq 0. \tag{2.31}$$

This completes the description of the two series $\{x_{m,n}\}$ and $\{\hat{x}_{m,n}\}$, which both are formal solutions of the equilibrium equations (2.7)-(2.9).

2.4. Main Theorem

This section is devoted to the derivation of the *Main Theorem*, which states that the equilibrium distribution $\{p_{m,n}\}$ is obtained by simply taking the sum of the two formal solutions $\{x_{m,n}\}$ and $\{\hat{x}_{m,n}\}$; see Theorem 2.1 at the end of this section. For the proof of this main result, some preliminary results are needed. This section starts with showing that all product forms which constitute the formal solutions are in the set P of appropriate solutions of (2.7) (up to now, this only has been verified for the initial product forms) and that the formal solutions are absolutely convergent in all states, except in the origin. For this, we shall refer to some results of Adan et al. [12]. After that, it is shown that a simple result for the formal solutions leads to the suggestion that a solution of all equilibrium equations may be obtained by taking the sum of the two formal solutions, which ultimately leads to the Main Theorem.

For the convergence of the formal solutions $\{x_{m,n}\}$ and $\{\hat{x}_{m,n}\}$, we need information about the limiting behavior of the α- and β-factors. We gather some results of the analysis presented in [12].

Lemma 2.3.

For the factors α_i, β_i, $\hat{\alpha}_i$ and $\hat{\beta}_i$, we find:

(i) $1 > \alpha_0 > \beta_0 > \alpha_1 > \beta_1 > \ldots \downarrow 0$;

(ii) $\dfrac{\alpha_{i+1}}{\beta_i} \to A_1$ *and* $\dfrac{\beta_i}{\alpha_i} \to \dfrac{1}{A_2}$ *as* $i \to \infty$;

(iii) $1 > \hat{\beta}_0 > \hat{\alpha}_0 > \hat{\beta}_1 > \hat{\alpha}_1 > \ldots \downarrow 0$;

(iv) $\dfrac{\hat{\beta}_{i+1}}{\hat{\alpha}_i} \to \dfrac{1}{A_2}$ *and* $\dfrac{\hat{\alpha}_i}{\hat{\beta}_i} \to A_1$ *as* $i \to \infty$;

Here, A_1 and A_2 are defined by

$$A_1 = \frac{(1-q_{0,0}) - \sqrt{(1-q_{0,0})^2 - 4q_{1,-1}q_{-1,1}}}{2q_{-1,1}},$$

$$A_2 = \frac{(1-q_{0,0}) + \sqrt{(1-q_{0,0})^2 - 4q_{1,-1}q_{-1,1}}}{2q_{-1,1}}.$$

Part (i) of this lemma is proved by first writing the quadratic equation (2.11), by which the α- and β-factors are defined, as a quadratic equation in $z = \beta/\alpha$ and then applying Rouché's theorem (see Titchmarch [71]). Part (ii) is found by first writing the roots of the quadratic equation in $z = \beta/\alpha$ as function of α and then letting $\alpha \to 0$ (see Lemma 6.1 of [12]). The parts (iii) and (iv) are proved along the same lines.

The parts (i) and (iii) of Lemma 2.3 show that all product forms which constitute the formal solutions, are members of the set P. Further, by Lemma 2.3,

$$\frac{(1-\alpha_{i+1})\alpha_{i+1}^m (1-\beta_{i+1})\beta_{i+1}^n}{(1-\alpha_i)\alpha_i^m (1-\beta_i)\beta_i^n} \to \left[\frac{A_1}{A_2}\right]^{m+n} \tag{2.32}$$

and

$$\frac{(1-\alpha_{i+2})\alpha_{i+2}^m (1-\beta_{i+1})\beta_{i+1}^n}{(1-\alpha_{i+1})\alpha_{i+1}^m (1-\beta_i)\beta_i^n} \to \left[\frac{A_1}{A_2}\right]^{m+n} \tag{2.33}$$

as $i \to \infty$. Elementary algebra shows that A_1 and A_2 are positive real-valued variables with $A_1 < 1 < A_2$. Therefore the limits in (2.32) and (2.33) are smaller than 1 for all states (m,n) with $m+n \geq 1$, which proves part (i) of the following lemma. Part (ii) of that lemma, which is needed in the second part of this section, is also easily proved by using Lemma 2.3.

Lemma 2.4.

For $\{x_{m,n}\}$, we have the following properties:

(i) The series

$$\sum_{i=0}^{\infty}(1-\alpha_i)\alpha_i^m (1-\beta_i)\beta_i^n \quad \textit{and} \quad \sum_{i=0}^{\infty}(1-\alpha_{i+1})\alpha_{i+1}^m (1-\beta_i)\beta_i^n$$

are absolutely convergent for all $m \geq 0$, $n \geq 0$ and $m+n \geq 1$. So, $\{x_{m,n}\}$ is well defined by (2.26) and (2.27) in all states except in the origin.

(ii) $$\sum_{\substack{m\geq 0,\, n\geq 0 \\ m+n\geq 1}} |x_{m,n}| < \infty .$$

The same results hold for $\{\hat{x}_{m,n}\}$.

By Lemma 2.4, $\{x_{m,n}\}$ and $\{\hat{x}_{m,n}\}$ satisfy the equilibrium equations for all states except for the states for which the equilibrium probability $p_{0,0}$ occurs in the corresponding equilibrium equations, i.e. except for the states (0,0), (1,0) and (0,1).

From the analysis in Adan et al. [12], we know that the equilibrium distribution is found by taking a linear combination of the formal solutions, of which the coefficients can be determined by substituting this linear combination in two of the three equilibrium equations for the states (0,0), (1,0) and (0,1). This is proved by analyzing the embedded process on the set of states where the formal solutions are absolutely convergent. For our problem, however, due to the projection property, it follows that we have to take the sum of the formal solutions and we can give an alternative proof to show that the equilibrium distribution is equal to this sum.

Due to the projection property, we found the formulae (2.26)-(2.29). Using these formulae, we easily see that

$$\sum_{n=0}^{\infty} x_{m,n} = (1-\alpha_0)\alpha_0^m \quad \text{for all } m \geq 1, \qquad \sum_{m=0}^{\infty} x_{m,n} = 0 \quad \text{for all } n \geq 1,$$

$$\sum_{n=0}^{\infty} \hat{x}_{m,n} = 0 \quad \text{for all } m \geq 1, \qquad \sum_{m=0}^{\infty} \hat{x}_{m,n} = (1-\hat{\beta}_0)\hat{\beta}_0^n \quad \text{for all } n \geq 1.$$

Since α_0 and $\hat{\beta}_0$ are equal to the parameters of the marginal distributions $\{p_k^{(1)}\}$ and $\{p_k^{(2)}\}$ (see also the paragraph right after the proof of Lemma 2.2), by defining $\{\hat{p}_{m,n}\}$ as the sum of the formal solutions, i.e.

$$\hat{p}_{m,n} := x_{m,n} + \hat{x}_{m,n} , \quad m \geq 0,\ n \geq 0,\ m+n \geq 1,$$

we obtain a solution for which

$$\sum_{n=0}^{\infty} \hat{p}_{m,n} = p_m^{(1)} \quad \text{for all } m \geq 1, \qquad \sum_{m=0}^{\infty} \hat{p}_{m,n} = p_n^{(2)} \quad \text{for all } n \geq 1. \tag{2.34}$$

To obtain a solution for which the probabilities add up to 1, we define $\hat{p}_{0,0}$ by

$$\hat{p}_{0,0} := 1 - \sum_{\substack{m \geq 0,\, n \geq 0 \\ m+n \geq 1}} \hat{p}_{m,n} ,$$

which is a correct definition due to Lemma 2.4. Now, rewriting $\hat{p}_{0,0}$ leads to

$$\hat{p}_{0,0} = (1-\alpha_0) - \sum_{n=1}^{\infty} \hat{p}_{0,n} = (1-\hat{\beta}_0) - \sum_{m=1}^{\infty} \hat{p}_{m,0} .$$

Hence the first equality in (2.34) also holds for $m=0$ and the second equality in (2.34) also holds for $n=0$. As a consequence, the marginal distributions of $\{\hat{p}_{m,n}\}$ are equal to the marginal distributions $\{p_m^{(1)}\}$ and $\{p_n^{(2)}\}$ of $\{p_{m,n}\}$:

$$\sum_{n=0}^{\infty} \hat{p}_{m,n} = p_m^{(1)} \quad \text{for all } m \geq 0, \qquad \sum_{m=0}^{\infty} \hat{p}_{m,n} = p_n^{(2)} \quad \text{for all } n \geq 0. \tag{2.35}$$

Since $\{\hat{p}_{m,n}\}$ is a linear combination of the formal solutions, we know that $\{\hat{p}_{m,n}\}$ satisfies the equilibrium equations in all states except in (0,0), (1,0) and (0,1). To show that $\{\hat{p}_{m,n}\}$ also satisfies the equations in these remaining states, we use (2.35) and the *balance principle*:

the stream out of a set M' = the stream into this set M', $M' \subset M$.

The balance principle for the set $M_1 = \{(m,n) \in M \mid m \geq 1\}$ gives the condition

$$(q_{-1,1} + q_{-1,0} + q_{-1,-1}) \sum_{n=0}^{\infty} p_{1,n} = q_{1,-1} \sum_{n=0}^{\infty} p_{0,n} .$$

By using (2.35) it is easily shown that $\{\hat{p}_{m,n}\}$ satisfies this condition, i.e. the balance principle for the set M_1. Further, $\{\hat{p}_{m,n}\}$ also satisfies the balance principle for the subset $M_2 = M_1 \setminus \{(1,0)\}$ of M_1, since it satisfies the balance principle (i.e. the equilibrium equation) for each state of this set. Hence, $\{\hat{p}_{m,n}\}$ also satisfies the balance principle for $M_1 \setminus M_2$, i.e. the equilibrium equation in (1,0). In the same way it is proved that $\{\hat{p}_{m,n}\}$ satisfies the equations in (0,1) and (0,0). So, we find that $\{\hat{p}_{m,n}\}$ satisfies all equilibrium equations. By Lemma 2.4, $\Sigma_{m,n \geq 0} |\hat{p}_{m,n}| < \infty$, and thus the equilibrium distribution $\{p_{m,n}\}$ may be obtained by normalizing the solution $\{\hat{p}_{m,n}\}$. Since, by the definition of $\hat{p}_{0,0}$, the probabilities $\hat{p}_{m,n}$ already add up to 1, we finally find that $\{p_{m,n}\}$ is equal to $\{\hat{p}_{m,n}\}$. This completes the proof of the Main Theorem.

Theorem 2.1. *(Main Theorem)*

The equilibrium distribution $\{p_{m,n}\}$ for a random walk of the class described in Section 2.2, is equal to the sum of two alternating series of pure two-dimensional product-form distributions:

$$p_{m,n} = \sum_{i=0}^{\infty}(1-\beta_i)\beta_i^n\,[(1-\alpha_i)\alpha_i^m-(1-\alpha_{i+1})\alpha_{i+1}^m] + \sum_{i=0}^{\infty}(1-\hat{\alpha}_i)\hat{\alpha}_i^m\,[(1-\hat{\beta}_i)\hat{\beta}_i^n-(1-\hat{\beta}_{i+1})\hat{\beta}_{i+1}^n]\,, \quad (m,n)\in M\backslash\{(0,0)\}, \tag{2.36}$$

$$p_{0,0} = 1 - \sum_{\substack{(m,n)\in M\\(m,n)\neq(0,0)}} p_{m,n}\,. \tag{2.37}$$

Here, the factors α_i, β_i, $\hat{\alpha}_i$ and $\hat{\beta}_i$ are defined as denoted at the end of Section 2.3.

2.5. Error bounds and numerical results

The analytic results for the equilibrium distribution $\{p_{m,n}\}$ make it possible to develop efficient numerical procedures for the computation of $\{p_{m,n}\}$. In this section, we first derive error bounds for the relevant series of product forms. After that, numerical results are presented for the symmetric 2×2 switch. Among others, we show that for most states only a few product forms are needed to obtain accurate approximations for the equilibrium probabilities $p_{m,n}$.

Suppose that we want to compute the equilibrium distribution within a given absolute or relative accuracy. For a state $(m,n)\neq(0,0)$, by (2.36), the equilibrium probability $p_{m,n}$ is equal to the sum of the series $x_{m,n}$ and $\hat{x}_{m,n}$ defined by (2.26)-(2.29). For the absolute difference between these series and their partial sums, we can derive tight error bounds by using the property that the sequences $\{\alpha_{i+1}/\beta_i\}$, $\{\beta_i/\alpha_i\}$, $\{\hat{\beta}_{i+1}/\hat{\alpha}_i\}$ and $\{\hat{\alpha}_i/\hat{\beta}_i\}$, of which all elements are positive and smaller than 1, and of which the limits are given by Lemma 2.3, are monotonously strictly decreasing (see Lemma 5.1 of [12]).

Let $(m,n)\in M\backslash\{(0,0)\}$ and let $x_{m,n}^{(k)}$ denote the partial sum consisting of the first $k\geq1$ product forms of $x_{m,n}$. Then

$$x_{m,n}^{(2k)} = \sum_{i=0}^{k-1}(1-\beta_i)\beta_i^n\,[(1-\alpha_i)\alpha_i^m-(1-\alpha_{i+1})\alpha_{i+1}^m]\,, \quad k\geq1,$$

$$x_{m,n}^{(2k+1)} = (1-\alpha_0)\alpha_0^m\,(1-\beta_0)\beta_0^n - \sum_{i=0}^{k-1}(1-\alpha_{i+1})\alpha_{i+1}^m\,[(1-\beta_i)\beta_i^n-(1-\beta_{i+1})\beta_{i+1}^n]\,, \quad k\geq0.$$

Since the sequences $\{\alpha_{i+1}/\beta_i\}$ and $\{\beta_i/\alpha_i\}$ are monotonously strictly decreasing, also the sequences $\{\alpha_{i+1}/\alpha_i\}$ and $\{\beta_{i+1}/\beta_i\}$ are monotonously strictly decreasing. We use this property to derive an error bound for the approximation of $x_{m,n}$ by a partial sum $x_{m,n}^{(2k)}$, where $k\geq1$. For the computation of the partial sum $x_{m,n}^{(2k)}$ itself, one needs the product factors $\alpha_0,\beta_0,\alpha_1,\ldots,\beta_{k-1},\alpha_k$. Suppose that, for the benefit of the error bound, we are willing to compute in advance the next product factor β_k. Then, after defining the factors x_α and x_β by $x_\alpha=(\alpha_k/\alpha_{k-1})^m$ and $x_\beta=(\beta_k/\beta_{k-1})^n$ (note that x_α and x_β are positive and smaller than 1), we

obtain the following error bound:

$$
\begin{aligned}
|x_{m,n} - x_{m,n}^{(2k)}| &= \left| \sum_{i=k}^{\infty} (1-\beta_i)\beta_i^n \left[(1-\alpha_i)\alpha_i^m - (1-\alpha_{i+1})\alpha_{i+1}^m\right] \right| \\
&\le \sum_{i=k}^{\infty} \left[\alpha_i^m \beta_i^n + \alpha_{i+1}^m \beta_i^n\right] \\
&= \alpha_k^m \beta_k^n \sum_{i=k}^{\infty} \left[\frac{\alpha_i^m \beta_i^n}{\alpha_k^m \beta_k^n} + \frac{\alpha_{i+1}^m \beta_i^n}{\alpha_k^m \beta_k^n}\right] \\
&\le \alpha_k^m \beta_k^n \sum_{i=k}^{\infty} \left[x_\alpha^{i-k} x_\beta^{i-k} + x_\alpha^{i-k+1} x_\beta^{i-k}\right] \\
&= \alpha_k^m \beta_k^n \frac{1+x_\alpha}{1-x_\alpha x_\beta} .
\end{aligned}
$$

Similarly, one can derive error bounds for the partial sums $x_{m,n}^{(2k+1)}$. Further, similar error bounds can be derived for the partial sums $\hat{x}_{m,n}^{(k)}$ of the series $\hat{x}_{m,n}$. This results in the following lemma.

Lemma 2.5

Let $(m,n) \in M \setminus \{(0,0)\}$. Then the following error bounds hold for the partial sums $x_{m,n}^{(k)}$ and $\hat{x}_{m,n}^{(k)}$ consisting of the first k product-form solutions of the series $x_{m,n}$ and $\hat{x}_{m,n}$ respectively (define $\beta_{-1} := 1$ and $\hat{\alpha}_{-1} := 1$ to let the bounds also be valid for $x_{m,n}^{(1)}$ and $\hat{x}_{m,n}^{(1)}$):

$$
\begin{aligned}
|x_{m,n} - x_{m,n}^{(2k)}| &\le \alpha_k^m \beta_k^n \frac{1+(\alpha_k/\alpha_{k-1})^m}{1-(\alpha_k/\alpha_{k-1})^m(\beta_k/\beta_{k-1})^n} && \text{for } k \ge 1, \\
|x_{m,n} - x_{m,n}^{(2k+1)}| &\le \alpha_{k+1}^m \beta_k^n \frac{1+(\beta_k/\beta_{k-1})^n}{1-(\alpha_{k+1}/\alpha_k)^m(\beta_k/\beta_{k-1})^n} && \text{for } k \ge 0, \\
|\hat{x}_{m,n} - \hat{x}_{m,n}^{(2k)}| &\le \hat{\alpha}_k^m \hat{\beta}_k^n \frac{1+(\hat{\beta}_k/\hat{\beta}_{k-1})^n}{1-(\hat{\alpha}_k/\hat{\alpha}_{k-1})^m(\hat{\beta}_k/\hat{\beta}_{k-1})^n} && \text{for } k \ge 1, \\
|\hat{x}_{m,n} - \hat{x}_{m,n}^{(2k+1)}| &\le \hat{\alpha}_k^m \hat{\beta}_{k+1}^n \frac{1+(\hat{\alpha}_k/\hat{\alpha}_{k-1})^n}{1-(\hat{\alpha}_k/\hat{\alpha}_{k-1})^m(\hat{\beta}_{k+1}/\hat{\beta}_k)^n} && \text{for } k \ge 0.
\end{aligned}
$$

Denote the error bounds for $x_{m,n}^{(k)}$ and $\hat{x}_{m,n}^{(k)}$ by $b_{m,n}^{(k)}$ and $\hat{b}_{m,n}^{(k)}$. These bounds tend to 0, exponentially fast, as $k \to \infty$. Further, for fixed k, these bounds decrease monotonously and exponentially fast to 0, as $m \to \infty$ and/or $n \to \infty$.

For the computation of the equilibrium probability $p_{m,n}$ for a state $(m,n) \neq (0,0)$ within a given *absolute* accuracy ε_{abs}, we propose to use the following procedure. Compute $x_{m,n}^{(1)}$ and $\hat{x}_{m,n}^{(1)}$, and the corresponding error bounds $b_{m,n}^{(1)}$ and $\hat{b}_{m,n}^{(1)}$. If $b_{m,n}^{(1)} + \hat{b}_{m,n}^{(1)} \le \varepsilon_{abs}$, then $x_{m,n}^{(1)} + \hat{x}_{m,n}^{(1)}$ approximates $p_{m,n}$ within the desired accuracy. Else, one continues by approximating the series $x_{m,n}$ and $\hat{x}_{m,n}$ more accurately, say with absolute accuracy ε_1 and ε_2, respectively. The parameters ε_1 and ε_2 must be defined such that $\varepsilon_1 > 0$, $\varepsilon_2 > 0$ and $\varepsilon_1 + \varepsilon_2 = \varepsilon_{abs}$. It seems reasonable to divide ε_{abs} over ε_1 and ε_2 proportional to the values for the error bounds $b_{m,n}^{(1)}$ and

$\hat{b}^{(1)}_{m,n}$; here, in order to avoid the situation that one series must be computed with a very small absolute accuracy and the other series only with a relatively very large absolute accuracy (this could occur in case m is very small and n very large, or vice versa), it must be provided that each ε_i gets at least 5% of ε_{abs}. Thus, we take

$$\varepsilon_1 = 0.05\cdot\varepsilon_{abs} \quad \text{and} \quad \varepsilon_2 = 0.95\cdot\varepsilon_{abs} \qquad \text{if } b^{(1)}_{m,n} \le 0.05\cdot(b^{(1)}_{m,1} + \hat{b}^{(1)}_{m,n});$$

$$\varepsilon_1 = 0.95\cdot\varepsilon_{abs} \quad \text{and} \quad \varepsilon_2 = 0.05\cdot\varepsilon_{abs} \qquad \text{if } \hat{b}^{(1)}_{m,n} \le 0.05\cdot(b^{(1)}_{m,1} + \hat{b}^{(1)}_{m,n});$$

$$\varepsilon_1 = \frac{b^{(1)}_{m,n}}{b^{(1)}_{m,n} + \hat{b}^{(1)}_{m,n}}\varepsilon_{abs} \quad \text{and} \quad \varepsilon_2 = \frac{\hat{b}^{(1)}_{m,n}}{b^{(1)}_{m,n} + \hat{b}^{(1)}_{m,n}}\varepsilon_{abs} \qquad \text{otherwise}.$$

The series $x_{m,n}$ is approximated within absolute accuracy ε_1, by computing the partial sums $x^{(k)}_{m,n}$ and the corresponding error bounds $b^{(k)}_{m,n}$ for $k = 2,3,\cdots$ until $b^{(k)}_{m,n} < \varepsilon_1$; and similarly for $\hat{x}_{m,n}$. This completes the description for the computation of $p_{m,n}$, $(m,n) \in M\backslash\{(0,0)\}$, within a given absolute accuracy. The equilibrium probability $p_{0,0}$ for the origin may be easily computed out of the equilibrium equation for this state (see (2.10)). Finally, we remark that a probability $p_{m,n}$ can be computed within a given *relative* accuracy by performing the above procedure for decreasing values of the absolute accuracy ε_{abs}.

Below, the proposed procedure is applied for the computation of the equilibrium distribution of the 2×2 switch. Based on this example, it is also shown that the explicit expressions for the equilibrium distribution $\{p_{m,n}\}$ may lead to similar expressions for the relevant performance measures.

Example 2.2: The 2×2 switch *(continued)*

The *symmetric 2×2 switch* is defined as a 2×2 switch with parameters

$$r_1 = r_2 = r\,, \quad \hat{r}_{1,1} = \hat{r}_{1,2} = \hat{r}_{2,1} = \hat{r}_{2,2} = \tfrac{1}{2}\,, \tag{2.38}$$

where r denotes the arrival rate of the Bernoulli streams for both types of arriving jobs. The parameter r is equal to the fraction of time units that each server works, and therefore r is called the workload. We assume that $0 < r < 1$ (the case $r = 1$ is excluded, since it corresponds to a non-ergodic system). Remark that the symmetry leads to some simplifications in the formulae for the equilibrium distribution $\{p_{m,n}\}$. For the symmetric 2×2 switch, the transition probabilities simplify to

$$q_{-1,1} = q_{1,-1} = \tfrac{1}{4}r^2\,, \quad q_{0,0} = \tfrac{1}{2}r^2\,, \quad q_{-1,0} = q_{0,-1} = r(1-r)\,, \quad q_{-1,-1} = (1-r)^2\,. \tag{2.39}$$

So, $q_{i,j} = q_{j,i}$ for all directions (i,j), and the equilibrium distribution will also be symmetric. Indeed, it is easily verified that in the symmetric case $\hat{\beta}_i = \alpha_i$ and $\hat{\alpha}_i = \beta_i$ for all $i \ge 0$, which implies that $\hat{x}_{m,n} = x_{n,m}$ for all $(m,n) \in M\backslash\{(0,0)\}$, and $p_{m,n} = p_{n,m}$ for all $(m,n) \in M$.

For the symmetric 2×2 switch, we used the procedure described in this section to compute the equilibrium probabilities $p_{m,n}$ for all states (m,n), $m+n \le 10$, within absolute accuracy $\varepsilon_{abs} = 10^{-6}$. In Table 2.1, we have depicted the values of the equilibrium probabilities $p_{m,n}$ for the case $r = 0.8$. In Table 2.2, for each state $(m,n) \neq (0,0)$, it is denoted how many product forms of the series $x_{m,n}$ and $\hat{x}_{m,n}$ were needed for the computation of $p_{m,n}$ within the desired accuracy. As we see, for all states which are not too close to the origin, only a few terms of the series $x_{m,n}$ and $\hat{x}_{m,n}$ are needed to obtain accurate approximations for $p_{m,n}$. Further, the results indicate that, for a workload $r = 0.8$, the sum

n ↑	$m=0$	1	2	3	4	5	6	7	8	9	10
10	0.0001										
9	0.0003	0.0000									
8	0.0007	0.0001	0.0000								
7	0.0017	0.0002	0.0000	0.0000							
6	0.0037	0.0005	0.0001	0.0000	0.0000						
5	0.0083	0.0011	0.0002	0.0000	0.0000	0.0000					
4	0.0187	0.0025	0.0004	0.0001	0.0000	0.0000	0.0000				
3	0.0415	0.0060	0.0010	0.0002	0.0001	0.0000	0.0000	0.0000			
2	0.0892	0.0157	0.0033	0.0010	0.0004	0.0002	0.0001	0.0000	0.0000		
1	0.1711	0.0495	0.0157	0.0060	0.0025	0.0011	0.0005	0.0002	0.0001	0.0000	
0	0.2201	0.1711	0.0892	0.0415	0.0187	0.0083	0.0037	0.0017	0.0007	0.0003	0.0001

$m \rightarrow$

Table 2.1. The equilibrium probabilities $p_{m,n}$ for all states (m,n) with $m+n \leq 10$ for the symmetric 2×2 switch with workload $r=0.8$.

n ↑	$m=0$	1	2	3	4	5	6	7	8	9	10
10	(1,1)										
9	(1,1)	(1,1)									
8	(1,1)	(1,1)	(1,1)								
7	(1,1)	(1,1)	(1,1)	(1,1)							
6	(2,1)	(1,1)	(1,1)	(1,1)	(1,1)						
5	(2,2)	(2,1)	(1,1)	(1,1)	(1,1)	(1,1)					
4	(2,3)	(2,2)	(1,1)	(1,1)	(1,1)	(1,1)	(1,1)				
3	(3,3)	(2,3)	(2,2)	(1,1)	(1,1)	(1,1)	(1,1)	(1,1)			
2	(4,5)	(3,3)	(2,2)	(2,2)	(1,1)	(1,1)	(1,1)	(1,1)	(1,1)		
1	(10,11)	(5,5)	(3,3)	(3,2)	(2,2)	(1,2)	(1,1)	(1,1)	(1,1)	(1,1)	
0	–	(11,10)	(5,4)	(3,3)	(3,2)	(2,2)	(1,2)	(1,1)	(1,1)	(1,1)	(1,1)

$m \rightarrow$

Table 2.2. The numbers of product forms of $x_{m,n}$ and $\hat{x}_{m,n}$ needed to compute the equilibrium probabilities $p_{m,n}$ for the symmetric 2×2 switch with workload $r=0.8$ within absolute accuracy $\varepsilon_{abs}=10^{-6}$.

$$x^{(1)}_{m,n} + \hat{x}^{(1)}_{m,n} = (1-\alpha_0)(1-\beta_0)\alpha_0^m\beta_0^n + (1-\hat{\alpha}_0)(1-\hat{\beta}_0)\hat{\alpha}_0^m\hat{\beta}_0^n$$

is already a sufficiently accurate approximation for $p_{m,n}$ for all states (m,n) with $m+n \geq 7$.

Except for a procedure for the computation of the equilibrium distribution $\{p_{m,n}\}$ itself, the explicit formulae for $\{p_{m,n}\}$ may also be exploited for the computation of queue lengths, for example. Let $\mathbf{L}_1$ and $\mathbf{L}_2$ denote the lengths of the queues at the servers 1 and 2, and let $\mathbf{L}=\mathbf{L}_1+\mathbf{L}_2$ denote the total number of jobs present at the beginning of a time unit. The distributions for $\mathbf{L}_1$ and $\mathbf{L}_2$ are given by $\{p^{(1)}_m\}$ and $\{p^{(2)}_n\}$, i.e. $\mathbf{L}_1$ and $\mathbf{L}_2$ have geometric distributions with parameters $\alpha_0 = q^{(1)}_1/q^{(1)}_{-1}$ and $\hat{\beta}_0 = q^{(2)}_1/q^{(2)}_{-1}$, respectively. Therefore, the mean, standard deviation and coefficient of variation of $\mathbf{L}_1$ and $\mathbf{L}_2$ are given by

$$\mathbb{E}\mathbf{L}_1 = \frac{\alpha_0}{1-\alpha_0}\,, \quad \sigma(\mathbf{L}_1) = \frac{\sqrt{\alpha_0}}{1-\alpha_0}\,, \quad cv(\mathbf{L}_1) = \frac{1}{\sqrt{\alpha_0}}\,,$$

$$\mathbb{E}\mathbf{L}_2 = \frac{\hat{\beta}_0}{1-\hat{\beta}_0}\,, \quad \sigma(\mathbf{L}_2) = \frac{\sqrt{\hat{\beta}_0}}{1-\hat{\beta}_0}\,, \quad cv(\mathbf{L}_2) = \frac{1}{\sqrt{\hat{\beta}_0}}\,.$$

For the corresponding quantities for $\mathbf{L}$, we find

$$\mathbb{E}\mathbf{L} = \mathbb{E}\mathbf{L}_1 + \mathbb{E}\mathbf{L}_2\,, \quad \sigma(\mathbf{L}) = \sqrt{\mathbb{E}\mathbf{L}^2 - (\mathbb{E}\mathbf{L})^2}\,, \quad cv(\mathbf{L}) = \sigma(\mathbf{L})/\mathbb{E}\mathbf{L}\,.$$

The joint queue length distribution $\{p_{m,n}\}$ is needed to compute

$$\begin{aligned}\mathbb{E}\mathbf{L}^2 &= \mathbb{E}(\mathbf{L}_1+\mathbf{L}_2)^2 = \mathbb{E}\mathbf{L}_1^2 + 2\,\mathbb{E}\{\mathbf{L}_1\mathbf{L}_2\} + \mathbb{E}\mathbf{L}_2^2 \\ &= \frac{\alpha_0(1+\alpha_0)}{(1-\alpha_0)^2} + 2\,\mathbb{E}\{\mathbf{L}_1\mathbf{L}_2\} + \frac{\hat{\beta}_0(1+\hat{\beta}_0)}{(1-\hat{\beta}_0)^2}\,,\end{aligned}$$

i.e. to compute

$$\begin{aligned}E\{\mathbf{L}_1\mathbf{L}_2\} &= \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} m\,n\,p_{m,n} \\ &= \sum_{i=0}^{\infty}\frac{\beta_i}{1-\beta_i}\Big[\frac{\alpha_i}{1-\alpha_i} - \frac{\alpha_{i+1}}{1-\alpha_{i+1}}\Big] + \sum_{i=0}^{\infty}\frac{\hat{\alpha}_i}{1-\hat{\alpha}_i}\Big[\frac{\hat{\beta}_i}{1-\hat{\beta}_i} - \frac{\hat{\beta}_{i+1}}{1-\hat{\beta}_{i+1}}\Big]\,. \qquad (2.40)\end{aligned}$$

For $\mathbb{E}\{\mathbf{L}_1\mathbf{L}_2\}$, one can derive similar error bounds and a similar numerical procedure as for the equilibrium probabilities $p_{m,n}$. The formula for the mean of the product of the queue lengths $\mathbf{L}_1$ and $\mathbf{L}_2$ enables us to compute also the coefficient of correlation of $\mathbf{L}_1$ and $\mathbf{L}_2$:

$$\rho(\mathbf{L}_1,\mathbf{L}_2) = \frac{E\{\mathbf{L}_1\mathbf{L}_2\} - E\mathbf{L}_1\,E\mathbf{L}_2}{\sigma(\mathbf{L}_1)\,\sigma(\mathbf{L}_2)}\,. \qquad (2.41)$$

In Table 2.3, numerical results for the symmetric 2×2 switch are presented for increasing values of r. In the second column, the numbers of terms of the two series in (2.39) needed to compute $\mathbb{E}\{\mathbf{L}_1\mathbf{L}_2\}$ within an absolute accuracy of 10^{-6} are given. Next, the relevant information for the length of a queue at each server (note that $\mathbf{L}_1$ and $\mathbf{L}_2$ are equally distributed in this case), and the total number of jobs $\mathbf{L}$ is given. After that, it is presented which values for the total number of jobs one would obtain in case we would have a system consisting of 2 independent queues; the corresponding random variable is denoted by $\mathbf{L}_{ind}$. Finally, the coefficient of correlation $\rho(\mathbf{L}_1,\mathbf{L}_2)$ is given. The results show that the higher the workload r, the more terms are needed for the computation of $\mathbb{E}\{\mathbf{L}_1\mathbf{L}_2\}$ (a similar behavior has been established for the number of terms needed for the computation of the equilibrium probabilities $p_{m,n}$). This is caused by the fact that the product factors are relatively large and decrease only slowly to 0 for high values of r. The results also show that the total number of jobs for the 2×2 switch has a smaller variability than the total number of jobs $\mathbf{L}_{ind}$ for a system with independent queues (note that the means of $\mathbf{L}$ and $\mathbf{L}_{ind}$ are the same). This is linked up with the negative correlation of the two queue lengths $\mathbf{L}_1$ and $\mathbf{L}_2$, which we have due to the negative coupling between the two streams of arriving jobs; there is a negative correlation of $-r/(2-r)$ between the number of jobs arriving at the two servers at the beginning of an arbitrary time unit.

Let $\mathbf{N}$ denote the number of non-empty queues at the beginning of a time unit and let $p(k)$ denote the probability that $\mathbf{N}$ equals k. Then

r	terms	$\mathbb{E}\mathbf{L}_i$	$\sigma(\mathbf{L}_i)$	$cv(\mathbf{L}_i)$	$\mathbb{E}\mathbf{L}$	$\sigma(\mathbf{L})$	$cv(\mathbf{L})$	$\mathbb{E}\mathbf{L}_{ind}$	$\sigma(\mathbf{L}_{ind})$	$cv(\mathbf{L}_{ind})$	$\rho(\mathbf{L}_1,\mathbf{L}_2)$
0.01	(1,1)	0.00	0.01	199.00	0.00	0.01	140.71	0.00	0.01	140.71	-0.000
0.2	(1,1)	0.01	0.11	9.00	0.03	0.16	6.33	0.03	0.16	6.36	-0.012
0.4	(1,1)	0.07	0.27	4.00	0.13	0.37	2.75	0.13	0.38	2.83	-0.057
0.6	(3,3)	0.23	0.53	2.33	0.45	0.69	1.52	0.45	0.74	1.65	-0.148
0.8	(5,5)	0.80	1.20	1.50	1.60	1.44	0.89	1.60	1.70	1.06	-0.282
0.95	(12,12)	4.51	4.99	1.11	9.03	5.51	0.61	9.03	7.05	0.78	-0.390
0.99	(30,30)	24.50	25.00	1.02	49.01	27.05	0.55	49.01	35.35	0.72	-0.415

Table 2.3. Some relevant performance measures for the symmetric 2×2 switch and the corresponding system consisting of 2 independent queues for varying values of the workload r.

$$p(0) = p_{0,0}\,, \quad p(1) = \sum_{m=1}^{\infty} p_{m,0} + \sum_{n=1}^{\infty} p_{0,n} = (1-\alpha_0)+(1-\hat{\beta}_0)-2p_{0,0}\,,$$

$$p(2) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} p_{m,n} = \sum_{i=0}^{\infty} \beta_i\,(\alpha_i-\alpha_{i+1}) + \sum_{i=0}^{\infty} \hat{\alpha}_i\,(\hat{\beta}_i-\hat{\beta}_{i+1})\,.$$

For the computation of $p(2)$, one can also use the formula $p(2)=1-p(0)-p(1)$. From the distribution of $\mathbf{N}$, one can easily compute the distribution $\{\hat{p}(k)\}$ of the number $\hat{\mathbf{N}}$ of working servers during a time unit. In Table 2.4, for the symmetric 2×2 switch, the distribution, mean, deviation and coefficient of correlation of $\hat{\mathbf{N}}$, and the corresponding quantities for the system consisting of independent queues, are given for increasing values of r; note that $\mathbb{E}\hat{\mathbf{N}}=2r$ for both systems. The results show that only for moderate values of r the negative correlation between the queue lengths causes a really smaller variability of the number of working servers for the 2×2 switch than for the system with independent queues. For high workloads, we obtain the same distributions and variabilities, since then in both systems the servers have to work almost all time units.

Let us finally pay some more attention to the coefficient of correlation $\rho(\mathbf{L}_1,\mathbf{L}_2)$. For the symmetric 2×2 switch, $\rho(\mathbf{L}_1,\mathbf{L}_2)$ is only a function of r, i.e. $\rho(\mathbf{L}_1,\mathbf{L}_2)=\rho(r)$. The function $\rho(r)$ is pictured in Figure 2.3. As we already learned from the results in Table 2.3, $\rho(r)$ is a negative, strictly decreasing function of r. Figure 2.3 shows that the negative correlation for the two queue lengths is very weak for low workloads r. It may easily be shown that $\rho(r)\sim-\frac{1}{4}r^2$ as $r\downarrow0$. For higher workloads r the negative correlation gets rather strong. The maximal strength of the correlation is reached for r close to 1: $\rho(r)\to-0.4203$ as $r\uparrow1$. Indeed, by exploiting the symmetry ($\hat{\beta}_i=\alpha_i$ and $\hat{\alpha}_i=\beta_i$ for all i) and using the formulae (2.22) and (2.23) for α_i and β_i, one can show that for $r\uparrow1$:

$$\alpha_i = 1-4(i+1)(2i+1)(1-r)+o(1-r)\,, \quad \beta_i = 1-4(i+1)(2i+3)(1-r)+o(1-r)\,.$$

Hence, from (2.40) and (2.41):

$$\lim_{r\uparrow1}\rho(r) = -1+2\sum_{i=0}^{\infty}\frac{1}{(i+1)(2i+3)}\left[\frac{1}{(i+1)(2i+1)}-\frac{1}{(i+2)(2i+3)}\right].$$

Rewriting, leads to

r	2×2 switch						independent queues					
	$\hat{p}(0)$	$\hat{p}(1)$	$\hat{p}(2)$	$I\!E\hat{N}$	$\sigma(\hat{N})$	$cv(\hat{N})$	$\hat{p}(0)$	$\hat{p}(1)$	$\hat{p}(2)$	$I\!E\hat{N}$	$\sigma(\hat{N})$	$cv(\hat{N})$
0.01	0.9801	0.0198	0.0001	0.02	0.141	7.027	0.9801	0.0198	0.0001	0.02	0.141	7.036
0.2	0.6242	0.3516	0.0242	0.4	0.537	1.343	0.6400	0.3200	0.0400	0.4	0.566	1.414
0.4	0.3151	0.5697	0.1151	0.8	0.625	0.781	0.3600	0.4800	0.1600	0.8	0.693	0.866
0.6	0.1025	0.5950	0.3025	1.2	0.604	0.503	0.1600	0.4800	0.3600	1.2	0.693	0.577
0.8	0.0088	0.3824	0.6088	1.6	0.508	0.317	0.0400	0.3200	0.6400	1.6	0.566	0.354
0.95	0.0000	0.1000	0.9000	1.9	0.300	0.158	0.0025	0.0950	0.9025	1.9	0.308	0.162
0.99	0.0000	0.0200	0.9800	1.98	0.140	0.071	0.0001	0.0198	0.9801	1.98	0.141	0.071

Table 2.4. The distribution of the number of working servers during a time unit for the symmetric 2×2 switch; the second part gives the distribution for independent queues.

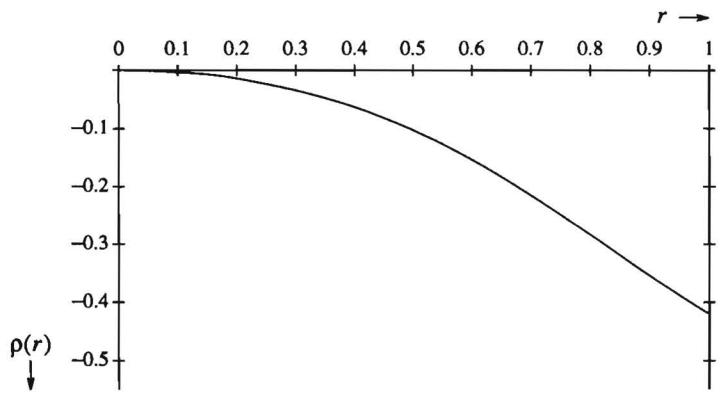


Figure 2.3. The coefficient of correlation $\rho(r)$ of the two queue lengths for the symmetric 2×2 switch.

$$\begin{aligned}
\lim_{r\uparrow 1}\rho(r) &= -1+2\sum_{k=1}^{\infty}\frac{1}{k^2(2k-1)(2k+1)}-2\sum_{k=1}^{\infty}\frac{1}{k(k+1)(2k+1)^2}\\
&= -1+8\sum_{k=1}^{\infty}\frac{1}{(2k-1)(2k+1)}-2\sum_{k=1}^{\infty}\frac{1}{k^2}-2\sum_{k=1}^{\infty}\frac{1}{k(k+1)}+8\sum_{k=1}^{\infty}\frac{1}{(2k+1)^2}\\
&= -1+4-\frac{\pi^2}{3}-2+8\left(\frac{\pi^2}{8}-1\right)\\
&= \frac{2}{3}\pi^2-7\,.
\end{aligned}$$

This leading term for the heavy traffic behavior of the queue length correlation coefficient also follows from Section 5 of Jaffe [47].

2.6. Complex-variable methods

As observed in the introduction of this chapter, Jaffe [46, 47] has analyzed the *symmetric* 2×2 switch by two different complex-variable methods, viz. the boundary value method and the uniformization technique. In the present section we briefly outline these two solutions, and we point out some differences and similarities with the results obtained by the compensation approach for the symmetric 2×2 switch.

In the symmetric case, the input parameters of the 2×2 switch are given by (2.38). In both complex-variable methods the first step is the introduction of the generating function

$$f(x,y) := \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} p_{m,n}\, x^m y^n\,, \quad |x|\le 1,\ |y|\le 1.$$

It follows from (2.7)-(2.10) and (2.39) that, for $|x|\le 1$, $|y|\le 1$, $f(x,y)$ satisfies the following functional equation:

$$\begin{aligned}(xy-\hat r(x,y))f(x,y) \;=\; &(y-1)\hat r(x,0)f(x,0)+(x-1)\hat r(0,y)f(0,y)\\ &+(x-1)(y-1)\hat r(0,0)f(0,0)\,,\end{aligned} \tag{2.42}$$

where

$$\hat r(x,y) := (1-r+\frac{r}{2}(x+y))^2\,.$$

Denote by S the complex curve $xy-\hat r(x,y)=0$ (the zeros of the 'kernel' of (2.42)), by D the interior of the unit circle, by $\overline{D}$ the closure of the unit circle; and, subsequently, separate the x- and y-parts of the right-hand side of (2.42) by defining

$$g(x) := \frac{\hat r(x,0)f(x,0)}{x-1}+\frac{1}{2}\hat r(0,0)f(0,0) = \frac{\hat r(0,x)f(0,x)}{x-1}+\frac{1}{2}\hat r(0,0)f(0,0)\,. \tag{2.43}$$

Then the boundedness of $f(x,y)$ in $\overline{D}^2$ implies that

$$g(x)+g(y) = 0\,, \quad (x,y)\in S\cap\overline{D}^2,\ x,y\ne 1, \tag{2.44}$$

whereas, as seen from (2.43), g has a simple pole at 1. Formula (2.44) is the fundamental equation in both complex-variable methods, which are successively discussed below.

A. The boundary value method

We outline the approach in Jaffe [46]. The boundary value method considers a suitable subset of $S\cap\overline{D}^2$, by taking $y=\bar x$. Formula (2.44) now reduces to

$$g(x)+g(\bar x) = 0\,, \quad x\in E\backslash\{1\}.$$

Here, E is the ellipse $\{x \mid |x|^2=\hat r(x,\bar x),\ x\in D\}$. Let ϕ, with inverse ψ, be the conformal mapping of the unit disk onto the region bounded by E, with normalization conditions $\phi(0)=r/(1+r)$, $\phi(1)=1$. Define $h(w):=g(\phi(w))$. We now obtain a 'boundary value problem with a pole' of an extremely simple form, cf. Section I.3.3 of [23], for $h(.)$ on the unit circle Γ:

$$\operatorname{Re} h(w) = 0, \quad w \in \Gamma \setminus \{1\},$$

$$\lim_{w \to 1} (w-1) h(w) = \frac{1-r}{\phi'(1)},$$

with $h(.)$ analytic on D and continuous on $\bar{D} \setminus \{1\}$. The solution of this boundary value problem is

$$h(w) = \frac{1}{2} \frac{1-r}{\phi'(1)} \frac{w+1}{w-1}, \quad w \in D,$$

which determines $g(x) = h(\psi(x))$; here, the conformal mapping $\psi(x)$ is explicitly expressed in the Jacobi elliptic (*sin am* or *sn*) function. Substitution of

$$g(u) = \frac{1}{2} \frac{1-r}{\phi'(1)} \frac{\psi(u)+1}{\psi(u)-1}$$

in (2.42) (for $u = x, y$) finally yields $f(x,y)$, for $|x| \le 1$, $|y| \le 1$:

$$f(x,y) = (1-r)\psi'(1) \frac{(x-1)(y-1)}{(\psi(x)-1)(\psi(y)-1)} \frac{\psi(x)\psi(y)-1}{xy - \hat{r}(x,y)}.$$

B. The uniformization technique

Starting point is again formula (2.44). Jaffe [47] exploits the following idea. Suppose that g is meromorphic, i.e., all its singularities are isolated poles. It follows from (2.43) that g has a simple pole at 1, with residue $1-r$. For points $(x,y) \in S$, a simple pole of $g(x)$ at x with known residue must be *compensated* by a simple pole of $g(y)$ at y, with residue being determined by (2.44). Starting from the pole at 1, one now iteratively determines a countable set of poles. Jaffe shows that g has simple poles at

$$b_m := \hat{\alpha} \lambda^m + \hat{\beta} \lambda^{-m} + \hat{\gamma}, \quad m = 0, 1, \cdots, \tag{2.45}$$

with residues

$$a_m := (-1)^m \sqrt{1-r^2}\, (\hat{\alpha} \lambda^m - \hat{\beta} \lambda^{-m}); \tag{2.46}$$

here

$$\hat{\alpha} := \frac{1+\sqrt{1-r^2}}{2(1+r)}, \quad \hat{\beta} := \frac{1-\sqrt{1-r^2}}{2(1+r)}, \quad \hat{\gamma} := \frac{r}{1+r}, \quad \lambda := \frac{\hat{\alpha}}{\hat{\beta}}.$$

Observe that $b_0 = 1 < b_1 < b_2 < \cdots$. For later reference we note that

$$b_1 = (\frac{2}{r} - 1)^2, \quad b_2 = (1 + \frac{2}{r} - \frac{4}{r^2})^2. \tag{2.47}$$

Jaffe derives (2.45) and (2.46) by first introducing linear coordinate changes for the hyperbola $xy - \hat{r}(x,y) = 0$:

$$x = \hat{\alpha} \hat{x} + \hat{\beta} \hat{y} + \hat{\gamma}, \quad y = \hat{\alpha} \hat{y} + \hat{\beta} \hat{x} + \hat{\gamma},$$

which transform S into $\hat{x}\hat{y} = 1$, or $\hat{y} = 1/\hat{x}$. Hence $\hat{x}$ is a *uniformizing variable* which parameterizes S. Jaffe shows that the transformed version of (2.44) gives rise to simple poles at $\hat{x} = \lambda^m$, $m = 0, 1, \cdots$, which implies that g has simple poles given by (2.45). He finally verifies that his initial assumption of g being a meromorphic function is indeed correct.

Comparison between the compensation approach and the uniformization technique

Using the compensation approach we have $\hat{\beta}_i = \alpha_i$ and $\hat{\alpha}_i = \beta_i$ for the completely symmetric case. Hence, from (2.36), we obtain the following generating function of $\{p_{m,n}\}$:

$$p_{0,0} + \sum_{i=0}^{\infty} \frac{1-\beta_i}{1-\beta_i y}\left[\frac{(1-\alpha_i)\alpha_i x}{1-\alpha_i x} - \frac{(1-\alpha_{i+1})\alpha_{i+1} x}{1-\alpha_{i+1} x} - (\alpha_i - \alpha_{i+1})\beta_i y\right]$$
$$+ \sum_{i=0}^{\infty} \frac{1-\beta_i}{1-\beta_i x}\left[\frac{(1-\alpha_i)\alpha_i y}{1-\alpha_i y} - \frac{(1-\alpha_{i+1})\alpha_{i+1} y}{1-\alpha_{i+1} y} - (\alpha_i - \alpha_{i+1})\beta_i x\right]. \tag{2.48}$$

Observe that this generating function, and the ones for $\{p_{m,0}\}$ and $\{p_{0,n}\}$, are meromorphic functions with simple poles at $1/\alpha_i$, $i = 0, 1, \cdots$ and at $1/\beta_i$, $i = 0, 1, \cdots$. We claim that the sequence $\{1/\alpha_0, 1/\beta_0, 1/\alpha_1, 1/\beta_1, \cdots\}$ corresponds to the sequence $\{b_1, b_2, b_3, b_4, \cdots\}$. From (2.12), it is seen that

$$\alpha_0 = \left(\frac{2}{r} - 1\right)^{-2}, \quad \beta_0 = \left(1 + \frac{2}{r} - \frac{4}{r^2}\right)^{-2}.$$

Comparison with (2.47) reveals that indeed $b_1 = 1/\alpha_0$ and $b_2 = 1/\beta_0$. In fact Jaffe [47] starts with $b_0 = 1$ and b_1, which corresponds to our observation, in the proof of Lemma 2.2, that the quadratic equation for α as given at the end of the proof, has two real solutions, viz. 1 and α_0. The successive α_i and β_i are determined from (2.22) and (2.23), what really amounts to finding product forms which satisfy the equilibrium equation (2.7) for the interior and which, together with a previous product form, satisfy one of the two equilibrium equations (2.8) and (2.9) for the boundaries. In terms of generating functions, this is translated into finding those zero tuples (x,y) of the 'kernel' $xy - \hat{r}(x,y) = 0$ that are related via (2.44) (note that (i) the kernel $xy - \hat{r}(x,y) = 0$ is completely determined by the behavior of the random walk in the interior; (ii) the right-hand side of (2.42) reflects the behavior of the random walk on the boundaries; and (iii) demanding that (2.44) holds for points (x,y) that are zeroes of the kernel corresponds to demanding that the equilibrium equations are satisfied both in the interior and on the boundaries). Remember that in the compensation approach each time a new term is added, to compensate an error on one of the boundaries; in terms of generating functions, this is translated into adding a new pole b_m to compensate a pole b_{m-1} in (2.44). The above reasoning implies the following:

(i) The mechanism to find β_i for given α_i (or α_{i+1} for given β_i) is equivalent with Jaffe's mechanism to find b_m given b_{m-1}, viz., by solving the equation $b_{m-1} y - \hat{r}(b_{m-1}, y) = 0$.

(ii) $b_{2m+1} = 1/\alpha_m$, $b_{2m+2} = 1/\beta_m$.

Hence we see that the generating function given by (2.48) has exactly the same (simple) poles as the generating function $f(x,y)$, and that in both approaches these poles, in increasing order of absolute value, are successively obtained from one another by compensating the effect of the preceding pole.

Comparison between the boundary value method and the uniformization technique

In the boundary value method, $g(x)$ has poles at the zeroes of $\psi(x) - 1 = 0$. The normalization condition $\phi(1) = 1$ for the conformal mapping implies $\psi(1) = 1$, so that $b_0 = 1$ is again found to be a pole of g. The periodic nature of the Jacobian elliptic function $\psi(.)$ subsequently leads to the sequence of poles $b_1, b_2, \cdots$.

From an analytic point of view, both complex-variable methods are for the present model of similar complexity (compared with the shortest queue problem and similar two-dimensional problems, one might say: of similar simplicity). They lead to different representations of the two-dimensional queue-length generating function. From a numerical point of view these representations can be exploited to obtain, e.g., queue length moments; however, the explicit representation obtained by the compensation approach seems more suitable for numerical calculations.

2.7. Conclusions

In this chapter, we have considered the class of two-dimensional, irreducible, positive recurrent, homogeneous, nearest-neighboring random walks with the projection property. This class is a subclass of the class of random walks studied by Adan et al. [12]. For a random walk of the considered class, according to [12], one can use the compensation approach to determine the equilibrium distribution if and only if for the states in the interior there are no transitions possible to the North, East and North-East. The study in this chapter has shown that if this condition is satisfied, then the equilibrium distribution is equal to the sum of two alternating series of pure product-form solutions of the equilibrium equation for the interior, and explicit formulae have been obtained for all product factors of the product-form solutions; see Theorem 2.1, where this main result has been stated.

In the last two sections of this chapter, some additional results have been presented. First of all, we have derived error bounds for the two series of product forms which constitute the equilibrium distribution. These error bounds have led to an efficient numerical procedure for the computation of the equilibrium distribution, and numerical results have been presented to show that for all states which are not too close to the origin, only a few product forms are needed to approximate the equilibrium probabilities sufficiently accurate. Further, it has been shown that the explicit formulae for the equilibrium distribution may be used to obtain explicit formulae and efficient numerical procedures for the relevant performance measures. Finally, based on the symmetric 2×2 switch, the compensation approach has been compared to the other methods available for the analysis of two-dimensional random walks. The compensation approach appears to have a striking resemblance to the uniformization technique. The product factors obtained by the compensation approach are the reciprocals of the poles of the generating functions for the equilibrium probabilities on the two axes found by applying the uniformization technique, and they are generated by a mechanism which is equivalent to the mechanism for the generation of the poles of the uniformization technique.

Chapter 3

The Equilibrium Distribution for a Class of Multi-Dimensional Random Walks

3.1. Introduction

This chapter is devoted to the application of the compensation approach to the class of N-dimensional, homogeneous, nearest-neighboring random walks with the projection property and with states $(m_1, \ldots, m_N)$, $m_i \in \mathbb{N}_0$ for all i. The *objective* is to generalize the main results which in the previous chapter have been derived for the case $N=2$. It will be established under which condition the compensation approach works for models with general $N \geq 2$. The condition appears to be rather simple. Because of the projection property, all transition probabilities/rates are uniquely determined by the transition probabilities/rates for the states in the interior. Let the probability/rate for a transition from an interior state into the direction $(t_1, \ldots, t_N)$ be denoted by $q_{t_1,\ldots,t_N}$. It will be shown that the compensation approach can be applied successfully under the following condition, which is a generalization of the condition for the two-dimensional case (see Assumption 2.2):

$$q_{t_1,\ldots,t_N} = 0 \quad \text{for all directions } (t_1, \ldots, t_N) \text{ with } t_i + t_j > 0 \text{ for some } i,j \in \{1, \ldots, N\},\ i \neq j. \tag{3.1}$$

This condition essentially restricts the applicability of the compensation approach for a random walk with dimension $N \geq 3$; in that case from a state in the interior only transitions are allowed to the state itself or to states closer to the origin (i.e. to states with a smaller total number of jobs, if the components m_i of the states $(m_1, \ldots, m_N)$ represent queue lengths). However, when the condition is satisfied, we obtain explicit results: the equilibrium distribution is equal to an alternating sum of infinitely many, pure product-form solutions of the equilibrium equation for the interior; and, similarly for all marginal distributions (see Section 3.8, Theorem 3.4). Note that, apart from the class of product-form networks (see Baskett et al. [15]), up to now no such explicit results have been derived for random walks/Markov processes with a three- or higher-dimensional state space being infinite in each dimension.

In the Sections 3.2-3.7, we present in detail the analysis for the three-dimensional case. Although the analysis will be rather long and complex, the main results, viz. the condition under which the compensation approach works and the explicit formulae which are obtained for the equilibrium distribution in case this condition is satisfied, will appear to be extremely compact and relatively simple. The analysis consists of two parts.

In the first part, consisting of the Sections 3.2-3.4, it is shown for the case $N=3$ that the condition stated in (3.1) is needed for the absolute convergence of the formal solutions, as constructed by the compensation approach. It will be indicated in Section 3.8, how the analysis leading to this result may be generalized to the case with an arbitrary $N \geq 2$. In the paper [78], on which this chapter is based in fact, it is shown how the analysis may be generalized to show that condition (3.1) is also needed necessary for random walks without the projection property.

In the second part, consisting of the Sections 3.5-3.7, it is shown for the class of three-dimensional random walks which satisfy condition (3.1), that the formal solutions constructed by the compensation approach are absolutely convergent, and that they may be used to obtain explicit formulae for the equilibrium distribution. This proves that condition (3.1) is also sufficient for the determination of the equilibrium distribution by the compensation approach. The explicit formulae for the equilibrium distribution are derived by using, among others, explicit formulae for the two-dimensional marginal distributions. Because of the projection property, these marginal distributions are the equilibrium distributions of two-dimensional random walks which explicitly can be solved by applying the theory of Chapter 2. As we will indicate in Section 3.8, the explicit formulae for the equilibrium distribution and the marginal distributions may be generalized to any $N \geq 2$ by using induction with respect to the dimension N. The induction step from dimension $N=2$ to dimension $N=3$, which in fact is described in this chapter, will appear to contain all elements required for the general induction step.

The main results of this chapter are summarized in the *Main Theorem* (i.e. Theorem 3.4) at the end of Section 3.8. This theorem will show that the infinitely many product-form solutions which constitute the equilibrium distribution of a random walk satisfying condition (3.1), are obtained from *(N-1)-fold trees* (i.e. trees, where at each node one parent splits into $N-1$ branches). The tree structure behind the relevant product forms will be investigated extensively in Chapter 4. In that chapter, it will be shown that the tree structure can be exploited to obtain error bounds and efficient numerical procedures for the computation of the equilibrium distribution and related quantities. Numerical results will be given for the $2 \times N$ switch (which satisfies condition (3.1)).

The organization of this chapter is as follows. In Section 3.2, we describe the class of N-dimensional random walks to which we want to apply the compensation approach, mainly for the case $N=3$. In Section 3.3, it is shown which type of formal solutions for the equilibrium equations is generated by the compensation approach for the three-dimensional case. The formal solutions are required to be absolutely convergent, which leads to condition (3.1) and to a reformulation of the formal solutions; this is the subject of Section 3.4. Next, in Section 3.5, condition (3.1) is shown to be also sufficient for the absolute convergence. Subsequently, explicit formulae for the equilibrium distribution are derived in Section 3.6, and alternative and more compact formulae for the equilibrium distribution, and its marginal distributions, are given in Section 3.7. In Section 3.8, the main results for the N-dimensional case are presented. Finally, in Section 3.9, the conclusions are presented.

3.2. The class of three-dimensional random walks

In this section, the class of N-dimensional, homogeneous, nearest-neighboring random walks with the projection property is described for the case $N=3$. The description is such that the definitions can easily be generalized to the case with general $N \geq 2$. Just as for the two-dimensional case, we may assume (w.l.o.g.) that we have discrete-time random walks. This assumption slightly simplifies the formulae, since a discrete-time random walk has the property that for each state the total probability of outgoing transitions adds up to 1. Nevertheless, the theory developed in this chapter also applies to continuous-time random walks.

For several queueing systems, the behavior is described by an N-dimensional random walk on states $(m_1, \ldots, m_N)$, where the components m_i denote queue lengths and are elements of the set $\mathbb{N}_0$ of nonnegative integers. For the class of three-dimensional random walks considered in this chapter, the state space M is assumed to be equal to

$$M = \{ (m,n,r) \mid m,n,r \in \mathbb{N}_0 \} .$$

The state space may be divided into a set of interior points and various sets of boundary points. Define

$$M_J = \{ (m_1,m_2,m_3) \in M \mid m_i > 0 \text{ for all } i \in J \text{ and } m_i = 0 \text{ for all } i \notin J \} , \quad J \subset I,$$

where $I := \{1,2,3\}$. Then M_I is the interior of M; $M_{\{1,2\}}$, $M_{\{1,3\}}$ and $M_{\{2,3\}}$ are the boundary planes; $M_{\{1\}}$, $M_{\{2\}}$ and $M_{\{3\}}$ are the axes; and, $M_\varnothing$ is the origin. The subscript indicates which of the components m_i are larger than zero; see Figure 3.1.

For the transitions and the corresponding probabilities, the same assumptions are made as for the case $N=2$ in the previous chapter. The main assumption concerns *homogeneity* in the transition probabilities. All states (m,n,r) with the same non-zero components, i.e. all states of the same set M_J, $J \subset I$, are assumed to have the same outgoing transition probabilities; these probabilities are denoted by variables $q^J_{t_1,t_2,t_3}$. Further, we assume that only transitions to *nearest neighbors* occur, and that the transition probabilities satisfy the *projection property*; these two assumptions are mainly made to simplify the analysis. Because of its complexity, the projection property is further explained in the next paragraph. After that, all three assumptions are formally described in Assumption 3.1.

In Section 2.2, the interpretation of the projection property has been given for the two-dimensional case; see also Figure 2.1. For the three-dimensional case, the projection property means the following. For the boundary plane $m=0$, i.e. for $M_{\{2,3\}}$, the probabilities $q^{\{2,3\}}_{t_2,t_3}$ are the same as the probabilities q^I_{1,t_2,t_3} for the interior, i.e. for M_I, and the probabilities $q^{\{2,3\}}_{0,t_2,t_3}$ are equal to the sums of the probabilities q^I_{0,t_2,t_3} and q^I_{-1,t_2,t_3}. So to speak, the set of transitions for $M_{\{2,3\}}$ is obtained by pushing the set of transitions for the interior against the boundary plane $m=0$. We say that the set of transitions for the boundary plane $M_{\{2,3\}}$ is the *projection* of the set of transitions for the interior M_I. Similarly, the sets of transitions for the boundary planes $n=0$ and $r=0$ (i.e. $M_{\{1,3\}}$ and $M_{\{1,2\}}$) are the projections of the set of transitions for the interior M_I. Just like for the origin in the two-dimensional case, for the axes and the origin the impact of the projection property is more complex. The set of transitions for the m-axis (i.e. $M_{\{1\}}$) is the projection of both the set of transitions for the boundary plane $n=0$ (i.e. $M_{\{1,3\}}$) and the set of transitions for the boundary plane $r=0$ (i.e. $M_{\{1,2\}}$); and

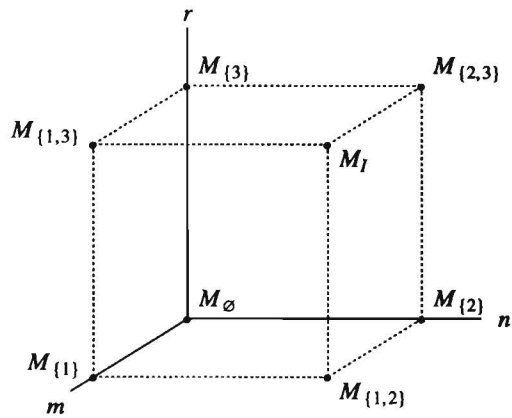


Figure 3.1. Eight states (m,n,r) of eight different subsets M_J of the state space M.

similarly for the sets of transitions for the n-axis and the r-axis. The set of transitions for the origin (i.e. $M_\varnothing$) is the projection of the set of transitions for the m-axis ($M_{\{1\}}$), the n-axis ($M_{\{2\}}$) as well as the r-axis ($M_{\{3\}}$). It may be verified that the projection property is satisfied if and only if the transition probabilities satisfy the equations as given in Assumption 3.1(iii).

Assumption 3.1.

(i) *For all states only transitions to nearest neighbors occur, i.e. for all $J \subset I$ and all states $(m,n,r) \in M_J$, it is only possible to make transitions to the states $(m+t_1, n+t_2, r+t_3)$ with $(t_1,t_2,t_3) \in T_J$, where*

$$T_J = \{ (t_1,t_2,t_3) \mid t_i \in \{-1,0,1\} \text{ for all } i \in J \text{ and } t_i \in \{0,1\} \text{ for all } i \notin J \}.$$

(ii) *Homogeneity: all states belonging to the same subset M_J, $J \subset I$, have the same outgoing transition probabilities; for all $J \subset I$, the probabilities for the transitions from a state $(m,n,r) \in M_J$ to the states $(m+t_1, n+t_2, r+t_3)$, $(t_1,t_2,t_3) \in T_J$, are denoted by variables $q^J_{t_1,t_2,t_3}$;*

(iii) *Projection property: the transition probabilities satisfy the following equations:*

$$q^J_{t_1,t_2,t_3} = \sum_{(u_1,u_2,u_3) \in U_J(t_1,t_2,t_3)} q^I_{u_1,u_2,u_3} \quad \text{for all } J \subset I \text{ and } (t_1,t_2,t_3) \in T_J,$$

where $U_J(t_1,t_2,t_3)$ is defined by

$$U_J(t_1,t_2,t_3) = \{ (u_1,u_2,u_3) \in T_I \mid u_i \in \{-1,0\} \text{ if } i \notin J \text{ and } t_i = 0;\ u_i = t_i \text{ else} \}.$$

It is clear that for a random walk with the projection property all transition probabilities are uniquely determined by the transition probabilities $q^I_{t_1,t_2,t_3}$ for the states in the interior of the state space. Because of the importance of these probabilities, in the sequel we shall omit the superscript I of $q^I_{t_1,t_2,t_3}$; similarly, we shall write T instead of T_I.

The properties stated in Assumption 3.1 are formulated in such a way that they can easily be generalized to the N-dimensional case. The first two properties are satisfied by several queueing systems, among others by the N-dimensional versions of the symmetric shortest queue system, the 2×2 switch and the fork and join model. The last two systems also satisfy the third property, i.e. the projection property. The symmetric shortest queue system violates the projection property (note that for the case $N=2$, this already has been verified in Section 2.2).

Example 3.1: The symmetric shortest queue system

The N-dimensional symmetric shortest queue system consists of N parallel, identical servers, where $N \geq 2$. All servers are assumed to have exponentially distributed service times with mean $1/\mu$. Jobs arrive at the system according to a Poisson stream with intensity λ, $0 < \lambda < N\mu$ (this implies the ergodicity of the system), and an arriving job always joins the shortest queue (ties are broken with equal probabilities). This system may be modeled by a continuous-time Markov process with states $(m_1, \ldots, m_N)$, where m_1 denotes the number of jobs at the shortest queue and m_i denotes the difference between the queue lengths of the i-th shortest queue and the $(i-1)$-th shortest queue (for all queue lengths the jobs in service are included). It may easily be verified that, for all $N \geq 2$, this system satisfies the Assumptions 3.1(i) and 3.1(ii), but violates Assumption 3.1(iii) (compare the transition rates for the subset $M_{\{2,3\}}$ to the transition rates for the interior M_I, for example). Nevertheless, the analysis presented in this chapter still is relevant for the symmetric shortest queue problem (one only obtains more complex formulae for the formal solutions generated by the compensation approach).

Example 3.2: The 2×N switch

The $2\times N$ switch is an extension of the 2×2 switch; see Example 2.2 in Section 2.2. The $2\times N$ switch is obtained in case one has N instead of 2 parallel servers. The behavior of this system is described by a discrete-time Markov process with states $(m_1, \ldots, m_N)$, where m_i denotes the number of waiting jobs at server i at the beginning of a time unit. It is left to the reader to verify that the $2\times N$ switch satisfies all three properties stated in Assumption 3.1.

Example 3.3: The fork and join model

Consider the fork and join model with N instead of 2 parallel servers; see Example 2.3 in Section 2.2. The behavior of the N-dimensional fork and join model is described by a continuous-time Markov process with states $(m_1, \ldots, m_N)$, where m_i denotes the number of unfinished subjobs at server i. Also for this system it may be shown that all three properties stated in Assumption 3.1 are satisfied.

Apart from satisfying Assumption 3.1, the random walks considered in this chapter are assumed to be *irreducible* and *positive recurrent* (= ergodic). The class of random walks which is obtained by these assumptions, is called the class of three-dimensional, irreducible, positive recurrent, homogeneous, nearest-neighboring random walks with the projection property. Each random walk of this class has an equilibrium distribution $\{p_{m,n,r}\}$, which is characterized as the unique normalized solution of the equilibrium equations. The equilibrium equations are given at the end of this section. After that, in the next section, we shall

try to solve them by applying the compensation approach. But first, we discuss the marginal distributions of $\{p_{m,n,r}\}$, which leads to some interesting properties.

Consider a three-dimensional, irreducible, positive recurrent, homogeneous, nearest-neighboring random walk with the projection property. Due to the projection property, we have the nice feature that the transitions for a subset of the components of the random walk are independent of the state of the whole system. This does not mean that the marginal distribution for a subset of the components is independent of the distribution for the other components, but it does mean that all marginal distributions of $\{p_{m,n,r}\}$ can be characterized as equilibrium distributions of lower-dimensional, homogeneous, nearest-neighboring random walks with the projection property.

Let us start with considering the one-dimensional marginal distributions $\{p_m^{(1)}\}$, $\{p_n^{(2)}\}$ and $\{p_r^{(3)}\}$ for the components m, n and r, respectively:

$$p_m^{(i)} = \sum_{\substack{(n_1,n_2,n_3)\in M \\ n_i=m}} p_{n_1,n_2,n_3}, \quad m \geq 0,\ i \in I. \tag{3.2}$$

Analyzing these distributions for the component random walks of the full random walk does not only lead to explicit formulae for these distributions, but it also leads to a simple, necessary and sufficient condition for the positive recurrence. Due to the projection property, for all states (m,n,r) with $m \geq 1$ the total probability for transitions to states $(m+t,n',r')$ equals $q_t^{(1)} = \sum_{(t,t_2,t_3)\in T} q_{t,t_2,t_3}$, where t is fixed and $t \in \{-1,0,1\}$, and for all states $(0,n,r)$ the total probability for transitions to states (t,n',r') equals $q_1^{(1)}$ for $t=1$ and $q_0^{(1)}+q_{-1}^{(1)}$ for $t=0$; similarly for the n- and r-direction. This shows that the distributions $\{p_m^{(i)}\}$ may be characterized as the equilibrium distributions of one-dimensional, homogeneous, nearest-neighboring random walks with the projection property; here, the transition probabilities for the interior are given by the variables $q_t^{(i)}$ defined by

$$q_t^{(i)} = \sum_{\substack{(s_1,s_2,s_3)\in T \\ s_i=t}} q_{s_1,s_2,s_3}, \quad t \in \{-1,0,1\},\ i \in I. \tag{3.3}$$

The full random walk will be positive recurrent if and only if all component random walks are positive recurrent, i.e. if and only if the component random walks have negative drifts. So, we obtain the following necessary and sufficient condition for the positive recurrence:

$$q_{-1}^{(i)} > q_1^{(i)} \quad \text{for all } i \in I. \tag{3.4}$$

Further, we obtain the following product-form (geometric) distributions for the one-dimensional marginal distributions $\{p_m^{(i)}\}$:

$$p_m^{(i)} = \left[1 - \frac{q_1^{(i)}}{q_{-1}^{(i)}}\right] \left[\frac{q_1^{(i)}}{q_{-1}^{(i)}}\right]^m, \quad m \geq 0,\ i \in I. \tag{3.5}$$

If $q_1^{(i)} = 0$ for some i, then all states (m_1,m_2,m_3) with $m_i > 0$ are transient and we can restrict ourselves to a lower-dimensional problem. The assumed irreducibility implies that this special case is excluded; so, $q_1^{(i)} > 0$ for all i. Together with (3.4), this leads to the property that

$$q_{-1}^{(i)} > q_1^{(i)} > 0 \quad \text{for all } i \in I. \tag{3.6}$$

Let us now consider the two-dimensional marginal distributions, which we denote by $\{p^{(2,3)}_{n,r}\}$, $\{p^{(1,3)}_{m,r}\}$ and $\{p^{(1,2)}_{m,n}\}$:

$$p^{(i,j)}_{m_1,m_2} = \sum_{\substack{(n_1,n_2,n_3)\in M \\ n_i=m_1,\ n_j=m_2}} p_{n_1,n_2,n_3}\,, \quad m_1,m_2 \geq 0,\ i,j\in I,\ i>j. \tag{3.7}$$

For all $i,j\in I$, $i>j$, $\{p^{(i,j)}_{m_1,m_2}\}$ is the equilibrium distribution of the two-dimensional random walk with the projection property for which the transition probabilities are given by

$$q^{(i,j)}_{t_1,t_2} = \sum_{\substack{(s_1,s_2,s_3)\in T \\ s_i=t_1,\ s_j=t_2}} q_{s_1,s_2,s_3}\,, \quad t_1,t_2\in\{-1,0,1\}. \tag{3.8}$$

Note that, according to the theory of Chapter 2, explicit formulae for the distributions $\{p^{(i,j)}_{m_1,m_2}\}$ can be obtained by applying the compensation approach if the transition probabilities $q^{(i,j)}_{t_1,t_2}$ satisfy Assumption 2.2, i.e. if

$$q^{(i,j)}_{0,1} = q^{(i,j)}_{1,0} = q^{(i,j)}_{1,1} = 0 \quad \text{for all } i,j\in I,\ i>j. \tag{3.9}$$

We finally give the equilibrium equations, which uniquely determine the equilibrium distribution $\{p_{m,n}\}$ and which will be tried to be solved by using the compensation approach. For the time being, we only need the equilibrium equations for the interior and the boundary planes $m=0$, $n=0$ and $r=0$, for which in the sequel we shall use the notations $M_{\Lambda\{1\}}$, $M_{\Lambda\{2\}}$ and $M_{\Lambda\{3\}}$ (instead of $M_{\{2,3\}}$, $M_{\{1,3\}}$ and $M_{\{1,2\}}$), since they indicate more explicitly which component must be equal to 0:

$$p_{m,n,r} = \sum_{(t_1,t_2,t_3)\in T} q_{t_1,t_2,t_3}\, p_{m-t_1,n-t_2,r-t_3}\,, \qquad (m,n,r)\in M_I, \tag{3.10}$$

$$p_{0,n,r} = \sum_{(-1,t_2,t_3)\in T} q_{-1,t_2,t_3}\, p_{1,n-t_2,r-t_3} + \sum_{(0,t_2,t_3)\in T} (q_{0,t_2,t_3}+q_{-1,t_2,t_3})\, p_{0,n-t_2,r-t_3}\,, \quad (0,n,r)\in M_{\Lambda\{1\}}, \tag{3.11}$$

$$p_{m,0,r} = \sum_{(t_1,-1,t_3)\in T} q_{t_1,-1,t_3}\, p_{m-t_1,1,r-t_3} + \sum_{(t_1,0,t_3)\in T} (q_{t_1,0,t_3}+q_{t_1,-1,t_3})\, p_{m-t_1,0,r-t_3}\,, \quad (m,0,r)\in M_{\Lambda\{2\}}, \tag{3.12}$$

$$p_{m,n,0} = \sum_{(t_1,t_2,-1)\in T} q_{t_1,t_2,-1}\, p_{m-t_1,n-t_2,1} + \sum_{(t_1,t_2,0)\in T} (q_{t_1,t_2,0}+q_{t_1,t_2,-1})\, p_{m-t_1,n-t_2,0}\,, \quad (m,n,0)\in M_{\Lambda\{3\}}. \tag{3.13}$$

3.3. The compensation approach

For a three-dimensional random walk of the class described in the previous section, the compensation approach generates whole networks of product-form solutions of the equilibrium equation (3.10) for the interior. Each network provides a formal solution of the equilibrium equations (3.10)-(3.13) for the interior and the three boundary planes, and is built up such that each product-form solution is connected to three other solutions (in fact, the structure of the network is such that we can make pairs of connected product-form solutions in three different ways, which is needed to show that the network of solutions satisfies the three equilibrium equations for the boundary planes). This section is devoted to the construction of the formal solutions by the compensation approach. Only in later sections, it will be shown under which conditions these formal solutions are well-defined and how they lead to a solution for *all* equilibrium equations.

The first step of the compensation approach consists of the characterization of appropriate product-form solutions of the equilibrium equation for the interior (see the main idea of the compensation approach as described in Section 1.3). Substituting $\alpha^m\beta^n\gamma^r$ in (3.10) and dividing both sides of the equation by $\alpha^{m-1}\beta^{n-1}\gamma^{r-1}$ leads to the following lemma:

Lemma 3.1.

The product form $\alpha^m\beta^n\gamma^r$ is a solution of the equilibrium equation (3.10) for the interior if and only if (α,β,γ) satisfies

$$\alpha\beta\gamma = \sum_{(t_1,t_2,t_3)\in T} q_{t_1,t_2,t_3}\,\alpha^{1-t_1}\beta^{1-t_2}\gamma^{1-t_3} \,. \tag{3.14}$$

Note that equation (3.14) is a quadratic equation in one variable in case two of the three factors α, β and γ are fixed.

Any linear combination $\Sigma_i c_i\alpha_i^m\beta_i^n\gamma_i^r$ consisting of solutions $(\alpha_i,\beta_i,\gamma_i)$ of (3.14) also satisfies (3.10). By the second part of the main idea of the compensation approach, we must try to build linear combinations which also satisfy the equilibrium equations (3.11)-(3.13) for the boundary planes. Product forms with one or more factors equal to zero lead to special, non-relevant cases, and, since later on the final solution has to be normalized, also product forms with one of the factors larger than or equal to 1 in absolute value are not relevant. Hence, we are only interested in solutions $(\alpha_i,\beta_i,\gamma_i)$ that are elements of

$$P = \{\,(\alpha,\beta,\gamma)\in\mathbb{C}^3 \mid (\alpha,\beta,\gamma) \text{ satisfies (3.14)},\ \alpha,\beta,\gamma\neq 0 \text{ and } |\alpha|,|\beta|,|\gamma| < 1\,\}\,.$$

We shall construct a linear combination consisting of solutions $(\alpha,\beta,\gamma)\in P$ by using the following compensation idea. Each error of a solution (α,β,γ) of equation (3.14) on one of the boundary planes (i.e. a violation of one of the equilibrium equations (3.11)-(3.13) for the boundary planes by the product-form solution $\alpha^m\beta^n\gamma^r$) may be compensated by adding another solution of (3.14). The addition of such a product-form solution is called a *compensation step*. We shall construct solutions of the equilibrium equations (3.10)-(3.14) by starting with a solution $(\alpha,\beta,\gamma)\in P$ and successively performing compensation steps to compensate errors of previous product-form solutions. Only later on, it will be checked whether the

product factors of the solutions added during the compensation steps are not equal to 0 and smaller than 1 in absolute value.

Let us start with the description of a compensation step on the boundary plane $m=0$ (i.e. on $M_{\Lambda\setminus\{1\}}$). Let $(\tilde\alpha,\tilde\beta,\tilde\gamma)$ satisfy (3.14), i.e. let $\tilde\alpha^m\tilde\beta^n\tilde\gamma^r$ be a solution of (3.10), and suppose that this product form violates the equilibrium equation (3.11) for the boundary plane $m=0$. To correct the error of $\tilde\alpha^m\tilde\beta^n\tilde\gamma^r$ on $m=0$, we add a compensation term $\hat a\alpha^m\beta^n\gamma^r$ such that (α,β,γ) satisfies (3.14) and $\tilde\alpha^m\tilde\beta^n\tilde\gamma^r+\hat a\alpha^m\beta^n\gamma^r$ satisfies (3.11). Substitution of this linear combination in (3.11) leads to the condition

$$K(\tilde\alpha,\tilde\beta,\tilde\gamma)\,\tilde\beta^{n-1}\tilde\gamma^{r-1}+\hat a\,K(\alpha,\beta,\gamma)\,\beta^{n-1}\gamma^{r-1}=0 \quad\text{for all } n\ge 1,\ r\ge 1, \tag{3.15}$$

where the function $K(\alpha,\beta,\gamma)$ is defined by

$$K(\alpha,\beta,\gamma)=\alpha\sum_{(-1,t_2,t_3)\in T}q_{-1,t_2,t_3}\,\beta^{1-t_2}\gamma^{1-t_3}+\sum_{(0,t_2,t_3)\in T}(q_{0,t_2,t_3}+q_{-1,t_2,t_3})\,\beta^{1-t_2}\gamma^{1-t_3}-\beta\gamma.$$

Here, $K(\alpha,\beta,\gamma)=0$ if and only if $\alpha^m\beta^n\gamma^r$ is a solution of (3.11). Because $\tilde\alpha^m\tilde\beta^n\tilde\gamma^r$ has been supposed to violate (3.11), $K(\tilde\alpha,\tilde\beta,\tilde\gamma)\neq 0$ and condition (3.15) forces us to take $\beta=\tilde\beta$ and $\gamma=\tilde\gamma$. Next, equation (3.14) is used for the choice of α. Substitution of $(\alpha,\beta,\gamma)=(\alpha,\tilde\beta,\tilde\gamma)$ into (3.14) and rearrangement of terms leads to the following quadratic equation for α:

$$\left[\sum_{(-1,t_2,t_3)\in T}q_{-1,t_2,t_3}\,\tilde\beta^{1-t_2}\tilde\gamma^{1-t_3}\right]\alpha^2-\left[\tilde\beta\tilde\gamma-\sum_{(0,t_2,t_3)\in T}q_{0,t_2,t_3}\,\tilde\beta^{1-t_2}\tilde\gamma^{1-t_3}\right]\alpha+\left[\sum_{(1,t_2,t_3)\in T}q_{1,t_2,t_3}\,\tilde\beta^{1-t_2}\tilde\gamma^{1-t_3}\right]=0. \tag{3.16}$$

The product factor $\tilde\alpha$ represents one root of this equation. Since choosing $\alpha=\tilde\alpha$ would not lead to compensation, we have to take α equal to the companion solution to $\tilde\alpha$ of (3.16), i.e.

$$\alpha=\hat\alpha=\frac{f_1(\tilde\beta,\tilde\gamma)}{\tilde\alpha},$$

where the function $f_1(\beta,\gamma)$ denotes the product of the roots α of the quadratic equation (3.14) for fixed β and γ:

$$f_1(\beta,\gamma)=\frac{\sum_{(1,t_2,t_3)\in T}q_{1,t_2,t_3}\,\beta^{1-t_2}\gamma^{1-t_3}}{\sum_{(-1,t_2,t_3)\in T}q_{-1,t_2,t_3}\,\beta^{1-t_2}\gamma^{1-t_3}}. \tag{3.17}$$

Finally, the factor $\hat a$ is chosen such that $\alpha^m\beta^n\gamma^r+\hat a\hat\alpha^m\beta^n\gamma^r$ satisfies (3.11). By (3.15), we obtain

$$\hat a=-\frac{K(\tilde\alpha,\tilde\beta,\tilde\gamma)}{K(\hat\alpha,\tilde\beta,\tilde\gamma)}.$$

Note that we would get $\hat a=0$ if $\tilde\alpha^m\tilde\beta^n\tilde\gamma^r$ already satisfied (3.11). By using the formula for the sum of the roots of the quadratic equation (3.16), we find

$$K(\tilde\alpha,\tilde\beta,\tilde\gamma)=(1-\hat\alpha)\sum_{(-1,t_2,t_3)\in T}q_{-1,t_2,t_3}\,\tilde\beta^{1-t_2}\tilde\gamma^{1-t_3},$$

$$K(\hat{\alpha},\tilde{\beta},\tilde{\gamma}) = (1-\alpha) \sum_{(-1,t_2,t_3)\in T} q_{-1,t_2,t_3}\, \tilde{\beta}^{1-t_2}\tilde{\gamma}^{1-t_3} ,$$

by which the expression for $\hat{a}$ may be rewritten to

$$\hat{a} = -\frac{1-\hat{\alpha}}{1-\tilde{\alpha}} .$$

The compensation step on the boundary plane $m=0$ fails if the denominator of $f_1(\tilde{\beta},\tilde{\gamma})$ vanishes (in that case (3.16) has only one solution). However, this is not very likely to occur, and therefore, for the time being, we shall neglect this special case (at the end of the next section, the formal solutions defined in this section are renovated, after which it is shown that this special case indeed does not occur).

For a compensation step on the boundary planes $n=0$ and $r=0$, similar results can be derived as for the compensation step on the boundary plane $m=0$. These results are summarized in the following lemma.

Lemma 3.2.

(i) Let (α,β,γ) *satisfy (3.14) and let* $a\in\mathbb{C}\backslash\{0\}$. *Then* $(\hat{\alpha},\beta,\gamma)$ *satisfies (3.14) and* $a\alpha^m\beta^n\gamma^r+\hat{a}\hat{\alpha}^m\beta^n\gamma^r$ *satisfies (3.11), if* $\hat{\alpha}$ *and* $\hat{a}$ *are taken equal to*

$$\hat{\alpha} = \frac{f_1(\beta,\gamma)}{\alpha} \quad \text{and} \quad \hat{a} = -\frac{1-\hat{\alpha}}{1-\alpha}\, a ,$$

where $f_1(\beta,\gamma)$ *is defined by (3.17);*

(ii) Let (α,β,γ) *satisfy (3.14) and let* $b\in\mathbb{C}\backslash\{0\}$. *Then* $(\alpha,\hat{\beta},\gamma)$ *satisfies (3.14) and* $b\alpha^m\beta^n\gamma^r+\hat{b}\alpha^m\hat{\beta}^n\gamma^r$ *satisfies (3.12), if* $\hat{\beta}$ *and* $\hat{b}$ *are taken equal to*

$$\hat{\beta} = \frac{f_2(\alpha,\gamma)}{\beta} \quad \text{and} \quad \hat{b} = -\frac{1-\hat{\beta}}{1-\beta}\, b ,$$

where $f_2(\alpha,\gamma)$ *is defined in the same way as* $f_1(\beta,\gamma)$, *but with the combinations* t,t_2,t_3 *replaced by* t_1,t,t_3 *for* $t=-1,1$ *and the powers* β^{1-t_2} *by* α^{1-t_1};

(iii) Let (α,β,γ) *satisfy (3.14) and let* $c\in\mathbb{C}\backslash\{0\}$. *Then* $(\alpha,\beta,\hat{\gamma})$ *satisfies (3.14) and* $c\alpha^m\beta^n\gamma^r+\hat{c}\alpha^m\beta^n\hat{\gamma}^r$ *satisfies (3.13), if* $\hat{\gamma}$ *and* $\hat{c}$ *are taken equal to*

$$\hat{\gamma} = \frac{f_3(\alpha,\beta)}{\gamma} \quad \text{and} \quad \hat{c} = -\frac{1-\hat{\gamma}}{1-\gamma}\, c ,$$

where $f_3(\alpha,\beta)$ *is defined in the same way as* $f_1(\beta,\gamma)$, *but with the combinations* t,t_2,t_3 *replaced by* t_1,t_2,t *for* $t=-1,1$ *and the powers* γ^{1-t_3} *by* α^{1-t_1}.

The Lemmas 3.1 and 3.2 provide the tools for the compensation approach to construct a solution of (3.10)-(3.13). Let $(\alpha,\beta,\gamma)\in P$, i.e. $\alpha^m\beta^n\gamma^r$ is a solution of (3.10). Most likely, this solution, which we call the *starting solution*, is not a solution of the equilibrium equations (3.11)-(3.13) for the boundary planes. Therefore *compensation terms* have to be added to correct the errors of $\alpha^m\beta^n\gamma^r$ on these boundary planes. To correct the error on the boundary plane $m=0$ for example, we have to add a product form $a_{(1)}\alpha_{(1)}^m\beta_{(1)}^n\gamma_{(1)}^r$ with $\beta_{(1)}=\beta$, $\gamma_{(1)}=\gamma$ and $\alpha_{(1)}$ and $a_{(1)}$ defined according to Lemma 3.2(i). Unfortunately, this compensation term introduces two new errors on the other two boundary planes. To compensate these new errors, which are hoped to be smaller than the initial error of $\alpha^m\beta^n\gamma^r$ on $m=0$, two more

compensation terms have to be added. To compensate the new error of the term $a_{(1)}\alpha_{(1)}^m\beta_{(1)}^n\gamma_{(1)}^r$ on $n=0$, we add a product form $a_{(1,2)}b_{(1,2)}\alpha_{(1,2)}^m\beta_{(1,2)}^n\gamma_{(1,2)}^r$ with $\alpha_{(1,2)}=\alpha_{(1)}$, $\gamma_{(1,2)}=\gamma_{(1)}$, $a_{(1,2)}=a_{(1)}$ and $\beta_{(1,2)}$ and $b_{(1,2)}$ defined according to Lemma 3.2(ii). To compensate the new error on $r=0$ a product form $a_{(1,3)}c_{(1,3)}\alpha_{(1,3)}^m\beta_{(1,3)}^n\gamma_{(1,3)}^r$ is added.

Continuing the above procedure leads to the generation of a tree or a network of product forms; see Figure 3.2. The product forms are labeled as follows. For each vector v out of the set

$$V = \{ (v_1, \ldots, v_l) \mid l \in \mathbb{N}_0, \text{ if } l \geq 1 \text{ then } v_1 \in I \text{ and } v_k \in I \backslash \{v_{k-1}\} \text{ for all } k \geq 2 \},$$

we get a product form $a_v b_v c_v \alpha_v^m \beta_v^n \gamma_v^r$. The empty vector $\varnothing$, which we get for $l=0$, is used as subscript for the starting solution. For all other elements $v=(v_1, \ldots, v_l) \in V\backslash\{\varnothing\}$ the product form $a_v b_v c_v \alpha_v^m \beta_v^n \gamma_v^r$ is the compensation term which compensates an error of $a_{p(v)} b_{p(v)} c_{p(v)} \alpha_{p(v)}^m \beta_{p(v)}^n \gamma_{p(v)}^r$, where $p(v)=(v_1, \ldots, v_{l-1})$ is the *parent* of v. The last component of v denotes on which boundary an error of $a_{p(v)} b_{p(v)} c_{p(v)} \alpha_{p(v)}^m \beta_{p(v)}^n \gamma_{p(v)}^r$ is compensated: on $m=0$ if $v_l=1$, on $n=0$ if $v_l=2$ and on $r=0$ if $v_l=3$. In Figure 3.2, the factors α, β and γ denote which new factor one gets for each compensation step. When compensating on $m=0$ we get a compensation term with a new α-factor, on $n=0$ we get a new β-factor and on $r=0$ we get a new γ-factor.

The sum of the starting solution $(\alpha,\beta,\gamma) \in P$ and all compensation terms is denoted by $x_{m,n,r}(\alpha,\beta,\gamma)$. So,

$$x_{m,n,r}(\alpha,\beta,\gamma) = \sum_{v \in V} a_v b_v c_v \alpha_v^m \beta_v^n \gamma_v^r , \tag{3.18}$$

where $\alpha_\varnothing=\alpha$, $\beta_\varnothing=\beta$, $\gamma_\varnothing=\gamma$ and for all $v \in V\backslash\{\varnothing\}$ we have (see Lemma 3.2 and the previous paragraph):

$$\beta_v=\beta_{p(v)}, \quad \gamma_v=\gamma_{p(v)}, \quad b_v=b_{p(v)}, \quad c_v=c_{p(v)},$$
$$\alpha_v = \frac{f_1(\beta_{p(v)},\gamma_{p(v)})}{\alpha_{p(v)}}, \quad a_v = -\frac{1-\alpha_v}{1-\alpha_{p(v)}} a_{p(v)} \quad \text{if } v_{l(v)}=1;$$

$$\alpha_v=\alpha_{p(v)}, \quad \gamma_v=\gamma_{p(v)}, \quad a_v=a_{p(v)}, \quad c_v=c_{p(v)},$$
$$\beta_v = \frac{f_2(\alpha_{p(v)},\gamma_{p(v)})}{\beta_{p(v)}}, \quad b_v = -\frac{1-\beta_v}{1-\beta_{p(v)}} b_{p(v)} \quad \text{if } v_{l(v)}=2;$$

$$\alpha_v=\alpha_{p(v)}, \quad \beta_v=\beta_{p(v)}, \quad a_v=a_{p(v)}, \quad b_v=b_{p(v)},$$
$$\gamma_v = \frac{f_3(\alpha_{p(v)},\beta_{p(v)})}{\gamma_{p(v)}}, \quad c_v = -\frac{1-\gamma_v}{1-\gamma_{p(v)}} c_{p(v)} \quad \text{if } v_{l(v)}=3.$$

Here, $l(v)$ is defined as the *length* (i.e. the number of components) of a vector $v \in V$ and $v_{l(v)}$ is the last component of v. For the initial factors $a_\varnothing$, $b_\varnothing$ and $c_\varnothing$ any choice is allowed. Define $a_\varnothing=1-\alpha$, $b_\varnothing=1-\beta$ and $c_\varnothing=1-\gamma$, then, by using induction with respect to $l(v)$, it may be shown that

$$a_v b_v c_v = (-1)^{l(v)}(1-\alpha_v)(1-\beta_v)(1-\gamma_v) \quad \text{for all } v \in V,$$

by which formula (3.18) simplifies to

$$x_{m,n,r}(\alpha,\beta,\gamma) = \sum_{v \in V} (-1)^{l(v)}(1-\alpha_v)\alpha_v^m (1-\beta_v)\beta_v^n (1-\gamma_v)\gamma_v^r . \tag{3.19}$$

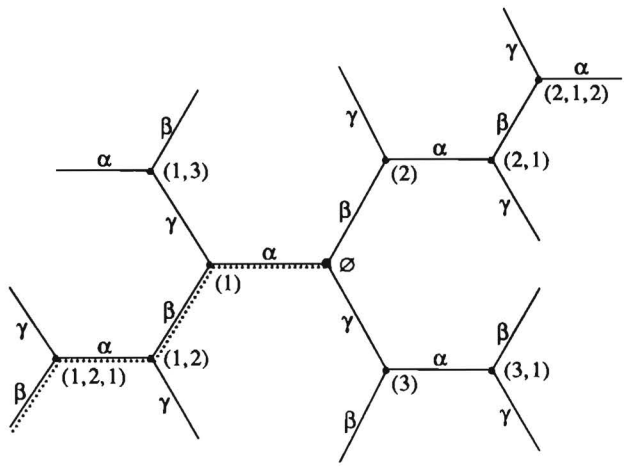


Figure 3.2. The construction process of a formal solution. Node v represents the product form $a_v b_v c_v \alpha_v^m \beta_v^n \gamma_v^r$.

So, each solution is a kind of alternating sum of pure product-form distributions.

For each $(\alpha,\beta,\gamma) \in P$, the solution $\{x_{m,n,r}(\alpha,\beta,\gamma)\}$ is well-defined by (3.19) if the sum in (3.19) converges absolutely, i.e. if

$$\sum_{v \in V} |(-1)^{l(v)}(1-\alpha_v)\alpha_v^m (1-\beta_v)\beta_v^n (1-\gamma_v)\gamma_v^r| < \infty. \tag{3.20}$$

In principle, this should hold for all states $(m,n,r) \in M$. Since we do not know for which starting solutions (α,β,γ), and for which states (m,n,r), condition (3.20) holds, we call each solution $\{x_{m,n,r}(\alpha,\beta,\gamma)\}$ a *formal* solution. If (3.20) holds, then $\{x_{m,n,r}(\alpha,\beta,\gamma)\}$ will be a solution of the equilibrium equations (3.10)-(3.13) for the interior and the boundary planes. Since each term of the sum in (3.19) is a solution of (3.10), it is obvious that the whole sum also satisfies (3.10). By taking connected pairs of product forms (see the network depicted in Figure 3.2), we find that $\{x_{m,n,r}(\alpha,\beta,\gamma)\}$ is also a solution of the equations for the boundary planes. If (3.20) holds, then $x_{m,n,r}(\alpha,\beta,\gamma)$ may be rewritten as a sum of pairs of product forms with the same β- and γ-factor (i.e. pairs of product forms which are connected by an α-edge in Figure 3.2),

$$x_{m,n,r}(\alpha,\beta,\gamma) = \sum_{\substack{v \in V \setminus \{\varnothing\} \\ v_{l(v)}=1}} (-1)^{l(v)} [(1-\alpha_{p(v)})\alpha_{p(v)}^m - (1-\alpha_v)\alpha_v^m](1-\beta_v)\beta_v^n (1-\gamma_v)\gamma_v^r, \tag{3.21}$$

from which it immediately follows that $\{x_{m,n,r}(\alpha,\beta,\gamma)\}$ is a solution of the equilibrium equation (3.11) for the boundary plane $m=0$, since each pair of product forms in the above sum is a solution of (3.11). By taking pairs with the same α-factor and γ-factor and pairs with the same α-factor and β-factor, it is readily seen that $\{x_{m,n,r}(\alpha,\beta,\gamma)\}$ satisfies the equilibrium equations (3.12) and (3.13) for the boundary planes $n=0$ and $r=0$. In the next section we shall investigate whether condition (3.20) is satisfied. We shall also investigate whether all

solutions $(\alpha_v, \beta_v, \gamma_v)$ of a formal solution are in the set P, especially whether

$$|\alpha_v| < 1, \quad |\beta_v| < 1, \quad |\gamma_v| < 1 \quad \text{for all } v \in V. \tag{3.22}$$

Since each starting solution is required to be an element of P, (3.22) is satisfied for $v = \varnothing$ by definition; for all other $v \in V$, this still must be verified.

3.4. Two necessary conditions

In this section we show that the requirements (3.20) and (3.22) lead to necessary conditions for the transition probabilities q_{t_1,t_2,t_3} in the interior of the state space and for the starting solution (α, β, γ). As we shall see, the resulting condition for the probabilities q_{t_1,t_2,t_3} is rather severe, but it is satisfied by the $2 \times N$ switch (the other two examples of queueing systems mentioned in Section 3.2 violate this condition). The resulting condition for the starting solution (α, β, γ) will lead to a small renovation of the definition of the formal solutions, but fortunately, it does not further restrict the applicability of the compensation approach.

In general, (almost) all terms of a formal solution $\{x_{m,n,r}(\alpha, \beta, \gamma)\}$ will be nonnull terms and in that case it is required that the terms get sufficiently small in absolute value for vectors $v \in V$ with large lengths $l(v)$. Define a *path* as a sequence $\{v^{(k)}\}$ of vectors of V with $v^{(0)} = \varnothing$ and $v^{(k)} \in O(v^{(k-1)})$ for all $k \geq 1$, where $O(v)$ is the *offspring* of a vector v:

$$O(v) = \{v' \in V \mid p(v') = v\}, \quad v \in V.$$

In Figure 3.2, the dotted line denotes an example of a path. Requirement (3.20) implies that

$$\sum_{k=0}^{\infty} |(-1)^{l(v)}(1-\alpha_{v^{(k)}})\alpha_{v^{(k)}}^{m}(1-\beta_{v^{(k)}})\beta_{v^{(k)}}^{n}(1-\gamma_{v^{(k)}})\gamma_{v^{(k)}}^{r}| < \infty \tag{3.23}$$

for all paths $\{v^{(k)}\}$. It is interesting to verify this condition for the two paths for which we have product forms with alternately new α- and β-factors (i.e. the paths with $v_{l(v)}^{(k)} \in \{1,2\}$ for all $k \geq 1$, see Figure 3.2). In general, for at least one of these two paths, all terms in the sum of (3.23) will be nonnull. Without loss of generality we may assume that this is at least the case for the path with $v^{(0)} = \varnothing$ and $v^{(k)} = (1,2,\ldots,1)$ if $k \geq 1$ and k odd, $v^{(k)} = (1,2,\ldots,1,2)$ if $k \geq 1$ and k even (which is precisely the path denoted by the dotted line in Figure 3.2). For this path condition (3.23) reduces to

$$\sum_{k=0}^{\infty} |(1-\alpha_{v^{(k)}})\alpha_{v^{(k)}}^{m}(1-\beta_{v^{(k)}})\beta_{v^{(k)}}^{n}| < \infty,$$

since $\gamma_{v^{(k)}} = \gamma$ for all k. Because we want this condition to hold for (almost) all m and n, it seems reasonable to require that

$$\alpha_{v^{(k)}} \to 0 \text{ and } \beta_{v^{(k)}} \to 0 \quad \text{as } k \to \infty \tag{3.24}$$

(alternative properties, such as $\alpha_{v^{(k)}} \to 1$ or $\beta_{v^{(k)}} \to 1$, are neglected, since it seems very unlikely that they can be satisfied). For odd k the factors $\alpha_{v^{(k-1)}}$ and $\alpha_{v^{(k)}}$ are the roots of the quadratic equation (3.14) (see also (3.16)) for fixed $\beta = \beta_{v^{(k-1)}}$ and γ and therefore

$$\alpha_{\nu^{(k-1)}}\alpha_{\nu^{(k)}} = f_1(\beta_{\nu^{(k-1)}},\gamma), \quad \alpha_{\nu^{(k-1)}} + \alpha_{\nu^{(k)}} = \frac{\beta_{\nu^{(k-1)}}\gamma - \sum_{(0,t_2,t_3)\in T} q_{0,t_2,t_3}\,\beta_{\nu^{(k-1)}}^{1-t_2}\gamma^{1-t_3}}{\sum_{(-1,t_2,t_3)\in T} q_{-1,t_2,t_3}\,\beta_{\nu^{(k-1)}}^{1-t_2}\gamma^{1-t_3}}.$$

According to (3.24), for both equations the parts on the left-hand side go to zero as $k\to\infty$, whereas the parts on the right-hand side go to

$$\frac{\sum_{(1,1,t_3)\in T} q_{1,1,t_3}\,\gamma^{1-t_3}}{\sum_{(-1,1,t_3)\in T} q_{-1,1,t_3}\,\gamma^{1-t_3}} \quad \text{and} \quad -\frac{\sum_{(0,1,t_3)\in T} q_{0,1,t_3}\,\gamma^{1-t_3}}{\sum_{(-1,1,t_3)\in T} q_{-1,1,t_3}\,\gamma^{1-t_3}}$$

respectively. Of course both sides of an equation have to go to the same limit and therefore (3.24) results in the condition that $q_{0,1,t_3} = q_{1,1,t_3} = 0$ for all t_3. In the same way, considering the sum and the product for $\beta_{\nu^{(k-1)}}$ and $\beta_{\nu^{(k)}}$ for even k, leads to the condition that $q_{1,0,t_3} = q_{1,1,t_3} = 0$ for all t_3. So, summarizing, in order to satisfy requirement (3.23) for the paths $\{\nu^{(k)}\}$ with $\nu_{(\nu)}^{(k)} \in \{1,2\}$ for all $k \geq 1$, it seems to be necessary that

$$q_{0,1,t_3} = q_{1,0,t_3} = q_{1,1,t_3} = 0 \text{ for all } t_3 \in \{-1,0,1\}.$$

Similar conditions are obtained by considering paths with $\nu_{(\nu)}^{(k)} \in \{1,3\}$ or $\nu_{(\nu)}^{(k)} \in \{2,3\}$ for all values of k. Combining these conditions results in the following necessary condition (cf. condition (3.1)), *which in the sequel (i.e. till the end of Section 3.6) is assumed to be satisfied:*

Condition 3.1. *(condition stemming from requirement (3.20))*

$$q_{t_1,t_2,t_3} = 0 \text{ for all transitions } (t_1,t_2,t_3)\in T \text{ with } t_i + t_j > 0 \text{ for some } i,j\in I,\ i\neq j. \quad (3.25)$$

Unfortunately, Condition 3.1 is rather severe: for all states in the interior, transitions may only have positive probabilities, if a positive step in one coordinate is always accompanied by negative steps in the other two coordinates, i.e. $q_{t_1,t_2,t_3} > 0$ and $t_i = 1$ for some $i\in I$ implies $t_j = -1$ for all $j\in I\backslash\{i\}$. In case the coordinates m, n and r represent queue lengths, Condition 3.1 implies that for all states (m,n,r) in the interior only transitions are possible to themselves or to states with a lower total number of jobs. Although Condition 3.1 only has been derived for the class of random walks described in Section 3.2, the condition is also necessary for random walks without the projection property (see also [78]); for these random walks the condition may be derived along the same lines (note that Condition 3.1 follows from a requirement with respect to the product factors α_ν, β_ν and γ_ν (see (3.24)), and that only the transition probabilities for the interior are involved in the definition of the product factors). The $2\times N$ switch with $N=3$ satisfies the condition, but the other two queueing problems mentioned in Section 3.2 violate the condition; this is verified below. As a consequence, we may conclude that the compensation approach possibly works for the 2×3 switch, whereas this method will not work for the other problems. For the 2×3 switch and the three-dimensional symmetric shortest queue system the same was already concluded in [77] from numerical results. Similar conclusions can be drawn for the N-dimensional case with $N>3$.

Example 3.1: The symmetric shortest queue system *(continued)*

In Section 3.2, we verified that this system violates the projection property. However, since Condition 3.1 is also necessary for random walks without this property, the condition still is relevant for the symmetric shortest queue system. For the three-dimensional case, the positive transition rates q_{t_1,t_2,t_3} for the interior are:

$$q_{1,-1,0} = 3\rho\,, \quad q_{-1,1,0} = q_{0,-1,1} = q_{0,0,-1} = 1\,.$$

The first rate is due to an arrival of a job and the other three rates stem from departures of jobs. Condition 3.1 is obviously violated and therefore we may conclude that, contrary to the two-dimensional case, the compensation approach will not work for the three-dimensional shortest queue system.

Example 3.2: The 2×N switch *(continued)*

For the 2×3 switch, the positive transition probabilities q_{t_1,t_2,t_3} for the interior are:

$$q_{1,-1,-1} = r_{1,1}r_{2,1}\,, \quad q_{-1,1,-1} = r_{1,2}r_{2,2}\,, \quad q_{-1,-1,1} = r_{1,3}r_{2,3}\,,$$

$$q_{-1,0,0} = r_{1,2}r_{2,3} + r_{1,3}r_{2,2}\,, \quad q_{0,-1,0} = r_{1,1}r_{2,3} + r_{1,3}r_{2,1}\,,$$

$$q_{0,0,-1} = r_{1,1}r_{2,2} + r_{1,2}r_{2,1}\,, \quad q_{-1,-1,0} = (1-r_1)r_{2,3} + (1-r_2)r_{1,3}\,,$$

$$q_{-1,0,-1} = (1-r_1)r_{2,2} + (1-r_2)r_{1,2}\,, \quad q_{0,-1,-1} = (1-r_1)r_{2,1} + (1-r_2)r_{1,1}\,,$$

$$q_{-1,-1,-1} = (1-r_1)(1-r_2)\,.$$

As we see, Condition 3.1 is satisfied for this system. This perfectly corresponds with the intuitive interpretation described above. If for the 2×3 switch at the beginning of a time unit all servers have jobs available (i.e. we are in a state of the interior), then at the next discrete time event we will have at least one job less in the system since three jobs are served and will leave the system while at most two jobs arrive.

Example 3.3: The fork and join model *(continued)*

For the three-dimensional version of the fork and join model we have

$$q_{1,1,1} = \lambda\,, \quad q_{-1,0,0} = \mu_1\,, \quad q_{0,-1,0} = \mu_2\,, \quad q_{0,0,-1} = \mu_3\,,$$

by which it is concluded that the compensation approach is not suitable for this case (as we saw in Section 2.2, the same holds for the two-dimensional fork and join model).

The assumption that in the sequel Condition 3.1 is satisfied brings on three interesting consequences. By Condition 3.1, we find

$$q_1^{(1)} = q_{1,-1,-1}\,, \quad q_1^{(2)} = q_{-1,1,-1}\,, \quad q_1^{(3)} = q_{-1,-1,1}\,,$$

which because of the assumed irreducibility implies that (see (3.6))

$$q_{1,-1,-1} > 0\,, \quad q_{-1,1,-1} > 0\,, \quad q_{-1,-1,1} > 0\,.$$

Since the positivity of these probabilities ensures the irreducibility, we find that, given Condition 3.1, the inequalities stated in (3.6) are necessary and sufficient for having an irreducible and positive recurrent random walk. Another consequence concerns the two-dimensional marginal distributions. Condition 3.1 appears to be equivalent to (3.9), and therefore the

assumption that this condition is satisfied implies that we can derive explicit formulae for the two-dimensional marginal distributions by applying the theory of Chapter 2. This will be exploited in Section 3.6. The third consequence concerns the considerable simplification that is obtained for the quadratic equation (3.14). Due to this simplification, some useful properties of the solutions of (3.14) can be proved by using *Rouché's Theorem* (cf. Titchmarsh [71]); see Lemma 3.3. These properties may be exploited for the derivation of a necessary condition for the starting solutions of the formal solutions, as we shall show after having proved Lemma 3.3.

Lemma 3.3.

(i) *For fixed* β *and* γ, $0<|\beta|<1$ *and* $0<|\gamma|<1$, *the quadratic equation (3.14) has exactly one root* α *with* $0<|\alpha|<C_1|\beta\gamma|$ *and* $C_1=q_1^{(1)}/q_{-1}^{(1)}<1$. *The second root* α, *which only exists if*

$$\sum_{(-1,t_2,t_3)\in T} q_{-1,t_2,t_3}\beta^{1-t_2}\gamma^{1-t_3}\neq 0,$$

satisfies $|\alpha|>|\beta\gamma|$.

(ii) *For fixed* α *and* γ, $0<|\alpha|<1$ *and* $0<|\gamma|<1$, *the quadratic equation (3.14) has exactly one root* β *with* $0<|\beta|<C_2|\alpha\gamma|$ *and* $C_2=q_1^{(2)}/q_{-1}^{(2)}<1$. *The second root* β, *which only exists if*

$$\sum_{(t_1,-1,t_3)\in T} q_{t_1,-1,t_3}\alpha^{1-t_1}\gamma^{1-t_3}\neq 0,$$

satisfies $|\beta|>|\alpha\gamma|$.

(iii) *For fixed* α *and* β, $0<|\alpha|<1$ *and* $0<|\beta|<1$, *the quadratic equation (3.14) has exactly one root* γ *with* $0<|\gamma|<C_3|\alpha\beta|$ *and* $C_3=q_1^{(3)}/q_{-1}^{(3)}<1$. *The second root* γ, *which only exists if*

$$\sum_{(t_1,t_2,-1)\in T} q_{t_1,t_2,-1}\alpha^{1-t_1}\beta^{1-t_2}\neq 0,$$

satisfies $|\gamma|>|\alpha\beta|$.

Proof.

We shall only prove part (i); the parts (ii) and (iii) may be proved along the same lines.

Consider the quadratic equation (3.14) for fixed β and γ, $0<|\beta|<1$ and $0<|\gamma|<1$. After rewriting (3.14) to (3.16), using (3.25), dividing by $\beta^2\gamma^2$ and substituting $z=\alpha/(\beta\gamma)$, we obtain the quadratic equation

$$\left(\sum_{(-1,t_2,t_3)\in T} q_{-1,t_2,t_3}\beta^{1-t_2}\gamma^{1-t_3}\right)z^2 - (1-[q_{0,0,0}+q_{0,-1,0}\beta+q_{0,0,-1}\gamma+q_{0,-1,-1}\beta\gamma])z+q_{1,-1,-1}=0. \tag{3.26}$$

Let $f(z)$ be the first term of the quadratic function in (3.26) and let $g(z)$ be the remaining part. Then we have the following bounds for $f(z)$ and $g(z)$ (for $z\neq 0$):

$$|f(z)|\leq\left(\sum_{(-1,t_2,t_3)\in T} q_{-1,t_2,t_3}|\beta^{1-t_2}\gamma^{1-t_3}|\right)|z|^2<q_{-1}^{(1)}|z|^2,$$

$$\begin{aligned} |g(z)| &\geq |z| - |[q_{0,0,0}+q_{0,-1,0}\beta+q_{0,0,-1}\gamma+q_{0,-1,-1}\beta\gamma]|\,|z| - q_{1,-1,-1} \\ &\geq (1-[q_{0,0,0}+q_{0,-1,0}+q_{0,0,-1}+q_{0,-1,-1}])|z| - q_{1,-1,-1} \\ &= (1-q_0^{(1)})|z| - q_1^{(1)} . \end{aligned}$$

As we see, these bounds only depend on the absolute value of z. For all z on the circle $C=\{z \mid |z|=r\}$ with radius $r>0$, we have $|f(z)|-|g(z)|<h(r)$, where

$$h(r) = q_{-1}^{(1)} r^2 - (1-q_0^{(1)})r + q_1^{(1)} .$$

Obviously, $|f(z)|<|g(z)|$ for all z on C, if r is chosen such that $h(r)\leq 0$. Since $h(r)$ is a convex quadratic function for which $h(0)=q_{1,-1,-1}>0$ and $h(1)=0$, the function $h(r)$ has two positive zeros, namely 1 and $r=q_1^{(1)}/q_{-1}^{(1)}$ (use the rule for the product of the two roots of a quadratic equation), and $h(r)\leq 0$ for all r in the closed interval between these two zeros. So, Rouché's theorem may be applied for all $r\in[C_1,1]$, where $C_1=q_1^{(1)}/q_{-1}^{(1)}$ (by (3.6), we know that $C_1<1$). This theorem tells that the number of solutions of (3.26) in the region $|z|<r$ is equal to the number of zeros of $g(z)$ in this region. The linear function $g(z)$ has one zero z_0, which is located in the region $|z|<C_1$, since

$$\begin{aligned} |z_0| &= \frac{q_{1,-1,-1}}{|1-[q_{0,0,0}+q_{0,-1,0}\beta+q_{0,0,-1}\gamma+q_{0,-1,-1}\beta\gamma]|} \\ &\leq \frac{q_{1,-1,-1}}{1-[q_{0,0,0}+q_{0,-1,0}+q_{0,0,-1}+q_{0,-1,-1}]} = \frac{q_1^{(1)}}{q_1^{(1)}+q_{-1}^{(1)}} < C_1 . \end{aligned}$$

As a result, applying Rouché's theorem for $r=C_1$ proves that (3.26) has exactly one solution with $|z|<C_1$, i.e. $|\alpha|<C_1|\beta\gamma|$ (since $q_{1,-1,-1}>0$, this solution is a nonnull solution, so we also know that $|\alpha|>0$). Next, applying Rouché's theorem for $r=1$ shows that if (3.26) has a second root, which is the case if and only if the coefficient of z^2 is not equal to zero, then this root must be in the region $|z|>1$, i.e. $|\alpha|>|\beta\gamma|$. □

Part (i) of Lemma 3.3 implies that if the quadratic equation (3.14) for fixed β and γ, $0<|\beta|<1$ and $0<|\gamma|<1$, has two roots α, then one root satisfies $0<|\alpha|<|\beta\gamma|$ and the other root satisfies $|\alpha|>|\beta\gamma|$; and similarly for the parts (ii) and (iii). Lemma 3.3 is used to prove Lemma 3.4, which states that for each relevant solution of the quadratic equation (3.14), i.e. for each $(\alpha,\beta,\gamma)\in P$, exactly one factor is smaller than the product of (some constant $C_i<1$ and) the other two factors. Next, with the help of the Lemmas 3.3 and 3.4, we will be able to define a path that always leads to a vector v for which one of the factors α_v, β_v and γ_v is larger than or equal to 1 in absolute value (i.e. for which (3.22) is violated).

Lemma 3.4.

Each solution $(\alpha,\beta,\gamma)\in P$ *possesses exactly one of the following three properties:*

(i) $|\alpha|<C_1|\beta\gamma|$, $|\beta|>|\alpha\gamma|$ *and* $|\gamma|>|\alpha\beta|$;

(ii) $|\beta|<C_2|\alpha\gamma|$, $|\alpha|>|\beta\gamma|$ *and* $|\gamma|>|\alpha\beta|$;

(iii) $|\gamma|<C_3|\alpha\beta|$, $|\alpha|>|\beta\gamma|$ *and* $|\beta|>|\alpha\gamma|$.

Proof.

By Lemma 3.3, each solution $(\alpha,\beta,\gamma)\in P$ satisfies

$$\begin{aligned}&(\ |\alpha|<C_1|\beta\gamma|\ \text{ or }\ |\alpha|>|\beta\gamma|\ )\\ &\text{and}\quad (\ |\beta|<C_2|\alpha\gamma|\ \text{ or }\ |\beta|>|\alpha\gamma|\ )\\ &\text{and}\quad (\ |\gamma|<C_3|\alpha\beta|\ \text{ or }\ |\gamma|>|\alpha\beta|\ ).\end{aligned}\tag{3.27}$$

Since $|\alpha|<C_1|\beta\gamma|$ implies that $|\alpha|<|\beta|$ and $|\alpha|<|\gamma|$, and similarly for $|\beta|<C_2|\alpha\gamma|$ and $|\gamma|<C_3|\alpha\beta|$, (α,β,γ) satisfies at most one of the '<'-inequalities in (3.27). Further, since (α,β,γ) is a solution of (3.14), (α,β,γ) satisfies

$$\begin{aligned}0 &= \left|\alpha\beta\gamma-\sum_{(t_1,t_2,t_3)\in T} q_{t_1,t_2,t_3}\,\alpha^{1-t_1}\beta^{1-t_2}\gamma^{1-t_3}\right|\\ &\geq |\alpha\beta\gamma|-\sum_{(t_1,t_2,t_3)\in T} q_{t_1,t_2,t_3}\,|\alpha^{1-t_1}\beta^{1-t_2}\gamma^{1-t_3}|\\ &= \sum_{(t_1,t_2,t_3)\in T} q_{t_1,t_2,t_3}\,(|\alpha\beta\gamma|-|\alpha^{1-t_1}\beta^{1-t_2}\gamma^{1-t_3}|)\\ &= q_{1,-1,-1}\,(|\alpha|-|\beta\gamma|)+q_{-1,1,-1}\,(|\beta|-|\alpha\gamma|)+q_{-1,-1,1}\,(|\gamma|-|\alpha\beta|)\\ &\quad+\sum_{t_1,t_2,t_3\in\{-1,0\}} q_{t_1,t_2,t_3}\,(|\alpha\beta\gamma|-|\alpha^{1-t_1}\beta^{1-t_2}\gamma^{1-t_3}|)\\ &\geq q_{1,-1,-1}\,(|\alpha|-|\beta\gamma|)+q_{-1,1,-1}\,(|\beta|-|\alpha\gamma|)+q_{-1,-1,1}\,(|\gamma|-|\alpha\beta|),\end{aligned}$$

which shows that (α,β,γ) cannot satisfy all three '>'-inequalities in (3.27). So, (α,β,γ) has to satisfy at least one of the '<'-inequalities. This proves that exactly one of the '<'-inequalities in (3.27) is satisfied, which completes the proof. □

Lemma 3.5.

For each starting solution $(\alpha,\beta,\gamma)\in P$, *there exists a vector* $v\in V$ *such that* $|\alpha_v|\geq 1$, $|\beta_v|\geq 1$ *or* $|\gamma_v|\geq 1$.

Proof.

Let $(\alpha,\beta,\gamma)\in P$ be a starting solution. Due to the properties stated in the Lemmas 3.3 and 3.4, we can construct a path $\{v^{(k)}\}$ for which the absolute values of the factors $\alpha_{v^{(k)}}$, $\beta_{v^{(k)}}$ and $\gamma_{v^{(k)}}$ are monotonously non-decreasing for increasing k. The path starts with the empty vector $\varnothing$, for which the corresponding solution $(\alpha_\varnothing,\beta_\varnothing,\gamma_\varnothing)=(\alpha,\beta,\gamma)$ is an element of P and thus satisfies exactly one of the three properties stated in Lemma 3.4. Suppose that property (i) is satisfied, i.e. $|\alpha_\varnothing|<C_1|\beta_\varnothing\gamma_\varnothing|$, $|\beta_\varnothing|>|\alpha_\varnothing\gamma_\varnothing|$ and $|\gamma_\varnothing|>|\alpha_\varnothing\beta_\varnothing|$. As we know, the vector (1) has the same β- and γ-factor as $\varnothing$, but a new α-factor $\alpha_{(1)}$, which is the companion solution to $\alpha_\varnothing$ of the quadratic equation (3.14) for fixed $\beta=\beta_\varnothing$ and $\gamma=\gamma_\varnothing$. By Lemma 3.3, $\alpha_{(1)}$ has to be the root which satisfies $|\alpha|>|\beta_\varnothing\gamma_\varnothing|$ and thus $\alpha_{(1)}$ is larger than $\alpha_\varnothing$ in absolute value. We find

$$|\alpha_{(1)}| > |\beta_{(1)}\gamma_{(1)}| = |\beta_{\varnothing}\gamma_{\varnothing}| > \frac{1}{C_1}|\alpha_{\varnothing}| = \frac{1}{C_1}|\alpha| .$$

If $|\alpha_{(1)}| < 1$, then $(\alpha_{(1)},\beta_{(1)},\gamma_{(1)})$ is also an element of P and therefore also satisfies one of the properties stated in Lemma 3.4. Since $|\alpha_{(1)}| > |\beta_{(1)}\gamma_{(1)}|$, it satisfies property (ii) or property (iii). Suppose that (ii) is satisfied, then it is useful to consider the vector (1,2). This vector has the same α- and γ-factor as (1), but a new and larger β-factor:

$$|\beta_{(1,2)}| > |\alpha_{(1,2)}\gamma_{(1,2)}| = |\alpha_{(1)}\gamma_{(1)}| > \frac{1}{C_2}|\beta_{(1)}| .$$

When comparing the factors of vector (1,2) with the factors of the starting solution, we find

$$|\alpha_{(1,2)}| = |\alpha_{(1)}| > \frac{1}{C_1}|\alpha| , \quad |\beta_{(1,2)}| > \frac{1}{C_2}|\beta_{(1)}| = \frac{1}{C_2}|\beta| , \quad |\gamma_{(1,2)}| = |\gamma_{(1)}| = |\gamma| .$$

If $|\beta_{(1,2)}| < 1$, then $(\alpha_{(1,2)},\beta_{(1,2)},\gamma_{(1,2)}) \in P$ and again the construction process may be continued.

In general we construct a path $\{\nu^{(k)}\}$ with $\nu^{(0)} = \varnothing$ and for all $k = 1,2,\cdots$ the vector $\nu^{(k)}$ is an element of the offspring of $\nu^{(k-1)}$, i.e. $p(\nu^{(k)}) = \nu^{(k-1)}$, and the last element $\nu^{(k)}_{l(\nu^{(k)})}$ of $\nu^{(k)}$ is taken equal to

$$\nu^{(k)}_{l(\nu^{(k)})} = \begin{cases} 1 & \text{if } |\alpha_{\nu^{(k-1)}}| < |\beta_{\nu^{(k-1)}}\gamma_{\nu^{(k-1)}}| ; \\ 2 & \text{if } |\beta_{\nu^{(k-1)}}| < |\alpha_{\nu^{(k-1)}}\gamma_{\nu^{(k-1)}}| ; \\ 3 & \text{if } |\gamma_{\nu^{(k-1)}}| < |\alpha_{\nu^{(k-1)}}\beta_{\nu^{(k-1)}}| . \end{cases}$$

Here, the construction process is stopped as soon as

$$|\alpha_{\nu^{(k)}}| \geq 1 , \quad |\beta_{\nu^{(k)}}| \geq 1 \text{ or } |\gamma_{\nu^{(k)}}| \geq 1 \tag{3.28}$$

for some $k \geq 1$. In that case $(\alpha_{\nu^{(k)}},\beta_{\nu^{(k)}},\gamma_{\nu^{(k)}})$ is not an element of P, by which the essential properties of Lemma 3.4 cannot be used anymore. To complete the proof of Lemma 3.5, it suffices to prove that (3.28) always occurs for some k.

For each vector $\nu^{(k)}$, $k \geq 1$, two of the factors $\alpha_{\nu^{(k)}}$, $\beta_{\nu^{(k)}}$ and $\gamma_{\nu^{(k)}}$ are equal to the corresponding factors for $\nu^{(k-1)}$, whereas the third factor is new and may be proved to be larger in absolute value (by using the Lemmas 3.3 and 3.4):

$$|\alpha_{\nu^{(k)}}| > \frac{1}{C_1}|\alpha_{\nu^{(k-1)}}| , \quad |\beta_{\nu^{(k)}}| = |\beta_{\nu^{(k-1)}}| , \quad |\gamma_{\nu^{(k)}}| = |\gamma_{\nu^{(k-1)}}| \quad \text{if } \nu^{(k)}_{l(\nu^{(k)})} = 1;$$

$$|\alpha_{\nu^{(k)}}| = |\alpha_{\nu^{(k-1)}}| , \quad |\beta_{\nu^{(k)}}| > \frac{1}{C_2}|\beta_{\nu^{(k-1)}}| , \quad |\gamma_{\nu^{(k)}}| = |\gamma_{\nu^{(k-1)}}| \quad \text{if } \nu^{(k)}_{l(\nu^{(k)})} = 2;$$

$$|\alpha_{\nu^{(k)}}| = |\alpha_{\nu^{(k-1)}}| , \quad |\beta_{\nu^{(k)}}| = |\beta_{\nu^{(k-1)}}| , \quad |\gamma_{\nu^{(k)}}| > \frac{1}{C_3}|\gamma_{\nu^{(k-1)}}| \quad \text{if } \nu^{(k)}_{l(\nu^{(k)})} = 3.$$

Let $n_i(k)$ denote the number of i-s in the sequence $\nu^{(1)}_{l(\nu^{(1)})}, \ldots, \nu^{(k)}_{l(\nu^{(k)})}$, i.e. in the vector $\nu^{(k)}$:

$$n_i(k) = |\{ l \mid l \in \{1,\ldots,k\} \text{ and } \nu^{(k)}_l = i \}| , \quad i \in I, k \geq 0.$$

Then, by induction, it is easily proved that

$$|\alpha_{\nu^{(k)}}| \geq \left[\frac{1}{C_1}\right]^{n_1(k)} |\alpha| , \quad |\beta_{\nu^{(k)}}| \geq \left[\frac{1}{C_2}\right]^{n_2(k)} |\beta| , \quad |\gamma_{\nu^{(k)}}| \geq \left[\frac{1}{C_3}\right]^{n_3(k)} |\gamma| \tag{3.29}$$

for all $k \geq 0$. Since all three constants C_i are smaller than 1 and $n_1(k)+n_2(k)+n_3(k)=k$ for all k, at least one of the right-hand sides in (3.29) has to become larger than or equal to 1 for some k, which proves that (3.28) always occurs for some k. □

Lemma 3.5 shows that the requirement stated in (3.22) is never satisfied by the formal solutions as defined in Section 3.3. However, (3.22) can be satisfied by modifying the formal solutions such that there is no longer a path leading to a vector v for which $|\alpha_v| \geq 1$, $|\beta_v| \geq 1$ or $|\gamma_v| \geq 1$. This is achieved by compensating the starting solution (α,β,γ) at only two of the three boundary planes: leave out the compensation at the boundary plane $m=0$ if $|\alpha| < |\beta\gamma|$, at the boundary plane $n=0$ if $|\beta| < |\alpha\gamma|$ and at $r=0$ if $|\gamma| < |\alpha\beta|$. In fact, the compensation at this third boundary plane is not needed if the starting solution already satisfies the equilibrium equation for this boundary plane. This results in the second necessary condition for a formal solution. Note that this condition is also sufficient to meet (3.22) (see also Lemma 3.6 below).

Condition 3.2 *(condition stemming from the requirements (3.20) and (3.22))*
A starting solution $(\alpha,\beta,\gamma) \in P$ also has to be a solution of the equilibrium equations for one of the boundary planes. It has to satisfy:

- *equilibrium equation (3.11) for the boundary plane $m=0$ if $|\alpha| < |\beta\gamma|$;*
- *equilibrium equation (3.12) for the boundary plane $n=0$ if $|\beta| < |\alpha\gamma|$;*
- *equilibrium equation (3.13) for the boundary plane $r=0$ if $|\gamma| < |\alpha\beta|$.*

From now on we are only interested in formal solutions $\{x_{m,n,r}(\alpha,\beta,\gamma)\}$ for which the starting solution (α,β,γ) satisfies Condition 3.2. All these starting solutions belong to one of the sets P_i, $i \in I$, where P_1 is defined as the set of appropriate starting solutions on the boundary plane $m=0$, i.e.

$$P_1 = \{ (\alpha,\beta,\gamma) \in P \mid \alpha^m\beta^n\gamma^r \text{ is also a solution of (3.11) and } |\alpha| < |\beta\gamma| \},$$

and P_2 and P_3 are defined as the sets of appropriate starting solutions on $n=0$ and $r=0$, respectively. Due to the projection property, there exists a nice characterization for the elements of the sets P_i; this characterization will be given in Section 3.6. For each starting solution $(\alpha,\beta,\gamma) \in P_i$, $i \in I$, the corresponding formal solution reduces to a *binary tree* of product forms and it is denoted by $\{x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)\}$, where the superscript (i) denotes on which boundary it starts:

$$x^{(i)}_{m,n,r}(\alpha,\beta,\gamma) = \sum_{v \in V_i} (-1)^{l(v)}(1-\alpha_v)\alpha_v^m\,(1-\beta_v)\beta_v^n\,(1-\gamma_v)\gamma_v^r \tag{3.30}$$

with

$$V_i = \{ (v_1,\ldots,v_l) \in V \mid \text{if } v \neq \varnothing \text{ then } v_1 \neq i \}.$$

The following lemma states that for each formal solution $\{x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)\}$ all factors α_v, β_v and γ_v are well-defined (i.e. one always has nonnull denominators for the terms $f_i(\cdot,\cdot)$ in the definitions of these factors) and all $(\alpha_v,\beta_v,\gamma_v)$ are elements of P. Part (i) of the following lemma may be easily proved with the help of Lemmas 3.3 and 3.4; the other parts follow from part (i).

Lemma 3.6.

Let $i \in I$ and $(\alpha,\beta,\gamma) \in P_i$. Then all product factors α_ν, β_ν and γ_ν of the formal solution $\{x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)\}$ are well-defined and they have the following properties:

(i) For all $\nu \in V_i$, we have

$$\begin{cases} 0 < |\alpha_\nu| < C_1 |\beta_\nu \gamma_\nu| & \text{if } \nu_{l(\nu)} = 1; \\ 0 < |\beta_\nu| < C_2 |\alpha_\nu \gamma_\nu| & \text{if } \nu_{l(\nu)} = 2; \\ 0 < |\gamma_\nu| < C_3 |\alpha_\nu \beta_\nu| & \text{if } \nu_{l(\nu)} = 3; \end{cases}$$

(ii) For each path $\{\nu^{(k)}\}$ in V_i, all three factors $\alpha_{\nu^{(k)}}$, $\beta_{\nu^{(k)}}$ and $\gamma_{\nu^{(k)}}$ are monotonously non-increasing in absolute value for increasing k;

(iii) $(\alpha_\nu,\beta_\nu,\gamma_\nu) \in P$ for all $\nu \in V_i$;

(iv) For each path $\{\nu^{(k)}\}$ in V_i, at least two of the three factors $\alpha_{\nu^{(k)}}$, $\beta_{\nu^{(k)}}$ and $\gamma_{\nu^{(k)}}$ tend exponentially fast to 0, as $k \to \infty$.

We end this section with a brief recapitulation of what we have found for the formal solutions generated by the compensation approach in the previous section. We have derived two conditions which are needed to satisfy the requirements stated in (3.20) and (3.22). The second condition has lead to the modified formal solutions $\{x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)\}$, which consist of well-defined solutions $(\alpha_\nu,\beta_\nu,\gamma_\nu) \in P$. The first condition, Condition 3.1, in principle has been obtained for the formal solutions $\{x_{m,n,r}\alpha,\beta,\gamma)\}$ defined in the previous section, but it is easily verified that this condition is also needed for the absolute convergence of the modified formal solutions $\{x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)\}$, i.e. for

$$\operatorname{abs}(x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)) = \sum_{\nu \in V_i} |(-1)^{l(\nu)}(1-\alpha_\nu)\alpha_\nu^m (1-\beta_\nu)\beta_\nu^n (1-\gamma_\nu)\gamma_\nu^r| < \infty. \qquad (3.31)$$

In the next section, the absolute convergence is further investigated (under the assumption that Condition 3.1 is satisfied). One of the difficulties we have to deal with is that each formal solution $\{x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)\}$ is a binary tree instead of a series, by which it is no longer sufficient for the convergence to prove that the ratio of two successive terms has a limit smaller than 1.

3.5. Absolute convergence of the formal solutions

This section is devoted to the proof of the following theorem, which holds under the assumption that Condition 3.1 is satisfied; from this theorem, it follows among others that the modified formal solutions $\{x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)\}$ converge absolutely in the larger part of the state space M.

Theorem 3.1.

For all $i \in I$ and $(\alpha,\beta,\gamma) \in P_i$:

(i) $x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)$ is absolutely convergent for all states $(m,n,r) \in M_c$, where

$$M_c = \{\, (m,n,r) \in M \mid (m,n,r) \in M_I \text{ or } (m,n,r) \in M_{\Lambda\{j\}} \text{ for some } j \in I \,\};$$

$$(ii)\quad \sum_{(m,n,r)\in M_c} |x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)| \le \sum_{(m,n,r)\in M_c} \mathrm{abs}(x^{(i)}_{m,n,r}(\alpha,\beta,\gamma))$$

$$\le \frac{\mathrm{abs}(x^{(i)}_{0,1,1}(\alpha,\beta,\gamma))}{(1-|\beta|)(1-|\gamma|)} + \frac{\mathrm{abs}(x^{(i)}_{1,0,1}(\alpha,\beta,\gamma))}{(1-|\alpha|)(1-|\gamma|)}$$

$$+ \frac{\mathrm{abs}(x^{(i)}_{1,1,0}(\alpha,\beta,\gamma))}{(1-|\alpha|)(1-|\beta|)} + \frac{\mathrm{abs}(x^{(i)}_{1,1,1}(\alpha,\beta,\gamma))}{(1-|\alpha|)(1-|\beta|)(1-|\gamma|)} < \infty.$$

Part (i) of this theorem states that $x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)$ is absolutely convergent for all states in the interior and on the boundary planes. The set of these states is called the *convergence region* M_c. It is easily shown that $x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)$ is not absolutely convergent, i.e. $\mathrm{abs}(x^{(i)}_{m,n,r}(\alpha,\beta,\gamma))$ diverges, on the axes and in the origin. For example, $x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)$ is shown to be not absolutely convergent for all states $(m,0,0)$ by considering the terms of a path $\{\nu^{(k)}\}$ in V_i with $\nu^{(k)}_{l(\nu)} \in \{2,3\}$ for all k. For this path $\alpha_{\nu^{(k)}} = \alpha$ for all k and $|\beta_{\nu^{(k)}}|$ and $|\gamma_{\nu^{(k)}}|$ decrease monotonously (see Lemma 3.6), so

$$\mathrm{abs}(x^{(i)}_{m,0,0}(\alpha,\beta,\gamma)) \ge \sum_{k=0}^{\infty} |(1-\alpha_{\nu^{(k)}})\alpha^m_{\nu^{(k)}}(1-\beta_{\nu^{(k)}})(1-\gamma_{\nu^{(k)}})|$$

$$\ge (1-|\beta|)(1-|\gamma|)\sum_{k=0}^{\infty} |(1-\alpha)\alpha^m| = \infty.$$

Part (ii) of Theorem 3.1 is needed in the next section; this part gives a useful upper bound for the summation of $\mathrm{abs}(x^{(i)}_{m,n,r}(\alpha,\beta,\gamma))$ over all $(m,n,r)\in M_c$ and it states that $\{x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)\}$ can be normalized. The properties stated in Lemma 3.6 constitute the basis of the proof of Theorem 3.1.

For the proof of Theorem 3.1, we shall use a recurrence relation for the sums $x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)$, which, as we know, are *binary trees* of product forms. Therefore, we temporarily have to extend the domains for the formal solutions $\{x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)\}$, which at the end of Section 3.4 only have been defined for starting solutions $(\alpha,\beta,\gamma)\in P_i$; see (3.30) and the definitions of the sets P_i. In the remainder of this section, we drop the condition that $(\alpha,\beta,\gamma)\in P_i$ has to satisfy the equilibrium equation for the i-th boundary plane and we let $\{x^{(1)}_{m,n,r}(\alpha,\beta,\gamma)\}$ be defined for all solutions $(\alpha,\beta,\gamma)\in P_i'$, where

$$P_1' = \{\,(\alpha,\beta,\gamma)\in P \mid |\alpha| < |\beta\gamma|\,\}$$

and P_2' and P_3' are defined similarly, but with the condition $|\alpha| < |\beta\gamma|$ replaced by $|\beta| < |\alpha\gamma|$ and $|\gamma| < |\alpha\beta|$, respectively. As one can easily verify, then the following recurrence relation holds for all $i\in I$ and $(\alpha,\beta,\gamma)\in P_i'$:

$$x^{(i)}_{m,n,r}(\alpha,\beta,\gamma) = (1-\alpha)\alpha^m(1-\beta)\beta^n(1-\gamma)\gamma^r - \sum_{\substack{\nu\in V_i \\ l(\nu)=1}} x^{(\nu_{l(\nu)})}_{m,n,r}(\alpha_\nu,\beta_\nu,\gamma_\nu). \tag{3.32}$$

We shall use this recurrence relation to prove Theorem 3.1 for all $i\in I$ and $(\alpha,\beta,\gamma)\in P_i'$. Remark that for all these (α,β,γ) the properties for the factors α_ν, β_ν and γ_ν given in Lemma 3.6 still hold; this lemma will be used to derive two preliminary results.

To prove the absolute convergence of a series, i.e. a 'one-fold' tree, it suffices to show that for all $k\ge 0$ the k-th term is in absolute value smaller than dC^k for some constants d and

C with $C < 1$; in that case the sum of all terms is smaller than $d/(1-C)$. The analogue of this concept for a binary tree is proving that for all $k \geq 0$ all terms at distance k from the root/origin are in absolute value smaller than dC^k for some constants d and C with $C < \frac{1}{2}$; in that case the sum of the terms at distance k is smaller than $2^k dC^k = d(2C)^k$ and the sum of all terms is bounded by $d/(1-2C)$. This concept is used to derive a bound for the binary trees $x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)$ with $|\alpha|, |\beta|, |\gamma| < \frac{1}{2}$.

Consider a formal solution $\{x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)\}$ with $i \in I$ and $(\alpha,\beta,\gamma) \in P'_i$. Let the real constant C satisfy $C \geq \max[|\alpha|, |\beta|, |\gamma|]$. Remark that if $i = 1$ then $|\alpha| < |\beta\gamma|$ and we have $\max[|\alpha|, |\beta|, |\gamma|] = \max[|\beta|, |\gamma|]$; and similarly for the cases $i = 2$ and $i = 3$. With the help of (i) and (ii) of Lemma 3.6 (property (ii) implies that $|\alpha_\nu|, |\beta_\nu|, |\gamma_\nu| < C$ for all ν) and by using induction with respect to $l(\nu)$ one can show that

$$|\beta_\nu\gamma_\nu| \leq C^{l(\nu)+2} \text{ if } \nu_{l(\nu)} = 1, \quad |\beta_\nu\gamma_\nu| \leq C^{l(\nu)+3} \text{ if } \nu_{l(\nu)} \in \Lambda\{1\};$$

$$|\alpha_\nu\gamma_\nu| \leq C^{l(\nu)+2} \text{ if } \nu_{l(\nu)} = 2, \quad |\alpha_\nu\gamma_\nu| \leq C^{l(\nu)+3} \text{ if } \nu_{l(\nu)} \in \Lambda\{2\};$$

$$|\alpha_\nu\beta_\nu| \leq C^{l(\nu)+2} \text{ if } \nu_{l(\nu)} = 3, \quad |\alpha_\nu\beta_\nu| \leq C^{l(\nu)+3} \text{ if } \nu_{l(\nu)} \in \Lambda\{3\}.$$

for all $\nu \in V_i$, where $\nu_{l(\nu)} := i$ for $\nu = \varnothing$. This implies that

$$|\beta_\nu\gamma_\nu|, |\alpha_\nu\gamma_\nu|, |\alpha_\nu\beta_\nu| \leq C^{l(\nu)+2} \text{ for all } \nu \in V_i. \tag{3.33}$$

As a consequence, $|\alpha_\nu^m \beta_\nu^n \gamma_\nu^r| \leq C^{l(\nu)+2}$ for all states (m,n,r) with at most one coordinate equal to zero, i.e. for all states $(m,n,r) \in M_c$, and, if $C < \frac{1}{2}$, then we find

$$\begin{aligned} \text{abs}(x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)) &= \sum_{\nu \in V_i} |(-1)^{l(\nu)}(1-\alpha_\nu)\alpha_\nu^m (1-\beta_\nu)\beta_\nu^n (1-\gamma_\nu)\gamma_\nu^r| \\ &\leq 2^3 \sum_{\nu \in V_i} |\alpha_\nu^m \beta_\nu^n \gamma_\nu^r| \leq 8 \sum_{\nu \in V_i} C^{l(\nu)+2} = \frac{8C^2}{1-2C}, \end{aligned}$$

which proves the following lemma.

Lemma 3.7.

Let $i \in I$ and $(\alpha,\beta,\gamma) \in P'_i$. Further, let the constant C satisfy $C \geq \max[|\alpha|, |\beta|, |\gamma|]$ and assume $C < \frac{1}{2}$. Then $x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)$ is absolutely convergent in all states $(m,n,r) \in M_c$ and

$$\text{abs}(x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)) \leq \frac{8C^2}{1-2C}.$$

The upper bound for $\text{abs}(x^{(i)}_{m,n,r}(\alpha,\beta,\gamma))$ given in Lemma 3.7, together with Lemma 3.6(i) and the recursion in (3.32), is used to prove the second preliminary result.

Lemma 3.8.

Let $i \in I$ and $(\alpha,\beta,\gamma) \in P'_i$. Further, assume that $\min[|\beta|, |\gamma|] < \frac{1}{2}$ if $i = 1$, $\min[|\alpha|, |\gamma|] < \frac{1}{2}$ if $i = 2$ and $\min[|\alpha|, |\beta|] < \frac{1}{2}$ if $i = 3$. Then $x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)$ is absolutely convergent in all states $(m,n,r) \in M_c$.

Proof.

It suffices to prove the lemma for a formal solution $x^{(1)}_{m,n,r}(\alpha,\beta,\gamma)$, i.e. for the case $i=1$. W.l.o.g. we may assume $|\beta| \le |\gamma|$; so $|\beta| < 1/2$. Define the path $\{\nu^{(k)}\}$ by $\nu^{(0)}=\varnothing$ and $\nu^{(k)}=(2,1,\ldots,2)$ if $k\ge 1$ and k odd, $\nu^{(k)}=(2,1,\ldots,2,1)$ if $k\ge 1$ and k even. Further let the vectors $w^{(k)}$ for all $k\ge 1$ be defined by $w^{(k)}=(2,1,\ldots,2,1,3)$ if k odd, $w^{(k)}=(2,1,\ldots,2,3)$ if k even ($w^{(k)}$ follows from $\nu^{(k-1)}$ by adding a 3). Then by using (3.32) one can show that

$$\operatorname{abs}(x^{(1)}_{m,n,r}(\alpha,\beta,\gamma)) = \sum_{k=0}^{\infty} |(1-\alpha_{\nu^{(k)}})\alpha^{m}_{\nu^{(k)}}(1-\beta_{\nu^{(k)}})\beta^{n}_{\nu^{(k)}}(1-\gamma_{\nu^{(k)}})\gamma^{r}_{\nu^{(k)}}| + \sum_{k=1}^{\infty} \operatorname{abs}(x^{(3)}_{m,n,r}(\alpha_{w^{(k)}},\beta_{w^{(k)}},\gamma_{w^{(k)}})) . \tag{3.34}$$

By using Lemma 3.6(i) and induction with respect to k it is shown that

$$|\alpha_{\nu^{(k)}}| \le |\beta|\,|\gamma|^{k+1} \quad \text{and} \quad |\beta_{\nu^{(k)}}| \le |\beta|\,|\gamma|^{k} \quad \text{if } k\ge 0 \text{ and } k \text{ even};$$

$$|\alpha_{\nu^{(k)}}| \le |\beta|\,|\gamma|^{k} \quad \text{and} \quad |\beta_{\nu^{(k)}}| \le |\beta|\,|\gamma|^{k+1} \quad \text{if } k\ge 0 \text{ and } k \text{ odd},$$

by which one can easily see that the first series on the right-hand side of (3.34) converges for all $(m,n,r)\in M_c$ (so, $m+n\ge 1$; further, note that $\gamma_{\nu^{(k)}}=\gamma$ for all k):

$$\sum_{k=0}^{\infty} |(1-\alpha_{\nu^{(k)}})\alpha^{m}_{\nu^{(k)}}(1-\beta_{\nu^{(k)}})\beta^{n}_{\nu^{(k)}}(1-\gamma_{\nu^{(k)}})\gamma^{r}_{\nu^{(k)}}|$$

$$= |1-\gamma|\,|\gamma|^{r} \sum_{k=0}^{\infty} |(1-\alpha_{\nu^{(k)}})\alpha^{m}_{\nu^{(k)}}(1-\beta_{\nu^{(k)}})\beta^{n}_{\nu^{(k)}}|$$

$$\le 8 \sum_{k=0}^{\infty} |\alpha^{m}_{\nu^{(k)}}\beta^{n}_{\nu^{(k)}}| \le 8 \sum_{k=0}^{\infty} (|\beta|\,|\gamma|^{k})^{m+n} \le 8 \sum_{k=0}^{\infty} |\beta|\,|\gamma|^{k} = \frac{8|\beta|}{1-|\gamma|} .$$

Since $\alpha_{w^{(k)}}=\alpha_{\nu^{(k-1)}}$ and $\beta_{w^{(k)}}=\beta_{\nu^{(k-1)}}$ for all $k\ge 1$, we have

$$\max[|\alpha_{w^{(k)}}|, |\beta_{w^{(k)}}|, |\gamma_{w^{(k)}}|] = \max[|\alpha_{w^{(k)}}|, |\beta_{w^{(k)}}|] \le |\beta|\,|\gamma|^{k-1}$$

for all $k\ge 1$. Combining this result with Lemma 3.7 shows that also the second series on the right-hand side of (3.34) converges for all $(m,n,r)\in M_c$:

$$\sum_{k=1}^{\infty} \operatorname{abs}(x^{(3)}_{m,n,r}(\alpha_{w^{(k)}},\beta_{w^{(k)}},\gamma_{w^{(k)}})) \le \sum_{k=1}^{\infty} \frac{8|\beta|\,|\gamma|^{k-1}}{1-2|\beta|\,|\gamma|^{k-1}}$$

$$\le \sum_{k=1}^{\infty} \frac{8|\beta|\,|\gamma|^{k-1}}{1-2|\beta|} = \frac{8|\beta|}{(1-2|\beta|)(1-|\gamma|)} .$$

As a result, for all $(m,n,r)\in M_c$ the sum $\operatorname{abs}(x^{(1)}_{m,n,r}(\alpha,\beta,\gamma))$ is finite, i.e. $x^{(1)}_{m,n,r}(\alpha,\beta,\gamma)$ converges absolutely. □

Proof of Theorem 3.1.

Now we are able to prove part (i) of Theorem 3.1. Let $i\in I$ and $(\alpha,\beta,\gamma)\in P'_i$. Define the constant C by $C=\max[|\alpha|,|\beta|,|\gamma|]$. Since $C<1$, there is an integer $k\ge 0$ such that $C^{1/2k+1} < 1/2$. By repeated application of (3.32), we get

$$\mathrm{abs}(x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)) = \sum_{\substack{\nu\in V_i \\ l(\nu)<k}} |(1-\alpha_\nu)\alpha_\nu^m (1-\beta_\nu)\beta_\nu^n (1-\gamma_\nu)\gamma_\nu^r| + \sum_{\substack{\nu\in V_i \\ l(\nu)=k}} \mathrm{abs}(x^{(\nu_{l(\nu)})}_{m,n,r}(\alpha,\beta,\gamma)) . \tag{3.35}$$

By (3.33), for all $\nu \in V_i$ with $l(\nu)=k$, we have $|\beta_\nu\gamma_\nu| \le C^{k+2}$ and therefore

$$\min[|\beta_\nu|, |\gamma_\nu|] \le \sqrt{C^{k+2}} = C^{\frac{1}{2}k+1} < \tfrac{1}{2};$$

and similarly for $\min[|\alpha_\nu|, |\gamma_\nu|]$ and $\min[|\alpha_\nu|, |\beta_\nu|]$. So, by Lemma 3.8, for all $(m,n,r) \in M_c$ all terms of the second sum in (3.35) converge. Since this sum, and also the first sum in (3.35), consists of only a finite number of terms, we may conclude that for all $(m,n,r) \in M_c$ the sum $\mathrm{abs}(x^{(i)}_{m,n,r}(\alpha,\beta,\gamma))$ is finite, i.e. $x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)$ converges absolutely. This completes the proof of Theorem 3.1(i).

Let us now prove the second part of Theorem 3.1. The validity of the first inequality is trivial and the third inequality immediately follows from part (i). The second inequality is proved as follows. From the definition of M_c, it follows that

$$\begin{aligned}\sum_{(m,n,r)\in M_c} &\mathrm{abs}(x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)) \\ &= \sum_{(0,n,r)\in M_{\Lambda(1)}} \mathrm{abs}(x^{(i)}_{0,n,r}(\alpha,\beta,\gamma)) + \sum_{(m,0,r)\in M_{\Lambda(2)}} \mathrm{abs}(x^{(i)}_{m,0,r}(\alpha,\beta,\gamma)) \\ &\quad + \sum_{(m,n,0)\in M_{\Lambda(3)}} \mathrm{abs}(x^{(i)}_{m,n,0}(\alpha,\beta,\gamma)) + \sum_{(m,n,r)\in M_f} \mathrm{abs}(x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)) .\end{aligned} \tag{3.36}$$

By using Lemma 3.6(ii), one easily derives the following bound for the first sum on the right-hand side of (3.36):

$$\begin{aligned}\sum_{(0,n,r)\in M_{\Lambda(1)}} \mathrm{abs}(x^{(i)}_{0,n,r}(\alpha,\beta,\gamma)) &= \sum_{n=1}^{\infty}\sum_{r=1}^{\infty}\sum_{\nu\in V_i} |(1-\alpha_\nu)(1-\beta_\nu)\beta_\nu^n (1-\gamma_\nu)\gamma_\nu^r| \\ = \sum_{\nu\in V_i} |1-\alpha_\nu|\,|1-\beta_\nu|\,\frac{|\beta_\nu|}{1-|\beta_\nu|}\,|1-\gamma_\nu|\,\frac{|\gamma_\nu|}{1-|\gamma_\nu|} &\le \frac{\mathrm{abs}(x^{(i)}_{0,1,1}(\alpha,\beta,\gamma))}{(1-|\beta|)(1-|\gamma|)} .\end{aligned} \tag{3.37}$$

In a similar way one derives bounds for the other sums on the right-hand side of (3.36), after which substitution of all these bounds in (3.36) completes the proof. □

3.6. The equilibrium distribution

The analysis in the Sections 3.3 and 3.4 has resulted in the definition of the formal solutions $\{x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)\}$ and in Condition 3.1, which has been shown to be needed for the absolute convergence of these formal solutions. In Section 3.5, it has been verified that, for the states in the convergence region M_c, this condition is also sufficient for the absolute convergence.

In this section, under the assumption that Condition 3.1 is satisfied, it will be investigated whether we can obtain the equilibrium distribution $\{p_{m,n,r}\}$ by using the formal solutions $\{x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)\}$. Since these formal solutions converge absolutely (i.e. are well-defined)

in all states of the convergence region M_c, which consists of the interior and the three boundary planes, each formal solution $\{x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)\}$ satisfies the equilibrium equations for the states in M_c which have no incoming transitions from states outside of M_c (see also the reasoning in the last paragraph of Section 3.3). As one can easily verify, the only states in M_c which have incoming transitions from states outside M_c, i.e. from states on the axes or from the origin, are the states $(m,0,1)$, $(m,1,0)$, $(0,n,1)$, $(1,n,0)$, $(0,1,r)$ and $(1,0,r)$. Let M_c' be the set of all states for which the equilibrium equations are satisfied by a formal solution. Then

$$M_c' = \{ (m_1,m_2,m_3) \in M \mid m_i + m_j \geq 2 \text{ for all } i,j \in I,\ i \neq j \}.$$

Since each formal solution satisfies the equilibrium equations for all states in M_c', also each linear combination of formal solutions satisfies the equations for this set M_c'. This gives us some freedom in finding a solution which also satisfies the equilibrium equations for the states outside of M_c', i.e. in finding the equilibrium distribution $\{p_{m,n,r}\}$. Now the question is which formal solutions should be linearly combined, or better, which starting solutions have to be selected.

The analysis in this section is built up as follows. First, due to Condition 3.1, we can present explicit expressions for the two-dimensional marginal distributions $\{p^{(i,j)}_{m_1,m_2}\}$. Next, we are able to derive a nice characterization for the starting solutions, from which we learn that each set P_i of starting solutions is uncountable. After that, we shall derive the explicit formula for the equilibrium distribution $\{p_{m,n,r}\}$, as stated in the *Main Theorem* at the end of this section. It will be shown that from the uncountable sets of candidates for starting solutions, only a countable number of starting solutions is needed to obtain a linear combination of formal solutions which also satisfies the equilibrium equations for the countable set $M \backslash M_c'$. For the selection of the appropriate candidates, and also for the choice of the coefficients of the linear combination, we shall use the explicit expressions for the two-dimensional marginal distributions $\{p^{(i,j)}_{m_1,m_2}\}$. In fact, it is at this point that induction has to be used to extend the expressions found for the equilibrium distribution $\{p_{m,n,r}\}$ to the N-dimensional case. Note that the problem of selecting the appropriate starting solutions did not appear in the two-dimensional case, where we obtained only a finite number of starting solutions, which all had to be used for the construction of the equilibrium distribution.

Explicit formulae for the two-dimensional marginal distributions $\{p^{(i,j)}_{m_1,m_2}\}$

In Section 3.2, we established that for all $i,j \in I$, $i < j$, the marginal distribution $\{p^{(i,j)}_{m_1,m_2}\}$ is the equilibrium distribution of the two-dimensional, irreducible, positive recurrent, homogeneous, nearest-neighboring random walk with the projection property and transition probabilities $q^{(i,j)}_{l_1,l_2}$ for the states in the interior. Since Condition 3.1 is assumed to be satisfied, we also satisfy the equivalent condition stated in (3.9), which implies that the two-dimensional marginal distributions can be determined by applying the compensation approach, as described in Chapter 2 for the two-dimensional case.

For the equilibrium distribution $\{p^{(2,3)}_{n,r}\}$ of the random walk with state space $\{ (n,r) \mid n,r \in \mathbb{N}_0 \}$ and transition probabilities $q^{(2,3)}_{l_1,l_2}$ for the states in the interior (see Figure 3.3), the use of the compensation approach leads to (see Theorem 2.1; the product factors are renumbered to let the expressions be better linked up with the expressions for the three-dimensional case):

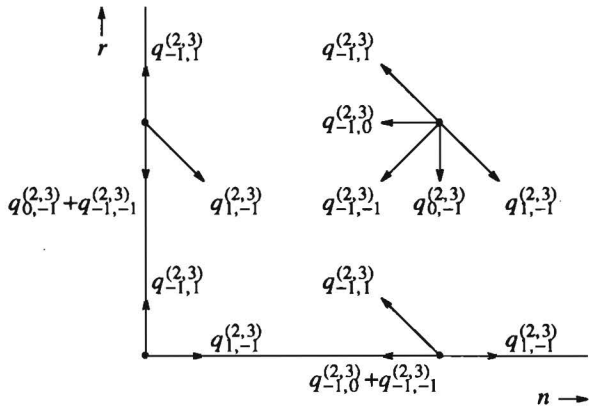


Figure 3.3. The transition probabilities for the random walk which describes the behavior for the components n and r; for all states the transitions to themselves have been left out.

$$p_{n,r}^{(2,3)} = \sum_{k=0}^{\infty} (-1)^k (1-\beta_k^{(1)})(\beta_k^{(1)})^n (1-\gamma_k^{(1)})(\gamma_k^{(1)})^r + \sum_{k=0}^{\infty} (-1)^k (1-\hat{\beta}_k^{(1)})(\hat{\beta}_k^{(1)})^n (1-\hat{\gamma}_k^{(1)})(\hat{\gamma}_k^{(1)})^r , \quad n,r \geq 0,\ n+r \geq 1, \tag{3.38}$$

$$p_{0,0}^{(2,3)} = 1 - \sum_{\substack{n,r \geq 0 \\ n+r \geq 1}} p_{n,r}^{(2,3)} . \tag{3.39}$$

Here, the first series in (3.38) is a formal solution starting on the horizontal boundary $r=0$ and it satisfies the following properties:

* All $(\beta_k^{(1)}, \gamma_k^{(1)})$ are real-valued solutions of the quadratic equation

$$\beta\gamma = \sum_{t_1,t_2 \in \{-1,0,1\}} q_{t_1,t_2}^{(2,3)} \beta^{1-t_1} \gamma^{1-t_2} , \tag{3.40}$$

  which is obtained after substituting the product form $\beta^n \gamma^r$ in the equilibrium equation for the interior. This quadratic equation is equivalent to the quadratic equation (3.14) for fixed $\alpha = 1$, which implies that the product of the roots β of (3.40) for fixed γ is equal to $f_2(1,\gamma)$ and the product of the roots γ of (3.40) for fixed β is equal to $f_3(1,\beta)$;

* $(\beta_0^{(1)}, \gamma_0^{(1)})$ is the unique solution of the equilibrium equations for the interior and the horizontal boundary $r=0$: $\beta_0^{(1)}$ is equal to the geometric factor of $\{p_n^{(2)}\}$, i.e. $\beta_0^{(1)} = q_1^{(2)}/q_{-1}^{(2)} = f_2(1,1)$, and $\gamma_0^{(1)}$ is the companion solution to 1 of the quadratic equation (3.40) for fixed $\beta = \beta_0^{(1)}$, i.e. $\gamma_0^{(1)} = f_3(1, \beta_0^{(1)})$;

* For all even k, the factors $\beta_{k+1}^{(1)}$ and $\gamma_{k+1}^{(1)}$ are chosen such that the sum of the terms with indices k and $k+1$ satisfies the equilibrium equations for the interior and the vertical boundary $n=0$: $\gamma_{k+1}^{(1)} = \gamma_k^{(1)}$ and $\beta_{k+1}^{(1)}$ is the companion solution to $\beta_k^{(1)}$ of (3.40) for fixed $\gamma = \gamma_k^{(1)}$, i.e. $\beta_{k+1}^{(1)} = f_2(1, \gamma_k^{(1)})/\beta_k^{(1)}$;

* For all odd k, the factors $\beta_{k+1}^{(1)}$ and $\gamma_{k+1}^{(1)}$ are chosen such that the sum of the terms with indices k and $k+1$ satisfies the equilibrium equations for the interior and the horizontal boundary $r=0$: $\beta_{k+1}^{(1)}=\beta_k^{(1)}$ and $\gamma_{k+1}^{(1)}$ is the companion solution to $\gamma_k^{(1)}$ of (3.40) for fixed $\beta=\beta_k^{(1)}$, i.e. $\gamma_{k+1}^{(1)}=f_3(1,\beta_k^{(1)})/\gamma_k^{(1)}$.

For the factors $\beta_k^{(1)}$ and $\gamma_k^{(1)}$, we have the property (see Lemma 2.3(i))

$$1>\beta_0^{(1)}>\gamma_0^{(1)}=\gamma_1^{(1)}>\beta_1^{(1)}=\beta_2^{(1)}>\gamma_2^{(1)}=\cdots$$

and the ratios $\beta_{k+1}^{(1)}/\gamma_k^{(1)}$ for even k and $\gamma_{k+1}^{(1)}/\beta_k^{(1)}$ for odd k decrease monotonously to

$$A_1=\frac{(1-q_{0,0}^{(2,3)})-\sqrt{(1-q_{0,0}^{(2,3)})^2-4q_{1,-1}^{(2,3)}q_{-1,1}^{(2,3)}}}{2q_{-1,1}^{(2,3)}}$$

and

$$A_2=\frac{(1-q_{0,0}^{(2,3)})+\sqrt{(1-q_{0,0}^{(2,3)})^2-4q_{1,-1}^{(2,3)}q_{-1,1}^{(2,3)}}}{2q_{-1,1}^{(2,3)}},$$

respectively (see Lemma 2.3(ii) and the remarks at the beginning of Section 2.5). As a result,

$$\frac{\beta_{k+2}^{(1)}}{\beta_k^{(1)}}\downarrow\frac{A_1}{A_2}\quad\text{and}\quad\frac{\gamma_{k+2}^{(1)}}{\gamma_k^{(1)}}\downarrow\frac{A_1}{A_2}\quad\text{as } k\to\infty. \tag{3.41}$$

Since $A_1/A_2<1$, the factors $\beta_k^{(1)}$ and $\gamma_k^{(1)}$ decrease exponentially fast to 0. Similar results hold for the second series in (3.38), which is a formal solution starting on the vertical boundary $n=0$. For this solution the factors $\hat\beta_k^{(1)}$ and $\hat\gamma_k^{(1)}$ are defined by $\hat\gamma_0^{(1)}=q_1^{(3)}/q_{-1}^{(3)}=f_3(1,1)$, $\hat\beta_0^{(1)}=f_2(1,\hat\gamma_0^{(1)})$ and for all $k\geq 0$:

$$\hat\beta_{k+1}^{(1)}=\hat\beta_k^{(1)},\ \hat\gamma_{k+1}^{(1)}=f_3(1,\hat\beta_k^{(1)})/\hat\gamma_k^{(1)}\quad\text{if } k \text{ is even;}$$
$$\hat\gamma_{k+1}^{(1)}=\hat\gamma_k^{(1)},\ \hat\beta_{k+1}^{(1)}=f_2(1,\hat\gamma_k^{(1)})/\hat\beta_k^{(1)}\quad\text{if } k \text{ is odd.}$$

For the other two two-dimensional marginal distributions $\{p_{m,r}^{(1,3)}\}$ and $\{p_{m,n}^{(1,2)}\}$, we obtain the following expressions. For $\{p_{m,r}^{(1,3)}\}$, we find

$$p_{m,r}^{(1,3)}=\sum_{k=0}^{\infty}(-1)^k(1-\alpha_k^{(2)})(\alpha_k^{(2)})^m(1-\gamma_k^{(2)})(\gamma_k^{(2)})^r$$
$$+\sum_{k=0}^{\infty}(-1)^k(1-\hat\alpha_k^{(2)})(\hat\alpha_k^{(2)})^m(1-\hat\gamma_k^{(2)})(\hat\gamma_k^{(2)})^r,\quad m,r\geq 0,\ m+r\geq 1,$$

where the first series represents the formal solution starting on the boundary $r=0$ and the second series is the formal solution starting on the boundary $m=0$. The factors $\alpha_k^{(2)}$, $\gamma_k^{(2)}$, $\hat\alpha_k^{(2)}$ and $\hat\gamma_k^{(2)}$ are defined by $\alpha_0^{(2)}=f_1(1,1)$, $\gamma_0^{(2)}=f_3(\alpha_0^{(2)},1)$, $\hat\gamma_0^{(2)}=f_3(1,1)$, $\hat\alpha_0^{(2)}=f_1(1,\hat\gamma_0^{(2)})$ and for all $k\geq 0$:

$$\gamma_{k+1}^{(2)}=\gamma_k^{(2)},\ \alpha_{k+1}^{(2)}=f_1(1,\gamma_k^{(2)})/\alpha_k^{(2)},\ \hat\alpha_{k+1}^{(2)}=\hat\alpha_k^{(2)},\ \hat\gamma_{k+1}^{(2)}=f_3(\hat\alpha_k^{(2)},1)/\hat\gamma_k^{(2)}\quad\text{if } k \text{ even;}$$
$$\alpha_{k+1}^{(2)}=\alpha_k^{(2)},\ \gamma_{k+1}^{(2)}=f_3(\alpha_k^{(2)},1)/\gamma_k^{(2)},\ \hat\gamma_{k+1}^{(2)}=\hat\gamma_k^{(2)},\ \hat\alpha_{k+1}^{(2)}=f_1(1,\hat\gamma_k^{(2)})/\hat\alpha_k^{(2)}\quad\text{if } k \text{ odd.}$$

Finally, for $\{p_{m,n}^{(1,2)}\}$, we get

$$p_{m,n}^{(1,2)}=\sum_{k=0}^{\infty}(-1)^k(1-\alpha_k^{(3)})(\alpha_k^{(3)})^m(1-\beta_k^{(3)})(\beta_k^{(3)})^n$$
$$+\sum_{k=0}^{\infty}(-1)^k(1-\hat\alpha_k^{(3)})(\hat\alpha_k^{(3)})^m(1-\hat\beta_k^{(3)})(\hat\beta_k^{(3)})^n,\quad m,n\geq 0,\ m+n\geq 1,$$

with $\alpha_0^{(3)} = f_1(1,1)$, $\beta_0^{(3)} = f_2(\alpha_0^{(3)},1)$, $\hat{\beta}_0^{(3)} = f_2(1,1)$, $\hat{\alpha}_0^{(3)} = f_1(\hat{\beta}_0^{(3)},1)$ and for all $k \geq 0$:

$$\beta_{k+1}^{(3)} = \beta_k^{(3)},\ \alpha_{k+1}^{(3)} = f_1(\beta_k^{(3)},1)/\alpha_k^{(3)},\ \hat{\alpha}_{k+1}^{(3)} = \hat{\alpha}_k^{(3)},\ \hat{\beta}_{k+1}^{(3)} = f_2(\hat{\alpha}_k^{(3)},1)/\hat{\beta}_k^{(3)} \quad \text{if } k \text{ even;}$$

$$\alpha_{k+1}^{(3)} = \alpha_k^{(3)},\ \beta_{k+1}^{(3)} = f_2(\alpha_k^{(3)},1)/\beta_k^{(3)},\ \hat{\beta}_{k+1}^{(3)} = \hat{\beta}_k^{(3)},\ \hat{\alpha}_{k+1}^{(3)} = f_1(\hat{\beta}_k^{(3)},1)/\hat{\alpha}_k^{(3)} \quad \text{if } k \text{ odd.}$$

Just like $p_{0,0}^{(2,3)}$, the remaining probabilities $p_{0,0}^{(1,3)}$ and $p_{0,0}^{(1,2)}$ follow from the normalization equation; cf. (3.39).

Characterization of the starting solutions $(\alpha,\beta,\gamma) \in P_i$

Due to the projection property, there exists a simple characterization for the starting solutions $(\alpha,\beta,\gamma) \in P_i$. Consider a solution $(\alpha,\beta,\gamma) \in P_1$, i.e. a starting solution on the boundary $m=0$. Such a solution has to satisfy the equilibrium equations (3.10) and (3.11), i.e. (α,β,γ) has to satisfy the quadratic equation (3.14) and the equation

$$\beta\gamma = \sum_{(-1,t_2,t_3)\in T} q_{-1,t_2,t_3}\, \alpha\beta^{1-t_2}\gamma^{1-t_3} + \sum_{(0,t_2,t_3)\in T} (q_{0,t_2,t_3} + q_{-1,t_2,t_3})\, \beta^{1-t_2}\gamma^{1-t_3}, \tag{3.42}$$

which is obtained by substituting the product form $\alpha^m\beta^n\gamma^r$ in (3.11) (see also the definition of $K(\alpha,\beta,\gamma)$ at the beginning of Section 3.3). Multiplying both sides of (3.42) by α and subtracting (3.42) from both sides of (3.14) leads to

$$0 = \sum_{(1,t_2,t_3)\in T} q_{1,t_2,t_3}\, \beta^{1-t_2}\gamma^{1-t_3} - \alpha \sum_{(-1,t_2,t_3)\in T} q_{-1,t_2,t_3}\, \beta^{1-t_2}\gamma^{1-t_3},$$

which shows that α has to be equal to $f_1(\beta,\gamma)$ (cf. (3.17)). Rewriting (3.14) to a quadratic equation in α (see (3.16)), dividing all terms by α and next substituting $\alpha = f_1(\beta,\gamma)$ shows that β and γ have to satisfy the equation

$$\beta\gamma = \sum_{(t_1,t_2,t_3)\in T} q_{t_1,t_2,t_3}\, \beta^{1-t_2}\gamma^{1-t_3}, \tag{3.43}$$

which is equivalent to (3.14) for fixed $\alpha = 1$. Finally, we have to evaluate the condition $0 < |\alpha| < |\beta\gamma|$. Let (β,γ) be a solution of (3.43) with $0 < |\beta| < 1$ and $0 < |\gamma| < 1$, then for these fixed β and γ the quadratic equation (3.14) has two solutions: $\alpha = 1$ and $\alpha = f_1(\beta,\gamma)$. Since $1 > |\beta\gamma|$, according to Lemma 3.3(i), the second root $\alpha = f_1(\beta,\gamma)$ satisfies $0 < |\alpha| < |\beta\gamma|$. This proves part (i) of the following lemma; the other two parts may be proved along the same lines.

Lemma 3.9

(i) (α,β,γ) *is a starting solution on the boundary plane* $m=0$, *i.e.* $(\alpha,\beta,\gamma) \in P_1$, *if and only if* (β,γ) *is a solution of the quadratic equation (3.14) for fixed* $\alpha = 1$ *and* α *is equal to* $\alpha = f_1(\beta,\gamma)$;

(ii) (α,β,γ) *is a starting solution on the boundary plane* $n=0$, *i.e.* $(\alpha,\beta,\gamma) \in P_2$, *if and only if* (α,γ) *is a solution of the quadratic equation (3.14) for fixed* $\beta = 1$ *and* β *is equal to* $\beta = f_2(\alpha,\gamma)$;

(iii) (α,β,γ) *is a starting solution on the boundary plane* $r=0$, *i.e.* $(\alpha,\beta,\gamma) \in P_3$, *if and only if* (α,β) *is a solution of the quadratic equation (3.14) for fixed* $\gamma = 1$ *and* γ *is equal to* $\gamma = f_3(\alpha,\beta)$.

Determination of the equilibrium distribution $\{p_{m,n,r}\}$

We now are able to derive the explicit formula for the equilibrium distribution $\{p_{m,n,r}\}$, as stated in the Main Theorem at the end of this section. It appears to be possible to obtain this distribution by choosing appropriate starting solutions $(\alpha,\beta,\gamma) \in P_i$ and linearly combining the corresponding formal solutions $x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)$. The idea for the appropriate selection of starting solutions is obtained from the explicit formulae for the two-dimensional marginal distributions and the characterization of the starting solutions, as described in Lemma 3.9.

When reading part (i) of Lemma 3.9, we realize the following. Since the quadratic equation (3.14) for fixed $\alpha = 1$ is equivalent to the quadratic equation for the two-dimensional random walk describing the behavior for the components n and r (see (3.40) and Figure 3.3), all product forms present in formula (3.38) for $\{p^{(2,3)}_{n,r}\}$ may be extended to starting solutions on the boundary plane $m = 0$. From this observation, we obtain the idea that the product forms of the marginal distributions $\{p^{(i,j)}_{m_1,m_2}\}$ have to be used in order to obtain the equilibrium distribution. This idea is strengthened by the remarkable property of the formal solutions $\{x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)\}$ described in the next paragraph.

Formula (3.30) for $\{x^{(i)}_{m,n,r}(\alpha,\beta,\gamma)\}$ shows that each formal solution is a kind of alternating sum of product-form distributions. As a consequence, for each formal solution two terms with the same values for two of the three factors α_ν, β_ν and γ_ν vanish when taking the summation over the coordinate belonging to the third factor. For example, for a formal solution $\{x^{(1)}_{m,n,r}(\alpha,\beta,\gamma)\}$ two terms with the same β- and γ-factor vanish if we take the summation of $x^{(1)}_{m,n,r}(\alpha,\beta,\gamma)$ over $m = 0$ to ∞, by which

$$\begin{aligned}\sum_{m=0}^{\infty} x^{(1)}_{m,n,r}(\alpha,\beta,\gamma) &= \sum_{m=0}^{\infty}\Bigg[(1-\alpha_\varnothing)\alpha_\varnothing^m\,(1-\beta_\varnothing)\beta_\varnothing^n\,(1-\gamma_\varnothing)\gamma_\varnothing^r \\ &\quad + \sum_{\substack{\nu\in V_1\setminus\{\varnothing\}\\ \nu_{l(\nu)}=1}} (-1)^{l(p(\nu))}\left[(1-\alpha_{p(\nu)})\alpha_{p(\nu)}^m-(1-\alpha_\nu)\alpha_\nu^m\right](1-\beta_{p(\nu)})\beta_{p(\nu)}^n\,(1-\gamma_{p(\nu)})\gamma_{p(\nu)}^r\Bigg] \\ &= (1-\beta)\beta^n\,(1-\gamma)\gamma^r \quad \text{for all } n,r\geq 1.\end{aligned}$$

Here, the last equality is found after changing summations, which is allowed by Theorem 3.1(ii). The first term of $x^{(1)}_{m,n,r}(\alpha,\beta,\gamma)$ does not vanish when summing over m, since this term does not have a companion term with the same β- and γ-factor. When summing over n, all terms have a companion term with the same α- and γ-factor, by which

$$\begin{aligned}&\sum_{n=0}^{\infty} x^{(1)}_{m,n,r}(\alpha,\beta,\gamma) \\ &\quad = \sum_{n=0}^{\infty}\sum_{\substack{\nu\in V_1\setminus\{\varnothing\}\\ \nu_{l(\nu)}=2}} (-1)^{l(p(\nu))}\left[(1-\beta_{p(\nu)})\beta_{p(\nu)}^n-(1-\beta_\nu)\beta_\nu^n\right](1-\alpha_{p(\nu)})\alpha_{p(\nu)}^n\,(1-\gamma_{p(\nu)})\gamma_{p(\nu)}^r \\ &\quad = 0 \quad \text{for all } m,r\geq 1;\end{aligned}$$

and similarly when summing over r. This proves Lemma 3.10 for $i = 1$; the cases with $i = 2$ and $i = 3$ are treated in a similar way.

Lemma 3.10

Let $i,j \in I$ and $(\alpha_1,\alpha_2,\alpha_3) \in P_i$. Then

$$\sum_{m_j=0}^{\infty} x^{(i)}_{m_1,m_2,m_3}(\alpha_1,\alpha_2,\alpha_3) = \begin{cases} \prod_{l \in I\setminus\{i\}} (1-\alpha_l)\alpha_l^{m_l} & \text{if } j=i; \\ 0 & \text{if } j \neq i, \end{cases}$$

for all $m_l \geq 1$, $l \in I\setminus\{j\}$.

Together with the expressions for the two-dimensional marginal distributions $\{p^{(i,j)}_{m_1,m_2}\}$, the results stated in the Lemmas 3.9 and 3.10 suggest which starting solutions have to be selected and how the coefficients of the linear combination of the corresponding formal solutions should be chosen to obtain the equilibrium distribution $\{p_{m,n,r}\}$.

Combining the results of Lemma 3.9(i) and Lemma 3.10 gives us the idea to define a linear combination of all formal solutions with starting solutions coming from the product forms in formula (3.38) for the marginal distribution $\{p^{(2,3)}_{n,r}\}$. Define $\alpha_k^{(1)} = f_1(\beta_k^{(1)},\gamma_k^{(1)})$ and $\hat{\alpha}_k^{(1)} = f_1(\hat{\beta}_k^{(1)},\hat{\gamma}_k^{(1)})$ for all $k \geq 0$, then, by Lemma 3.9(i), all solutions $(\alpha_k^{(1)},\beta_k^{(1)},\gamma_k^{(1)})$ and $(\hat{\alpha}_k^{(1)},\hat{\beta}_k^{(1)},\hat{\gamma}_k^{(1)})$ are starting solutions on the boundary plane $m=0$. Next, defining

$$y^{(1)}_{m,n,r} = \sum_{k=0}^{\infty} (-1)^k x^{(1)}_{m,n,r}(\alpha_k^{(1)},\beta_k^{(1)},\gamma_k^{(1)}) + \sum_{k=0}^{\infty} (-1)^k x^{(1)}_{m,n,r}(\hat{\alpha}_k^{(1)},\hat{\beta}_k^{(1)},\hat{\gamma}_k^{(1)}), \quad (m,n,r) \in M_c,$$

gives us a solution $\{y^{(1)}_{m,n,r}\}$, for which, by Lemma 3.10,

$$\begin{aligned} \sum_{m=0}^{\infty} y^{(1)}_{m,n,r} &= \sum_{k=0}^{\infty} (-1)^k \sum_{m=0}^{\infty} x^{(1)}_{m,n,r}(\alpha_k^{(1)},\beta_k^{(1)},\gamma_k^{(1)}) + \sum_{k=0}^{\infty} (-1)^k \sum_{m=0}^{\infty} x^{(1)}_{m,n,r}(\hat{\alpha}_k^{(1)},\hat{\beta}_k^{(1)},\hat{\gamma}_k^{(1)}) \\ &= p^{(2,3)}_{n,r} \quad \text{for all } n,r \geq 1. \end{aligned}$$

So, when summing $\{y^{(1)}_{m,n,r}\}$ over the m-component, one gets the marginal distribution for the other two components. This indicates that we are on the right track with our search for $\{p_{m,n,r}\}$, since this is a property which is satisfied by the equilibrium distribution $\{p_{m,n,r}\}$ by definition (see the definition (3.7) of the marginal distributions $\{p^{(i,j)}_{m_1,m_2}\}$). Summing $\{y^{(1)}_{m,n,r}\}$ over n or r leads to:

$$\sum_{n=0}^{\infty} y^{(1)}_{m,n,r} = 0 \quad \text{for all } m,r \geq 1, \qquad \sum_{r=0}^{\infty} y^{(1)}_{m,n,r} = 0 \quad \text{for all } m,n \geq 1.$$

As we see, in this case the result is 0 instead of a marginal probability; the marginal probabilities $p^{(1,3)}_{m,r}$ and $p^{(1,2)}_{m,n}$ will have to be obtained from linear combinations of formal solutions $\{x^{(2)}_{m,n,r}(\alpha,\beta,\gamma)\}$ and $\{x^{(3)}_{m,n,r}(\alpha,\beta,\gamma)\}$.

The definition and properties for $\{y^{(1)}_{m,n,r}\}$ are easily extended to solutions $\{y^{(i)}_{m,n,r}\}$, $i \in I$. Let

$$\begin{aligned} &\alpha_k^{(1)} = f_1(\beta_k^{(1)},\gamma_k^{(1)}), \quad \hat{\alpha}_k^{(1)} = f_1(\hat{\beta}_k^{(1)},\hat{\gamma}_k^{(1)}); \\ &\beta_k^{(2)} = f_2(\alpha_k^{(2)},\gamma_k^{(2)}), \quad \hat{\beta}_k^{(2)} = f_2(\hat{\alpha}_k^{(2)},\hat{\gamma}_k^{(2)}); \\ &\gamma_k^{(3)} = f_3(\alpha_k^{(3)},\beta_k^{(3)}), \quad \hat{\gamma}_k^{(3)} = f_3(\hat{\alpha}_k^{(3)},\hat{\beta}_k^{(3)}), \end{aligned}$$

for all $k \geq 0$ and let the solutions $\{y^{(i)}_{m,n,r}\}$, $i \in I$, be defined by

$$y^{(i)}_{m,n,r} = \sum_{k=0}^{\infty} (-1)^k x^{(i)}_{m,n,r}(\alpha^{(i)}_k, \beta^{(i)}_k, \gamma^{(i)}_k) + \sum_{k=0}^{\infty} (-1)^k x^{(i)}_{m,n,r}(\hat{\alpha}^{(i)}_k, \hat{\beta}^{(i)}_k, \hat{\gamma}^{(i)}_k), \quad (m,n,r) \in M_c.$$

Then for all $i,j \in I$, $k,l \in I \setminus \{i\}$, $k < l$, and all $m_k, m_l \geq 1$, it holds that

$$\sum_{m_j=0}^{\infty} y^{(i)}_{m_1,m_2,m_3} = \begin{cases} p^{(k,l)}_{m_k,m_l} & \text{if } j = i; \\ 0 & \text{if } j \neq i. \end{cases} \tag{3.44}$$

Obviously, the solution $\{y_{m,n,r}\}$ defined as the sum of the solutions $\{y^{(i)}_{m,n,r}\}$, i.e.

$$y_{m,n,r} = \sum_{i \in I} y^{(i)}_{m,n,r}, \quad (m,n,r) \in M_c,$$

satisfies the desired property: for all $i \in I$, $k,l \in I \setminus \{i\}$, $k < l$, it holds that

$$\sum_{m_i=0}^{\infty} y_{m_1,m_2,m_3} = p^{(k,l)}_{m_k,m_l} \quad \text{for all } m_k, m_l \geq 1. \tag{3.45}$$

Before continuing, we remark that the solution $\{y_{m,n,r}\}$, so far being defined for all states (m,n,r) in the convergence region M_c, is well-defined, since all six series constituting $\{y_{m,n,r}\}$ are absolutely convergent for all states in M_c:

$$\sum_{k=0}^{\infty} |x^{(i)}_{m,n,r}(\alpha^{(i)}_k, \beta^{(i)}_k, \gamma^{(i)}_k)| < \infty \quad \text{and}$$

$$\sum_{k=0}^{\infty} |x^{(i)}_{m,n,r}(\hat{\alpha}^{(i)}_k, \hat{\beta}^{(i)}_k, \hat{\gamma}^{(i)}_k)| < \infty, \quad (m,n,r) \in M_c, \ i \in I. \tag{3.46}$$

For $\Sigma_{k=0}^{\infty} x^{(1)}_{m,n,r}(\alpha^{(1)}_k, \beta^{(1)}_k, \gamma^{(1)}_k)$, the absolute convergence is proved by using the bound given in Lemma 3.7 and the property that the factors $\beta^{(1)}_k$ and $\gamma^{(1)}_k$ decrease monotonously and exponentially fast (see (3.41)); and similarly for the other series. Further, we have to remark that the properties stated in (3.44) and (3.45) have been derived after changing summations; this was allowed, since

$$\sum_{(m,n,r) \in M_c} \sum_{k=0}^{\infty} |x^{(i)}_{m,n,r}(\alpha^{(i)}_k, \beta^{(i)}_k, \gamma^{(i)}_k)| < \infty \quad \text{and}$$

$$\sum_{(m,n,r) \in M_c} \sum_{k=0}^{\infty} |x^{(i)}_{m,n,r}(\hat{\alpha}^{(i)}_k, \hat{\beta}^{(i)}_k, \hat{\gamma}^{(i)}_k)| < \infty, \quad i \in I, \tag{3.47}$$

which is proved by using (3.46), Theorem 3.1(ii) and the property that all factors of the starting solutions $(\alpha^{(i)}_k, \beta^{(i)}_k, \gamma^{(i)}_k)$ and $(\hat{\alpha}^{(i)}_k, \hat{\beta}^{(i)}_k, \hat{\gamma}^{(i)}_k)$ decrease monotonously and exponentially fast.

The solution $\{y_{m,n,r}\}$ defined for all states $(m,n,r) \in M_c$ up to now, satisfies two properties. In the first place, since $\{y_{m,n,r}\}$ is a linear combination of formal solutions, $\{y_{m,n,r}\}$ satisfies the equilibrium equations for all states $(m,n,r) \in M_c'$. Secondly, $\{y_{m,n,r}\}$ satisfies (3.45). Now, define $y_{m,n,r}$ on the m-axis by

$$y_{m,0,0} = p^{(1)}_m - \sum_{\substack{n,r \geq 0 \\ n+r \geq 1}} y_{m,n,r} \quad \text{for all } m \geq 1,$$

and similarly for the n-axis and r-axis. Finally, to obtain a solution for which the probabilities add up to 1, we define $y_{0,0,0}$ by

$$y_{0,0,0} = 1 - \sum_{(m,n,r) \in M \setminus \{(0,0,0)\}} y_{m,n,r}$$

(use (3.47) to show the correctness of these definitions, i.e. to show that the series at the right-hand sides are absolutely convergent). Then $\{y_{m,n,r}\}$ may be shown to satisfy (3.45) also for $m_k=0$ and/or $m_l=0$ (see Lemma 3.11), after which we are able to show that $\{y_{m,n,r}\}$ also satisfies the equilibrium equations for the states outside M_c'. This will lead to the conclusion that the solution $\{y_{m,n,r}\}$, for which the components $y_{m,n,r}$ already add up to 1, is equal to the equilibrium distribution $\{p_{m,n,r}\}$.

Lemma 3.11

Let $i \in I$ and $k,l \in I\backslash\{i\}$, $k<l$. Then

$$\sum_{m_i=0}^{\infty} y_{m_1,m_2,m_3} = p_{m_k,m_l}^{(k,l)} \quad \textit{for all } m_k,m_l \geq 0.$$

Proof.

The result stated in (3.45) is extended in two steps. In the first step (3.45) is extended to

$$\sum_{m_i=0}^{\infty} y_{m_1,m_2,m_3} = p_{m_k,m_l}^{(k,l)} \quad \text{for all } m_k,m_l \geq 0,\ m_k+m_l \geq 1, \tag{3.48}$$

where $i \in I$ and $k,l \in I\backslash\{i\}$, $k<l$. This extension is proved by rewriting the expressions for $y_{m,n,r}$ on the axes. For example, for the case $i=1$, so $k=2$ and $l=3$, we may rewrite $y_{0,n,0}$ for all $n \geq 1$ as (use (3.45))

$$y_{0,n,0} = p_n^{(2)} - \sum_{\substack{m,r\geq 0\\ m+r\geq 1}} y_{m,n,r} = p_n^{(2)} - \sum_{m=1}^{\infty} y_{m,n,0} - \sum_{r=1}^{\infty}\sum_{m=0}^{\infty} y_{m,n,r}$$

$$= p_n^{(2)} - \sum_{m=1}^{\infty} y_{m,n,0} - \sum_{r=1}^{\infty} p_{n,r}^{(2,3)} = p_{n,0}^{(2,3)} - \sum_{m=1}^{\infty} y_{m,n,0},$$

which proves that $\Sigma_{m=0}^{\infty} y_{m,n,0} = p_{n,0}^{(2,3)}$ for all $n \geq 1$; rewriting $y_{0,0,r}$ for all $r \geq 1$ proves the extension of (3.45) for the case $n=0$ and $r \geq 1$. In the second step (3.48) is extended to the result stated in Lemma 3.11; this extension is proved by rewriting $y_{0,0,0}$. □

To show that $\{y_{m,n,r}\}$ also satisfies the equilibrium equations for the states outside M_c', we shall use the *balance principle*:

$$\text{the stream out of a set } M' \;=\; \text{the stream into this set } M', \quad M' \subset M. \tag{3.49}$$

Obviously, for a subset M' consisting of a single state the balance principle is equivalent to the equilibrium equation for that state. Therefore $\{y_{m,n,r}\}$ satisfies (3.49) for all states of M_c'. Further, by Lemma 3.11, $\{y_{m,n,r}\}$ satisfies (3.49) for all subsets of the form

$$M' = \{(n_1,n_2,n_3) \in M \mid n_i \geq 0 \text{ and } n_j = m_j \text{ for all } j \in I\backslash\{i\}\}, \tag{3.50}$$

where $i \in I$ and $m_j \geq 0$, $j \in I\backslash\{i\}$, since for such a subset the balance principle is equivalent to the equilibrium equation in the state (m_k,m_l) of the two-dimensional marginal random walk describing the behavior for the components m_k and m_l, $k,l \in I\backslash\{i\}$, $k<l$. For example, for the subset $M' = \{(m,n,r) \mid m \geq 0\}$ with fixed $n,r \geq 1$ the balance principle is equivalent to (take the sum of (3.10) over $m \geq 1$ and add (3.11), after having replaced $q_{0,i_2,i_3}^{\{1\}}$ by $q_{0,i_2,i_3}+q_{-1,i_2,i_3}$)

$$p_{n,r}^{(2,3)} = \sum_{t_1,t_2 \in \{-1,0,1\}} q_{t_2,t_3}^{(2,3)} \, p_{n-t_2,r-t_3}^{(2,3)} \, ,$$

which is the equilibrium equation for the state (n,r) of the random walk describing the behavior for the last two components (see Figure 3.3). Now, by considering differences $M_1 \backslash M_2$ with $M_2 \subset M_1$ of the sets given in (3.50) and sets consisting of states of M_c', $\{y_{m,n,r}\}$ may be shown to satisfy also the equilibrium equations for the states outside M_c'. For example, for all $m \geq 1$, (3.49) is satisfied for the set $M_1 = \{(m,n,1) \mid n \geq 0\}$ (see (3.50)), and (3.49) is satisfied for the set $M_2 = \{(m,n,1) \mid n \geq 1\}$, since $\{y_{m,n,r}\}$ satisfies the balance principle for each state of this set. Therefore, $\{y_{m,n,r}\}$ also satisfies the balance principle (3.49) for $M_1 \backslash M_2 = \{(m,0,1)\}$ (since (3.49) for $M_1 \backslash M_2$ is obtained by subtracting (3.49) for M_2 from (3.49) for M_1). This proves that $\{y_{m,n,r}\}$ also satisfies the equilibrium equations for the states $(m,0,1)$, $m \geq 1$. One can easily check that all other states outside M_c' may be treated in a similar way. Hence we may conclude that $\{y_{m,n,r}\}$ satisfies all equilibrium equations. By using (3.47), it may be shown that $\Sigma_{(m,n,r) \in M} |y_{m,n,r}| < \infty$, and thus the equilibrium distribution $\{p_{m,n,r}\}$ may be obtained by normalizing the solution $\{y_{m,n,r}\}$. Since, by the definition for $y_{0,0,0}$, the $y_{m,n,r}$ already add up to 1, we finally find that $\{p_{m,n,r}\}$ is equal to $\{y_{m,n,r}\}$.

In Theorem 3.2, we have summarized the main results which follow from the analysis in the Sections 3.3-3.6 for the class of three-dimensional random walks described in Section 3.2.

Theorem 3.2 *(Main Theorem for the case $N=3$)*

The equilibrium distribution $\{p_{m,n,r}\}$ for a random walk of the class described in Section 3.2, can be determined by the compensation approach if Condition 3.1 is satisfied. If this condition is satisfied, then the equilibrium distribution $\{p_{m,n,r}\}$ is equal to the sum of six alternating series of alternating binary trees of pure product-form distributions:

$$p_{m,n,r} = \sum_{i=1}^{3} \left[\sum_{k=0}^{\infty} (-1)^k x_{m,n,r}^{(i)}(\alpha_k^{(i)}, \beta_k^{(i)}, \gamma_k^{(i)}) + \sum_{k=0}^{\infty} (-1)^k x_{m,n,r}^{(i)}(\hat{\alpha}_k^{(i)}, \hat{\beta}_k^{(i)}, \hat{\gamma}_k^{(i)}) \right] ,$$

$$(m,n,r) \in M_1 \cup M_{\Lambda\{1\}} \cup M_{\Lambda\{2\}} \cup M_{\Lambda\{3\}} , \quad (3.51)$$

$$p_{m,0,0} = \left[1 - \frac{q_1^{(1)}}{q_{-1}^{(1)}} \right] \left[\frac{q_1^{(1)}}{q_{-1}^{(1)}} \right]^m - \sum_{\substack{n,r \geq 0 \\ n+r \geq 1}} p_{m,n,r} \, , \quad (m,0,0) \in M_{\{1\}} , \quad (3.52)$$

$$p_{0,n,0} = \left[1 - \frac{q_1^{(2)}}{q_{-1}^{(2)}} \right] \left[\frac{q_1^{(2)}}{q_{-1}^{(2)}} \right]^n - \sum_{\substack{m,r \geq 0 \\ m+r \geq 1}} p_{m,n,r} \, , \quad (0,n,0) \in M_{\{2\}} , \quad (3.53)$$

$$p_{0,0,r} = \left[1 - \frac{q_1^{(3)}}{q_{-1}^{(3)}} \right] \left[\frac{q_1^{(3)}}{q_{-1}^{(3)}} \right]^r - \sum_{\substack{m,n \geq 0 \\ m+n \geq 1}} p_{m,n,r} \, , \quad (0,0,r) \in M_{\{3\}} , \quad (3.54)$$

$$p_{0,0,0} = 1 - \sum_{(m,n,r) \in M \backslash \{(0,0,0)\}} p_{m,n,r} \, . \quad (3.55)$$

Here, the factors $\alpha_k^{(i)}$, $\beta_k^{(i)}$, $\gamma_k^{(i)}$, $\hat{\alpha}_k^{(i)}$, $\hat{\beta}_k^{(i)}$ and $\hat{\gamma}_k^{(i)}$ are defined as described in this section, and the sums $x_{m,n,r}^{(i)}(\cdot,\cdot,\cdot)$ are defined as described in the Sections 3.3 and 3.4.

3.7. Reformulation of Theorem 3.2

In this section, we show that, by studying the definitions of the product-form solutions which constitute the equilibrium distribution, the formula for the equilibrium probabilities $p_{m,n,r}$ for the interior and the boundary planes may be rewritten to a considerably more compact and simpler formula; and, similarly for the marginal distributions. The alternative formulae which we obtain have the advantage that, as we shall see in the next section, they can be easily generalized to the N-dimensional case.

Looking in less detail, we can say that the Main Theorem states that the equilibrium distribution $\{p_{m,n,r}\}$ is equal to one alternating sum of product-form distributions. All these product forms, and also all product forms appearing in the formulae for the marginal distributions, are solutions of the equilibrium equation for the interior, i.e. of the quadratic equation (3.14). By studying the definitions of all product factors, it is easily verified that all these product forms are obtained by taking the trivial solution $(\alpha,\beta,\gamma)=(1,1,1)$ of (3.14) and generating new solutions of (3.14) by letting one factor free each time. The tree of solutions which we obtain in this way is depicted in Figure 3.4. Using this tree of product forms enables us to give more compact formulae for $\{p_{m,n,r}\}$, and also for its marginal distributions.

Let V be the set of vectors given in Section 3.3. Define $(\alpha_\varnothing,\beta_\varnothing,\gamma_\varnothing)=(1,1,1)$ and let for all other vectors $v \in V$ the factors of $(\alpha_v,\beta_v,\gamma_v)$ be defined by

$$\alpha_v = f_1(\beta_{p(v)},\gamma_{p(v)})/\alpha_{p(v)}\,, \quad \beta_v = \beta_{p(v)}\,, \quad \gamma_v = \gamma_{p(v)} \quad \text{if } v_{l(v)}=1;$$

$$\alpha_v = \alpha_{p(v)}\,, \quad \beta_v = f_2(\alpha_{p(v)},\gamma_{p(v)})/\beta_{p(v)}\,, \quad \gamma_v = \gamma_{p(v)} \quad \text{if } v_{l(v)}=2;$$

$$\alpha_v = \alpha_{p(v)}\,, \quad \beta_v = \beta_{p(v)}\,, \quad \gamma_v = f_3(\alpha_{p(v)},\beta_{p(v)})/\gamma_{p(v)} \quad \text{if } v_{l(v)}=3.$$

Then the set of all solutions depicted in Figure 3.4 is given by

$$P^* = \{\,(\alpha_v,\beta_v,\gamma_v) \mid v \in V\,\}\,.$$

For each solution $(\alpha_v,\beta_v,\gamma_v)$ in this set, all factors are real numbers in the interval $(0,1]$ and therefore P^* may be partitioned into the subsets

$$P_J^* = \{\,(\alpha_1,\alpha_2,\alpha_3)\in P^* \mid \alpha_i<1 \text{ for all } i\in J \text{ and } \alpha_i=1 \text{ for all } i\notin J\,\}\,, \quad J\subset I.$$

For all $J\subset I$, $J\neq\varnothing$, it may be shown that the set P_J^* is equal to the set of product-form solutions needed for describing the equilibrium behavior of the components $i\in J$.

As one can easily check, for the marginal distribution $\{p_m^{(1)}\}$ only the unique solution $(\alpha_v,\beta_v,\gamma_v)\in P^*$ with $\alpha_v<1$ and $\beta_v=\gamma_v=1$ is needed; and similarly for $\{p_n^{(2)}\}$ and $\{p_r^{(3)}\}$. Considering the formula for $\{p_{m,n}^{(1,2)}\}$ shows that $\{p_{m,n}^{(1,2)}\}$ consists of the product forms $\pm(1-\alpha_v)\alpha_v^m(1-\beta_v)\beta_v^n$ with $(\alpha_v,\beta_v,\gamma_v)\in P^*$ and $\alpha_v<1$, $\beta_v<1$ and $\gamma_v=1$, where the sign depends on the distance between the node v and the node $\varnothing$; and similarly for $\{p_{m,r}^{(1,3)}\}$ and $\{p_{m,n}^{(1,2)}\}$. Finally, for the equilibrium distribution $\{p_{m,n,r}\}$, all solutions $(\alpha_v,\beta_v,\gamma_v)\in P^*$ with $\alpha_v<1$, $\beta_v<1$ and $\gamma_v<1$ are needed. These observations lead to the compact and relatively simple formulae stated in the following theorem.

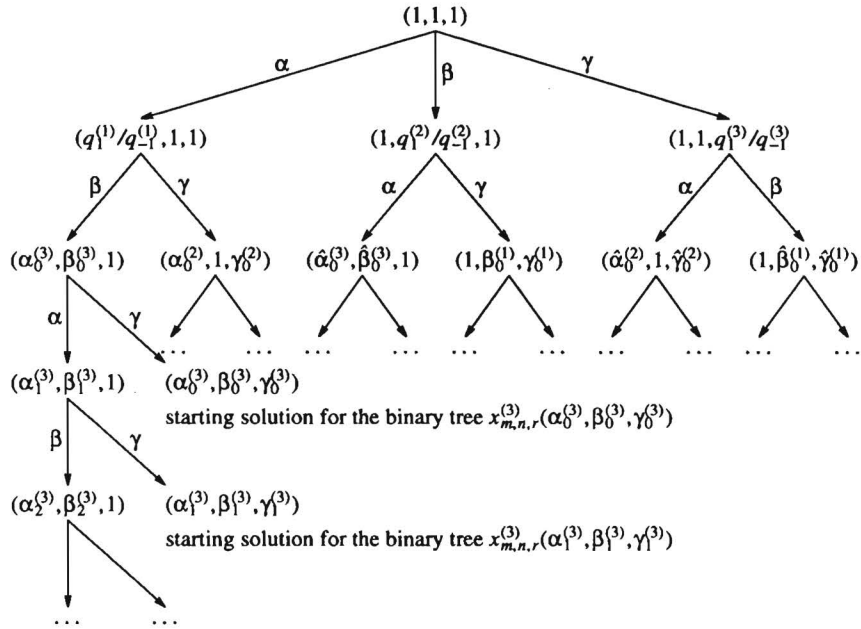


Figure 3.4. All relevant solutions of (3.14) needed for the equilibrium distribution $\{p_{m,n,r}\}$ and all its marginal distributions.

Theorem 3.3 *(Reformulation of Theorem 3.2)*

The equilibrium distribution $\{p_{m,n,r}\}$ for a random walk of the class described in Section 3.2, can be determined by the compensation approach if Condition 3.1 is satisfied. If this condition is satisfied, then the equilibrium distribution and its marginal distributions are given by:

$$p_m^{(1)} = \sum_{(\alpha_v,\beta_v,\gamma_v)\in P^*_{\{1\}}} (-1)^{l(v)-1}(1-\alpha_v)\alpha_v^m\,, \quad m\geq 0, \tag{3.56}$$

and similarly for $\{p_n^{(2)}\}$ and $\{p_r^{(3)}\}$;

$$p_{m,n}^{(1,2)} = \sum_{(\alpha_v,\beta_v,\gamma_v)\in P^*_{\{1,2\}}} (-1)^{l(v)-2}(1-\alpha_v)\alpha_v^m(1-\beta_v)\beta_v^n\,, \quad m,n\geq 0,\ m+n\geq 1, \tag{3.57}$$

$$p_{0,0}^{(1,2)} = 1 - \sum_{\substack{m,n\geq 0\\ m+n\geq 1}} p_{m,n}^{(1,2)}\,,$$

and similarly for $\{p_{m,r}^{(1,3)}\}$ and $\{p_{m,n}^{(1,2)}\}$;

$$p_{m,n,r} = \sum_{(\alpha_v,\beta_v,\gamma_v)\in P_I^*} (-1)^{l(v)-3}(1-\alpha_v)\alpha_v^m(1-\beta_v)\beta_v^n(1-\gamma_v)\gamma_v^r\,,$$

$$(m,n,r)\in M_I\cup M_{\Lambda\{1\}}\cup M_{\Lambda\{2\}}\cup M_{\Lambda\{3\}}\,, \tag{3.58}$$

and the equilibrium probabilities $p_{m,n,r}$ for the states in the interior and on the boundary planes follow from (3.52)-(3.55).

3.8. N-dimensional random walks

In this section, the main results derived in the Sections 3.2-3.7 for a class of three-dimensional random walks are generalized to the corresponding class of N-dimensional random walks, where $N \geq 2$. As the reader can easily check, the results appear to hold also for the case $N = 1$.

Consider an N-dimensional, irreducible, positive recurrent, homogeneous, nearest-neighboring random walk with the projection property on the set of states

$$M = \{ (m_1, \ldots, m_N) \mid m_i \in \mathbb{N}_0 \text{ for all } i \in I \},$$

where $N \geq 2$ and $I := \{1, \ldots, N\}$. For such a random walk, all transition probabilities/rates are uniquely determined by the transition probabilities/rates for the interior of the state space. Let the set of feasible transitions for the interior states be given by

$$T = \{ (t_1, \ldots, t_N) \mid t_i \in \{-1, 0, 1\} \text{ for all } i \in I \},$$

and let the corresponding transition probabilities/rates be denoted by $q_{t_1, \ldots, t_N}$. W.l.o.g., we may assume that we have discrete time, which implies that $\Sigma_{(t_1, \ldots, t_N) \in T} q_{t_1, \ldots, t_N} = 1$. Finally, let the equilibrium distribution be denoted by $\{p_{m_1, \ldots, m_N}\}$.

Because of the projection property, all marginal distributions of $\{p_{m_1, \ldots, m_N}\}$ can be characterized as the equilibrium distributions of lower-dimensional, homogeneous, nearest-neighboring random walks with the projection property. The one-dimensional marginal distribution $\{p_m^{(i)}\}$ for the component m_i, $i \in I$, which is defined by

$$p_m^{(i)} = \sum_{\substack{(n_1, \ldots, n_N) \in M \\ n_i = m}} p_{n_1, \ldots, n_N}, \quad m \geq 0, \tag{3.59}$$

is equal to the equilibrium distribution of the one-dimensional random walk with transition probabilities

$$q_t^{(i)} = \sum_{\substack{(s_1, \ldots, s_N) \in T \\ s_i = t}} q_{s_1, \ldots, s_N}, \quad t \in \{-1, 0, 1\}. \tag{3.60}$$

The full random walk will be positive recurrent if and only if all component random walks have negative drifts, i.e. if and only if $q_{-1}^{(i)} > q_1^{(i)}$ for all $i \in I$. Further, the irreducibility implies that both $q_{-1}^{(i)}$ and $q_1^{(i)}$ are positive. This leads to the property that

$$q_{-1}^{(i)} > q_1^{(i)} > 0 \quad \text{for all } i \in I. \tag{3.61}$$

To determine the equilibrium distribution $\{p_{m_1, \ldots, m_N}\}$, one can use the compensation approach, which tries to construct a solution of all equilibrium equations by linearly combining product-form solutions which satisfy the equilibrium equation for the interior:

$$p_{m_1, \ldots, m_N} = \sum_{(t_1, \ldots, t_N) \in T} q_{t_1, \ldots, t_N} p_{m_1 - t_1, \ldots, m_N - t_N}, \quad m_i > 0 \text{ for all } i \in I. \tag{3.62}$$

By substituting the product form $\Pi_{i=1}^N \alpha_i^{m_i}$ into this equation, it follows that this equation is satisfied if and only if $(\alpha_1, \ldots, \alpha_N)$ satisfies the quadratic equation

$$\prod_{i=1}^{N} \alpha_i = \sum_{(t_1,\ldots,t_N)\in T} q_{t_1,\ldots,t_N} \prod_{i=1}^{N} \alpha_i^{1-t_i} . \tag{3.63}$$

The application of the compensation approach leads to the construction of $(N-1)$-fold trees of product-form solutions of (3.63), which are called formal solutions. Each formal solution consists of a starting term, which is required to satisfy also the equilibrium equation for one of the boundary planes, and a countable set of compensation terms, which all correct an error of a previous term at one of the boundary planes. For the absolute convergence of the formal solutions, it appears to be needed that the following condition is satisfied:

$$q_{t_1,\ldots,t_N} = 0 \quad \text{for all transitions } (t_1,\ldots,t_N)\in T \text{ with } t_i+t_j>0 \text{ for some } i,j\in I,\ i\neq j. \tag{3.64}$$

This may be shown by generalizing the analysis presented in the Sections 3.3 and 3.4.

Together with (3.61), condition (3.64) implies that the probabilities $q_{t_1,\ldots,t_N}$ for all transitions with $t_i=1$ for some $i\in I$ and $t_j=-1$ for all $j\in I\backslash\{i\}$ must be positive. Given that (3.64) is satisfied, (3.61) represents a condition which is necessary and sufficient for the irreducibility and positive recurrence of the full random walk. Further, condition (3.64) itself is necessary and sufficient for the property that all marginal distributions can be determined by the compensation approach.

It is obvious that condition (3.64) essentially restricts the applicability of the compensation approach, especially for $N\geq 3$. However, if the condition is satisfied, then the formal solutions are absolutely convergent, at least in all states $(m_1,\ldots,m_N)$ with $m_i\geq 0$ for all $i\in I$ and $m_i=0$ for at most one i, and we obtain very explicit expressions for the equilibrium distribution and all its marginal distributions. These explicit expressions are given in the next paragraph. The absolute convergence of the formal solutions may be proved by generalizing the analysis of Section 3.5. The explicit expressions for the equilibrium and the marginal distributions are proved by induction with respect to the dimension N. The initial step, i.e. the case $N=2$, has been treated in Chapter 2. For the induction step from dimension N to dimension $N+1$, one can generalize the analysis of the Sections 3.6 and 3.7, where the step from $N=2$ to $N=3$ has been treated.

Suppose that condition (3.64) is satisfied. Then the equilibrium distribution and all its marginal distributions are equal to alternating sums of pure product-form distributions, where all product forms are obtained by taking the trivial solution $(1,\ldots,1)$ of the quadratic equation (3.63) and generating new solutions of (3.63) by letting one factor free each time. Let the set of vectors V be defined by

$$V = \{(v_1,\ldots,v_l) \mid l\in\mathbb{N}_0,\ \text{if } l\geq 1 \text{ then } v_1\in I \text{ and } v_k\in I\backslash\{v_{k-1}\} \text{ for all } k\geq 2\}.$$

Next, define $(\alpha_{1,\varnothing},\ldots,\alpha_{N,\varnothing})=(1,\ldots,1)$ and let for all other vectors $v\in V$ the factors of $(\alpha_{1,v},\ldots,\alpha_{N,v})$ be defined by

$$\alpha_{i,v} = \begin{cases} f_i(\alpha_{1,p(v)},\ldots,\alpha_{i-1,p(v)},\alpha_{i+1,p(v)},\ldots,\alpha_{N,p(v)})/\alpha_{i,p(v)} & \text{if } v_{l(v)}=i; \\ \alpha_{i,p(v)} & \text{if } v_{l(v)}\neq i, \end{cases}$$

where $p(v)$ and $l(v)$ are the parent and the length of a vector v and the function $f_i(\alpha_1,\ldots,\alpha_{i-1},\alpha_{i+1},\ldots,\alpha_N)$ denotes the product of the two roots of the quadratic equation (3.63) for fixed $\alpha_1,\ldots,\alpha_{i-1},\alpha_{i+1},\ldots,\alpha_N$. Then the set of all relevant product-form solutions is given by

$$P^* = \{ (\alpha_{1,v}, \ldots, \alpha_{N,v}) \mid v \in V \}.$$

It can be shown that the factors of all solutions $(\alpha_{1,v}, \ldots, \alpha_{N,v})$ of this set are real numbers in the interval $(0,1]$. So, P^* can be partitioned into the subsets

$$P_J^* = \{ (\alpha_1, \ldots, \alpha_N) \in P^* \mid \alpha_i < 1 \text{ for all } i \in J \text{ and } \alpha_i = 1 \text{ for all } i \notin J \}, \quad J \subset I.$$

The solutions in a set P_J^* are precisely the ones needed to describe the equilibrium behavior of the components belonging to J.

Theorem 3.4 *(Main Theorem for the case with general $N \geq 2$)*

The equilibrium distribution for an N-dimensional, irreducible, positive recurrent, homogeneous, nearest-neighboring random walk with the projection property on the states $(m_1, \ldots, m_N)$, where $N \geq 2$ and $m_i \in \mathbb{N}_0$ for all i, can be determined by the compensation approach if condition (3.64) is satisfied. If this condition is satisfied, then the following formula is found for the equilibrium distribution and the marginal distributions. Let $\{p_{m_1,\ldots,m_L}^{(j_1,\ldots,j_L)}\}$ be the equilibrium (marginal) distribution for the components $j_1, \ldots, j_L$, where $1 \leq L \leq N$ and $1 \leq j_1 < \ldots < j_L \leq N$, and let $J = \{j_1, \ldots, j_N\}$, then

$$p_{m_1,\ldots,m_L}^{(j_1,\ldots,j_L)} = \sum_{(\alpha_{1,v},\ldots,\alpha_{N,v}) \in P_J^*} (-1)^{l(v)-L} \prod_{i=1}^{L} (1-\alpha_{j_i,v}) \alpha_{j_i,v}^{m_i} \tag{3.65}$$

for all $(m_1, \ldots, m_L)$ with $m_i \geq 0$ for all $i = 1, \ldots, L$ and $m_i = 0$ for at most one i.

Remark that according to the notation used in this theorem the distribution $\{p_{m_1,\ldots,m_N}\}$ for the full random walk is denoted by $\{p_{m_1,\ldots,m_N}^{(1,\ldots,N)}\}$. Further, note that all equilibrium probabilities $p_{m_1,\ldots,m_L}^{(j_1,\ldots,j_L)}$ for the states for which formula (3.65) does not hold, may be determined with the help of the marginal distributions of $\{p_{m_1,\ldots,m_L}^{(j_1,\ldots,j_L)}\}$. Of course, they may also be determined with the help of the equilibrium equations of the random walk for the components $j_1, \ldots, j_L$.

A thorough analysis of the structure behind the solutions $(\alpha_{1,v}, \ldots, \alpha_{N,v}) \in P^*$, which in fact are obtained from one large tree consisting of a root and N, $(N-1)$-fold subtrees, will be required for the computational aspects of the determination of the equilibrium distribution and related quantities. This structure will be investigated in Chapter 4 and leads to efficient numerical procedures. These procedures may be used for a performance analysis of the $2 \times N$ switch, which is in the considered class of random walks and satisfies condition (3.64) (for all $N \geq 2$).

3.9. Conclusions

We have applied the compensation approach to the class of N-dimensional, irreducible, positive recurrent, homogeneous, nearest-neighboring random walks with the projection property on the states $(m_1, \ldots, m_N)$, where $N \geq 2$ and $m_i \in \mathbb{N}_0$ for all i. It has been shown that the compensation approach works for a random walk of this class if and only if no transitions can be made from the states in the interior into directions which for some pair of components m_i and m_j of the state enlarge the distance to the origin, i.e. into directions $(t_1, \ldots, t_N)$ with

$t_i+t_j > 0$ for some $i,j \in \{1,\ldots,N\}$, $i \neq j$. We believe that this condition is also necessary for random walks which do not satisfy the projection property and for random walks for which the transitions are not restricted to nearest neighbors (see also Chapter 8).

If the compensation approach works for a random walk of the considered class, then the equilibrium distribution and all marginal distributions are equal to alternating sums of infinitely many, pure product-form solutions of the equilibrium equation for the states in the interior of the state space. Simple, recursive formulae are available for the determination of the required product-form solutions. These product-form solutions are in fact obtained from one large tree, which appears to have an interesting, geometric structure; this structure will be investigated in Chapter 4, and, among others, it will lead to efficient numerical procedures for the computation of the equilibrium distribution and related quantities.

Chapter 4

The Equilibrium Distribution for a Class of Multi-Dimensional Random Walks: Structure Analysis

4.1. Introduction

In the previous chapter, we have obtained explicit expressions for the equilibrium distribution and the marginal distributions for the class of N-dimensional, irreducible, positive recurrent, homogeneous, nearest-neighboring random walks which satisfy the projection property and condition (3.64) under which the compensation approach works; see formula (3.65) stated in Theorem 3.4. These expressions show that these distributions are equal to alternating sums of infinitely many product-form distributions. The *aim* of the present chapter is to gain insight into the *structure* of these sums, and, moreover, to use this insight for obtaining *error bounds* for the approximation of these infinite sums by finite (partial) sums.

Since the marginal distributions are the equilibrium distributions of lower-dimensional random walks, it suffices to pay attention to the equilibrium distribution of the full random walk. For the equilibrium distribution $\{p_{m_1,\ldots,m_N}\}$, by (3.65), we have the formula

$$p_{m_1,\ldots,m_N} = \sum_{(\alpha_{1,\nu},\ldots,\alpha_{N,\nu})\in P^*} (-1)^{l(\nu)-N} \prod_{i=1}^{N} (1-\alpha_{i,\nu})\alpha_{i,\nu}^{m_i}, \quad (m_1,\ldots,m_N)\in M_c, \tag{4.1}$$

where

$$M_c = \{\,(m_1,\ldots,m_N)\in M \mid m_i = 0 \text{ for at most one } i\in I\,\}$$

is the convergence region, which contains all states where the formal solutions constructed by the compensation approach are absolutely convergent (note that it is allowed to sum over all solutions $(\alpha_{1,\nu},\ldots,\alpha_{N,\nu})\in P^*$ instead of over the solutions of the subset P_I^* consisting of all $(\alpha_{1,\nu},\ldots,\alpha_{N,\nu})\in P^*$ with $\alpha_{i,\nu}<1$ for all $i\in I$, since all $(\alpha_{1,\nu},\ldots,\alpha_{N,\nu})\in P^*\backslash P_I^*$ have a contribution equal to 0). The solutions $(\alpha_{1,\nu},\ldots,\alpha_{N,\nu})\in P^*$, and therefore also the terms of the infinite sum on the right-hand side of formula (4.1), correspond to the nodes

$$\nu\in V = \{\,(\nu_1,\ldots,\nu_l) \mid l\in\mathbb{N}_0,\ \text{if } l\geq 1 \text{ then } \nu_1\in I \text{ and } \nu_k\in I\backslash\{\nu_{k-1}\} \text{ for all } k\geq 2\,\}$$

of the tree for which the root $\varnothing$ has N successors (viz. $(1),\ldots,(N)$), while all other nodes

$v=(v_1,\ldots,v_l)\in V\backslash\{\varnothing\}$ have $N-1$ successors (viz. the nodes $(v_1,\ldots,v_l,v_{l+1})$ with $v_{l+1}\in I\backslash\{v_l\}$).

By studying terms corresponding to pairs of connected nodes, it will be shown that, in essence, for each node $v=(v_1,\ldots,v_l)\in V\backslash\{\varnothing\}$ the contribution of its corresponding term only depends on the contribution of the term corresponding to its parent $p(v)=(v_1,\ldots,v_{l-1})$ and the value of its last component v_l. Such a behavior is typical for so-called *geometric trees*. By exploiting the geometric behavior, we can define error bounds for the approximation of an equilibrium probability by finite sums in terms of $(N-1)$-fold, geometric trees, i.e. for geometric trees for which each node has $N-1$ successors (more specifically, these $(N-1)$-fold, geometric trees serve as upper bounds for the subtrees of terms which are deleted when approximating the infinite sum on the right-hand side of (4.1) by a finite sum).

An attractive analysis of the geometric trees constitutes the *core* of this chapter. This analysis is based on matrix algebra, and it will lead to an explicit condition under which the sum of all terms of a geometric tree converges, and to an explicit formula for the sum itself. At the end of this chapter, it is shown that these explicit results can be exploited in an efficient numerical procedure for the computation of the equilibrium probabilities $p_{m_1,\ldots,m_N}$ in the convergence region M_c (the probabilities for the states outside this region may be computed with the help of appropriately chosen equilibrium equations).

The error bounds which we derive from the geometric trees and the efficient numerical procedure that is developed in this chapter, will appear to be generalizations of the error bounds and the efficient numerical procedure described in Section 2.5 for the case $N=2$. In case $N=2$, the sum on the right-hand side of (4.1) reduces to the sum of two alternating series of product-form solutions (cf. formula (2.36) in Theorem 2.1) and the $(N-1)$-fold, geometric trees reduce to series with exponentially (geometrically) decreasing terms; hence, the case $N=2$ constitutes a relatively easy case.

Let us finally present the outline of this chapter. In Section 4.2, we derive error bounds in terms of geometric trees for the approximation of the equilibrium probabilities $p_{m_1,\ldots,m_N}$ in the convergence region M_c by finite (partial) sums. Next, in Section 4.3, the attractive analysis of the geometric trees is presented. After that, in the Sections 4.4 and 4.5, we describe three numerical procedures for the computation of the equilibrium probabilities, and we present some numerical results for the equilibrium distribution and related quantities for the symmetric 2×3 switch. The conclusions are given in Section 4.6.

4.2. Error bounds

This section is devoted to the approximation of the infinite sum on the right-hand side of formula (4.1) for the equilibrium probabilities $p_{m_1,\ldots,m_N}$ for the states $(m_1,\ldots,m_N)$ in the convergence region M_c. We shall first describe the type of finite sums which are used to approximate the infinite sum, for which the terms are obtained from the tree with nodes $v\in V$. Next, we shall focus on error bounds for these approximations, i.e. on upper bounds for the terms which are not included in the finite sums. By investigating the product factors $\alpha_{i,v}$ of the solutions $(\alpha_{1,v},\ldots,\alpha_{N,v})\in P^*$, we will be able to obtain explicit upper bounds in terms of so-called geometric trees; see Lemma 4.2 at the end of this section. It is recalled that explicit expressions for the geometric trees itself are derived in the next section.

Let $(m_1,\ldots,m_N)\in M_c$ and consider formula (4.1) for the corresponding equilibrium probability $p_{m_1,\ldots,m_N}$. The terms of the infinite sum on the right-hand side of this formula correspond to the nodes $v\in V$ of the tree emerging from the definition of the solutions $(\alpha_{1,v},\ldots,\alpha_{N,v})\in P^*$. Therefore, formula (4.1) is equivalent to

$$p_{m_1,\ldots,m_N} = \sum_{v\in V} t_{m_1,\ldots,m_N}(v)\,, \tag{4.2}$$

where

$$t_{m_1,\ldots,m_N}(v) = (-1)^{l(v)-N}\prod_{i=1}^{N}(1-\alpha_{i,v})\alpha_{i,v}^{m_i} \quad \text{for all } v\in V.$$

As already noted in the introduction of this chapter, the tree with the nodes $v\in V$ has a very regular structure. The root $\varnothing$ has N successors, viz. $(1),\ldots,(N)$, and each vector $v\in V\backslash\{\varnothing\}$ has $N-1$ successors. So, from this point of view, $p_{m_1,\ldots,m_N}$ is equal to the sum of the N sums of terms $t_{m_1,\ldots,m_N}(v)$ over the nodes of the $(N-1)$-fold (sub)trees with roots $(1),\ldots,(N)$ (note that the term for the root $\varnothing$ may be neglected in this reasoning, since $t_{m_1,\ldots,m_N}(\varnothing)=0$). This property tells that in the case $N=2$ the equilibrium probabilities p_{m_1,m_2} are equal to the sum of two series (cf. formula (2.36) in Theorem 2.1); in the case $N=3$ the equilibrium probabilities p_{m_1,m_2,m_3} are equal to the sum of the 3 sums over the nodes of the binary trees starting at the nodes (1), (2) and (3) (the nodes with solutions $(q_1^{(1)}/q_{-1}^{(1)},1,1)$, $(1,q_1^{(2)}/q_{-1}^{(2)},1)$ and $(1,1,q_1^{(3)}/q_{-1}^{(3)})$; see Figure 3.4 in Section 3.7).

Obviously, the infinite sum on the right-hand side of formula (4.2) cannot be computed exactly. But, it may be approximated by finite sums of the terms $t_{m_1,\ldots,m_N}(v)$. We shall now describe the *type of finite sums* which we use. Our choice has been guided by the type of finite sums which we used for the two-dimensional case in Section 2.5.

In the two-dimensional case, we have approximated the two relevant series by partial sums, which is equivalent to saying that an equilibrium probability p_{m_1,m_2}, $(m_1,m_2)\in M_c$, is approximated by finite sums $\Sigma_{v\in V'} t_{m_1,m_2}(v)$, where the set $V'\subset V$ is of the form

$$V' = \{\varnothing\}\cup\{\,v\in V \mid 1\le l(v)\le k_1 \text{ and } v_1=2\,\}\cup\{\,v\in V\mid 1\le l(v)\le k_2 \text{ and } v_1=1\,\},$$

and k_1 and k_2 are fixed, positive integers; the nodes $v\in V$ with $1\le l(v)\le k_1$ and $v_1=2$ correspond to the first k_1 product forms in the first series on the right-hand side of formula (2.36) and the nodes $v\in V$ with $1\le l(v)\le k_2$ and $v_1=1$ correspond to the first k_2 product forms in the second series on the right-hand side of formula (2.36). Note that the first terms of both series are always included in the approximation. In the N-dimensional case, we shall use finite sums which constitute a direct generalization of the sums $\Sigma_{v\in V'} t_{m_1,m_2}(v)$ with V' of the above form. In the N-dimensional case, we approximate the equilibrium probability $p_{m_1,\ldots,m_N}$ for some state $(m_1,\ldots,m_N)\in M_c$ by finite sums $\Sigma_{v\in V'} t_{m_1,\ldots,m_N}(v)$, where $V'\subset V$ is required to satisfy the following conditions:

$$\{\varnothing,(1),\ldots,(N)\}\subset V', \tag{4.3}$$

$$\text{if } v=(v_1,\ldots,v_{l(v)})\in V',\ v\neq\varnothing,\ \text{then } (v_1,\ldots,v_k)\in V' \text{ for all } k=1,\ldots,l(v)-1\,, \tag{4.4}$$

$$\text{if } v\in V', \text{ then } O(v)\subset V' \text{ or } O(v)\cap V'=\varnothing\,, \tag{4.5}$$

where $O(v)=\{\,w\in V\mid p(w)=v\,\}$ denotes the offspring of a node $v\in V$.

Condition (4.3) states that the root $\varnothing$ of the whole tree with nodes $v \in V$ and the roots $(1), \ldots, (N)$ of the $(N-1)$-fold subtrees into which the whole tree can be divided, must be included in the finite set V'. Condition (4.4) states that a node v can only be contained in V', if all nodes on the path from the root $\varnothing$ to v are already contained in V'. This is a sensible condition for two reasons. First of all, since the higher a node v is in the tree, the larger its contribution $t_{m_1,\ldots,m_N}(v)$ to $p_{m_1,\ldots,m_N}$ will be in general (this follows from the property that for each path $\{v^{(k)}\}$ in V all factors of the solutions $(\alpha_{1,v^{(k)}}, \ldots, \alpha_{N,v^{(k)}})$ are monotonously non-increasing for increasing k; cf. Lemma 3.6(ii)). Secondly, because of the recursive formulae for the product-form solutions, the contribution $t_{m_1,\ldots,m_N}(v)$ of a node v can only be computed after having computed the product factors $\alpha_{1,w}, \ldots, \alpha_{N,w}$ for all nodes w on the path to v; therefore for each node w on the path to v the computation of the contribution $t_{m_1,\ldots,m_N}(w)$ requires only little extra computational effort if it is known that v will be included in V'. By condition (4.5), if one successor $w \in O(v)$ of an element $v \in V'$ is included in V', then the whole offspring $O(v)$ of v must be included in V'. This condition is mainly introduced to simplify the derivation of an error bound for the approximation of an equilibrium probability by a finite sum.

Suppose that a finite sum $\Sigma_{v \in V'} t_{m_1,\ldots,m_N}(v)$ is used as an approximation for the equilibrium probability $p_{m_1,\ldots,m_N}$, where $(m_1, \ldots, m_N) \in M_c$ and V' is a finite subset of nodes $v \in V$ which satisfies the conditions (4.3)-(4.5). Then the tree of nodes $v \in V'$ is obtained by deleting all nodes below the nodes v of the set

$$B(V') = \{ v \in V' \mid O(v) \cap V' = \varnothing \},$$

i.e. by deleting all nodes of the subtree

$$S(v) = \{ w \in V \mid l(w) \geq l(v) \text{ and } w_k = v_k \text{ for all } k = 1, \ldots, l(v) \},$$

except the root v itself, for each node $v \in B(V')$. This implies that $V \backslash V' = \cup_{v \in B(V')} S(v) \backslash \{v\}$, and the absolute error of the approximation by the finite sum $\Sigma_{v \in V'} t_{m_1,\ldots,m_N}(v)$ is bounded by

$$\left| p_{m_1,\ldots,m_N} - \sum_{v \in V'} t_{m_1,\ldots,m_N}(v) \right| \leq \sum_{v \in B(V')} b_{m_1,\ldots,m_N}(v), \tag{4.6}$$

where the upper bound $b_{m_1,\ldots,m_N}(v)$ must be defined such that

$$\left| \sum_{w \in S(v) \backslash \{v\}} t_{m_1,\ldots,m_N}(w) \right| \leq b_{m_1,\ldots,m_N}(v) \quad \text{for all } v \in V \backslash \{\varnothing\}; \tag{4.7}$$

note that it is not needed to define an upper bound $b_{m_1,\ldots,m_N}(v)$ for $v = \varnothing$, since, by (4.3), the node $v = \varnothing$ is never contained in $B(V')$. Since for all $v \in V \backslash \{\varnothing\}$, the subtree $S(v)$ is an $(N-1)$-fold tree, in the remainder of this section we can perform a uniform analysis in order to find an appropriate definition for the upper bounds $b_{m_1,\ldots,m_N}(v)$.

The upper bounds $b_{m_1,\ldots,m_N}(v)$ must be defined such that they satisfy condition (4.7). Further they must be computable and it is desired that they are tight. Since

$$| t_{m_1,\ldots,m_N}(v) | \leq \prod_{i=1}^{N} \alpha_{i,v}^{m_i} \quad \text{for all } v \in V,$$

the variables $\hat{b}_{m_1,\ldots,m_N}(v)$ defined by

$$\hat{b}_{m_1,\ldots,m_N}(v) = \sum_{w \in S(v)\setminus\{v\}} \prod_{i=1}^{N} \alpha_{i,w}^{m_i} \quad \text{for all } v \in V\setminus\{\varnothing\}, \tag{4.8}$$

satisfy condition (4.7). Besides, these upper bounds will be tight for nodes v for which the factors $\alpha_{1,v},\ldots,\alpha_{N,v}$ are small, i.e. for most nodes v at large distances from the root $\varnothing$. Although the upper bounds $\hat{b}_{m_1,\ldots,m_N}(v)$ are not computable, they still are useful. Below, we will find that for each node $v \in V\setminus\{\varnothing\}$, the value of the term $\Pi_{i=1}^{N} \alpha_{i,v}^{m_i}$ mainly depends on the value of the term $\Pi_{i=1}^{N} \alpha_{i,p(v)}^{m_i}$ for its parent $p(v)$ and the value of the last component $v_{l(v)}$ of v. This geometric behavior is established by considering ratios of the terms $\Pi_{i=1}^{N} \alpha_{i,v}^{m_i}$ for pairs of connected nodes $v \in V$, and it can be exploited to derive tight and computable upper bounds $b_{m_1,\ldots,m_N}(v)$ in terms of geometric trees for the initial upper bounds $\hat{b}_{m_1,\ldots,m_N}(v)$.

The ratio of the term corresponding to a node $v \in V\setminus\{\varnothing\}$ and the term corresponding to its parent $p(v)$ is given by

$$\frac{\Pi_{i=1}^{N} \alpha_{i,v}^{m_i}}{\Pi_{i=1}^{N} \alpha_{i,p(v)}^{m_i}} = \frac{\alpha_{k,v}^{m_k}}{\alpha_{k,p(v)}^{m_k}} = [h_k(\alpha_{1,v},\ldots,\alpha_{k-1,v},\alpha_{k+1,v},\ldots,\alpha_{N,v})]^{m_k}, \tag{4.9}$$

where $k = v_{l(v)}$ and the function $h_k(\cdot)$ is defined as the ratio of the smallest root α_k and the largest root α_k of the quadratic equation (3.63) for fixed $\alpha_1,\ldots,\alpha_{k-1},\alpha_{k+1},\ldots,\alpha_N$; let for each $k \in I$, the function $h_k(\cdot)$ be defined for all $(\alpha_1,\ldots,\alpha_{k-1},\alpha_{k+1},\ldots,\alpha_N) \in (0,1]^{N-1}$. Now, we first derive an expression for $h_k(\cdot)$, and next we will prove a useful property for the functions $h_k(\cdot)$.

Let $k \in I$ and $\alpha_j \in (0,1]$ for all $j \in I\setminus\{k\}$. We are interested in the roots α_k of the quadratic equation (3.63), which we rewrite to

$$\left[\prod_{i\in I\setminus\{k\}} \alpha_i\right] \alpha_k = \sum_{t=-1}^{1} \sum_{\substack{(t_1,\ldots,t_N)\in T \\ t_k = t}} q_{t_1,\ldots,t_N} \left[\prod_{i\in I\setminus\{k\}} \alpha_i^{1-t_i}\right] \alpha_k^{1-t}.$$

This equation is equivalent to the equation

$$a_1\alpha_k^2 - a_2\left[\prod_{i\in I\setminus\{k\}} \alpha_i\right]\alpha_k + a_3\left[\prod_{i\in I\setminus\{k\}} \alpha_i\right]^2 = 0 \tag{4.10}$$

with coefficients

$$a_1 = \sum_{\substack{(t_1,\ldots,t_N)\in T \\ t_k=-1}} q_{t_1,\ldots,t_N} \prod_{i\in I\setminus\{k\}} \alpha_i^{1-t_i},$$

$$a_2 = 1 - \sum_{\substack{(t_1,\ldots,t_N)\in T \\ t_k=0}} q_{t_1,\ldots,t_N} \prod_{i\in I\setminus\{k\}} \alpha_i^{-t_i},$$

$$a_3 = q_1^{(k)},$$

and, by using the assumption that condition (3.64) is satisfied, the coefficients a_i may be shown to be real-valued and positive (to prove that $a_2 > 0$, one has to use that, by (3.64), for all positive rates $q_{t_1,\ldots,t_N}$ with $t_k = 0$ all other coordinates t_i are ≤ 0). Finally, (4.10) is rewritten to

$$a_1 z_k^2 - a_2 z_k + a_3 = 0 \tag{4.11}$$

with

$$z_k = \alpha_k \left[\prod_{i \in \Lambda\{k\}} \alpha_i \right]^{-1} .$$

The function on the left-hand side of equation (4.11) is a quadratic function, which is >0 for $z_k = 0$ and ≤ 0 for $z_k = q_1^{(k)}/q_{-1}^{(k)} < 1$ and for $z_k = 1$ (use that $a_1 \leq q_{-1}^{(k)}$, $a_2 \geq 1 - q_0^{(k)}$ and $a_3 = q_1^{(k)}$) and which tends to ∞ as $z_k \to \infty$. As a consequence, (4.11) has two real-valued, positive roots: one root $z_k^{(1)} \leq q_1^{(k)}/q_{-1}^{(k)}$ and one root $z_k^{(2)} \geq 1$ (cf. Lemma 3.3). Let the smaller and the larger root of (4.10) be denoted by $\alpha_k^{(1)}$ and $\alpha_k^{(2)}$, respectively, then we find

$$h_k(\alpha_1, \ldots, \alpha_{k-1}, \alpha_{k+1}, \ldots, \alpha_N) = \frac{\alpha_k^{(1)}}{\alpha_k^{(2)}} = \frac{z_k^{(1)}}{z_k^{(2)}} = \frac{A_1}{A_2} \tag{4.12}$$

with

$$A_1 = a_2 - \sqrt{D}\,, \quad A_2 = a_2 + \sqrt{D}\,, \quad D = a_2^2 - 4a_1a_3\,.$$

From the properties stated above for the roots $z_k^{(1)}$ and $z_k^{(2)}$ of (4.11), we know that the discriminant D is positive, $A_2 > A_1 > 0$ and $0 < h_k(\cdot) \leq q_1^{(k)}/q_{-1}^{(k)} < 1$.

By using expression (4.12), we are able to prove that the function $h_k(\cdot)$ is monotonously non-decreasing in each argument; see Lemma 4.1. So, $h_k(\cdot)$ reaches its maximum in case $\alpha_i = 1$ for all $i \in \Lambda\{k\}$; and, as it is easily verified, the maximum value is equal to $h_k(1, \ldots, 1) = q_1^{(k)}/q_{-1}^{(k)}$. It is noted that, along the same lines as in the proof of Lemma 4.1, one could prove that $h_k(\cdot)$ is also convex in each argument. Finally, we state that $h_k(\cdot) \to 0$ as $\alpha_i \to 0$ for all $i \in \Lambda\{k\}$ in case $N \geq 3$; for $N = 2$ the functions $h_1(\alpha_2)$ and $h_2(\alpha_1)$ appear to tend to positive constants as their arguments tend to 0.

Lemma 4.1.

For all $k \in I$ and $l \in \Lambda\{k\}$, the function $h_k(\alpha_1, \ldots, \alpha_{k-1}, \alpha_{k+1}, \ldots, \alpha_N)$ is monotonously non-decreasing as a function of $\alpha_l \in (0,1]$, where $\alpha_i \in (0,1]$ for all $i \in \Lambda\{k,l\}$.

Proof.

The lemma is proved by showing that the partial derivative of $h_k(\cdot) = A_1/A_2$ with respect to α_l is larger than or equal to 0. For the derivatives of the variables a_i we find

$$\frac{\partial}{\partial \alpha_l}\{a_1\} = \sum_{\substack{(t_1, \ldots, t_N) \in T \\ t_k = -1,\ t_l = 0}} q_{t_1, \ldots, t_N} \prod_{i \in \Lambda\{k,l\}} \alpha_i^{1-t_i} + 2\alpha_l \sum_{\substack{(t_1, \ldots, t_N) \in T \\ t_k = -1,\ t_l = -1}} q_{t_1, \ldots, t_N} \prod_{i \in \Lambda\{k,l\}} \alpha_i^{1-t_i}$$

$$\geq 0\,,$$

$$\frac{\partial}{\partial \alpha_l}\{a_2\} = - \sum_{\substack{(t_1, \ldots, t_N) \in T \\ t_k = 0,\ t_l = -1}} q_{t_1, \ldots, t_N} \prod_{i \in \Lambda\{k,l\}} \alpha_i^{-t_i} \leq 0\,,$$

$$\frac{\partial}{\partial \alpha_l}\{a_3\} = 0\,.$$

These (in)equalities are used to show that (note that all variables a_i, D and A_i are positive)

$$\frac{\partial}{\partial\alpha_l}\{A_1\}A_2 - \frac{\partial}{\partial\alpha_l}\{A_2\}A_1$$
$$= \left[\frac{\partial}{\partial\alpha_l}\{a_2\} - \frac{1}{2\sqrt{D}}\frac{\partial}{\partial\alpha_l}\{D\}\right](a_2+\sqrt{D}) - \left[\frac{\partial}{\partial\alpha_l}\{a_2\} + \frac{1}{2\sqrt{D}}\frac{\partial}{\partial\alpha_l}\{D\}\right](a_2-\sqrt{D})$$
$$= \frac{1}{\sqrt{D}}\left[2D\frac{\partial}{\partial\alpha_l}\{a_2\} - a_2\frac{\partial}{\partial\alpha_l}\{D\}\right]$$
$$= \frac{1}{\sqrt{D}}\left[2(a_2^2-4a_1a_3)\frac{\partial}{\partial\alpha_l}\{a_2\} - a_2(2a_2\frac{\partial}{\partial\alpha_l}\{a_2\} - 4a_3\frac{\partial}{\partial\alpha_l}\{a_1\})\right]$$
$$= \frac{4a_3}{\sqrt{D}}\left[a_2\frac{\partial}{\partial\alpha_l}\{a_1\} - 2a_1\frac{\partial}{\partial\alpha_l}\{a_2\}\right] \geq 0,$$

by which also

$$\frac{\partial}{\partial\alpha_l}\{h_k(\cdot)\} = \frac{\partial}{\partial\alpha_l}\left\{\frac{A_1}{A_2}\right\} = \left[\frac{\partial}{\partial\alpha_l}\{A_1\}A_2 - \frac{\partial}{\partial\alpha_l}\{A_2\}A_1\right]A_2^{-2} \geq 0. \qquad \square$$

By exploiting the property that for each path $\{v^{(k)}\}$ in V the factors $\alpha_{1,v^{(k)}}, \ldots, \alpha_{N,v^{(k)}}$ are non-increasing for increasing k and the non-decreasing behavior of the functions $h_k(\cdot)$, we now obtain the upper bounds $b_{m_1,\ldots,m_N}(v)$ for the initial upper bounds $\hat{b}_{m_1,\ldots,m_N}(v)$.

Let $v \in V\backslash\{\varnothing\}$ and consider the upper bound $\hat{b}_{m_1,\ldots,m_N}(v)$ and the terms $\Pi_{i=1}^N \alpha_{i,w}^{m_i}$, $w \in S(v)$. For each path $\{w^{(k)}\}$ in $S(v)$ all factors $\alpha_{1,w^{(k)}}, \ldots, \alpha_{N,w^{(k)}}$ are monotonously non-increasing (cf. Lemma 3.6(ii); note that this property also follows from the fact that $h_k(\cdot)$ always takes a value in the interval $(0,1)$). Further, by Lemma 4.1, the functions $h_k(\cdot)$ are monotonously non-decreasing in all arguments. So, for all $w \in S(v)\backslash\{v\}$, we find

$$\frac{\Pi_{i=1}^N \alpha_{i,w}^{m_i}}{\Pi_{i=1}^N \alpha_{i,p(w)}^{m_i}} = [h_k(\alpha_{1,w}, \ldots, \alpha_{k-1,w}, \alpha_{k+1,w}, \ldots, \alpha_{N,w})]^{m_k}$$
$$\leq [h_k(\alpha_{1,v}, \ldots, \alpha_{k-1,v}, \alpha_{k+1,v}, \ldots, \alpha_{N,v})]^{m_k},$$

where $k = w_{l(w)}$. Define the factors x_k by

$$x_k = [h_k(\alpha_{1,v}, \ldots, \alpha_{k-1,v}, \alpha_{k+1,v}, \ldots, \alpha_{N,v})]^{m_k}, \quad k \in I, \tag{4.13}$$

and let the terms y_w, $w \in S(v)$, be defined as follows. Let $y_v := 1$ and let

$$y_w := x_k y_{p(w)} \text{ with } k = w_{l(w)}, \quad w \in S(v)\backslash\{v\}.$$

Then, by using induction with respect to $l(w)$, one may show that

$$\frac{\Pi_{i=1}^N \alpha_{i,w}^{m_i}}{\Pi_{i=1}^N \alpha_{i,v}^{m_i}} \leq y_w \quad \text{for all } w \in S(v).$$

This leads to the following upper bound for $\hat{b}_{m_1,\ldots,m_N}(v)$:

$$\hat{b}_{m_1,\ldots,m_N}(v) \leq \prod_{i=1}^N \alpha_{i,v}^{m_i} \sum_{w \in S(v)\backslash\{v\}} y_w, \quad v \in V\backslash\{\varnothing\}. \tag{4.14}$$

As one can easily see, the values of the terms y_w only depend on the factors x_k. The sum of all y_w further only depends on the value of the last component $\nu_{l(\nu)}$ of the starting point ν of the $(N-1)$-fold subtree $S(\nu)$; the value of $\nu_{l(\nu)}$ determines which factor x_k is not used when computing the terms y_w for the successors $w \in O(\nu)$, and similarly for the successors of the successors $w \in O(\nu)$, and so on. Let the sum of all terms y_w of the subtree $S(\nu)$ be denoted by $G^{(k)}(x_1, \ldots, x_N)$, i.e. let

$$G^{(k)}(x_1, \ldots, x_N) = \sum_{w \in S(\nu)} y_w , \tag{4.15}$$

where $\nu \neq \varnothing$ and $k = \nu_{l(\nu)}$. Then $G^{(k)}(x_1, \ldots, x_N)$ is the sum of the terms y_w over the nodes w of an $(N-1)$-fold tree, where for each path $\{w^{(k)}\}$ in $S(\nu)$ the terms $y_{w^{(k)}}$ have a geometrically decreasing behavior that is determined by the values of the last components of the nodes $w^{(k)}$. This tree is also called a *geometric tree*, and the notation $G^{(k)}(x_1, \ldots, x_N)$ is also used to refer to this tree itself.

The results stated in (4.14) and (4.15) show that upper bounds $b_{m_1, \ldots, m_N}(\nu)$ for the initial upper bounds $\hat{b}_{m_1, \ldots, m_N}(\nu)$ are obtained by defining

$$b_{m_1, \ldots, m_N}(\nu) = \prod_{i=1}^{N} \alpha_{i,\nu}^{m_i} [G^{(\nu_{l(\nu)})}(x_1, \ldots, x_N) - 1] , \quad \nu \in V \backslash \{\varnothing\}, \tag{4.16}$$

where the factors x_k are given by (4.13). It is easily seen that by this definition the bounds $b_{m_1, \ldots, m_N}(\nu)$ satisfy the condition stated in (4.7). The upper bounds $b_{m_1, \ldots, m_N}(\nu)$ are expected to be tight bounds for most nodes at sufficiently large distances from the root $\varnothing$. Since a simple, explicit formula can be derived for the sums $G^{(k)}(x_1, \ldots, x_N)$, as we shall see in the next section (see Theorem 4.1), the bounds $b_{m_1, \ldots, m_N}(\nu)$ are also computable. So, apart from the analysis of the geometric trees, this completes the derivation of appropriate upper bounds for the sums as given on the left-hand side of inequality (4.7), and therefore also our search for appropriate error bounds for the approximation of the equilibrium probabilities $p_{m_1, \ldots, m_N}$ in the convergence region M_c by finite sums $\Sigma_{\nu \in V'} t_{m_1, \ldots, m_N}(\nu)$. In the following lemma, the main results found in this section are recapitulated.

Lemma 4.2.

Let $(m_1, \ldots, m_N) \in M_c$ and let $V' \subset V$ be a finite subset which satisfies the conditions (4.3)-(4.5). Then an upper bound for the absolute error of the approximation of the equilibrium probability $p_{m_1, \ldots, m_N}$ by the finite sum $\Sigma_{\nu \in V'} t_{m_1, \ldots, m_N}(\nu)$ is given by (4.6). The bounds $b_{m_1, \ldots, m_N}(\nu)$, which occur in this upper bound, are defined by (4.16), and the factors x_k and the sums $G^{(\nu_{l(\nu)})}(x_1, \ldots, x_N)$ are given by (4.13) and (4.15).

It is noted that the geometric trees $G^{(k)}(x_1, \ldots, x_N)$ can easily be analyzed for the case $N = 2$. In that case,

$$G^{(1)}(x_1, x_2) = 1 + x_2 + x_1 x_2 + x_1 x_2^2 + \cdots = (1 + x_2) \sum_{i=0}^{\infty} (x_1 x_2)^i ,$$

which shows that $G^{(1)}(x_1, x_2)$ converges if and only if $x_1 x_2 < 1$, and if this condition is satisfied, then

$$G^{(1)}(x_1, x_2) = \frac{1 + x_2}{1 - x_1 x_2} ; \tag{4.17}$$

and similarly for $G^{(2)}(x_1,x_2)$. By using these results, it may be verified that for the two-dimensional case the error bounds as described in this section are equivalent to the error bounds which we have derived in Section 2.5.

4.3. Analysis of the geometric trees

This section is devoted to the *sums/geometric trees* $G^{(k)}(x_1,\ldots,x_N)$ (recall that the notation $G^{(k)}(x_1,\ldots,x_N)$ is used for the sum of the terms y_w over the nodes $w \in S(v)$, see formula (4.15), as well as the corresponding geometric tree itself). We want to derive an explicit formula for the sums $G^{(k)}(x_1,\ldots,x_N)$ in order to complete the derivation of the error bounds given in the previous section. We shall first derive a matrix formula for these sums; see formula (4.21). This formula contains a matrix sum $\Sigma_{d=0}^{\infty} X^d$, where X is a special, nonnegative, squared matrix of order N. By using the Perron-Frobenius Theorem and exploiting the special structure of the matrix X, we will be able to derive three necessary and sufficient conditions for the convergence of a geometric tree, i.e. for the finiteness of the corresponding sum. The last condition is a simple and explicit condition which is appropriate for a quick verification of the convergence of a geometric tree. Finally, by using Cramer's Rule, among others, we obtain an explicit formula for the sum of all terms of a convergent geometric tree.

Before we start with deriving a matrix formula for the sums $G^{(k)}(x_1,\ldots,x_N)$, we first simplify the definition given by (4.15). A sum $G^{(k)}(x_1,\ldots,x_N)$ does not depend on the whole vector v, but only on the value of the last component $v_{l(v)}$, which is denoted by the k. Therefore, instead of the whole vectors w of the subtree $S(v)$, it suffices to use the tails $(w_{l(v)+1},\ldots,w_{l(w)})$ as subindices for the terms y_w. We then obtain the following definition. Let $k \in I$ and let $x_1,\ldots,x_N$ be positive real-valued variables, i.e. $x_i \in (0,\infty)$ for all $i \in I$, then the sum $G^{(k)}(x_1,\ldots,x_N)$ is defined by

$$G^{(k)}(x_1,\ldots,x_N) = \sum_{v \in V_k} y_v ,$$

where

$$V_k = \{ (v_1,\ldots,v_l) \in V \mid \text{if } v \neq \varnothing \text{ then } v_1 \neq k \}$$

and the terms y_v are defined by $y_{\varnothing} := 1$ and

$$y_v := x_i\, y_{p(v)} \quad \text{with} \quad i = v_{l(v)}$$

for all $v \in V_k \backslash \{\varnothing\}$. The vectors $v \in V_k$ are the nodes of an $(N-1)$-fold, geometric tree with corresponding terms y_v, which are also called the *weights*. The notation $G^{(k)}(x_1,\ldots,x_N)$ is also used to refer to this geometric tree itself. In Figure 4.1, we have depicted an example of a geometric tree.

To find an expression for the total weight $G^{(k)}(x_1,\ldots,x_N)$, we define $g_i(d)$ as the sum of the weights y_v over all nodes v with $l(v)=d$ (i.e. at *depth* d) and with $v_{l(v)}=i$:

$$g_i(d) := \sum_{\substack{v \in V_k \\ l(v)=d,\ v_{l(v)}=i}} y_v , \quad d \geq 1,\ i \in I;$$

further, $g_i(0)$ is defined by $g_i(0) := 0$ for all $i \in I \backslash \{k\}$ and $g_k(0) := 1$ (since for the root $v = \varnothing$ the

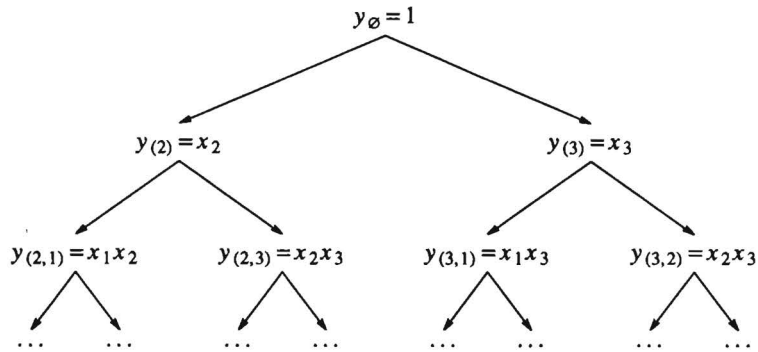


Figure 4.1. The geometric tree $G^{(k)}(x_1, \ldots, x_N)$ of terms y_v for $N=3$ and $k=1$.

k denotes which factor x_k is not used for the computation of the weights for its successors). Due to the relatively simple definition of the terms y_v, $g_i(d)$ can be expressed as a function of the sums $g_{i-1}(d)$ of weights at depth $d-1$:

$$g_i(d) = x_i \sum_{j \in \Lambda\{i\}} g_j(d-1), \quad d \geq 1, \ i \in I.$$

Writing this recurrence relation in matrix notation leads to

$$g(d) = g(d-1)X, \quad d \geq 1, \tag{4.18}$$

where $g(d) = (g_1(d), \ldots, g_N(d))$ for all $d \geq 0$ and the matrix X is defined by

$$X = \begin{pmatrix} 0 & x_2 & x_3 & \cdots & x_{N-1} & x_N \\ x_1 & 0 & x_3 & \cdots & x_{N-1} & x_N \\ x_1 & x_2 & 0 & \cdots & x_{N-1} & x_N \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ x_1 & x_2 & x_3 & \cdots & 0 & x_N \\ x_1 & x_2 & x_3 & \cdots & x_{N-1} & 0 \end{pmatrix}. \tag{4.19}$$

By (4.18), we find

$$g(d) = g(0)X^d = e_k X^d, \tag{4.20}$$

where e_k is the k-th unit vector, i.e. $e_k = (0, \ldots, 0, 1, 0, \ldots, 0)$ with the 1 on the k-th position. The total weight at depth d is given by $g(d)e^T$, where $e = (1, \ldots, 1)$. Finally, by taking the sum of $g(d)e^T$ over all $d \geq 0$ and substituting (4.20), we find the matrix formula

$$G^{(k)}(x_1, \ldots, x_N) = e_k \left[\sum_{d=0}^{\infty} X^d \right] e^T. \tag{4.21}$$

This formula shows that the sum $G^{(k)}(x_1, \ldots, x_N)$ is equal to the sum of the elements in the k-th row of the matrix sum $\Sigma_{d=0}^{\infty} X^d$, and it constitutes the basis of the analysis in the remainder of this section.

Before we derive an explicit, closed-form formula for the sum of all terms of a geometric tree, we first focus on conditions for the convergence of a geometric tree. A simple condition which is sufficient for the convergence, can easily be given; see Remark 4.1 at the end of this section. However, we are interested in conditions which are necessary and sufficient. Three of such conditions are given by Lemma 4.3.

Lemma 4.3.

Let $k \in I$, $x_i \in (0,\infty)$ for all $i \in I$, and let the nonnegative matrix X be defined by (4.19). Then the following four conditions are equivalent:

(i) $G^{(k)}(x_1, \ldots, x_N)$ *is convergent;*

(ii) $\Sigma_{d=0}^{\infty} X^d < \infty$;

(iii) $\rho(X) < 1$;

(iv) $det(I_N - X) > 0$,

where the notations $\rho(\cdot)$ and $det(\cdot)$ are used to denote the spectral radius and the determinant of a matrix. The squared matrix I_N denotes the $N \times N$ unit matrix.

We prove Lemma 4.3 by showing that condition (ii) is equivalent to (i), that (iii) is equivalent to (ii), and that (iv) is equivalent to (iii). The equivalence between (i) and (ii) is shown by using the matrix formula (4.21). The equivalence between (ii) and (iii) is shown for an arbitrary, nonnegative, squared matrix by using the Perron-Frobenius Theorem. To prove the equivalence between (iii) and (iv), we among others derive an explicit formula for the characteristic polynomial $\det(\lambda I_N - X)$. Due to this explicit formula, condition (iv) is appropriate for quickly verifying whether a geometric tree converges, or not.

It is easily verified that for $N=2$ the convergence conditions (ii)-(iv) of Lemma 4.3 reduce to the condition $x_1 x_2 < 1$, which we derived at the end of the previous section. Further, it is noted that, after the proof of Lemma 4.3, for convergent geometric trees the matrix formula (4.21) can be transformed into the explicit, closed-form formula as given in Theorem 4.1 by substituting the formula $\Sigma_{d=0}^{\infty} X^d = (I_N - X)^{-1}$, which holds if condition (ii) is satisfied, into the formula (4.21) and next explicitly determining the inverse $(I_N - X)^{-1}$.

Proof of the equivalence between the conditions (i) and (ii) of Lemma 4.3

By the matrix formula (4.21), a geometric tree $G^{(k)}(x_1, \ldots, x_N)$ converges if and only if the elements in the k-th row of the matrix sum $\Sigma_{d=0}^{\infty} X^d$ are finite. From this property it immediately follows that condition (ii), which states that all elements of the matrix sum $\Sigma_{d=0}^{\infty} X^d$ are finite, is sufficient for the convergence of a geometric tree $G^{(k)}(x_1, \ldots, x_N)$. That condition (ii) is also necessary is shown by also using the recurrence relation

$$G^{(k)}(x_1, \ldots, x_N) = 1 + \sum_{i \in I \setminus \{k\}} x_i \, G^{(i)}(x_1, \ldots, x_N), \tag{4.22}$$

which follows from the property that $G^{(k)}(x_1, \ldots, x_N)$ is equal to the sum of the contribution $y_v = 1$ of the root $v = \varnothing$ and the contributions of the $N-1$ subtrees starting at the nodes $v \in V_k$ with $l(v) = 1$. By combining the matrix formula (4.21) and the recurrence relation (4.22), we obtain the property that if a geometric tree $G^{(k)}(x_1, \ldots, x_N)$ converges, then not only the elements in the k-th row of the matrix sum $\Sigma_{d=0}^{\infty} X^d$ are finite, but also the elements in the other rows with indices $i \in I \setminus \{k\}$ are finite (otherwise it would hold that $G^{(i)}(x_1, \ldots, x_N) = \infty$ for

one of the indices $i \in I\backslash\{k\}$). This completes the proof of the equivalence between the conditions (i) and (ii) of Lemma 4.3.

Proof of the equivalence between the conditions (ii) and (iii) of Lemma 4.3

The equivalence between the conditions (ii) and (iii) is proved for an arbitrary, nonnegative, $N \times N$ matrix A by using the *Perron-Frobenius Theorem* for so-called irreducible, nonnegative matrices (see Seneta [64], Theorem 1.5). Suppose that A is irreducible (for completeness, we note that our matrix X has this property). Then, by the Perron-Frobenius Theorem, the spectral radius $\rho(A)$ itself is a positive, real-valued eigenvalue of A, and the left and right eigenvectors associated with the eigenvalue $\rho(A)$ are strictly positive. Let $\hat{y}$ be a strictly positive left eigenvector associated with $\lambda = \rho(A)$. Then

$$\hat{y}\left[\sum_{d=0}^{\infty} A^d\right] = \sum_{d=0}^{\infty} \hat{y} A^d = \sum_{d=0}^{\infty} \lambda^d \hat{y} = \left[\sum_{d=0}^{\infty} \lambda^d\right] \hat{y} ,$$

and thus, $\Sigma_{d=0}^{\infty} A^d < \infty$, if and only if $\Sigma_{d=0}^{\infty} \lambda^d < \infty$, i.e. if and only if $\lambda = \rho(A) < 1$. This completes the proof for an irreducible matrix A. For a matrix A which is not irreducible, the proof that $\Sigma_{d=0}^{\infty} A^d < \infty$ if and only if $\rho(A) < 1$, is given as follows: the 'only if'-part follows from the Perron-Frobenius Theorem for an arbitrary nonnegative matrix (see [64], Exercise 1.12); the 'if'-part may easily be shown by exploiting that $\rho(A) < 1$ implies that there exists a slightly modified matrix $\hat{A}$ such that $\hat{A}$ is irreducible, $A \leq \hat{A}$ and $\rho(\hat{A}) < 1$.

Proof of the equivalence between the conditions (iii) and (iv) of Lemma 4.3

By using two elementary properties, we can prove for an arbitrary $N \times N$ matrix A with real-valued elements that

$$\rho(A) < 1 \Rightarrow \det(I_N - A) > 0 . \tag{4.23}$$

The first property we need tells that all eigenvalues of a squared $N \times N$ matrix A are zero points of the characteristic equation $\det(\lambda I_N - A) = 0$, and vice versa. The second property we need is the property that the coefficient of λ^N of the characteristic polynomial $\det(\lambda I_N - A)$, which is a polynomial in λ of degree N, is equal to 1. This second property implies that $\det(\lambda I_N - A) \to \infty$, as $\lambda \to \infty$. So, if $\det(\lambda I_N - A) \leq 0$ for $\lambda = 1$, then the equation $\det(\lambda I_N - A) = 0$ has a real-valued root in the interval $[1, \infty)$, and, by the first property, we find that $\rho(A) \geq 1$. This proves (4.23), and it implies that condition (iv) of Lemma 4.3 is necessary for having a spectral radius $\rho(X) < 1$ for a matrix X defined by (4.19). To prove that condition (iv) is also sufficient for this, we shall first derive an explicit expression for the determinant $\det(\lambda I_N - X)$ (see Lemma 4.4), and subsequently, by using Rouché's Theorem, it is shown that all roots of the characteristic equation $\det(\lambda I_N - X) = 0$ are lying inside the unit disk if $\det(\lambda I_N - X) > 0$ for $\lambda = 1$ (it is noted that an alternative proof may be given by using the explicit expression for $\det(\lambda I_N - X)$ with $\lambda = 1$ and applying Theorem 2.2 of [64]).

Lemma 4.4.

Let the variables x_i be arbitrary (possibly complex) variables for all $i \in I$, and let the matrix X be defined by (4.19). Then

$$det(\lambda I_N - X) = \lambda^N - \sum_{i=2}^{N} (i-1) \lambda^{N-i} \sum_{\substack{J \subset I \\ |J| = i}} \prod_{j \in J} x_j , \quad \lambda \in \mathbb{C}. \tag{4.24}$$

Proof.

Formula (4.24) is derived by using the definition of the determinant. In this proof the matrix $\lambda I_N - X$ is denoted by $A = (a_{i,j})$. By definition, we have

$$\det(A) = \sum_{\sigma \in S_N} s(\sigma) \prod_{i \in I} a_{i,\sigma(i)} , \tag{4.25}$$

where S_N denotes the set of all permutations of $I = \{1, \ldots, N\}$, and $s(\sigma)$ denotes a sign function of S_N, i.e. $s: S_N \to \{-1, 1\}$. A permutation $\sigma \in S_N$ may be denoted by a vector $(\sigma(1), \ldots, \sigma(N))$, where $\sigma(i)$ denotes the image of i. Here, the elements $\sigma(1), \ldots, \sigma(N)$ have to constitute a sequence of N distinct numbers out of the set $I = \{1, \ldots, N\}$. The permutation which depicts each element of $\{1, \ldots, N\}$ to itself is denoted by $(1, \ldots, N)$ and it is known that each permutation $(\sigma(1), \ldots, \sigma(N)) \in S_N$ may be obtained by repeatedly interchanging two elements of the sequence $1, \ldots, N$. Therefore we can divide the set S_N into two disjunct subsets: the set E_N of even permutations, which may be obtained out of $(1, \ldots, N)$ by performing an even number of interchanges of two elements, and the set F_N of odd permutations, which are obtained by an odd number of interchanges. The sign function $s(\sigma)$ is defined as follows:

$$s(\sigma) = \begin{cases} 1 & \text{if } \sigma \in E_N \text{ (i.e. } \sigma \text{ is even);} \\ -1 & \text{if } \sigma \in F_N \text{ (i.e. } \sigma \text{ is odd).} \end{cases}$$

Let us consider the contribution of $s(\sigma) \prod_{i \in I} a_{i,\sigma(i)}$ of an arbitrary permutation $\sigma \in S_N$. By the definition of the matrices X and A, we find that $a_{i,\sigma(i)} = \lambda$ if $\sigma(i) = i$ and $a_{i,\sigma(i)} = -x_{\sigma(i)}$ if $\sigma(i) \neq i$. Further, the value of $\prod_{i \in I} a_{i,\sigma(i)}$ appears to be the same for all permutations for which the same elements of $\{1, \ldots, N\}$ are depicted to themselves. Let for all $J \subset I$, $S_N(J) \subset S_N$ denote the set of permutations which depict all elements $i \notin J$ to themselves and all elements $i \in J$ to other elements, then formula (4.25) may be rewritten to

$$\begin{aligned} \det(A) &= \sum_{J \subset I} \sum_{\sigma \in S_N(J)} s(\sigma) \lambda^{N-|J|} \prod_{j \in J} (-x_{\sigma(j)}) = \sum_{J \subset I} \sum_{\sigma \in S_N(J)} s(\sigma) \lambda^{N-|J|} \prod_{j \in J} (-x_j) \\ &= \sum_{i=0}^{N} \lambda^{N-i} (-1)^i \sum_{\substack{J \subset I \\ |J| = i}} \left[\sum_{\sigma \in S_N(J)} s(\sigma) \right] \prod_{j \in J} x_j . \end{aligned}$$

Let S_i', $i \geq 1$, denote the set of permutations of $\{1, \ldots, i\}$ for which no element is depicted to itself, then for each J with $|J| = i$ each permutation $\sigma \in S_N(J)$ is equivalent to a permutation $\sigma' \in S_i'$ and vice versa. Further, it is obvious that the permutation $(1, \ldots, N)$ is the only element belonging to $S_N(\varnothing)$, and we find

$$\begin{aligned} \det(A) &= \lambda^N + \sum_{i=1}^{N} \lambda^{N-i} (-1)^i \sum_{\substack{J \subset I \\ |J| = i}} \left[\sum_{\sigma' \in S_i'} s(\sigma') \right] \prod_{j \in J} x_j \\ &= \lambda^N + \sum_{i=1}^{N} \lambda^{N-i} (-1)^i (|E_i'| - |F_i'|) \sum_{\substack{J \subset I \\ |J| = i}} \prod_{j \in J} x_j , \end{aligned} \tag{4.26}$$

where $E_i' \subset S_i'$ denotes the set of even permutations of $\{1, \ldots, i\}$ which depict no element to itself, and $F_i' \subset S_i'$ denotes the corresponding subset of odd permutations. Now, the only problem left is counting the elements of E_i' and F_i' for $i = 1, \ldots, N$.

For small i, $|E_i'|$ and $|F_i'|$ are easily verified by checking the whole set S_i of permutations of $\{1,\ldots,i\}$. For $i=1$ and $i=2$, we find

$$S_1 = \{(1)\}\,, \quad |E_1'| = 0\,, \quad |F_1'| = 0\,;$$

$$S_2 = \{(1,2),(2,1)\}\,, \quad |E_2'| = 0\,, \quad |F_2'| = 1\,.$$

For larger i we can derive simple recurrence relations for $|E_i'|$ and $|F_i'|$. Let $i \geq 3$ and let us consider the permutations $\sigma \in E_i'$. Each of these permutations satisfies exactly one of the following two properties:

(i) $\sigma(1) = j$, where $j \in \{2,\ldots,i\}$, and $\sigma(j) = 1$;

(ii) $\sigma(1) = j$, where $j \in \{2,\ldots,i\}$, and $\sigma(j) \neq 1$.

If $\sigma \in E_i'$ satisfies (i), then $(\sigma(2),\ldots,\sigma(j-1),\sigma(j+1),\ldots,\sigma(i))$ is an odd permutation of $\{2,\ldots,j-1,j+1,\ldots,i\}$ which depicts no element to itself. As a consequence, for each j there are $|F_{i-2}'|$ permutations $\sigma \in E_i'$ satisfying (i); and, the total number of permutations $\sigma \in E_i'$ satisfying (i) is equal to $(i-1)|F_{i-2}'|$, since there are $i-1$ values possible for j. If a permutation $\sigma \in E_i'$ satisfies (ii), then $\sigma(2),\ldots,\sigma(i)$ is a sequence of distinct numbers of $\{1,\ldots,i\}\backslash\{j\}$. Replacing the 1 in this sequence by j gives a sequence $\sigma'(2),\ldots,\sigma'(i)$ of distinct numbers of $\{2,\ldots,i\}$, and, if $\sigma \in E_i'$, then $(\sigma'(2),\ldots,\sigma'(i))$ is an odd permutation of $\{2,\ldots,i\}$ which depicts no element to itself. Therefore, the total number of permutations $\sigma \in E_i'$ satisfying (ii) is equal to $(i-1)|F_{i-1}'|$. Adding this number to the total number of permutations satisfying (i) shows that

$$|E_i'| = (i-1)(|F_{i-1}'| + |F_{i-2}'|)\,, \quad i \geq 3.$$

Similarly, one may show that

$$|F_i'| = (i-1)(|E_{i-1}'| + |E_{i-2}'|)\,, \quad i \geq 3.$$

The recurrence relations for $|E_i'|$ and $|F_i'|$ are used to prove by induction with respect to i that

$$|E_i'| - |F_i'| = (-1)^{i+1}(i-1)\,, \quad i \geq 1. \tag{4.27}$$

It is easily verified that (4.27) holds for $i=1$ and $i=2$. Now let $i \geq 3$ and suppose that (4.27) holds for $1,\ldots,i-1$, then

$$\begin{aligned} |E_i'| - |F_i'| &= (i-1)(|F_{i-1}'| + |F_{i-2}'|) - (i-1)(|E_{i-1}'| + |E_{i-2}'|) \\ &= -(i-1)[(|E_{i-1}'| - |F_{i-1}'|) + (|E_{i-2}'| - |F_{i-2}'|)] \\ &= -(i-1)[(-1)^i(i-2) + (-1)^{i-1}(i-3)] \\ &= (-1)^{i+1}(i-1)\,, \end{aligned}$$

by which (4.27) also holds for i.

Finally, substituting (4.27) into (4.26) completes the proof. □

Let $x_i \in (0,\infty)$ for all $i \in I$ and let the matrix X be defined by (4.19). Further, assume that $\det(I_N - X) > 0$, which means that the characteristic polynomial $\hat{h}(\lambda) := \det(\lambda I_N - X)$ is positive for $\lambda = 1$. By using formula (4.24) and Rouché's Theorem, we can show that the assumption $\hat{h}(1) > 0$ implies that all (possibly complex) zero points of the characteristic polynomial $\hat{h}(\lambda)$ are smaller than 1 in modulus, i.e. that $\rho(X) < 1$. Define $\hat{g}(\lambda) := \lambda^N$ for all $\lambda \in \mathbb{C}$ and let $\hat{f}(\lambda)$

be equal to the remaining part of $\hat{h}(\lambda)$, i.e. (see (4.24))

$$\hat{f}(\lambda) := -\sum_{i=2}^{N} (i-1)\lambda^{N-i} \sum_{\substack{J\subset I \\ |J|=i}} \prod_{j\in J} x_j .$$

Then for all $\lambda \in \mathbb{C}$ with $|\lambda| = 1$, we have

$$|\hat{f}(\lambda)| \leq \sum_{i=2}^{N} (i-1)\,|\lambda|^{N-i} \sum_{\substack{J\subset I \\ |J|=i}} \prod_{j\in J} x_j = \sum_{i=2}^{N} (i-1) \sum_{\substack{J\subset I \\ |J|=i}} \prod_{j\in J} x_j = 1-\hat{h}(1) < 1 ,$$

$$|\hat{g}(\lambda)| = 1 .$$

So, $|\hat{f}(\lambda)| < |\hat{g}(\lambda)|$ for all λ with $|\lambda| = 1$, and therefore, by *Rouché's Theorem* (see Titchmarsh [71]), $\hat{h}(\lambda) = \hat{f}(\lambda) + \hat{g}(\lambda)$ has the same number of zeros inside the unit disk as $\hat{g}(\lambda)$, where all zeros have to be counted according to their multiplicity. Since $\hat{g}(\lambda)$ has N zeros inside the unit disk ($\lambda = 0$ with multiplicity N), this means that also all zeros of $\hat{h}(\lambda)$ are lying inside the unit disk. This completes the proof of the equivalence between the conditions (iii) and (iv) of Lemma 4.3 and of Lemma 4.3 itself.

By condition (iv) of Lemma 4.3 and the explicit formula (4.24) for $\det(\lambda I_N - X)$, we find the following simple, necessary and sufficient condition for the convergence of a geometric tree $G^{(k)}(x_1, \ldots, x_N)$: the expression on the right-hand side of equation (4.24) must be positive for $\lambda = 1$; see the first part of Theorem 4.1. Further, if it is known that a geometric tree converges, then, by condition (ii) of Lemma 4.3, $\Sigma_{d=0}^{\infty} X^d = (I_N - X)^{-1}$ (the sequence $\{\Sigma_{d=0}^{l} X^d\}_{l\geq 0}$ is monotonously non-decreasing and, by condition (ii) of Lemma 4.3, also bounded; therefore the sequence $\{\Sigma_{d=0}^{l} X^d\}_{l\geq 0}$ converges, and its limit may be shown to be equal to $(I_N - X)^{-1}$), and substituting this result into the matrix formula (4.21) leads to

$$G^{(k)}(x_1, \ldots, x_N) = e_k (I_N - X)^{-1} e^T . \tag{4.28}$$

This formula shows that for a convergent geometric tree, the sum $G^{(k)}(x_1, \ldots, x_N)$ of all terms y_v is equal to the k-th row sum of the inverse $(I_N - X)^{-1}$. An explicit, closed-form formula for the sum $G^{(k)}(x_1, \ldots, x_N)$ is obtained by determining first the inverse $(I_N - X)^{-1}$ itself (by applying Cramer' rule and using Lemma 4.4) and subsequently its k-th row sum. This is shown in the proof of Theorem 4.1, in which the main results for the geometric trees are summarized. It is noted that formula (4.30) for the case $N = 2$ and $k = 1$ is equivalent to formula (4.17), which we found at the end of the previous section.

Theorem 4.1

Let $k \in I$ and $x_i \in (0, \infty)$ for all $i \in I$. Then the geometric tree $G^{(k)}(x_1, \ldots, x_N)$ converges if and only if

$$D(x_1, \ldots, x_N) = 1 - \sum_{i=2}^{N} (i-1) \sum_{\substack{J\subset I \\ |J|=i}} \prod_{j\in J} x_j > 0 . \tag{4.29}$$

If this condition is satisfied, then

$$G^{(k)}(x_1, \ldots, x_N) = \frac{1}{D(x_1, \ldots, x_N)} \prod_{i\in I\setminus\{k\}} (1+x_i) . \tag{4.30}$$

Proof.

The first part follows from the reasoning given in the paragraph prior to the theorem.

For the second part of the theorem we first have to derive the inverse of $I_N - X$. Denote $I_N - X$ by A and $(I_N - X)^{-1}$ by $B = (b_{i,j})$. Further, assume that condition (4.29) is satisfied. So, $\det(A) = D(x_1, \ldots, x_N) \neq 0$, and therefore B is the unique solution of $BA = I_N$. As a consequence, the i-th row b_i of B is the unique solution of $b_i A = e_i$, where e_i is the i-th unit row vector, and, by using *Cramer's Rule*, we find that

$$b_{i,j} = \frac{(-1)^{i+j} \det(A_{j,i})}{\det(A)} = \frac{(-1)^{i+j} \det(A_{j,i})}{D(x_1, \ldots, x_N)}, \quad i,j \in I, \tag{4.31}$$

where $A_{j,i}$ is the (j,i)-th minor matrix of A. For all i and j the matrix $A_{j,i}$ is a squared matrix of order $N-1$ which is obtained by deleting the j-th row and the i-th column of A.

Up to multiplying factors, the determinants $\det(A_{j,i})$ appear to be equal to determinants of matrices of the type $I_{N-1} - X'$, where the squared matrix X' is of the same type as the matrix X given by (4.19), but of order $N-1$ and with factors $x'_1, \ldots, x'_{N-1}$:

$$X' = \begin{pmatrix} 0 & x'_2 & \cdots & x'_{N-2} & x'_{N-1} \\ x'_1 & 0 & \cdots & x'_{N-2} & x'_{N-1} \\ \vdots & \vdots & & \vdots & \vdots \\ x'_1 & x'_2 & \cdots & 0 & x'_{N-1} \\ x'_1 & x'_2 & \cdots & x'_{N-2} & 0 \end{pmatrix}.$$

Define $D'(x'_1, \ldots, x'_{N-1}) := \det(I_{N-1} - X')$, then, by applying Lemma 4.4 for a squared matrix of order $N-1$, we obtain

$$D'(x'_1, \ldots, x'_{N-1}) = 1 - \sum_{i=2}^{N-1} (i-1) \sum_{\substack{J \subset I' \\ |J| = i}} \prod_{j \in J} x_j ,$$

where $I' := \{1, \ldots, N-1\}$ (note that, if $N=2$, then $X' = (0)$ and $D'(x'_1) = 1$). This explicit formula leads to explicit formulae for the determinants $\det(A_{i,j})$.

Let $i,j \in I$. If $j = i$, then it is easily verified that

$$\det(A_{j,i}) = \det(A_{i,i}) = D'(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N). \tag{4.32}$$

If $j \neq i$, then, by dividing the elements of the j-th column by $-x_j$ and rearranging columns, one may show that

$$\det(A_{j,i}) = \begin{cases} -x_j (-1)^{i-j-1} D'(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{i-1}, -1, x_{i+1}, \ldots, x_N) & \text{if } j < i; \\ -x_j (-1)^{i-j-1} D'(x_1, \ldots, x_{i-1}, -1, x_{i+1}, \ldots, x_{j-1}, x_{j+1}, \ldots, x_N) & \text{if } j > i. \end{cases}$$

For arbitrary factors $x'_1, \ldots, x'_{N-1}$ with $x'_m = -1$ for some $m \in I'$, we find

$$\begin{aligned} D'(x'_1, \ldots, x'_{N-1}) &= 1 - \sum_{n=2}^{N-1} (n-1) \left(\sum_{\substack{J \subset I' \\ |J| = n,\ m \notin J}} \prod_{l \in J} x'_l + \sum_{\substack{J \subset I' \\ |J| = n,\ m \in J}} \prod_{l \in J} x'_l \right) \\ &= 1 - \sum_{n=2}^{N-2} (n-1) \sum_{\substack{J \subset I' \setminus \{m\} \\ |J| = n}} \prod_{l \in J} x'_l - x'_m \sum_{n=2}^{N-1} (n-1) \sum_{\substack{J \subset I' \setminus \{m\} \\ |J| = n-1}} \prod_{l \in J} x'_l \end{aligned}$$

$$= 1 - \sum_{n=1}^{N-2}(n-1) \sum_{\substack{J\subset I'\setminus\{m\} \\ |J|=n}} \prod_{l\in J} x_l' + \sum_{n=1}^{N-2} n \sum_{\substack{J\subset I'\setminus\{m\} \\ |J|=n}} \prod_{l\in J} x_l'$$

$$= \sum_{n=0}^{N-2} \sum_{\substack{J\subset I'\setminus\{m\} \\ |J|=n}} \prod_{l\in J} x_l'$$

$$= \prod_{l\in I'\setminus\{m\}} (1+x_l'),$$

by which

$$\det(A_{j,i}) = (-1)^{i-j} x_j \prod_{l\in\Lambda\setminus\{i,j\}} (1+x_l), \quad j\neq i. \tag{4.33}$$

Substituting (4.32) and (4.33) into (4.31) completes the proof of the following formula for the elements $b_{i,j}$ of the inverse $B=(I_N-X)^{-1}$:

$$b_{i,j} = \begin{cases} \dfrac{D'(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_N)}{D(x_1,\ldots,x_N)} & \text{if } j=i; \\ \dfrac{1}{D(x_1,\ldots,x_N)}\, x_j \displaystyle\prod_{l\in\Lambda\setminus\{i,j\}} (1+x_l) & \text{if } j\neq i. \end{cases}$$

To find the explicit formula (4.30) for $G^{(k)}(x_1,\ldots,x_N)$, by (4.28), we have to compute the k-th row sum of B. We find

$$G^{(k)}(x_1,\ldots,x_N) = \sum_{j=1}^{N} b_{k,j}$$

$$= \frac{1}{D(x_1,\ldots,x_N)} \left[D'(x_1,\ldots,x_{k-1},x_{k+1},\ldots,x_N) + \sum_{j\in\Lambda\setminus\{k\}} x_j \prod_{l\in\Lambda\setminus\{k,j\}} (1+x_l) \right]. \tag{4.34}$$

The second term of the part in parentheses may be rewritten to

$$\begin{aligned} \sum_{j\in\Lambda\setminus\{k\}} x_j \prod_{l\in\Lambda\setminus\{k,j\}} (1+x_l) &= \sum_{j\in\Lambda\setminus\{k\}} \prod_{l\in\Lambda\setminus\{k\}} (1+x_l) - \sum_{j\in\Lambda\setminus\{k\}} \prod_{l\in\Lambda\setminus\{k,j\}} (1+x_l) \\ &= (N-1) \sum_{i=0}^{N-1} \sum_{\substack{J\subset\Lambda\setminus\{k\} \\ |J|=i}} \prod_{j\in J} x_j - \sum_{i=0}^{N-2} (N-i-1) \sum_{\substack{J\subset\Lambda\setminus\{k\} \\ |J|=i}} \prod_{j\in J} x_j \\ &= \sum_{i=1}^{N-1} i \sum_{\substack{J\subset\Lambda\setminus\{k\} \\ |J|=i}} \prod_{j\in J} x_j, \end{aligned} \tag{4.35}$$

where the second equality follows from

$$\sum_{j\in\Lambda\setminus\{k\}} \prod_{l\in\Lambda\setminus\{k,j\}} (1+x_l) = \sum_{i=0}^{N-2} (N-i-1) \sum_{\substack{J\subset\Lambda\setminus\{k\} \\ |J|=i}} \prod_{j\in J} x_j,$$

which is easily verified by working out the expression on the left-hand side and counting for each $J\subset\Lambda\setminus\{k\}$ the number of times that the term $\Pi_{j\in J} x_j$ occurs. Subsequently, by using the formula for $D'(x_1',\ldots,x_{N-1}')$ and (4.35), it is shown that

$$
\begin{aligned}
&D'(x_1,\ldots,x_{k-1},x_{k+1},\ldots,x_N) + \sum_{j\in\Lambda\{k\}} x_j \prod_{l\in\Lambda\{k,j\}} (1+x_l) \\
&\quad = 1 - \sum_{i=2}^{N-1} (i-1) \sum_{\substack{J\subset\Lambda\{k\} \\ |J|=i}} \prod_{j\in J} x_j + \sum_{i=1}^{N-1} i \sum_{\substack{J\subset\Lambda\{k\} \\ |J|=i}} \prod_{j\in J} x_j \\
&\quad = \sum_{i=0}^{N-1} \sum_{\substack{J\subset\Lambda\{k\} \\ |J|=i}} \prod_{j\in J} x_j \\
&\quad = \prod_{j\in\Lambda\{k\}} (1+x_j) .
\end{aligned}
$$

Finally, substituting this result into (4.34) completes the proof of the second part of Theorem 4.1. □

Remark 4.1. *(a sufficient condition for the convergence of a geometric tree)*

There is one situation for which one can immediately see that a geometric tree $G^{(k)}(x_1,\ldots,x_N)$ converges. It directly follows from the definition that $G^{(k)}(x_1,\ldots,x_N)$ converges if

$$\sum_{i\in\Lambda\{k\}} x_i < 1 \quad \text{for all } k\in I, \tag{4.36}$$

i.e. if for each node v the sum of the weights of its successors is smaller than the weight of v itself; in that case the total weight is bounded by

$$G^{(k)}(x_1,\ldots,x_N) = \sum_{d=0}^{\infty} g(d)e^T \le \sum_{d=0}^{\infty} \left[\max_{k\in I} \sum_{i\in\Lambda\{k\}} x_i \right]^d < \infty .$$

It is noted that, since according to a corollary of the Perron-Frobenius Theorem (see Theorem 1.5 and Corollary 1 of Theorem 1.1 of [64]), the spectral radius of an irreducible, nonnegative matrix is bounded by its maximal row sum (and also by its maximal column sum), condition (4.36) implies that $\rho(X) < 1$. This shows that condition (4.36) is a stronger version of condition (iii) of Lemma 4.3.

Remark 4.2. *(geometric trees with arbitrary real-valued factors x_i)*

The whole analysis in this section for the geometric trees $G^{(k)}(x_1,\ldots,x_N)$ has been restricted to the case in which all real-valued factors x_i are positive. The reason is that this case suffices for obtaining error bounds for the approximation of equilibrium probabilities by finite sums; see Lemma 4.2. However, the geometric trees $G^{(k)}(x_1,\ldots,x_N)$ can also be defined for the case with arbitrary real-valued factors x_i. Cases with zero factors x_i may be reduced to 'lower-dimensional' cases. For the case with $x_i \in \mathbb{R}\backslash\{0\}$ for all $i\in I$, the main results derived in this section may be generalized as follows. For the sum $G^{(k)}(x_1,\ldots,x_N)$ of all terms y_v we still obtain the matrix formula (4.21). For the absolute sum of all terms y_v, we obtain the expression

$$\sum_{v\in V_k} |y_v| = e_k \left[\sum_{d=0}^{\infty} |X|^d \right] e^T ,$$

where $|X|$ is identical to the matrix X given by (4.19), but with the factors $x_1,\ldots,x_N$

replaced by their absolute values $|x_1|, \ldots, |x_N|$. By applying Lemma 4.3, we find that the following four conditions are equivalent: (i) $G^{(k)}(x_1, \ldots, x_N)$ is absolutely convergent; (ii) $\Sigma_{d=0}^{\infty} |X|^d < \infty$; (iii) $\rho(|X|) < 1$; (iv) $\det(I_N - |X|) > 0$. The fourth condition can easily be verified by using the formula given in Lemma 4.4. The third condition implies that $\rho(X) < 1$. This follows from the property that for each matrix A the spectral radius $\rho(A)$ is smaller than or equal to the spectral radius $\rho(|A|)$ of the matrix $|A|$ obtained from A by taking the absolute value for each element (this property may be proved by using the formula

$$\rho(A) = \lim_{n \to \infty} \|A^n\|^{1/n},$$

as given by Dunford and Schwartz [27], p. 567, and showing that $\|A^n\| \leq \||A|^n\|$ for all n). Next, the property $\rho(X) < 1$ implies that the inverse $(I_N - X)^{-1}$ exists and is equal to $\Sigma_{d=0}^{\infty} X^d$ (use Jordan's canonical form to show that $X^d \to 0$ as $d \to \infty$, and apply Lemma B.1 of Seneta [64]). As a consequence, we find that the formulae (4.28) and (4.30) also hold for the case with $x_i \in \mathbb{R} \backslash \{0\}$ for all $i \in I$. So, we obtain the following generalization of Theorem 4.1 (it is noted that although the cases with factors $x_i = 0$ reduce to 'lower-dimensional' cases, the same main results are still valid for these cases):

Let $k \in I$, $x_i \in \mathbb{R}$ for all $i \in I$ and let the function $D(x_1, \ldots, x_N)$ be defined by (4.29). Then the geometric tree $G^{(k)}(x_1, \ldots, x_N)$ converges absolutely if and only if $D(|x_1|, \ldots, |x_N|) > 0$. And, if this condition is satisfied, then for the sum $G^{(k)}(x_1, \ldots, x_N)$ the formula stated in (4.30) is obtained.

4.4. Three procedures for the computation of the equilibrium distribution

The explicit results which we found in the previous section for the sums/geometric trees $G^{(k)}(x_1, \ldots, x_N)$ (see Theorem 4.1), enable us to use the error bounds derived in Section 4.2 (see Lemma 4.2) in numerical procedures for the computation of an equilibrium probability $p_{m_1, \ldots, m_N}$ for a state $(m_1, \ldots, m_N)$ in the convergence region M_c (note that for the states outside this region, the equilibrium probabilities can be computed from appropriately chosen equilibrium equations). In this section, we will describe three numerical procedures for the computation of an equilibrium probability within an arbitrary desired accuracy. In one of these procedures empirical upper bounds for the errors of the approximations by finite sums are used, whereas in the other two procedures the hard upper bounds of Section 4.2 are exploited to decide when the computation process can be stopped. The procedure with the empirical upper bounds is a simple procedure which serves as a reference for the evaluation of the quality of the other two procedures.

Let $(m_1, \ldots, m_N) \in M_c$ and suppose that the equilibrium probability $p_{m_1, \ldots, m_N}$ has to be computed within a given absolute or relative accuracy. The probability $p_{m_1, \ldots, m_N}$ is equal to the infinite sum on the right-hand side of formula (4.2), and appropriate approximations for $p_{m_1, \ldots, m_N}$ are given by finite sums $\Sigma_{v \in V'} t_{m_1, \ldots, m_N}(v)$, where V' is a finite subset of V and satisfies the conditions (4.3)-(4.5). Since it is expected that, in general, the contribution of the terms $t_{m_1, \ldots, m_N}(v)$ is the largest for the nodes $v \in V$ which are close to the root $\varnothing$, it seems to be appropriate to take V' equal to the subsets

$$V(d) = \{ v \in V \mid l(v) \leq d \}, \quad d \in \mathbb{N}.$$

A node $v \in V$ is said to be lying at a *depth* $l(v)$ in the tree of all nodes $v \in V$, and a subset $V(d)$ consists of all nodes at the depths $0, \ldots, d$. Let the corresponding approximations for $p_{m_1,\ldots,m_N}$ be denoted by

$$p_{m_1,\ldots,m_N}(d) = \sum_{v \in V(d)} t_{m_1,\ldots,m_N}(v). \tag{4.37}$$

It is obvious that the approximations $\{p_{m_1,\ldots,m_N}(d)\}_{d \in \mathbb{N}}$ constitute a sequence which converges to $p_{m_1,\ldots,m_N}$ as d tends to infinity. Therefore the probability $p_{m_1,\ldots,m_N}$ may be determined by successively computing $p_{m_1,\ldots,m_N}(d)$ for $d = 1,2,\cdots$. This computing process should be stopped as soon as for some d the relative or absolute error of $p_{m_1,\ldots,m_N}(d)$ with respect to $p_{m_1,\ldots,m_N}$ is sufficiently small.

Whether the error of an approximation $p_{m_1,\ldots,m_N}(d)$ is sufficiently small can be established by using the error bounds derived in Section 4.2. Applying Lemma 4.2 shows that

$$|p_{m_1,\ldots,m_N} - p_{m_1,\ldots,m_N}(d)| \leq \sum_{\substack{v \in V \\ l(v)=d}} b_{m_1,\ldots,m_N}(v), \tag{4.38}$$

where the bounds $b_{m_1,\ldots,m_N}(v)$ are defined by (4.16); here, a bound $b_{m_1,\ldots,m_N}(v)$ is said to be equal to ∞ in case the geometric tree $G^{(v_{l(v)})}(x_1,\ldots,x_N)$ on the right-hand side of (4.16) is not convergent. Note that from this upper bound for the absolute error of $p_{m_1,\ldots,m_N}(d)$, one can also obtain an upper bound for the relative error. The upper bounds given by (4.38) are used in a procedure which is called the *simple procedure with hard bounds*. In this procedure, the computation of the finite sums $p_{m_1,\ldots,m_N}(d)$ is stopped as soon as it can be guaranteed on the basis of the hard upper bounds given by (4.38) that $p_{m_1,\ldots,m_N}(d)$ approximates $p_{m_1,\ldots,m_N}$ within the given accuracy.

Instead of using the hard upper bounds given by (4.38), we can also use empirical upper bounds, which are obtained as follows. The approximations $p_{m_1,\ldots,m_N}(d)$ are the partial sums of the convergent and alternating series $\Sigma_{k=0}^{\infty} a_{m_1,\ldots,m_N}(k)$, with

$$a_{m_1,\ldots,m_N}(k) = \sum_{\substack{v \in V \\ l(v)=k}} t_{m_1,\ldots,m_N}(v), \quad k \geq 0$$

(that this series is alternating follows from the property that for all v all factors $\alpha_{1,v},\ldots,\alpha_{N,v}$ are in the interval $(0,1]$). The convergence of the series $\Sigma_{k=0}^{\infty} a_{m_1,\ldots,m_N}(k)$ implies that $a_{m_1,\ldots,m_N}(k) \to 0$ as $k \to \infty$. Further, it seems to be reasonable to assume that, at least in the long run, the terms $a_{m_1,\ldots,m_N}(k)$ are monotonously non-increasing in absolute value. Assume that

$$|a_{m_1,\ldots,m_N}(k+1)| \leq |a_{m_1,\ldots,m_N}(k)| \quad \text{for all } k \geq d_0,$$

where $d_0 \in \mathbb{N}_0$. Then, since the absolute value of an alternating series for which the terms are monotonously non-increasing in absolute value, is bounded by the absolute value of the first term, we find that

$$|p_{m_1,\ldots,m_N} - p_{m_1,\ldots,m_N}(d)| = \left| \sum_{k=d+1}^{\infty} a_{m_1,\ldots,m_N}(k) \right| \leq |a_{m_1,\ldots,m_N}(d)| \tag{4.39}$$

for all $d \geq d_0$; note that $a_{m_1,\ldots,m_N}(d)$ denotes the difference between $p_{m_1,\ldots,m_N}(d)$ and $p_{m_1,\ldots,m_N}(d-1)$. On this result, we base the empirical upper bound

$$\begin{cases} |a_{m_1,\ldots,m_N}(d)| & \text{if } |a_{m_1,\ldots,m_N}(d)| < |a_{m_1,\ldots,m_N}(d-1)| \,; \\ \infty & \text{otherwise} \end{cases} \tag{4.40}$$

for the absolute errors of the approximations $p_{m_1,\ldots,m_N}(d)$ with respect to $p_{m_1,\ldots,m_N}$. This bound is used in the so-called *simple procedure with empirical bounds*. It is noted that by definition the lowest value for d at which this procedure stops is larger than or equal to $N+1$, since $a_{m_1,\ldots,m_N}(d)=0$ for all $d=0,\ldots,N-1$ (which is due to the property that for all $v \in V$ with $l(v) \leq N-1$ at least one of the factors $\alpha_{1,v},\ldots,\alpha_{N,v}$ is equal to 1).

In both simple procedures, the selection of new nodes $v \in V$, which are added to the set V' in order to obtain a better approximation, is based on their distances $l(v)$ to the root $\varnothing$. In the *sophisticated procedure*, which is the third and last procedure we describe, new nodes $v \in V$ are selected in a more sophisticated way. This procedure computes the sum on the right-hand side of (4.2) within a given absolute accuracy by truncating subtrees at different depths. The procedure keeps track of a set W to of nodes where the computing process still has to be continued. This set is initiated by the set of nodes at depth 1. For each node one stores the solution $(\alpha_{1,v},\ldots,\alpha_{N,v})$ and the bound $b_{m_1,\ldots,m_N}(v)$ given by (4.16), i.e. the upper bound for the sum of the weights of all nodes, except v itself, of the subtree starting at v (here, again the bound $b_{m_1,\ldots,m_N}(v)$ is said to be equal to ∞ if the geometric tree on the right-hand side of (4.16) is not convergent). Further, we store the absolute accuracy with which the subtree starting at v has to be computed. In the initialization step the initial allowed inaccuracy is divided among the N nodes at depth 1 proportional to their values of $b_{m_1,\ldots,m_N}(v)$ (provided that each node gets at least a given, fixed percentage; take 5% in case $N=3$, for example). In each next step, one selects a node v from the set W and computes the contribution of the corresponding term. Subsequently, it is checked whether the subtree starting at v may be truncated below v, i.e. whether $b_{m_1,\ldots,m_N}(v)$ is smaller than or equal to the inaccuracy allocated to v. If so, then one can continue with another element of W, otherwise one first has to add the $N-1$ successors of v to the set W. Here, again the inaccuracy allocated to v is divided among its successors proportional to their values of $b_{m_1,\ldots,m_N}(v)$ (provided that each successor gets at least the given, fixed percentage). The procedure ends as soon as the set W is empty. Remark that an equilibrium probability $p_{m_1,\ldots,m_N}$ for a state in the convergence region M_c may be computed within a given *relative* accuracy by applying the sophisticated procedure for decreasing values of the allowed absolute inaccuracy.

We have applied all three procedures to the *symmetric 2×3 switch*, which is modeled as a three-dimensional random walk (so $N=3$) which satisfies all required properties. The symmetric 2×3 switch is defined as the 2×3 switch (see Example 2.2 of Chapter 2 and Example 3.2 of Chapter 3) with equal arrival rates $r_1=r_2=r$, where $r \in (0,1]$, and $\hat{r}_{j,i}=1/3$ for all $j=1,2$ and $i \in I$. In Table 4.1, the three procedures are compared on the basis of their performance for the computation of the equilibrium probability $p_{0,1,1}$. For all cases an absolute accuracy of 10^{-6} has been required. In the first two columns the range of chosen values of r and the corresponding values of $p_{0,1,1}$ are depicted. In the fourth column the number of computed terms of the sum in (4.2) is given, and in the fifth column the number of computed

procedure	r	$p_{0,1,1}$	depth	terms	relevant terms	error bound	absolute error
simple procedure with empirical bounds	0.01	0.000000	4	46	24	$6.9 \cdot 10^{-25}$	0
	0.2	0.000000	4	46	24	$1.0 \cdot 10^{-11}$	$2.1 \cdot 10^{-16}$
	0.4	0.000020	4	46	24	$1.6 \cdot 10^{-8}$	$6.1 \cdot 10^{-12}$
	0.6	0.000323	5	94	66	$3.4 \cdot 10^{-9}$	$7.6 \cdot 10^{-12}$
	0.8	0.002593	5	94	66	$4.2 \cdot 10^{-7}$	$3.4 \cdot 10^{-9}$
	1.0	0.013901	6	190	156	$6.4 \cdot 10^{-7}$	$1.6 \cdot 10^{-8}$
simple procedure with hard bounds	0.01	0.000000	3	22	6	$2.8 \cdot 10^{-15}$	$6.9 \cdot 10^{-25}$
	0.2	0.000000	3	22	6	$2.1 \cdot 10^{-7}$	$1.0 \cdot 10^{-11}$
	0.4	0.000020	4	46	24	$3.1 \cdot 10^{-7}$	$6.1 \cdot 10^{-12}$
	0.6	0.000323	5	94	66	$5.0 \cdot 10^{-7}$	$7.6 \cdot 10^{-12}$
	0.8	0.002593	7	382	342	$1.1 \cdot 10^{-7}$	$2.7 \cdot 10^{-13}$
	1.0	0.013901	9	1534	1482	$1.5 \cdot 10^{-7}$	$3.3 \cdot 10^{-13}$
sophisticated procedure	0.01	0.000000	2	10	0	$2.5 \cdot 10^{-10}$	$2.8 \cdot 10^{-15}$
	0.2	0.000000	3	14	2	$6.5 \cdot 10^{-7}$	$2.3 \cdot 10^{-7}$
	0.4	0.000020	5	34	12	$2.7 \cdot 10^{-8}$	$1.6 \cdot 10^{-8}$
	0.6	0.000323	6	54	26	$4.8 \cdot 10^{-8}$	$1.3 \cdot 10^{-9}$
	0.8	0.002593	7	78	46	$3.7 \cdot 10^{-7}$	$1.6 \cdot 10^{-7}$
	1.0	0.013901	9	134	92	$5.5 \cdot 10^{-7}$	$2.2 \cdot 10^{-7}$

Table 4.1. Performance characteristics for the computation of $p_{0,1,1}$ within absolute accuracy 10^{-6} for the symmetric 2×3 switch.

relevant terms, i.e. the number of computed terms for which $\alpha_{1,v}, \alpha_{2,v}, \alpha_{3,v} < 1$, is depicted. The maximum depth reached during the computing process, i.e. the maximum length of the indices v of the computed terms, can be found in the fifth column. For both simple procedures, this value is equal to the smallest d for which $p_{0,1,1}(d)$ approximates $p_{0,1,1}$ within the required accuracy. The sixth column gives the upper bound for the absolute accuracy with which $p_{0,1,1}$ has been computed. Of course, for the simple procedures these values are equal to the bounds given by (4.38) and (4.40), respectively, with d equal to the depth depicted in the third column. For the sophisticated procedure this value is equal to the bound $b_{m_1,m_2,m_3}(v)$ summed up over all v where subtrees have been truncated. Finally, in the last column the absolute accuracy itself has been depicted. These values have been computed after having determined $p_{0,1,1}$ with a higher absolute accuracy.

Table 4.1 shows that for the simple procedure with hard bounds more (relevant) terms have to be computed than for the sophisticated procedure, especially for high values of r, i.e. for high workloads. This seems to be caused by the roughness of the upper bound used for the absolute accuracy (compare the values in the last two columns), which is mainly due to the use of the bound $b_{m_1,m_2,m_3}(v)$ at nodes v for which at least one of the factors of $(\alpha_{1,v}, \alpha_{2,v}, \alpha_{3,v})$ is equal to 1. Subtrees starting at such nodes v have infinitely many nodes w with $\alpha_{1,w}$, $\alpha_{2,w}$ or $\alpha_{3,w}$ equal to 1. Since the contribution of these terms is equal to 0 for

p_{m_1,m_2,m_3}, but equal to $\Pi_{i=1}^{3} \alpha_{i,w}^{m_i}$ for $\hat{b}_{m_1,m_2,m_3}(v)$ (see (4.8)), and therefore even larger for $b_{m_1,m_2,m_3}(v)$ (since the bounds $b_{m_1,\ldots,m_N}(v)$ also are upper bounds for $\hat{b}_{m_1,\ldots,m_N}(v)$, see (4.14)-(4.16)), the use of the upper bound $b_{m_1,m_2,m_3}(v)$ causes a large gap between the guaranteed upper bound for the absolute accuracy and the absolute accuracy itself, especially when we are at a node v at a low depth (i.e. with a small length $l(v)$). The sophisticated procedure overcomes the problem caused by the roughness of the bounds $b_{m_1,m_2,m_3}(v)$ for nodes v with at least one of the factors $\alpha_{i,v}$ equal to 1, by going deeper in the tree at such nodes than at other nodes. As a consequence, the sophisticated procedure performs much better than the simple procedure with hard bounds. The results in Table 4.1 also show that in the simple procedure with empirical bounds the empirical upper bound given by (4.40) appears to serve as a good upper bound for the absolute accuracy of $p_{m_1,m_2,m_3}(d)$. Therefore also this procedure is an appropriate procedure for the computation of the equilibrium probabilities.

It is obvious that all three procedures are efficient. Nevertheless, we do see the following ordering. The sophisticated procedure performs the best, and, in general, the simple procedure with empirical bounds performs better than the simple procedure with hard bounds (however, for small values of r, in the simple procedure with hard bounds the computing process may be stopped for $d=3$, while, by definition, the value at which the simple procedure with empirical bounds may stop always is larger than or equal to 4). From this it may be concluded that the sophisticated analysis which has led to the error bounds and the explicit expressions for them, must be accompanied by a sophisticated use of these error bounds in order to obtain a better procedure than the simple procedure with the simple empirical bounds given by (4.40).

4.5. Comparison of the 2×3 switch to a system with independent servers

Except for the equilibrium distribution, the procedures described in the previous section may also be used for other quantities. For the 2×3 switch, for example, we can also determine moments of queue lengths (however, note that for the moments $\mathbb{E}\mathbf{L}_1^{k_1}\mathbf{L}_2^{k_2}\mathbf{L}_3^{k_3}$, where $\mathbf{L}_i$ denotes the length of the queue at server i, it suffices to analyze a 2×2 switch in case one or more of the powers k_i are equal to 0) and the distribution of the number $\mathbf{N}$ of non-empty queues at the beginning of a time unit. Denote the probability that $\mathbf{N}$ equals k by $p(k)$. By using (4.1), we find that

$$p(3) = \sum_{m_1=1}^{\infty} \sum_{m_2=1}^{\infty} \sum_{m_3=1}^{\infty} p_{m_1,m_2,m_3}$$

$$= \sum_{v \in V} 1_{\{\alpha_{1,v}\alpha_{2,v}\alpha_{3,v}<1\}} (-1)^{l(v)-3} \alpha_{1,v}\alpha_{2,v}\alpha_{3,v}\,, \tag{4.41}$$

where $1_{\{\alpha_{1,v}\alpha_{2,v}\alpha_{3,v}<1\}}$ is equal to 1 if all three factors $\alpha_{1,v},\alpha_{2,v},\alpha_{3,v}$ are smaller than 1 and equal to 0 otherwise. A similar expression may be found for $p(2)$, while $p(0)=p_{0,0,0}$ and $p(1)$ follows from the property that the probabilities $p(i)$ add up to 1. The sum in (4.41) may be computed in the same way as the sums for the equilibrium probabilities given in (4.2) and we may even use the same error bounds.

system	r	$\hat{p}(0)$	$\hat{p}(1)$	$\hat{p}(2)$	$\hat{p}(3)$	$\mu(\hat{\mathbf{N}})$	$\sigma(\hat{\mathbf{N}})$	$vc(\hat{\mathbf{N}})$
2×3	0.01	0.9801	0.0198	0.0001	0.0000	0.02	0.140	7.024
switch	0.2	0.6302	0.3397	0.0299	0.0002	0.4	0.548	1.371
	0.4	0.3345	0.5336	0.1294	0.0025	0.8	0.659	0.823
	0.6	0.1302	0.5549	0.2997	0.0152	1.2	0.671	0.559
	0.8	0.0245	0.4091	0.5084	0.0580	1.6	0.636	0.398
	1.0	0.0000	0.1732	0.6536	0.1732	2.0	0.589	0.294
indepen-	0.01	0.9801	0.0198	0.0001	0.0000	0.02	0.141	7.047
dent	0.2	0.6510	0.3004	0.0462	0.0024	0.4	0.589	1.472
servers	0.4	0.3944	0.4302	0.1564	0.0190	0.8	0.766	0.957
	0.6	0.2160	0.4320	0.2880	0.0640	1.2	0.849	0.707
	0.8	0.1016	0.3485	0.3982	0.1517	1.6	0.864	0.540
	1.0	0.0370	0.2222	0.4445	0.2963	2.0	0.816	0.408

Table 4.2. The distribution of the number of working servers during a time unit for the symmetric 2×3 switch; the second part gives the distribution for independent servers.

From the distribution of $\mathbf{N}$, one can easily compute the distribution of the number $\hat{\mathbf{N}}$ of working servers during a time unit. In Table 4.2, this distribution ($\hat{p}(k)$ denotes the probability that $\hat{\mathbf{N}}$ equals k), and also its first moment, standard deviation and coefficient of variation, are given for the symmetric 2×3 switch. In the second part the same is depicted for the corresponding system with independent servers, i.e. the system consisting of three, parallel servers where each server has two Bernoulli streams of arriving jobs with rate $r/3$. The results in Table 4.2 show that, for all r, the 2×3 switch has a smaller variability in the number of working servers than the system with independent servers, which, of course, is due to the (negative) correlation between the streams of arriving jobs. For high workloads r this correlation has a considerable impact, while for low workloads r the impact is almost negligible.

4.6. Conclusions

In this chapter, we have performed a structure analysis of the equilibrium distribution for the class of multi-dimensional random walks for which explicit results were obtained in the previous chapter by applying the compensation approach. For the infinite, alternating sum of product-form solutions which describes the equilibrium behavior of a random walk of this class, we have established a behavior which is typical for so-called geometric trees. An extensive analysis for these geometric trees has led to explicit upper bounds for the absolute errors of the approximations of the equilibrium probabilities by finite, alternating sums of product-form solutions. The error bounds have been exploited in efficient numerical procedures for the computation of the equilibrium distribution and related quantities within a given accuracy, and numerical results have been presented for the symmetric 2×3 switch.

Chapter 5

The Precedence Relation Method for Deriving Flexible Bound Models

5.1. Introduction

As we have stated in the introductory chapter, the objective of this monograph is the development of methods for the analysis of queueing systems for which the behavior is described by multi-dimensional random walks/Markov processes on state spaces which are infinite in each component. The Chapters 2-4 have been devoted to the compensation approach. By using this approach, we have been able to derive explicit formulae for the equilibrium distribution for a restricted class of multi-dimensional random walks. Besides the class of product-form networks (see Baskett et. al. [15]), up to now, this class is the only other class of multi-dimensional random walks which can be solved analytically. This indicates that in general it will be hard to solve a multi-dimensional problem in an analytical way, and therefore it is desired to have alternative methods. One alternative method is the power-series algorithm (see Blanc [18]), with which the equilibrium distribution and the corresponding relevant performance measures can be determined within a given accuracy, provided that the corresponding requirements with respect to the computational efforts are met. Another alternative is constituted by the use of *solvable truncation models* which can approximate the original model or Markov process as accurately as desired. This latter property may be obtained by defining the truncated state space such that its size depends on one or more truncation parameters. Truncation models with this property are called *flexible* truncation models. Truncation models in fact lead to approximations for the equilibrium distribution of the original model, and therefore also to approximations for the relevant performance measures. Of a particular interest are flexible truncation models which produce *bounds* for the relevant performance measures. Such models are also called *flexible bound models*. The second part of this monograph, consisting of the Chapter 5-7, will be devoted to a systematic method, called *precedence relation method*, with which such models may be derived. This method will be developed in this chapter, and after that, in the Chapter 6 and 7, it will be applied to two particular queueing systems, viz. the Symmetric Shortest Queue System (SSQS), with $N \geq 2$ servers, and to the generalization of this system as described in Section 1.2.

The precedence relation method can be used for proving monotonicity results between performance measures of two Markovian systems, where the state space of one system is a subset of the state space of the other system. These monotonicity results are derived by comparing the costs in the two corresponding Markov cost models. The *main idea* of the

precedence relation method is that this comparison may be simplified by first deriving *precedence pairs* for the Markov cost model with the larger state space, i.e. by first proving for pairs of states of the Markov cost model with the larger state space that they satisfy a certain *precedence relation* which denotes that the first state of a pair is more attractive with respect to the costs than the second state.

In this chapter, we shall mainly focus on the comparison of performance measures for truncation models to the corresponding performance measures for the original model. Due to the introduction of the precedence relation, we will find extremely simple, sufficient conditions for obtaining truncation models which produce lower and upper bounds for the relevant performance measures of a given original model. This directly leads to a simple method with which we can *derive* (or *construct*) bound models, and especially *flexible* bound models, which can be used to approximate the original model as accurately as desired. The method for the derivation of the flexible bound models will also be called the precedence relation method. An attractive property is that, after having derived the precedence pairs for the original model, this method may easily and quickly produce a whole set of sensible flexible bound models.

In Section 1.4 of the introductory chapter, on the basis of two flexible truncation for the SSQS with $N=2$ servers, we have described globally how the precedence relation method *proves* that a particular (flexible) truncation model is a bound model. In this chapter, the two-dimensional SSQS is used to illustrate how the precedence relation method *derives* flexible bound models. We will obtain *six*, solvable, flexible bound models: four lower bound models and two upper bound models. Among these bound models are the two models presented in Section 1.4, and also the truncation models given in the papers by Conolly [24] and Rao and Posner [60]. In Chapter 6 (see also [2]), the precedence relation method is applied to the N-dimensional SSQS with general $N \geq 2$; in that chapter we will generalize the two bound models of Section 1.4 (which for the two-dimensional case will appear to produce the tightest bounds for the mean waiting time, see Section 5.5). The present list of systems for which the precedence relation method has appeared to lead to appropriate, flexible bound models, further contains the shortest queue system with a job-dependent parallelism (which has been described in Section 1.2 and which will be treated in Chapter 7), and the symmetric *longest* queue system (see [1]); it is our conviction that this list can be extended with several other systems.

For both the precedence relation method as used for proving monotonicity results and the precedence relation method as used for deriving flexible bound models, there exist some related methods in the literature. For proving monotonicity results between Markovian (queueing) systems, there are basically four methods available, of which the sample path technique seems to be the best known one; for a short overview of these methods, see Van Dijk and Van der Wal [76]. The precedence relation method originates from the technique used in the papers by Van der Wal [72], Van Dijk and Van der Wal [76], and Van Dijk and Lamond [75]. A typical property of the precedence relation method is the explicitly defined precedence relation for the states of the original model. The strength of the precedence relation method, as used for proving monotonicity results, is that in general, it will lead to simpler and intuitively clearer proofs than other methods (it seems that many monotonicity results which can be proved by the precedence relation method, may also be proved by using other methods). A more important feature of the precedence relation method is that it can be used

for *deriving* monotonicity results.

As a method for deriving flexible bound models, the precedence relation method is related to the methods developed by Van Dijk [74] and Stepanov [66]. In his book [74], Van Dijk advocates the use of a method for modifying an original non-product-form system into product-form systems (by repairing station balance, for example) in order to obtain a first indication on the orders of magnitude of the relevant performance measures (such indications may be useful for measures like blocking and loss probabilities in communication systems, which in many situations are hard to determine). He further claims that in general it will be intuitively clear whether a modification leads to lower or upper bounds; and, in several cases, formal proofs of the bounds can be given by using the technique of the papers [72, 75, 76]. In [66] (see also [65, 67]), for a number of queueing systems with repeated calls, Stepanov also describes an approach for deriving flexible bound/truncation models. His approach bears some similarity with the precedence relation method in the sense that it also constructs bound models by redirecting transitions ending in states outside the truncated state space to more/less attractive states inside the truncated state space. However, in his approach the concept of 'attractiveness' is only based on intuitive arguments. The way in which formal proofs are given (see [67]) appears to be totally different (and much more complicated).

This chapter is organized as follows. In Section 5.2, we describe a general original Markovian system, for which we want to derive flexible bound models. Next, the precedence relation method is presented in Section 5.3, and it is shown how truncation models should be defined in order to lead to lower and upper bounds for the relevant performance measure(s) of the original model. In Section 5.4, an extensive treatment is given of the derivation of precedence pairs, which turns out to be the essential step of the precedence relation method. Section 5.5 is devoted to a discussion on the quality of the flexible bound models derived by the precedence relation method; at the end of th section, we shall discuss some other monotonicity results than between original and truncation models, which can be proved by using the precedence relation method. Finally, Section 5.6 is devoted to the conclusions.

5.2. The original model

In general, the relevant performance measures of a given Markovian queueing system may be determined by defining appropriate Markov cost models and computing the average costs. This property is exploited by the precedence relation method, which in the next section is developed for the comparison between the average costs of an original Markov cost model and the average costs in a related truncation model. In this section, we describe the original Markov cost model. To ensure that in the next section simple comparisons can be made between the so-called t-period costs for different (starting) states, we assume that we have *discrete* time. Note that this assumption can be made w.l.o.g., since (under some mild conditions) continuous-time Markov processes may be transformed to equivalent discrete-time Markov processes by using the uniformization technique (as described in Tijms [70], for example).

Consider a discrete-time, irreducible and positive recurrent Markov process with a state space M consisting of N-dimensional vectors $m=(m_1,\ldots,m_N)$. Let the transition probabilities be given by $q_{m,n}$ and let $\{p_m\}$ denote the equilibrium distribution, which is characterized as the unique normalized solution of the equilibrium equations. Suppose that *direct costs* $c(m)$ are incurred for each period that the Markov process is in state m. Then the corresponding *average costs* g per period are given by

$$g = \sum_{m \in M} p_m c(m); \tag{5.1}$$

note that the average costs do not depend on the starting state of the Markov process. In the next paragraph, it is indicated how to define the direct costs $c(m)$ such that the average costs g are equal to a certain performance measure of interest.

Suppose that we have a Markov process which describes the behavior of a queueing system consisting of N queues, and that each component m_i of a state describes the length of a particular queue. In that case, the average costs g are equal to the k-th moment of the total number of jobs $\mathbf{L}$ in the system, if $c(m)$ is defined by $c(m)=(\Sigma_{i=1}^N m_i)^k$; and, all probabilities $\mathbb{P}(\mathbf{L}\geq l)$ are obtained by letting $c(m)$ be equal to 1 if $\Sigma_{i=1}^N m_i \geq l$, and to 0 otherwise. Similarly one may obtain information on particular queues, or on shortest or longest queues. Note that for most performance measures it is possible to define the cost function $c(m)$ such that $c(m)$ is non-decreasing in each component.

Imagine that we want to determine (analytically or numerically) the average costs g in the original model, but that it is not possible (or not attractive) to use formula (5.1), since it is not possible (or very hard) to determine the equilibrium distribution $\{p_m\}$. Then it is quite customary to use truncation models to obtain some valuable information on g. In the next section, we shall investigate how truncation models should be defined in order to obtain lower or upper bounds for g. But first we discuss a system which is appropriate for serving as an example of the original model.

Example 5.1: The symmetric shortest queue system

Consider the symmetric shortest queue system, as described in the previous chapters. For completeness, we repeat the description. The symmetric shortest queue system (SSQS) consists of $N\geq 2$ parallel servers, which all have their own queue. Jobs arrive according to a Poisson stream with intensity $\lambda>0$, and an arriving job always joins the shortest queue (ties are broken with equal probabilities). All service times are exponentially distributed with parameter $\mu>0$. For simplicity, we assume that time is scaled such that $\lambda+N\mu=1$. In order to have an ergodic system, the workload $\rho=\lambda/(N\mu)$ is assumed to be smaller than 1.

Since we want to use the SSQS as an illustration model for the precedence relation method, in this chapter the SSQS is modeled as a discrete-time instead of a continuous-time Markov process. Further, we choose a slightly different state description which will appear to be more appropriate for the present analysis. The SSQS is modeled as follows.

Assume that the servers always work, but that a service completion is only accompanied by a departure of a job if there is a job present in the corresponding queue (note that this is equivalent to uniformizing the time intervals between jump moments, as done by the uniformization technique). Then the behavior of the system may be described by the discrete-time Markov process on the time instants right after job arrivals and service completions, and

with states $(m_1, \ldots, m_N)$, where m_i denotes the length of the i-th shortest queue.

The most interesting performance measure for the SSQS is the mean W of the *normalized* waiting time. The normalized waiting time is defined as the ratio of the waiting time and the mean service time of a job $(= 1/\mu)$, and has the attractive property that it only depends on N and ρ. By Little's formula, we find that W is equal to

$$W = \frac{L_w}{N\rho} , \tag{5.2}$$

where L_w denotes the mean of the total number of waiting jobs in the system. As a consequence, it suffices to determine L_w. To ensure that L_w is equal to the average costs g, we define the direct costs $c(m)$ by

$$c(m) = \sum_{i=1}^{N} \max\{ m_i - 1, 0 \} . \tag{5.3}$$

For the SSQS with $N \geq 3$, it is not known how to determine the equilibrium distribution by an analytical or a standard numerical method. Therefore we would like to derive solvable bound models in order to obtain information on L_w, i.e. on the average costs g. For the sake of clarity, in this chapter the various bound models will be described for the case with $N = 2$ servers; however, all bound models may also be defined for general N. Note that for $N = 2$ the state space is given by

$$M = \{ m \mid m = (m_1, m_2) \text{ with } 0 \leq m_1 \leq m_2 \} ,$$

where m_1 and m_2 denote the lengths of the shortest and the longest queue, the transition rates $q_{m,n}$ are as depicted in Figure 5.1, and the direct costs are given by

$$c(m) = c(m_1, m_2) = \max\{ m_1 - 1, 0 \} + \max\{ m_2 - 1, 0 \} , \quad m \in M. \tag{5.4}$$

5.3. The precedence relation method

This section contains the core of this chapter. It is devoted to the description of the precedence relation method, which in principle is a method for the comparison of the average costs in two related Markov cost models, of which the state space of one model is a subset of the state space of the other model. We shall mainly focus on the comparison between the original model of the previous section and truncation models. The main result is that simple, sufficient conditions are found for obtaining truncation models which produce bounds for the average costs g in the original model. This result enables us to present a simple method for the derivation of flexible bound models.

Let us consider a truncation model of the original model described in the previous section. A truncation model is obtained by first defining a truncated state space $M' \subset M$ (M' is usually defined such that it contains the states where most of the probability mass is present), and next modifying the transitions of the original model such that the states outside M' become transient (initially, all transition probabilities $q'_{m,n}$ for the truncation model are taken equal to the transition probabilities $q_{m,n}$ for the original model). This means that each transition starting in a state m inside M' and ending in a state n outside M', must be *redirected* to a

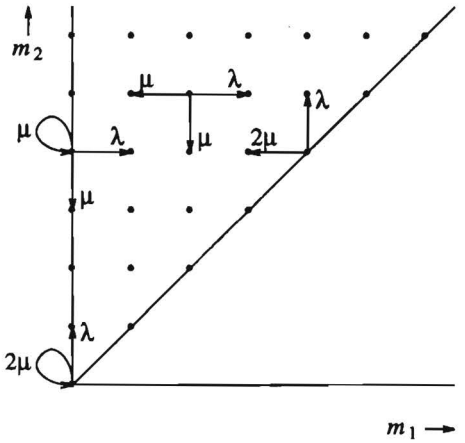


Figure 5.1. The transition probabilities for the discrete-time Markov process for the symmetric shortest queue system with $N=2$ servers.

state $n'(m,n)$ inside M' (the probability $q'_{m,n}$ is set equal to 0 and $q'_{m,n'(m,n)}$ is increased by $q_{m,n}$).

The truncation model is assumed to be irreducible and positive recurrent, and its equilibrium distribution is denoted by $\{p'_m\}$. Further, the direct costs $c'(m)$ in the truncation model are assumed to be equal to the direct costs $c(m)$ in the original model, i.e. $c'(m)=c(m)$ for all $m \in M'$. Finally, the average costs in the truncation model are denoted by g':

$$g' = \sum_{m \in M'} p'_m c(m) . \tag{5.5}$$

Suppose that the truncation model is expected to lead to a bound for the average costs g in the original model, say to a *lower* bound:

$$g' \leq g . \tag{5.6}$$

To prove this monotonicity result between the original and truncation model, we use the so-called t-period cost functions for both models.

Let $v_t(m)$ and $v'_t(m)$ denote the expected t-period costs in the original and truncation model, respectively. This means that $v_t(m)$ denotes the expected costs in the original model in the next t periods when starting in state $m \in M$; so,

$$v_0(m) = 0 \quad \text{for all } m \in M, \tag{5.7}$$

and for all $t \geq 0$, we have the recurrence relation

$$v_{t+1}(m) = c(m) + \sum_{n \in M} q_{m,n} v_t(n) \quad \text{for all } m \in M; \tag{5.8}$$

and, similarly for the costs $v'_t(m)$ in the truncation model. Because of the assumed irreducibility in both models, we have the property that

$$g = \lim_{t\to\infty} \frac{v_t(m)}{t} \quad \text{and} \quad g' = \lim_{t\to\infty} \frac{v'_t(m')}{t}, \tag{5.9}$$

where m and m' may be arbitrary states in M and M', respectively. As a consequence, to prove (5.6), it suffices to show that for some $m \in M$ and $m' \in M'$,

$$v'_t(m') \leq v_t(m) \quad \text{for all } t \geq 0. \tag{5.10}$$

Because of the resemblance between both models, it seems reasonable to try to prove that (5.10) holds for some states $m \in M$ and $m' \in M'$ with $m = m'$; further, if it holds for some state $m \in M'$ that $v'_t(m) \leq v_t(m)$ for all $t \geq 0$, then probably this also holds for all other states of M'. Therefore, we shall focus on trying to prove (5.6) by showing that

$$v'_t(m) \leq v_t(m) \quad \text{for all } m \in M' \text{ and } t \geq 0. \tag{5.11}$$

The inequalities stated in (5.11) may be proved by using the *precedence relation method*. The main idea of this method is that the comparison of the t-period costs $v'_t(m)$ in the truncation model to the corresponding t-period costs $v_t(m)$ in the original model may be simplified by first performing a preliminary step, in which on the basis of a *precedence relation* for the t-period costs $v_t(m)$ an ordering for the states of the original model is derived. The precedence relation method consists of the following two steps:

1. Derive a set P consisting of *precedence pairs* (m,n) of states $m,n \in M$, which satisfy the *precedence relation*

   $$v_t(m) \leq v_t(n) \quad \text{for all } t \geq 0. \tag{5.12}$$

   This relation states that in the original model, state m *has precedence over* state n with respect to the t-period costs, or equivalently, state m is *more attractive* than n, or n is *less attractive* than m;

2. Exploit the precedence pairs derived in step 1 to show that (5.11) holds.

Both steps are further explained below.

Step 1 is treated in detail in the next section. Here, it suffices to make some general remarks. Step 1 may be performed by first defining a set P which is expected to consist of precedence pairs, and next proving by induction with respect to t that the precedence relation (5.12) is satisfied for all pairs (m,n) of this set P. Since $v_1(m) = c(m)$ for all $m \in M$, all pairs $(m,n) \in P$ must satisfy the condition that $c(m) \leq c(n)$, i.e. all pairs $(m,n) \in P$ must at least be precedence pairs for the direct costs $c(m)$. So, if the direct costs $c(m)$ are explicitly defined for a given Markov cost model, then the values for $c(m)$ indicate which precedence pairs may (or, better: cannot) be derived. The definition of P may further be based on intuitive and/or numerical insight obtained by computing (and comparing) the values of the t-period costs $v_t(m)$ for some small t. Typical precedence pairs that can be derived in case the states $m \in M$ represent queue lengths, are pairs of the type $(m, m+e_i)$, where e_i is the i-th unity vector.

Step 2 appears to be quite simple; contrary to step 1, it can be performed in general terms. It appears that the inequalities stated in (5.11) can easily be proved by induction with respect to t, if the truncation model satisfies the following simple condition:

> for all $m \in M'$ and $n \in M \backslash M'$ with $q_{m,n} > 0$, the state $n'(m,n)$ to which the transition from m to n has been redirected, is more attractive (has smaller t-period costs) than the state n, i.e. for $n'(m,n)$ and n it holds that $(n'(m,n),n) \in P$.

The inequalities stated in (5.11) hold for $t=0$ by definition (see (5.7)). For all other $t \geq 0$, they follow from the induction step, which reads as follows (note that the condition for the redirections is needed for the second inequality in the following derivation):

$$\begin{aligned} v'_{t+1}(m) &= c(m) + \sum_{\substack{n \in M' \\ q_{m,n}>0}} q_{m,n}\, v'_t(n) + \sum_{\substack{n \in M \setminus M' \\ q_{m,n}>0}} q_{m,n}\, v'_t(n'(m,n)) \\ &\leq c(m) + \sum_{\substack{n \in M' \\ q_{m,n}>0}} q_{m,n}\, v_t(n) + \sum_{\substack{n \in M \setminus M' \\ q_{m,n}>0}} q_{m,n}\, v_t(n'(m,n)) , \\ &\leq c(m) + \sum_{\substack{n \in M' \\ q_{m,n}>0}} q_{m,n}\, v_t(n) + \sum_{\substack{n \in M \setminus M' \\ q_{m,n}>0}} q_{m,n}\, v_t(n) , \\ &= v_{t+1}(m) , \quad m \in M'. \end{aligned}$$

This completes the description of the precedence relation method as used for proving the monotonicity result stated in (5.6).

An important result of the analysis above is that we have found a simple, sufficient condition under which a truncation model can be guaranteed to lead to a lower bound for the average costs g in the original model. A similar condition must be satisfied in order to obtain a truncation model which leads to an upper bound for g. This leads to the following *conclusion* (note that for the derivation of these conditions, we did not use the positive recurrence of the original and truncation model):

> *The average costs g' of an irreducible truncation model constitute a lower bound (upper bound) for the average costs g in the corresponding, irreducible, original Markov cost model, if the truncation model has been constructed such that each transition starting in a state m inside the truncated state space M' and ending in a state n outside M', has been redirected to a state $n'(m,n) \in M'$ which, according to certain precedence pairs that can be derived for the original model, is more attractive (less attractive) than the state n.*

The result stated above provides us with a simple method for the *derivation* (or construction) of bound models, and especially for the derivation of *flexible* bound models, with which the original model can be approximated as accurately as desired; flexible bound models may be obtained by defining truncation models with a flexible truncated state space which depends on one or more truncation parameters. This method is also called the precedence relation method, or, more specifically, the *precedence relation method for deriving flexible bound models*. It consists of the following two steps (note that step 2 is a constructive step in this case):

1. The derivation of a set P of precedence pairs for the original model, i.e. the derivation of a set P consisting of pairs (m,n) of states $m,n \in M$ which satisfy (5.12);
2. The definition of flexible lower and upper bound models: to obtain a flexible lower (upper) bound model, first a flexible truncated state space M' must be defined, and next each transition from a state $m \in M'$ to a state $n \in M \setminus M'$ must be redirected to a state $n'(m,n) \in M'$ which, according to the precedence pairs derived in step 1, is more (less) attractive than the state n.

It must be noted that it is not possible to obtain a lower (upper) bound model for each choice of the truncated state space M', since there do not always have to be more (less) attractive states inside M', to which the transitions ending in the states outside M' can be redirected (for example, if we have only precedence pairs of the type $(m, m+e_i)$, then it may be impossible to obtain an upper bound model for a finite truncated state space M' consisting of the states $(m_1, \ldots, m_N)$ with the smallest components $m_1, \ldots, m_N$).

An attractive property of the precedence relation method is that, once the set of precedence pairs has been derived, a whole set of flexible bound models can be obtained. Further, it is obvious, that the best truncation models can be derived, if in step 1 as many precedence pairs are derived as possible. In fact, it will mainly depend on the precedence pairs derived in step 1, whether it is possible to obtain bound models which lead to tight bounds. This shows the essence of step 1, and therefore we will extensively treat this step in the next section.

From a practical point of view, it is mainly interesting to use the precedence relation method for the derivation of flexible bound models which are *solvable*, i.e. bound models for which the equilibrium distribution $\{p_m\}$ can be determined analytically or in an efficient way by a standard numerical technique, and for which the bound g' can be determined by using formula (5.5). Below, by applying the precedence relation method, we will derive *six*, solvable, flexible bound models for the two-dimensional SSQS.

Example 5.1: The symmetric shortest queue system *(continued)*

For the SSQS with two servers and the cost function $c(m)$ defined by (5.4), i.e. the cost function $c(m)$ which ensures that the corresponding average costs g are equal to the mean number of waiting jobs L_w, one can prove that for all $t \geq 0$,

$$v_t(m_1,m_2) \leq v_t(m_1+1,m_2) \quad \text{for all } (m_1,m_2) \in M,\ m_1 < m_2; \tag{5.13}$$

$$v_t(m_1,m_2) \leq v_t(m_1,m_2+1) \quad \text{for all } (m_1,m_2) \in M; \tag{5.14}$$

$$v_t(m_1,m_2) \leq v_t(m_1-1,m_2+1) \quad \text{for all } (m_1,m_2) \in M,\ m_1 > 0. \tag{5.15}$$

The inequalities in (5.13) and (5.14) state that it is more attractive to be or to start in a state with one job less at one of the two servers. The inequalities in (5.15) state that it is more attractive to be in a state with *more balance*, i.e. a state with a smaller difference between both queue lengths; the intuitive explanation of these inequalities is that from states with more balance, it is less likely to arrive in one of the so-called 'bad' states $(0,m_2)$, $m_2 \geq 2$, which correspond to the situation that one server is idle while there still are waiting jobs at the other server. Note that the inequalities in (5.13)-(5.15) also hold for the cost function $c(m)$ as defined by (5.4). All these inequalities may be proved by induction with respect to t (see [2]). Alternative proofs will be given in the next section. In Figure 5.2, the ordering for the states $(m_1,m_2) \in M$, as obtained from (5.13)-(5.15), has been depicted graphically.

Let the set P consist of all precedence pairs corresponding to (5.13)-(5.15), i.e.

$$\begin{aligned} P = &\{ ((m_1,m_2),(m_1+1,m_2)) \mid 0 \leq m_1 < m_2 \} \\ &\cup \{ ((m_1,m_2),(m_1,m_2+1)) \mid 0 \leq m_1 \leq m_2 \} \\ &\cup \{ ((m_1,m_2),(m_1-1,m_2+1)) \mid 0 < m_1 \leq m_2 \}. \end{aligned} \tag{5.16}$$

Note that, since the binary operator 'has precedence over' is reflexive and transitive (i.e. m

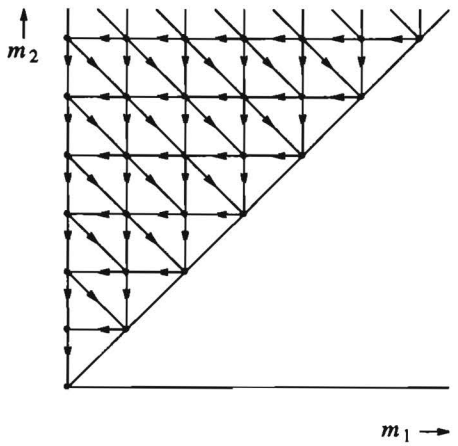


Figure 5.2. A graphical representation of the precedence pairs for the symmetric shortest queue system with two queues and the cost function given by (5.4). Each arrow points to a more attractive state.

has precedence over m for all $m \in M$, and, if m has precedence over n and n has precedence over r for some $m,n,r \in M$, then m has precedence over r), this set P may be completed to the following set P^* of precedence pairs:

$$P^* = \{ (m,n) \mid m=(m_1,m_2) \in M,\ n=(n_1,n_2) \in M,$$
$$m_1+m_2 \leq n_1+n_2 \text{ and } m_2 \leq n_2 \}. \tag{5.17}$$

When defining bound models, it seems sensible to choose the truncated state space such that it contains the states where most of the probability mass is present, since then it may be expected that the behavior of the modified model closely approximates the behavior of the original model. In Table 1.1, it has been shown how the probability mass is distributed among the different states for the case $\rho = 0.6$ (note that in Chapter 1, a slightly different state description has been used). The probability mass is concentrated around the origin and, due to the shortest queue routing, around the other states corresponding to situations with equal queue lengths. Good bound models should contain these states. We shall consider the following *six*, flexible bound models, which are depicted in Figure 5.3:

* *Finite Buffers (FB):*
  The simplest, modified model is obtained by truncating all states for which the number of jobs at the longest queue exceeds some threshold parameter $T \geq 1$. The truncated state space is equal to

  $$M' = \{ (m_1,m_2) \mid 0 \leq m_1 \leq m_2 \leq T \}$$

  and the only transition from a state inside M' to a state outside M' is the transition from (T,T) to $(T,T+1)$, which is caused by a new arriving customer. This transition is redirected to (T,T), which means that the new customer is rejected. As one can easily

see, this modified model is equivalent to a shortest queue system with finite buffers of size $T-1$, and therefore this model will be called the Finite Buffers (FB) model. Since the transition to $(T,T+1)$ is redirected to a more attractive state according to the precedence pairs, the FB model gives a *lower bound* for the average costs g, i.e. for the mean L_w of the total number of waiting jobs in the system.

* *Central Buffer (CB):*
A more sophisticated, modified model is obtained, when we add the states around the diagonal to the truncated state space. We define

$$M' = \{ (m_1,m_2) \mid 0 \le m_1 \le m_2 \le T \} \cup \{ (m_1,m_2) \mid T \le m_1 \le m_2 \text{ and } m_2 \le m_1+1 \},$$

where T is a threshold parameter again, $T \ge 1$. For all $m_1 \ge T$, the transition pointing from state (m_1,m_1+1) to (m_1-1,m_1+1), which is due to a service completion at the shortest queue, is redirected to the more attractive state (m_1,m_1). This model is a *lower bound* model and it is equivalent to the symmetric shortest queue system with finite local buffers of size $T-1$ at each server and an infinite central buffer in front of the local buffers. This model is called the Central Buffer (CB) model. In the CB model, each arriving job is assumed to join one of the queues at the servers, if not all local buffers are full, otherwise the job queues up in the central buffer. Further, jobs in the central buffer are immediately sent to a local buffer if a place becomes available there due to a service completion. It is easily checked that the CB system may be modeled by the modified Markov process which we just described. Note that for $T=1$, the CB model reduces to the $M \mid M \mid 2$ queueing system.

* *Threshold Jockeying (TJ):*
Models for which the equilibrium distribution has a matrix-geometric form, may be obtained by truncating all states for which the imbalance, i.e. the difference between the longest and the shortest queue length, exceeds some prescribed maximum level $T \ge 1$. Then the truncated state space is equal to

$$M' = \{ (m_1,m_2) \mid 0 \le m_1 \le m_2 \text{ and } m_2 \le m_1+T \}$$

and the only transitions from states of M' to states outside M' are the transitions from the states (m_1,m_1+T) to (m_1-1,m_1+T), where $m_1 \ge 1$. In the Threshold Jockeying (TJ) model these transitions, which are due to a service completion at the shortest queue, are redirected to the more attractive states (m_1,m_1+T-1). The physical interpretation is that a job jockeys from the longest queue to the shortest queue in case the difference between the longest and the shortest queue exceeds T. Also the TJ model is a *lower bound* model. Although for $T=1$, the TJ model is not identical to the $M \mid M \mid 2$ queueing system (= CB model for $T=1$), it does lead to an equivalent Markov process and therefore to the same values/behavior for several performance measures, among which the total number of waiting jobs in the system.

* *One Infinite Buffer (OIB):*
Another model for which the equilibrium distribution has a matrix-geometric form, is obtained by truncating all states for which m_1 exceeds some threshold parameter $T \ge 1$. Define

$$M' = \{ (m_1,m_2) \mid 0 \le m_1 \le m_2 \text{ and } m_1 \le T \},$$

and, for all $m_2 \ge T+1$, redirect the transition pointing from state (m_1,m_2) to (m_1+1,m_2),

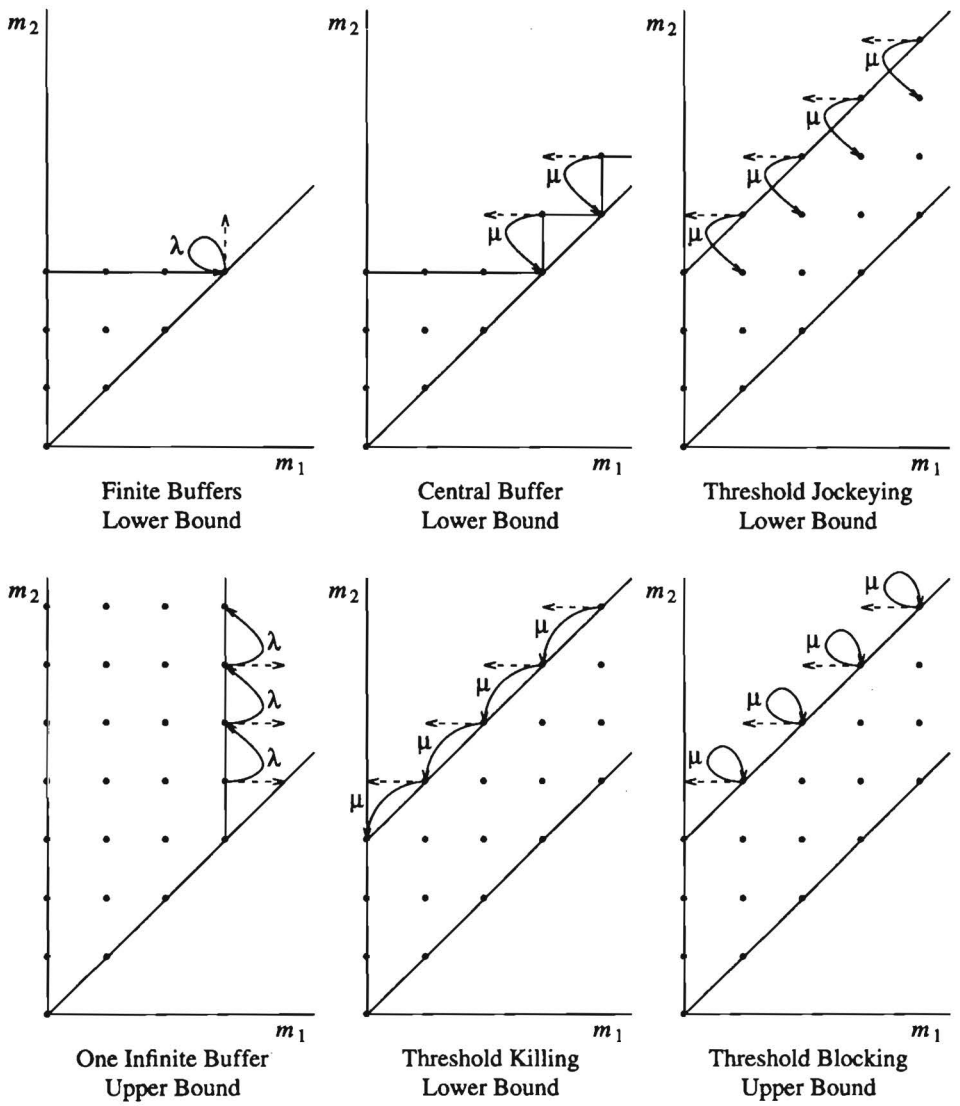


Figure 5.3. Six flexible bound models, all with threshold parameter $T=3$, for the symmetric shortest queue system with two queues and the cost function given by (5.4). The dashed arrows denote transitions of the original model to states outside the truncated state space and the corresponding uninterrupted arrows show how these transitions are modified.

which is due to an arrival of a new job, to the less attractive state (m_1,m_2+1). This modified model will give an *upper bound* for $g=L_w$, and it appears to be equivalent to a shortest queue system for which one buffer has infinite size and the other buffer has finite size $T-1$. Therefore, this model is called the One Infinite Buffer (OIB) model. In this model, each arriving job joins the shortest queue if there is a place free, otherwise the infinite size queue is joined.

* *Threshold Killing (TK):*
The last two bound models both have the same truncated state space as the TJ model, but the transitions to states outside the truncated state space are redirected in a different way. Let the truncated state space be defined by

$$M' = \{ (m_1,m_2) \mid 0 \le m_1 \le m_2 \text{ and } m_2 \le m_1+T \}$$

and let the transitions from the states (m_1,m_1+T) to (m_1-1,m_1+T), where $m_1 \ge 1$, be redirected to more attractive states (m_1-1,m_1+T-1). This means that, if a service completion at the shortest queue causes a too large difference between the longest and the shortest queue, then this service completion is accompanied by a destruction or killing of one job at the longest queue. Therefore, this model, which gives a *lower bound*, is called the Threshold Killing (TK) model.

* *Threshold Blocking (TB):*
For the last model the truncated state space is also defined by

$$M' = \{ (m_1,m_2) \mid 0 \le m_1 \le m_2 \text{ and } m_2 \le m_1+T \},$$

but for all $m_1 \ge 1$, the transition from the state (m_1,m_1+T) to state (m_1-1,m_1+T) is redirected to the state (m_1,m_1+T) itself, which is less attractive than (m_1-1,m_1+T). This means that, if the difference between the longest and the shortest queue has already reached its maximum value T, then a service completion at the shortest queue is not accompanied by a departure and the job in service has to be served once more. It is easily seen that (because of the memory-less property of the exponential service times) this is equivalent to letting the server at the shortest queue be blocked as long as the difference between the longest and the shortest queue is equal to its maximum value T. This model is an *upper bound* model and it is called the Threshold Blocking (TB) model.

This completes the description of the six, flexible bound models.

For each of the six bound models, we denote the resulting bound for L_w by $L_w^{tr}(T)$, where *tr* is FB, CB, TJ, OIB, TK or TB. The corresponding bound $W_{tr}(T)$ for the mean normalized waiting time W is defined by $W_{tr}(T)=L_w^{tr}(T)/(2\rho)$ (cf. (5.2)); by applying Little's formula, for each *tr* model, it follows that by this definition $W_{tr}(T)$ is precisely the mean normalized waiting time in the *tr* model.

All six, flexible bound models have been defined such that they are expected to give tight bounds, especially for large T, and that they can be solved numerically. For the FB model the state space is finite, by which the equilibrium distribution can be computed by applying some standard method, for example the method of successive substitutions. The CB model may be solved by the same method after having embedded the Markov process on the set of states (m_1,m_2) with $0 \le m_1 \le m_2 \le T$. The other four models may be solved by applying the matrix-geometric approach as described in Neuts [58].

It must be noted that, besides the bounds for L_w and W, the described bound models also lead to bounds for several other performance measures of the SSQS. The pairs of the set P defined by (5.16), and therefore also all pairs of the corresponding completed set P^* given by (5.17), can be proved to be precedence pairs for the whole class of cost functions $c(m)$ which satisfy the condition that $c(m) \leq c(n)$ for all $(m,n) \in P$ (this class appears to be the class of weak Schur convex functions; see Hordijk and Koole [44], p. 507). This implies that for each of the bound models, the total number of waiting jobs is *stochastically smaller/larger* than in the original SSQS; and similarly for the total number of jobs (waiting and in service) and for the length of the longest queue. For the FB, TK and TB model, even more monotonicity results hold. To show that these three models are bound models, in fact only the precedence pairs corresponding to the inequalities in (5.13) and (5.14) are needed. These inequalities can be proved to hold for all cost functions $c(m)$ which are non-decreasing in each component. As a result, for the FB, TK and TB model, we find among others that also the length of the shortest queue is stochastically smaller/larger than in the original SSQS; this is an extra interesting property, since, for the original SSQS, a stochastically smaller/larger distribution for the length of the shortest queue can be exploited to obtain a stochastically smaller/larger distribution for the normalized waiting time of the original SSQS.

Further, it is noted that since, as we already observed, the CB model with $T=1$ is identical to the $M \mid M \mid 2$ queueing system, some interesting monotonicity properties between the SSQS and the $M \mid M \mid 2$ queue are obtained as a side result. The above analysis has formally shown that the SSQS performs worse than (or at most equally good as) the corresponding $M \mid M \mid 2$ queue with respect to the total number of (waiting) jobs (which is stochastically larger than for the $M \mid M \mid 2$ queue) and the mean normalized waiting time. Intuitively, it is clear that this worse performance of the SSQS is caused by the existence of the so-called 'bad' states in the SSQS (i.e. by the situations in which one server is idle, while there still are jobs waiting at the other server).

Finally, it is noted that some of the described bound models have been studied in the literature; for the sake of completeness, we repeat that all six bound models described here, and all corresponding results, can be generalized to the case with general $N \geq 2$. The FB model, with $N=2$ servers, has been studied by Conolly [24] and the OIB model, with $N=2$ servers, has been analyzed by Rao and Posner [60]. Further, the TJ model has been studied by Gertsbakh [35] and Adan et al. [9] for the case with $N=2$ servers, and by Adan et al. [13] for the case with $N \geq 2$ servers; for the restricted case with $T=1$, the TJ model has been studied by Haight [38] for $N=2$ and by Disney and Mitchell [26], Elsayed and Bastani [28], Kao and Lin [48] and Zhao and Grassmann [83] for $N \geq 2$. In some of these papers, the model under consideration was mainly introduced to approximate the original shortest queue system (with possibly non-homogeneous servers). But, in none of them, it has been proved that (at least in the symmetric case) the model under consideration produces bounds for the main performance measures of the original shortest queue system. With the help of the precedence relation method, we have given one proof for all truncation models together. Recently, in [2], the precedence relation method has been used for proving that the TJ model and the TB model are bound models.

Remark 5.1. *(on the precedence relation method for proving monotonicity results)*

Up to now, we only have described the precedence relation method for the comparison between an original Markov cost model and a truncation model. However, if the second model is not a truncation model, but another Markov cost model with state space $M' \subset M$ ($M' = M$ is allowed), then the precedence relation method still may work for proving that the average costs g' in the second model constitute a bound for the average costs g in the original model. The only difference is that we obtain new conditions which must ensure that step 2 again can be performed in general terms by using induction with respect to t. It can be proved that $g' \leq g$, if the second model satisfies the following conditions:

(i) $c'(m) \leq c(m)$ for all $m \in M'$;

(ii) if $v_t(m) \leq v_t(n)$ for all $(m,n) \in P^*$ and some fixed $t \geq 0$, then

$$\sum_{\substack{n \in M' \\ q'_{m,n} > 0}} q'_{m,n}\, v'_t(m) \leq \sum_{\substack{n \in M \\ q_{m,n} > 0}} q_{m,n}\, v_t(m) \quad \text{for all } m \in M'.$$

The first condition states that the direct costs $c'(m)$ in the second model may not exceed the direct costs $c(m)$ in the original model. The second condition states that in the second model for each state $m \in M'$ the expected future costs must be smaller than in the original model; this condition is satisfied if, according to the precedence pairs derived in step 1, in the second model the states to which transitions are made from an arbitrary state $m \in M'$ are more attractive than the corresponding states in the original model. The conditions which are sufficient for showing that $g' \geq g$ are obtained by replacing the '$\leq$'-signs by '$\geq$'-signs in the above two conditions. The precedence relation method, in its more generalized form as described here, may be used, for example, for some of the monotonicity results mentioned at the end of Section 5.5, and for proving the intuitively clear result that the SSQS performs better than or at least equally good as the corresponding system consisting of two independent $M \mid M \mid 1$ queues (the SSQS has a stochastically smaller number of (waiting) jobs and a smaller or equal mean normalized waiting time).

Remark 5.2. *(on the precedence relation (5.12))*

That the precedence relation method leads to simple and clear proofs for monotonicity results, and that, furthermore, this method can be used for deriving monotonicity results, is mainly due to the introduction of the precedence relation (5.12) for the t-period costs in the original model. An interesting observation is that a similar relation has proved its usefulness for proving and deriving optimal strategies (optimal routing and service strategies, for example) in certain Markov decision problems; see Hordijk and Koole [43, 44], and the references therein.

5.4. The derivation of precedence pairs

In the previous section, we have seen that the first step of the precedence relation method, consisting of the derivation of precedence pairs for the original model, is essential for obtaining flexible bound models which might lead to tight bounds for the relevant performance measures of the original model. Therefore, this section will be devoted to a detailed treatment of this first step. First, for a given set P which is conjectured to consist of precedence pairs, we shall derive two conditions which are sufficient to confirm that P indeed consists of

precedence pairs. One of these two conditions concerns the direct costs $c(m)$ and can easily be verified. The other condition is more complicated. Among others, we shall show that its satisfiability is related to the existence of feasible flows in the corresponding *transportation problems*. The *theory of network flow problems* will be applied to obtain a simple, necessary and sufficient condition under which the feasible flows exist; this simple condition is hoped to lead to extra insight into the question which precedence pairs can be derived for a given Markov cost model.

Consider the original Markov cost model described in Section 5.2, and let P consist of pairs of states (m,n), $m,n \in M$, for which m is expected to have precedence over n; recall that the definition of the set P may be based on intuitive and/or numerical insight obtained by computing the t-period cost functions $v_t(m)$ for some small t. Then we must prove that

$$v_t(m) \leq v_t(n) \quad \text{for all } (m,n) \in P \text{ and } t \geq 0. \tag{5.18}$$

Since $v_1(m) = c(m)$ for all $m \in M$, it immediately follows that (5.18) can only hold if

$$c(m) \leq c(n) \quad \text{for all } (m,n) \in P, \tag{5.19}$$

i.e. if the pairs $(m,n) \in P$ are at least precedence pairs for the direct costs $c(m)$. By trying to prove (5.18) by using induction with respect to t, we shall obtain a second condition, which together with the necessary condition (5.19) is sufficient for proving (5.18).

Since $v_0(m) = 0$ for all $m \in M$, (5.18) trivially holds for $t = 0$. Now consider the induction step. Assume that (5.18) is satisfied for some, arbitrary, fixed $t \geq 0$. Then, by the recurrence relation stated in (5.8), (5.18) also holds for $t+1$, if

$$c(m) + \sum_{\substack{r \in M \\ q_{m,r} > 0}} q_{m,r}\, v_t(r) \leq c(n) + \sum_{\substack{r \in M \\ q_{n,r} > 0}} q_{n,r}\, v_t(r) \quad \text{for all } (m,n) \in P.$$

Assume that the necessary condition (5.19) is already satisfied, then the induction step can be made if, under the assumption that (5.18) holds for some fixed t, we can show that for all $(m,n) \in P$,

$$\sum_{\substack{r \in M \\ q_{m,r} > 0}} q_{m,r}\, v_t(r) \leq \sum_{\substack{r \in M \\ q_{n,r} > 0}} q_{n,r}\, v_t(r), \tag{5.20}$$

i.e. that for all $(m,n) \in P$, the expected future costs for state m are smaller than for state n. In the next paragraph, this condition is slightly refined.

The second condition states that the inequality in (5.20) must be shown to hold for all $(m,n) \in P$, under the assumption that $v_t(m) \leq v_t(n)$ for all $(m,n) \in P$, where $t \geq 0$ is fixed. However, before starting to prove (5.20) for all $(m,n) \in P$, by using the reflexivity and transitivity of the precedence relation as defined by (5.12), the available information may be completed to $v_t(m) \leq v_t(n)$ for all $(m,n) \in P^*$, where P^* is defined by

$$P^* = \{ (m,n) \mid m = n,\ (m,n) \in P \text{ or for some } l \geq 1 \text{ there exist states } r^1, \ldots, r^l \text{ such that } (m,r^1), (r^1,r^2), \ldots, (r^l,n) \in P \};$$

P^* is called *the completed set corresponding to* P. If (5.18) has been proved for all $(m,n) \in P$, then we also know that m has precedence over n for all $(m,n) \in P^*$. Note that choosing another P for which the corresponding completed set P^* is the same, leads to the same result.

Therefore, to minimize the work to be done, it is sensible to reduce P as much as possible under the restriction that P^* remains the same. The following lemma gives the precise conditions under which it can be guaranteed that the set P indeed consists of precedence pairs.

Lemma 5.1.

A set P of pairs of states (m,n) of the state space M, and also the corresponding completed set P^*, can be guaranteed to consist of precedence pairs if the following two conditions are satisfied:

(i) $c(m) \leq c(n)$ *for all* $(m,n) \in P$;

(ii) if $v_t(m) \leq v_t(n)$ *for all* $(m,n) \in P^*$ *and some fixed* $t \geq 0$, *then (5.20) holds for all* $(m,n) \in P$.

In the remainder of this section, it is assumed that some set P consisting of pairs (m,n), $m,n \in M$, is given. Suppose that we want to apply Lemma 5.1 to prove that this set P and also the corresponding completed set P^* consist of precedence pairs. If the direct costs $c(m)$ are explicitly defined, then it can easily be verified whether condition (i) is satisfied; if they are not explicitly defined, then condition (i) imposes a restriction on the cost function $c(m)$. The verification of condition (ii) requires much more work. In the remaining part of this section, we will derive two stronger conditions for condition (ii), which both can be verified rather easily.

We first derive a stronger condition for condition (ii) of Lemma 5.1 for the special, but relevant case, in which the original Markov cost model, for which we want to derive the precedence pairs, satisfies the following property:

> for all states $m \in M$ the outgoing transitions are caused by the same events $e \in E$ (such as job arrivals and service completions), where E is a finite or countable set, and for all states $m \in M$ these events $e \in E$ occur with the same probabilities q_e.

It is noted that this property is satisfied by many queueing systems, among which the SSQS. If this property is satisfied, then the expected costs as given on both sides of inequality (5.20), may be rewritten as

$$\sum_{\substack{r \in M \\ q_{m,r} > 0}} q_{m,r}\, v_t(r) = \sum_{e \in E} q_e\, v_t(r_e(m)), \quad m \in M,$$

where $r_e(m)$ denotes the state to which a transition is made if the Markov process is in state m and event e occurs (it is noted that for a given state $m \in M$ and different events $e \in M$, the states $r_e(m)$ may be the same). Consequently, condition (ii) of Lemma 5.1 simplifies to (let P be a set consisting of pairs of states (m,n), $m,n \in M$, and let P^* be the corresponding completed set):

> if $v_t(m) \leq v_t(n)$ for all $(m,n) \in P^*$ and some fixed $t \geq 0$, then
>
> $$\sum_{e \in E} q_e\, v_t(r_e(m)) \leq \sum_{e \in E} q_e\, v_t(r_e(n)) \quad \text{for all } (m,n) \in P.$$

This condition obviously is satisfied, if for all $(m,n) \in P$ the states $r_e(m)$ to which transitions are made from state m are more attractive than the corresponding states $r_e(n)$ to which

transitions are made from state n, i.e. if $(r_e(m)),r_e(n)) \in P^*$ for all $e \in E$ and $(m,n) \in P$. This results in the following lemma.

Lemma 5.2. *(on condition (ii) of Lemma 5.1)*

Condition (ii) of Lemma 5.1 is satisfied if:

a. there is a discrete distribution $\{q_e\}_{e \in E}$ on a finite or denumerable set of events E such that in the original Markov cost model, for all states $m \in M$ transitions are made from state m to certain states $r_e(m)$ according to the probabilities q_e, $e \in E$;

b. $(r_e(m),r_e(n)) \in P^*$ for all $e \in E$ and $(m,n) \in P$.

Example 5.1: The symmetric shortest queue system *(continued)*

We apply the Lemmas 5.1 and 5.2, to prove the inequalities stated in (5.13)-(5.15), i.e. to prove that the set P given by (5.16), and also the corresponding completed set P^* given by (5.17), consist of precedence pairs. It is easily verified that all pairs $(m,n) \in P$ are precedence pairs for the cost function $c(m)$ defined by (5.4); so, condition (i) of Lemma 5.1 is satisfied. Condition (ii) of Lemma 5.1 may be shown to be satisfied by exploiting Lemma 5.2.

In the two-dimensional SSQS, for each state $m \in M$, the outgoing transitions are caused by a job arrival, a service completion at the shortest queue or a service completion at the longest queue, and they occur with probabilities λ, μ and μ, respectively. Let $E=\{0,1,2\}$, $q_0=\lambda$, $q_1=\mu$ and $q_2=\mu$. Then we have the property that from each state $m \in M$ a transition is made to state

$$r_0(m) = r_0(m_1,m_2) = \begin{cases} (m_1+1,m_2) & \text{if } 0 \le m_1 < m_2; \\ (m_1,m_1+1) & \text{if } 0 \le m_1 = m_2 \end{cases}$$

with probability λ, to state

$$r_1(m) = r_1(m_1,m_2) = (\max\{m_1-1,0\},m_2)$$

with probability μ, and to state

$$r_2(m) = r_2(m_1,m_2) = \begin{cases} (m_1,m_2-1) & \text{if } 0 \le m_1 < m_2; \\ (\max\{m_1-1,0\},m_1) & \text{if } 0 \le m_1 = m_2 \end{cases}$$

with probability μ. This means that we satisfy condition a. of Lemma 5.2, and thus, to show that condition (ii) of Lemma 5.1 is satisfied, it further suffices to prove that

$$(r_e(m),r_e(n)) \in P^* \quad \text{for all } e \in E \text{ and } (m,n) \in P. \tag{5.21}$$

Let us first prove (5.21) for the pairs $(m,n) \in P$ of the first type, i.e. for the pairs (m,n) with $m=(m_1,m_2)$, $n=(m_1+1,m_2)$, $m_2 \ge 1$ and $0 \le m_1 \le m_2-1$. In this case, we find

$$r_0(m) = (m_1+1,m_2), \qquad r_0(n) = \begin{cases} (m_1+2,m_2) & \text{if } m_1 < m_2-1; \\ (m_1+1,m_2+1) & \text{if } m_1 = m_2-1, \end{cases}$$

$$r_1(m) = (\max\{m_1-1,0\},m_2), \quad r_1(n) = (m_1,m_2),$$

$$r_2(m) = (m_1,m_2-1), \qquad r_2(n) = \begin{cases} (m_1+1,m_2-1) & \text{if } m_1 < m_2-1; \\ (m_1,m_2) & \text{if } m_1 = m_2-1, \end{cases}$$

and it is obvious that $(r_e(m), r_e(n)) \in P^*$ for all three elements $e \in E$. The proof of (5.21) for the pairs $(m,n) \in P$ of the two other types is similar and is left to the reader.

The example shows that Lemma 5.2 may lead to a rather simple proof of the satisfiability of condition (ii) of Lemma 5.1. However, Lemma 5.2 cannot be applied, if the original Markov cost model does not have the property stated in condition a. Therefore, we now return to the question how the satisfiability of condition (ii) of Lemma 5.1 can be proved in general.

Let $t \geq 0$ be fixed, assume that $v_t(m') \leq v_t(n')$ for all $(m',n') \in P^*$, and let (m,n) be an arbitrary, fixed element of P. Then condition (ii) of Lemma 5.1 states that it must be proved that inequality (5.20) holds. The only way to prove this inequality seems to consist of combining $v_t(r)$'s for states r for which $q_{m,r} > 0$ with $v_t(\hat{r})$'s for states $\hat{r}$ for which $q_{n,\hat{r}} > 0$. For such a combination we know that $v_t(r) - v_t(\hat{r}) \leq 0$ if $(r,\hat{r}) \in P^*$; if $(r,\hat{r}) \notin P^*$, then nothing is known about $v_t(r) - v_t(\hat{r})$. Each combination should get some weight $x_{r,\hat{r}}$. Further, for each r with $q_{m,r} > 0$ the total weight $\Sigma_{\hat{r}}\, x_{r,\hat{r}}$ should be equal to $q_{m,r}$, and for each $\hat{r}$ with $q_{n,\hat{r}} > 0$ the total weight $\Sigma_r\, x_{r,\hat{r}}$ should be equal to $q_{n,\hat{r}}$. What we in fact obtain is the following *transportation problem*, which is known to be a special case of a *feasible flow problem* (see for example Ahuja et al. [14], p. 7 and 169).

Let V_1 denote all states r for which $q_{m,r} > 0$, give each state a weight $a_r = q_{m,r}$, and renumber the states from $1, \ldots, |V_1|$. Let V_2 denote all states r for which $q_{n,r} > 0$, give each state a weight $b_r = q_{n,r}$, and renumber these states from $|V_1|+1, \ldots, |V_1|+|V_2|$. The states in V_1 are called the *supply nodes* and the states in V_2 are the *demand nodes*. Note that a state can be in V_1 as well as in V_2 (this keeps the notations simple); the renumbering indicates that in such a case the state in V_1 will be considered to be different from the same state in V_2. Since the total supply $\Sigma_{i \in V_1}\, a_i$ is equal to the total demand $\Sigma_{j \in V_2}\, b_j$ (they both are equal to 1), there exist nonnegative variables $x_{i,j}$, with $i \in V_1$ and $j \in V_2$, which satisfy

$$\sum_{j \in V_2} x_{i,j} = a_i \quad \text{for all } i \in V_1\,, \qquad \sum_{i \in V_1} x_{i,j} = b_j \quad \text{for all } j \in V_2\,. \tag{5.22}$$

Each nonnegative solution $\{x_{i,j}\}$ of (5.22) is called an allocation or a *flow*. For each flow, we obtain that

$$\begin{aligned} \sum_{\substack{r \in M \\ q_{m,r} > 0}} q_{m,r}\, v_t(r) - \sum_{\substack{r \in M \\ q_{n,r} > 0}} q_{n,r}\, v_t(r) &= \sum_{i \in V_1} a_i\, v_t(i) - \sum_{j \in V_2} b_j\, v_t(j) \\ &= \sum_{i \in V_1} \sum_{j \in V_2} x_{i,j}\, v_t(i) - \sum_{j \in V_2} \sum_{i \in V_1} x_{i,j}\, v_t(j) \\ &= \sum_{i \in V_1} \sum_{j \in V_2} x_{i,j}\, [v_t(i) - v_t(j)]\,. \end{aligned} \tag{5.23}$$

For each flow, the variable $x_{i,j}$ denotes the weight for the difference $v_t(i) - v_t(j)$, which is only known to be ≤ 0 for pairs $(i,j) \in P^*$. Define the set of arcs A by

$$A = \{\, (i,j) \mid i \in V_1,\ j \in V_2 \text{ and } (i,j) \in P^* \,\}\,.$$

If we can find a nonnegative solution $\{x_{i,j}\}$ for which $x_{i,j} = 0$ *for all arcs* (i,j) *which are not in the set* A, then the inequality stated in (5.20) can easily be proved (use (5.23)):

$$\sum_{\substack{r\in M\\ q_{m,r}>0}} q_{m,r}\, v_t(r) - \sum_{\substack{r\in M\\ q_{n,r}>0}} q_{n,r}\, v_t(r) = \sum_{(i,j)\in A} x_{i,j}\,[v_t(i)-v_t(j)] \le 0\,;$$

if such a solution cannot be found, then (5.20) cannot be proved along this way. This completes the proof of the following lemma.

Lemma 5.3. *(on condition (ii) of Lemma 5.1)*

Let $t \ge 0$, $(m,n) \in P$, and further suppose that $v_t(m) \le v_t(n)$ for all $(m,n) \in P^*$. Then inequality (5.20) holds, if for the corresponding transportation problem there exists a nonnegative solution $\{x_{i,j}\}_{(i,j)\in A}$ which satisfies the following equations:

$$\sum_{\substack{j\in V_2\\ (i,j)\in A}} x_{i,j} = a_i \quad \text{for all } i \in V_1\,, \qquad \sum_{\substack{i\in V_1\\ (i,j)\in A}} x_{i,j} = b_j \quad \text{for all } j \in V_2\,. \tag{5.24}$$

A nonnegative solution satisfying these equations will be called a feasible flow.

Example 5.1: The symmetric shortest queue system *(continued)*

We now apply the Lemmas 5.1 and 5.3 to prove that the set P given by (5.16) and the corresponding completed set P^* given by (5.17) consist of precedence pairs. We have already seen that all pairs $(m,n) \in P$ are precedence pairs for the cost function $c(m)$ defined by (5.4), and thus, according to Lemma 5.1, it remains to prove (5.20) for all $(m,n) \in P$ and some, arbitrary, fixed $t \ge 0$, while it is given that $v_t(m) \le v_t(n)$ for all $(m,n) \in P^*$.

Let us first prove (5.20) for the pairs $(m,n) \in P$ of the first type, i.e. for the pairs (m,n) with $m = (m_1, m_2)$, $n = (m_1+1, m_2)$, $m_2 \ge 1$ and $0 \le m_1 \le m_2 - 1$. To prove (5.20), we have to compare the outgoing transitions for the states m and n. Because of the homogeneity in the Markov process (see Figure 5.1), we can partition the state space M into a finite number of subsets consisting of states which all have the same outgoing transitions:

$$M_I = \{\,(m_1,m_2) \mid 0 < m_1 < m_2\,\} \qquad \text{(interior)}$$

$$M_V = \{\,(0,m_2) \mid m_2 > 0\,\} \qquad \text{(vertical boundary)}$$

$$M_D = \{\,(m_1,m_1) \mid m_1 > 0\,\} \qquad \text{(diagonal)}$$

$$M_O = \{\,(0,0)\,\} \qquad \text{(origin)}$$

To prove (5.20) for all pairs $(m,n) \in P$ of the first type, by Lemma 5.3, it suffices to define the corresponding transportation problems and to show that they have feasible flows. Due to the homogeneity, we only have to distinguish 4 situations (see Figure 5.1):

1. $m \in M_I$ and $n \in M_I$ (i.e. $m_2 \ge 3$ and $1 \le m_1 \le m_2 - 2$):

   In this case, the supplying and demanding nodes are given by

   $$V_1 = \{\,1 = (m_1-1, m_2),\ 2 = (m_1, m_2-1),\ 3 = (m_1+1, m_2)\,\},$$

   $$V_2 = \{\,4 = (m_1, m_2),\ 5 = (m_1+1, m_2-1),\ 6 = (m_1+2, m_2)\,\},$$

   the supplies and demands are given by

   $$(a_1, a_2, a_3) = (b_4, b_5, b_6) = (\mu, \mu, \lambda),$$

   and for the set of arcs we find

$A = \{ (1,4), (1,6), (2,4), (2,5), (2,6), (3,6) \}$.

For this instance, the solution $\{x_{i,j}\}_{(i,j)\in A}$ with

$x_{1,4} = x_{2,5} = \mu$, $x_{3,6} = \lambda$, $x_{1,6} = x_{2,4} = x_{2,6} = 0$,

appears to be a feasible flow, which completes the proof for this situation.

2. $m \in M_V$ and $n \in M_I$ (i.e. $m_2 \geq 2$ and $m_1 = 0$):

   In this case, we have

   $V_1 = \{ 1=(0,m_2), 2=(0,m_2-1), 3=(1,m_2) \}$,

   $V_2 = \{ 4=(0,m_2), 5=(1,m_2-1), 6=(2,m_2) \}$,

   and the suplies a_i, demands b_j and set of arcs A are the same as for the previous case. As a consequence, the flow of the previous case is also feasible for this case.

3. $m \in M_I$ and $n \in M_D$ (i.e. $m_2 \geq 2$ and $m_1 = m_2 - 1$):

   In this situation, we have

   $V_1 = \{ 1=(m_2-2,m_2), 2=(m_2-1,m_2-1), 3=(m_2,m_2) \}$,

   $V_2 = \{ 4=(m_2-1,m_2), 5=(m_2,m_2+1) \}$,

   $(a_1,a_2,a_3) = (\mu,\mu,\lambda)$, $(b_4,b_5) = (2\mu,\lambda)$,

   $A = \{ (1,4), (1,5), (2,4), (2,5), (3,5) \}$,

   and a feasible flow is given by the variables

   $x_{1,4} = x_{2,4} = \mu$, $x_{3,5} = \lambda$, $x_{1,5} = x_{2,5} = 0$.

4. $m \in M_V$ and $n \in M_D$ (i.e. $m_2 = 1$ and $m_1 = 0$):

   In this situation, we have

   $V_1 = \{ 1=(0,1), 2=(0,0), 3=(1,1) \}$, $V_2 = \{ 4=(0,1), 5=(1,2) \}$,

   and the remaining variables for the transportation problem are the same as in case 3, and consequently the same feasible flow can be given.

This completes the proof of (5.20) for all pairs $(m,n) \in P$ of the first type. The work to be done for the pairs of the other two types is similar and is left to the reader.

From the example we learn the following. First of all, also Lemma 5.3 can easily be applied to prove the satisfiability of condition (ii) of Lemma 5.1. Secondly, in case one has homogeneity in a Markov model (which usually is the case for queueing models), then, to prove (5.20) for a large or infinite set P, it possibly suffices to solve (i.e. to find feasible flows for) only a small number of corresponding transportation problems.

When trying to prove that condition (ii) of Lemma 5.1 is satisfied, Lemma 5.3 is useful, if all required feasible flows can be found. However, if for some transportation problem, the required feasible flow cannot be found, then it is desirable to have a tool for showing why such a feasible flow cannot be found. For that purpose, we can present a simple condition which is necessary and sufficient for the existence of a feasible flow.

Consider the transportation problem mentioned in Lemma 5.3. For all subsets $U \subset V_1$ consisting of supply nodes, we let the set $\delta^+(U) = \{ j \in V_2 \mid (i,j) \in A \text{ for some } i \in U \}$ denote

the demand nodes to which the supplies of the nodes $i \in U$ may be transported according to the set of arcs A. It is obvious that for the existence of a feasible flow, it is required that, for each $U \subset V_1$, the total supply $\Sigma_{i \in U} a_i$ of the nodes of U does not exceed the maximum amount $\Sigma_{j \in \delta^+(U)} b_j$ that can be received by the nodes of $\delta^+(U)$. This shows the necessity of the condition stated in Lemma 5.4. From this lemma, it follows that this condition is also sufficient. The lemma may be proved by first transforming the transportation problem to a *maximum-flow problem* and next applying the well-known *max-flow min-cut theorem* (see [14], p. 185), or alternatively by directly applying Theorem 6.12 of Ahuja et al. [14]. Note that an equivalent, necessary and sufficient condition can be obtained by interchanging the roles of the supply nodes and the demand nodes.

Lemma 5.4.

There exists a feasible flow for the transportation problem mentioned in Lemma 5.3 if and only if for all $U \subset V_1$,

$$\sum_{i \in U} a_i \leq \sum_{j \in \delta^+(U)} b_j , \tag{5.25}$$

where $\delta^+(U) = \{ j \in V_2 \mid (i,j) \in A \text{ for some } i \in U \}$.

Example 5.1: The symmetric shortest queue system *(continued)*

It is easily verified that condition (5.25) is satisfied for the transportation problems which are obtained for the inequalities stated in (5.13)-(5.15), which we have proved earlier in this section by giving an example of a feasible flow for each transportation problem.

Example 5.2: A homogeneous, nearest-neighboring random walk

Apart from showing why in some cases a required feasible flow does not exist, Lemma 5.4 is also useful for deriving conditions for the existence of certain precedence pairs for a two-dimensional, homogeneous, nearest-neighboring random walk as depicted in Figure 1.6. For many queueing systems which can be modeled as such a random walk, the coordinates of the states (m_1,m_2) represent queue lengths, and it seems reasonable to expect that each state m has precedence over all states which are larger or at least equal in each coordinate. Applying Lemma 5.4 immediately results in the conditions under which this is true.

Suppose we want to prove that the set P^* given by

$$P^* = \{ (m,n) \mid m = (m_1,m_2) \in M , \; n = (n_1,n_2) \in M , \; m_1 \leq n_1 \text{ and } m_2 \leq n_2 \}$$

consists of precedence pairs; for this P^*, the set P may be taken equal to

$$P = \{ (m , m+e_1) \mid m \in M \} \cup \{ (m , m+e_2) \mid m \in M \} ,$$

where $e_1 = (1,0)$ and $e_2 = (0,1)$. Let us investigate which conditions have to be satisfied for proving that P^* indeed consists of precedence pairs.

According to Lemma 5.1, in the first place, all pairs $(m,n) \in P^*$ have to be precedence pairs for the cost function $c(m)$. Next, we have to prove (5.20) for all $(m,n) \in P$ and some fixed $t \geq 0$, where it is known that $v_t(m) \leq v_t(n)$ for all $(m,n) \in P^*$. Suppose we want to prove (5.20) for some $m \in M$ and $n = m+e_1 \in M$. We have to distinguish 4 situations. If m is an element of the interior or the horizontal boundary, then all inequalities given by (5.25) reduce to

trivialities, which implies that the required feasible flows exist, and (5.20) holds. If m is an element of the vertical boundary, then defining the corresponding transportation problem and applying the Lemmas 5.3 and 5.4 shows that (5.20) can be guaranteed to hold if the following inequalities are satisfied:

$$\nu_{1,1} \le q_{0,1}+q_{1,1}\,, \quad \nu_{0,1}+\nu_{1,1} \le q_{-1,1}+q_{0,1}+q_{1,1}\,,$$

$$\nu_{1,0}+\nu_{1,1} \le q_{0,0}+q_{0,1}+q_{1,0}+q_{1,1}\,, \quad \nu_{0,1}+\nu_{1,0}+\nu_{1,1} \le q_{-1,1}+q_{0,0}+q_{0,1}+q_{1,0}+q_{1,1}\,,$$

$$\nu_{0,0}+\nu_{0,1}+\nu_{1,0}+\nu_{1,1} \le q_{-1,0}+q_{-1,1}+q_{0,0}+q_{0,1}+q_{1,0}+q_{1,1}\,,$$

$$\nu_{1,-1}+\nu_{1,0}+\nu_{1,1} \le q_{0,-1}+q_{0,0}+q_{0,1}+q_{1,-1}+q_{1,0}+q_{1,1}\,,$$

$$\nu_{0,1}+\nu_{1,-1}+\nu_{1,0}+\nu_{1,1} \le q_{-1,1}+q_{0,-1}+q_{0,0}+q_{0,1}+q_{1,-1}+q_{1,0}+q_{1,1}\,,$$

$$\nu_{0,0}+\nu_{0,1}+\nu_{1,-1}+\nu_{1,0}+\nu_{1,1} \le q_{-1,0}+q_{-1,1}+q_{0,-1}+q_{0,0}+q_{0,1}+q_{1,-1}+q_{1,0}+q_{1,1}\,.$$

These inequalities state that it may not be much easier to go to the north and/or east when starting in a state at the vertical axis than when starting in a state in the interior. Similar inequalities for the rates $o_{i,j}$ and $h_{i,j}$ are obtained for the case that m is equal to the origin. Further, similar conditions are required for proving (5.18) for some $m \in M$ and $n = m+e_2 \in M$.

All conditions are easily seen to be satisfied if we have a random walk with the *projection property* (as defined in Chapter 2). This explicit result can easily be generalized to the N-dimensional case (for the proof, it is advised to exploit Lemma 5.2). Since the projection property is satisfied for several queueing systems, this confirms that in many cases precedence pairs of the type $(m, m+e_i)$ can be derived.

5.5. On the quality of flexible bound models

The previous sections have been devoted to the precedence relation method, which mainly has been developed for the derivation of flexible bound models for a given, original Markov cost model. In this section, we shall focus on the usefulness of the produced flexible bound models for the determination of the average costs in the original model.

The flexible bound models derived by the precedence relation method are truncation models with the following two properties:

(i) They produce lower and upper bounds for the average costs g in the original Markov cost model;

(ii) They have state spaces depending on one or more truncation parameters which enable us to approximate the original model as accurately as desired.

However, in order to be appropriate for the determination of the average costs g in the original model, the flexible bound/truncation models must also have the following two properties:

(iii) They are *solvable* (analytically or numerically);

(iv) They lead to *tight* bounds (compared to the computational effort which is required for solving the bound models).

These two properties are further discussed below.

Property (iii) states that a truncation model must be solvable, which means that it must be possible to determine its equilibrium distribution $\{p'_m\}$, and thus, by using (5.5), also the

corresponding average costs g', in an analytical way or by an efficient numerical method. In the most favorable case, a truncation model can be solved purely analytically, and then it might be possible that the average costs g in the original model can be determined also purely analytically by writing them as a limit of the average costs g' in the truncation model. However, in general, we will already be satisfied if a truncation model can be solved sufficiently efficient by some other method; for example, by the matrix-geometric approach or by a standard numerical technique for the determination of the equilibrium distribution of a Markov process with a finite state space. In that case, the determination of the bound produced by the truncation model requires a certain computational effort, and the corresponding computation times usually will strongly depend on the values of the truncation parameters. It is obvious that then property (iv) becomes important.

The higher the computational effort required to solve a truncation model, the more important it is that the truncation model leads to tight bounds for already small sizes of the truncated state space (which is determined by the truncation parameters). This is reflected by property (iv), which states that a truncation model must lead to a tight bound compared to the computational effort required to solve the model. The tightness of a truncation model will mainly be determined by:

- the fraction of periods that a redirection occurs in the truncation model;
- the impact of each redirection.

From this observation, we obtain a couple of rules of thumb for how a truncation model should be defined. The fraction of periods that a redirection occurs may be kept low by defining the truncated state space such that it contains the states where most of the probability mass is present. Further, the way in which the transitions to states outside the truncated state space are redirected influences the impact of the redirections. The impact of a redirection from a state n outside the truncated state space to a state n' inside the truncated state space consists of the *direct* impact on the costs, for which the difference between $c(n)$ and $c(n')$ is an appropriate measure, and the *indirect* impact on the costs in later periods. For a particular model, one usually will be able to estimate the impact of a redirection by using its physical interpretation. In general we can say that it is sensible to redirect to a state n' which is rather close to n according to the precedence pairs and for which the difference of $c(n')$ with respect to $c(n)$ is as small as possible.

The properties (iii) and (iv) determine the *quality* of a flexible bound model. Models of a sufficiently high quality can be used to determine the average costs g of the original model. In general, an appropriate way for the determination of the average costs g, within a given absolute or relative accuracy, is constituted by solving a lower and an upper bound model for increasing sizes of the truncated state space; here, the mean of a lower bound and an upper bound and half of their difference may serve as an approximation for g and an upper bound for the absolute inaccuracy of this approximation. To minimize the total computational effort, it is important to select the lower bound model and the upper bound model which lead to the tightest bounds compared to the computational efforts needed for solving them. For this selection, it is required to have some information on the tightness of the different bound models. This information may be obtained from: (i) intuitive insight, (ii) numerical results for small instances, and (iii) monotonicity results between the different bound models, which also can be derived by using the precedence relation method.

Below, we shall investigate the quality of the six, solvable, flexible bound models, which we have derived for the symmetric shortest queue system in Section 5.3. Numerical results will indicate which ordering there exists between them with respect to the tightness. At the end of this section, in Remark 5.3, we shall explain that a large part of this ordering can be proved by using the predence relation method, and that this method can also be used to prove some other monotonicity results that are suggested by the numerical results.

Example 5.1: The symmetric shortest queue model *(continued)*

In Section 5.3, by using the precedence relation method, we have derived six flexible truncation models which lead to bounds for the mean L_w of the total number of waiting jobs and the mean normalized waiting time W for the SSQS with two servers. We now determine the bounds $W_{tr}(T)$, and compare the gaps between them and W itself for varying values of the threshold (truncation) parameter T and the workload ρ. It will also be investigated to what extent these gaps can be explained by the fraction of time that a redirection occurs and the impact of each redirection. Since for the same value of the threshold (truncation) parameter T, all six bound models require comparable computation times for solving them, the results obtained for the tightness of the bounds $W_{tr}(T)$, can be used to establish which bound models are the best.

Before we explain the numerical results presented in the Tables 5.1 and 5.2, we have to introduce some notation. For each bound model, the fraction of periods that a redirection occurs is denoted by p_{rd} and the direct impact of a redirection is denoted by i_{rd}. For the FB model, p_{rd} is equal to the fraction of time $p^{FB}_{(T,T)}$ that we are in state (T,T), multiplied by the probability λ for making the transition from (T,T) to $(T,T+1)$ and back to (T,T). The fractions for the other models are derived similarly. For the FB model, and also for the TK model, each redirection physically means that one job is removed from the system, and the direct impact is said to be equal to $i_{rd}=-1$. For the TB model each redirection is equivalent to keeping one job extra in the system, and $i_{rd}=1$. For these models, each redirection also has a considerable indirect impact, but for this indirect impact it is difficult to give a good measure. For the CB, the TJ and the OIB model, each redirection physically means that a job is moved from one queue to the other queue, and obviously has no direct impact, i.e. $i_{rd}=0$. The quality of these models is determined by the indirect impact of the redirections, which make it less or more easy to visit one of the 'bad' states $(0,m_2)$, $m_2 \geq 2$.

In Table 5.1, numerical results are presented for a series of examples with fixed $\rho=0.9$ and varying values of the threshold parameter T. For this value of ρ, the mean normalized waiting time of the original model is equal to $W=4.475$. This value may be computed by solving a lower bound model and an upper bound model for a large value of T (or e.g. by using the compensation approach). For each bound model, in the first column the absolute difference $\Delta_{abs}=W_{tr}(T)-W$ is given. To explain the value of the absolute difference Δ_{abs}, in the second and third column the values of p_{rd} and i_{rd} are denoted. In case for some bound model each redirection would have the same (direct and indirect) impact on the average number of jobs in the system, and therefore also on the mean normalized waiting time, Δ_{abs} would be equal to p_{rd} multiplied by some constant c_{rd}; $c_{rd}=\Delta_{abs}/p_{rd}$ represents the average impact of a redirection and is denoted in the fourth column. In Table 5.2, the same information is given for a series of examples with varying values of ρ and fixed $T=3$. Note that, due to the destruction of capacity, the TB model may be non-ergodic for large values of ρ and/or

T	FB model				CB model				TJ model			
	Δ_{abs}	p_{rd}	i_{rd}	c_{rd}	Δ_{abs}	p_{rd}	i_{rd}	c_{rd}	Δ_{abs}	p_{rd}	i_{rd}	c_{rd}
1	-4.475	0.173613	-1	-25.8	-0.212	0.106284	0	-1.99	-0.212	0.106284	0	-1.99
2	-4.096	0.092218	-1	-44.4	-0.093	0.088178	0	-1.05	-0.093	0.023205	0	-4.00
3	-3.682	0.057333	-1	-64.2	-0.040	0.072213	0	-0.558	-0.037	0.006109	0	-6.01
4	-3.285	0.038810	-1	-84.6	-0.018	0.058784	0	-0.298	-0.013	0.001679	0	-8.02
5	-2.916	0.027668	-1	-105	-0.008	0.047722	0	-0.159	-0.005	0.000467	0	-10.0
6	-2.577	0.020404	-1	-126	-0.003	0.038694	0	-0.0854	-0.002	0.000130	0	-12.0
7	-2.269	0.015403	-1	-147	-0.001	0.031357	0	-0.0458	-0.001	0.000036	0	-14.0
8	-1.991	0.011822	-1	-168	-0.001	0.025404	0	-0.0246	-0.000	0.000010	0	-15.8

T	OIB model				TK model				TB model			
	Δ_{abs}	p_{rd}	i_{rd}	c_{rd}	Δ_{abs}	p_{rd}	i_{rd}	c_{rd}	Δ_{abs}	p_{rd}	i_{rd}	c_{rd}
1	∞	–	0	–	-3.420	0.078785	-1	-43.4	∞	–	1	–
2	18.000	0.234974	0	76.6	-1.844	0.022728	-1	-81.1	25.719	0.041851	1	615
3	2.261	0.163541	0	13.8	-0.761	0.006855	-1	-111	1.500	0.009438	1	159
4	0.738	0.122032	0	6.04	-0.263	0.002012	-1	-131	0.349	0.002425	1	144
5	0.288	0.094547	0	3.05	-0.083	0.000575	-1	-145	0.096	0.000656	1	147
6	0.121	0.074774	0	1.61	-0.025	0.000162	-1	-156	0.028	0.000181	1	154
7	0.052	0.059805	0	0.871	-0.008	0.000046	-1	-167	0.008	0.000050	1	162
8	0.023	0.048122	0	0.474	-0.002	0.000013	-1	-178	0.002	0.000014	1	171

Table 5.1. Overview of the quality of the six bound models for the symmetric shortest queue system with 2 servers and workload $\rho = 0.9$. For this workload the mean normalized waiting time of the original model is equal to $W = 4.475$.

small values of T, in which case $L_w^{TB}(T) = W_{TB}(T) = \Delta_{abs} = \infty$; similarly for the OIB model, where the 'bad' states are more often visited than in the original model.

The numerical results show that for each bound model the quality of the bound for W, which is measured by $|\Delta_{abs}|$, increases for increasing values of T (see Table 5.1) and decreases for increasing values of ρ (see Table 5.2). This behavior is partly explained by the behavior of p_{rd}. Further, studying the values of c_{rd}, shows that in general $|c_{rd}|$ is larger for the models for which the direct impact i_{rd} is equal to -1 or 1 than for the models for which $i_{rd} = 0$. This confirms the thought that both the fraction of periods in which a redirection occurs and the direct impact of a redirection will give an indication for the quality of a bound model. The diversity of the values of c_{rd} for varying values of T and ρ shows that p_{rd} and i_{rd} surely do not give more than an indication.

The numerical results also indicate that for a fixed T, the TJ model always produces the best lower bound for W and the TB almost always produces the best upper bound (for the case $T = 2$ in Table 5.1, the bound produced by the OIB model is slightly better than the bound produced by the TB model, but they both are worse). Since for the same value of T, all six bound models require comparable computation times for being solved, this implies that the TJ model and the TB model are the most appropriate models for being used in an algorithm for

ρ	W	FB model				CB model				TJ model			
		Δ_{abs}	p_{rd}	i_{rd}	c_{rd}	Δ_{abs}	p_{rd}	i_{rd}	c_{rd}	Δ_{abs}	p_{rd}	i_{rd}	c_{rd}
0.2	0.066	-0.000	0.000015	-1	-21.6	-0.000	0.000008	0	-3.89	-0.000	0.000002	0	-16.4
0.4	0.259	-0.019	0.001121	-1	-16.9	-0.001	0.000666	0	-1.60	-0.001	0.000097	0	-9.13
0.6	0.682	-0.198	0.010084	-1	-19.7	-0.007	0.007573	0	-0.929	-0.006	0.000873	0	-6.96
0.8	1.956	-1.246	0.036737	-1	-33.9	-0.025	0.038345	0	-0.641	-0.022	0.003581	0	-6.15
0.9	4.475	-3.682	0.057333	-1	-64.2	-0.040	0.072213	0	-0.558	-0.037	0.006109	0	-6.01
0.95	9.487	-8.661	0.069069	-1	-125	-0.051	0.095903	0	-0.527	-0.046	0.007733	0	-5.99
0.98	24.494	-23.653	0.076536	-1	-309	-0.057	0.112668	0	-0.510	-0.053	0.008831	0	-5.99
0.99	49.497	-48.650	0.079071	-1	-615	-0.060	0.118715	0	-0.505	-0.055	0.009218	0	-6.02

ρ	W	OIB model				TK model				TB model			
		Δ_{abs}	p_{rd}	i_{rd}	c_{rd}	Δ_{abs}	p_{rd}	i_{rd}	c_{rd}	Δ_{abs}	p_{rd}	i_{rd}	c_{rd}
0.2	0.066	0.000	0.000003	0	4.45	-0.000	0.000002	-1	-19.2	0.000	0.000003	1	3.20
0.4	0.259	0.001	0.000605	0	2.33	-0.001	0.000100	-1	-14.1	0.001	0.000164	1	5.65
0.6	0.682	0.023	0.010734	0	2.13	-0.016	0.000932	-1	-16.0	0.017	0.001403	1	11.8
0.8	1.956	0.309	0.075453	0	4.10	-0.153	0.003956	-1	-38.8	0.227	0.005577	1	40.6
0.9	4.475	2.261	0.163541	0	13.8	-0.761	0.006855	-1	-111	1.500	0.009438	1	159
0.95	9.487	23.842	0.232074	0	103	-2.909	0.008739	-1	-333	10.345	0.011920	1	868
0.98	24.494	∞	–	0	–	-13.125	0.009798	-1	-1310	∞	–	1	–
0.99	49.497	∞	–	0	–	-34.685	0.009798	-1	-3312	∞	–	1	–

Table 5.2. Overview of the quality of the six bound models for the symmetric shortest queue system with 2 servers. For all cases the parameter T has been taken equal to $T = 3$.

the determination of W within a given accuracy. The TJ and TB model may be expected to produce also the best bounds for the SSQS with $N \geq 2$ servers; therefore, in the next chapter, we shall generalize precisely these two bound models.

Remark 5.3. *(more results which can be proved by the precedence relation method)*
To show the diversity of the monotonicity results that can be derived/proved by the precedence relation method, below we list all results that can be derived by the precedence relation method for the mean normalized waiting time W in the SSQS and the bounds $W_{tr}(T)$ produced by the six bound models. It will be obvious that equivalent results hold for L_w and the bounds $L_w(T)$; further, several monotonicity properties will also hold for other variables and performance measures (see also the remarks at the end of Example 5.1 in Section 5.3). Among the results below are almost all monotonicity properties suggested by the numerical results in the Tables 5.1 and 5.2. We distinguish three types of results.

The first type of results concerns monotonicity results for varying values of T. We already proved that each of the six defined truncation models leads to bounds for W:

$W_{FB}(T) \leq W$, $W_{CB}(T) \leq W$, $W_{TJ}(T) \leq W$,

$W_{OIB}(T) \geq W$, $W_{TK}(T) \leq W$, $W_{TB}(T) \geq W$ for all $T \geq 1$.

For each of the bound models, by considering the bound model with threshold parameter T as a truncation model of the bound model with threshold parameter $T+1$, it moreover can be shown that the quality of the bounds is non-decreasing for increasing values of T:

$$W_{FB}(T) \le W_{FB}(T+1), \quad W_{CB}(T) \le W_{CB}(T+1), \quad W_{TJ}(T) \le W_{TJ}(T+1),$$

$$W_{OIB}(T) \ge W_{OIB}(T+1), \quad W_{TK}(T) \le W_{TK}(T+1), \quad W_{TB}(T) \ge W_{TB}(T+1) \quad \text{for all } T \ge 1.$$

Since further the bounds $W_{lr}(T)$ will tend to W as $T \to \infty$, we thus obtain that

$$W_{FB}(T) \uparrow W, \quad W_{CB}(T) \uparrow W, \quad W_{TJ}(T) \uparrow W,$$

$$W_{OIB}(T) \downarrow W, \quad W_{TK}(T) \uparrow W, \quad W_{TB}(T) \downarrow W \quad \text{as } T \to \infty.$$

Secondly, some monotonicity results between the different bound models can be proved. By exploiting that the FB model represents a truncation model of all three other lower bound models, it can be shown that

$$W_{FB}(T) \le W_{CB}(T), \quad W_{FB}(T) \le W_{TJ}(T), \quad W_{FB}(T) \le W_{TK}(T) \quad \text{for all } T \ge 1.$$

Further, by comparing the CB model to the TJ model, we find that

$$W_{CB}(T) \le W_{TJ}(T) \quad \text{for all } T \ge 1.$$

Subsequently, by using the precedence relation method as described in Remark 5.1, it can be proved that

$$W_{TK}(T) \le W_{TJ}(T) \quad \text{for all } T \ge 1.$$

It appears that, except for $T=1$, it cannot be proved that the CB model produces a better lower bound than the TK model, at least not by using the precedence relation method (for the proof, precedence pairs of the type $(m, m-e_1+e_2)$ would be needed for the TK model, but the precedence relation does not seem to hold for such pairs). From the above results, we obtain the following ordering between the four lower bound models:

$$W_{FB}(T) \le \min\{W_{CB}(T), W_{TK}(T)\} \le \max\{W_{CB}(T), W_{TK}(T)\} \le W_{TJ}(T) \quad \text{for all } T \ge 1.$$

This ordering proves that the TJ model always leads to the best lower bound.

The third and last type of results concerns intuitively trivial monotonicity results for varying values of the workload ρ. By using the precedence relation method as described in Remark 5.1, it can be shown that the average costs in the Markov cost model for the SSQS are larger than or equal to the corresponding costs in the modified system which is obtained by replacing the arrival intensity/probability λ by a smaller value $\lambda' = \lambda - \varepsilon$, $0 < \varepsilon \le \lambda$, and for all $m \in M$ adding a transition from state m to m itself with probability ε (for the states $m = (0, m_2)$, $m_2 \ge 0$, this means that the transition probability for the transition from m to m itself is increased by ε). Since the latter system corresponds to a SSQS with a smaller workload, this proves that the mean normalized waiting time W is non-decreasing as a function of ρ. Similarly, it can be shown that for each of the bound models, the bound $W_{lr}(T)$ is non-decreasing as a function of ρ.

5.6. Conclusions

This chapter has been devoted to the development of the so-called precedence relation method, which in principle is a method for proving a monotonicity result between the average costs of two, discrete-time, irreducible Markov cost models, where the state space of one model is a subset of the state space of the other model. We have mainly focused on how the method can be used for deriving flexible bound models for a given, original Markov cost model. As such, the method consists of a preliminary step, in which so-called precedence pairs for the original model are derived, and a second step, in which several, flexible lower and upper bound models may be defined; in this second step, a flexible lower (upper) bound model is obtained by first defining a truncated state space M' with a flexible size and next redirecting all transitions ending in states outside M' to states inside M', which are more (less) attractive according to the precedence pairs derived for the original model. The derivation of the precedence pairs in the preliminary step is essential for obtaining flexible bound models which accurately approximate the original model for already small sizes of the truncated state space, and therefore this step has extensively been treated in a separate section.

The practical use of the precedence relation method is that its application may lead to flexible bound models which can be solve efficiently and which produce tight bounds for the relevant performance measures of a given original Markovian (queueing) system. Solving such flexible bound models for increasing sizes of the truncated state space constitutes an exact method for the determination of the relevant performance measures of the original system, which may be very useful if the original system itself is very hard to solve or even not solvable at all.

The precedence relation method, as used for deriving flexible bound models, can easily and quickly be applied to several Markovian (queueing) systems. In this chapter, the method has been applied to the symmetric shortest queue system with two parallel servers, mainly to illustrate how the method works for a particular model. In the next two chapters, the method will be applied to the N-dimensional symmetric shortest queue system and to a generalized system, called the shortest queue system with a job-dependent parallelism.

Chapter 6

Flexible Bound Models for the Symmetric Shortest Queue System

6.1. Introduction

The *Symmetric Shortest Queue System (SSQS)*, with $N \geq 2$ parallel servers, a Poisson arrival process and exponentially distributed service times, is a classical queueing system, which, after its introduction by Haight [38] in 1958, has been studied extensively in the literature. Despite of its resemblance to the relatively simple $M \mid M \mid N$ queueing system, the determination of the equilibrium distribution and the relevant performance measures for the SSQS appears to be a hard problem. Since many methods have been applied to the SSQS, this system seems to be an appropriate system for testing new methods. The *objective* of this chapter is to investigate how well the SSQS can be analyzed by using *flexible bound models* derived by the *precedence relation method.*

That the SSQS is hard to solve is expressed, among others, by the fact that analytical results for the SSQS have only been obtained for the case with $N=2$ servers. Analytical results for mainly the generating function of the equilibrium distribution of the two-dimensional SSQS, have been obtained by Kingman [49] and Flatto and McKean [33], who developed the uniformization technique, and by Cohen and Boxma [23] (see also Fayolle [29] and Iasnogorodski [45]), who applied the boundary value method. By applying the compensation approach, Adan et al. [8] have generated explicit formulae for the equilibrium distribution itself (see also Section 1.3). This approach has also led to explicit formulae for the equilibrium distribution of the *asymmetric* shortest queue system, which is obtained if the servers work at different speeds, and for the system with Erlang distributed service times and shortest expected delay routing, which is more sensible in that case; see [6, 10]. Knessl et al. [50] have derived asymptotic expressions for the stationary queue length distribution.

On the two-dimensional SSQS, also several numerical studies have appeared in the literature. In Schassberger [62], an iteration method has been used to obtain approximations for the queue length distribution. Further, Foschini and Salz [34] presented heavy traffic diffusion approximations, and, by using linear programming, Halfin [39] obtained upper and lower bounds for the mean and the distribution of the total number of jobs in the system, and thus also for the mean waiting time. Zhao and Grassmann [84] derived a numerically stable algorithm for the computation of the queue length probabilities by exploiting some results of Flatto and McKean [33].

Other numerical studies concern flexible truncation models, with which the original SSQS can be approximated as accurately as desired. For the case with $N=2$ servers, such models have been derived by Grassmann [36], Gertsbakh [35], Conolly [24], Rao and Posner [60], and Adan et al. [9]. For each of these models, one can determine the equilibrium distribution by using a standard numerical technique or by applying the matrix-geometric approach, as described by Neuts [58], and therefore these models may be exploited in numerical procedures for the determination of the relevant performance measures for the original SSQS within an arbitrary, given accuracy. For each of the flexible truncation models, the model itself and the analysis can easily be generalized to the case with general $N \geq 2$. The model studied by Gertsbakh [35] and Adan et al. [9], is called the Threshold Jockeying (TJ) model, and its generalization has been described in [13]. For a fixed threshold parameter equal to 1, this generalized TJ model reduces to a non-flexible truncation model for which the behavior of the total number of (waiting) jobs is identical to the behavior of this quantity in the corresponding $M \mid M \mid N$ queueing system; this non-flexible truncation model has been studied by Disney and Mitchell [26], Elsayed and Bastani [28], Kao and Lin [48], and Zhao and Grassmann [83], and for $N=2$ also by Haight [38]. It is noted that, by using the precedence relation method, the TJ model as well as the models studied by Conolly [24] and Rao and Posner [60] may be proved to lead to bounds for some relevant performance measures, among which the mean waiting time (see the remarks at the end of Example 5.1 in Section 5.3).

Some recent studies have led to the most interesting results for the SSQS with N servers and general $N \geq 2$. Blanc [17] applied the power-series algorithm to the asymmetric shortest queue system, with which all equilibrium probabilities as well as all relevant performance measures may be computed (within a given accuracy). Lui et al. [54] (see also [53]) developed two flexible approximation models for a generalization of the SSQS, viz. the asymmetric system with servers working at different speeds and with shortest expected delay routing instead of shortest queue routing. One of these two approximation models is a kind of truncation model with some extra states added to it and has been proved to lead to a lower bound for the mean response time, and thus also for the mean waiting time, while the other model is a pure truncation model, which produces an upper bound for the mean response/waiting time. Both in Blanc [17] and in Lui et al. [54], numerical results have been presented for the SSQS. By using a numerical procedure based on their flexible lower bound models, Lui et al. [54] have been able to determine quite accurately the mean response/waiting time for systems with up to $N=8$ servers and workloads up to $\rho=0.9$ (for the instance with $N=8$ and $\rho=0.9$, they reach an absolute accuracy of 0.008 for the mean *normalized* waiting time, i.e. for the mean of the waiting time normalized by the mean service time); in Blanc [17], for the SSQS with workload $\rho=0.9$ numerical results are presented for systems with up to $N=12$ servers (for $N=12$ and $\rho=0.9$, Blanc reaches an absolute accuracy of about $0.5 \cdot 10^{-4}$ for the mean normalized waiting time).

Finally, there is a recent study by Nelson and Philips [57] (see also [56]), who have developed two simple, closed-form approximation formulae for the mean waiting time for the SSQS: one approximation formula for the case with generally distributed interarrival and service times and an improved formula for the case with exponential interarrival and service times. They show among others that for all instances with up to $N=8$ servers (and workloads ρ up to about 0.98), the improved approximation formula for the pure exponential case leads to relative errors smaller than 0.5 percent for the mean *response* time (note that somewhat larger errors would have been obtained for the mean waiting time).

In this chapter, we apply the precedence relation method to the SSQS with $N \geq 2$ servers. We derive two flexible truncation models, of which one model leads to lower bounds for the mean normalized waiting time and the other model produces upper bounds. These two flexible bound models are generalizations of the TJ model and the TB model, as described in Section 1.4 and in the previous chapter. Due to the structure of the transitions in the two flexible bound models, they both can be solved very efficiently by the matrix-geometric approach, and therefore they lead to an efficient numerical procedure for the determination of the mean normalized waiting time within a given, absolute or relative accuracy. As a result, we will be able to determine the mean normalized waiting time within an absolute accuracy of 0.005 for systems with up to $N = 50$ servers and workloads up to $\rho = 0.95$ (it is noted that for all computations, a standard SUN workstation is used; further, to make a rough comparison to Blanc's results, it is noted that for an absolute accuracy of $0.5 \cdot 10^{-4}$ and a workload of $\rho = 0.9$, we can determine the mean normalized waiting time also for systems with up to $N = 50$ servers).

The contents of this chapter will largely coincide with the contents of the paper [2], but there are some differences: an alternative proof of the bounds, a slightly different TJ model and a slightly different bound produced by the TB model (see also Remark 6.2).

The analysis of this chapter is started with a description of the model for the SSQS; see Section 6.2. Next, the precedence relation method is applied in Section 6.3 and the solution of the bound models is described in Section 6.4. In Section 6.5, some numerical results are given. The conclusions are drawn in Section 6.6.

6.2. Model

In this section, we present a discrete-time, irreducible Markov cost model, which describes the behavior of the SSQS with $N \geq 2$ servers. Further, it is denoted how to choose the direct costs such that the corresponding average costs can be exploited to obtain the mean normalized waiting time for the SSQS. The presented model is identical to the model described in Example 5.1 of Section 5.2.

We first describe the SSQS itself. The SSQS consists of $N \geq 2$ parallel servers, which all have their own queue. Jobs arrive according to a Poisson stream with intensity $\lambda > 0$, and an arriving job always joins the shortest queue (ties are broken with equal probabilities). All service times are exponentially distributed with parameter $\mu > 0$. For simplicity, we assume that time is scaled such that $\lambda + N\mu = 1$. In order to have an ergodic system, the workload $\rho = \lambda/(N\mu)$ is assumed to be smaller than 1.

For the application of the precedence relation method, it is required that the SSQS is modeled by a discrete-time, irreducible Markov process. Assume that the servers always work, but that a service completion is only accompanied by a departure of a job if there is a job present in the corresponding queue. Then the behavior of the system may be described by the discrete-time, irreducible Markov process on the time instants right after job arrivals and service completions, and with states $(m_1, \ldots, m_N)$, where m_i denotes the length of the i-th shortest queue, $i \in I := \{1, \ldots, N\}$. This Markov process has a state space

$$M = \{ m \mid m = (m_1, \ldots, m_N) \text{ with } 0 \leq m_1 \leq \cdots \leq m_N \},$$

and it has the property that for all states $m \in M$ the outgoing transitions are caused by the same events $k \in E := \{0\} \cup I$, where event 0 refers to an arrival of a new job and for all $k \in I$ the event k refers to a service completion at the server corresponding to the k-th component m_k of the state m. For all $m \in M$ and $k \in E$, let $r_k(m)$ denote the state to which a transition is made if the Markov process is in state m and event k occurs (note that, if $m_k = m_l$ for two different $k,l \in I$, then the states $r_k(m)$ and $r_l(m)$ are the same). The corresponding transition probabilities are denoted by q_k: $q_0 = \lambda$ and $q_k = \mu$ for all $k \in I$.

The direct costs in our Markov cost model are given by an arbitrary function $c(m)$ and the corresponding average costs are denoted by g. We finally show how to choose the direct costs $c(m)$ such that the mean W of the *normalized* waiting time can be obtained from g. The normalized waiting time is defined as the ratio of the waiting time and the mean service time of a job ($= 1/\mu$), and has the attractive property that it only depends on N and ρ. By Little's formula, we find that W is equal to

$$W = \frac{L_w}{N\rho} , \tag{6.1}$$

where L_w denotes the mean of the total number of waiting jobs in the system. So, it suffices to determine L_w. It is ensured that the average costs g are equal to this measure L_w, if the direct costs $c(m)$ are taken equal to

$$c(m) = \sum_{i=1}^{N} \max\{ m_i - 1, 0 \} . \tag{6.2}$$

This completes the description of the discrete-time, irreducible Markov cost model.

6.3. Application of the precedence relation method

In this section, we apply the precedence relation method in order to obtain one flexible lower bound model and one flexible upper bound model. The two models which we derive are generalizations of the TJ and TB model, which in the previous chapter were derived for the SSQS with $N = 2$ servers. The reason that we derive generalizations of precisely these two models is that they have appeared to lead to the best lower and upper bounds for the mean normalized waiting time for the case $N = 2$ (see Section 5.5) and that they may be expected to lead also to the best lower and upper bounds if $N = 3$.

The first step of the precedence relation method consists of the derivation of precedence pairs (m,n), $m,n \in M$, for the original Markov cost model. For a cost function $c(m)$ as given in (6.2), we expect that we can prove precedence pairs of the type $(m, m+e_i)$ and of the type $(m, m+e_i-e_j)$ with $i > j$, where e_i denotes the i-th unity vector. This means that we expect that it is more attractive to be in a state with one job less at one of the servers, and that it is more attractive to be in a state with *more balance*. Define

$$\begin{aligned} P = & \{ (m,n) \mid m,n \in M \text{ and } n = m + e_i \text{ for some } i \in I \} \\ & \cup \{ (m,n) \mid m,n \in M \text{ and } n = m + e_i - e_j \text{ for some } i,j \in I,\ i > j \} , \end{aligned} \tag{6.3}$$

and let P^* denote the corresponding completed set:

$$P^* = \{ (m,n) \mid m=(m_1,\ldots,m_N)\in M,\ n=(n_1,\ldots,n_N)\in M \text{ and } \Sigma_{i=k}^{N} m_i \le \Sigma_{i=k}^{N} n_i \text{ for all } k\in I \}. \tag{6.4}$$

From the following lemma, which will be proved by applying the Lemmas 5.1 and 5.2, it follows that P and P^* consist of precedence pairs for all cost functions for which the pairs of P represent precedence pairs; the direct costs $c(m)$ as given in (6.2) satisfy this latter condition.

Lemma 6.1.

The set P defined by (6.3) and the corresponding completed set P^* given by (6.4) consist of precedence pairs for all cost functions $c(m)$ for which $c(m)\le c(n)$ for all $(m,n)\in P$.

Proof.

Applying the Lemmas 5.1 and 5.2 immediately shows that it suffices to prove that

$$(r_k(m),r_k(n))\in P^* \quad \text{for all } k\in E \text{ and } (m,n)\in P, \tag{6.5}$$

i.e. that for all $(m,n)\in P$ and $k\in E$, the state $r_k(m)$ to which a transition is made from state m if event k occurs is more attractive according to the set P^* than the corresponding state $r_k(n)$ for state n.

For the proof of (6.5) we introduce the notation $\le_{pr}$: we say that $m\le_{pr} n$ if state m has precedence over state n according to the set P^*, i.e. if $(m,n)\in P^*$. We first prove (6.5) for a pair $(m,n)\in P$ of the first type, i.e. for a pair $(m,n)\in P$ with $m,n\in M$ and $n=m+e_i$ for some $i\in I$. Note that $i\in\{ j \mid j=N \text{ or } (j\in I\backslash\{N\} \text{ and } m_j<m_{j+1}) \}$. Let j be the smallest index of this set. Then we have $j\le i$, and $r_0(m)=m+e_j$. To prove that $(r_0(m),r_0(n))\in P^*$, i.e. $r_0(m)\le_{pr} r_0(n)$, we distinguish two cases:

- $j=i$: Then we find $r_0(m)=m+e_i=n\le_{pr} r_0(n)$.
- $j<i$: Then $r_0(n)=n+e_j$, and we find $r_0(m)=m+e_j\le_{pr} n+e_j=r_0(n)$.

To prove that $(r_k(m),r_k(n))\in P^*$, i.e. that $r_k(m)\le_{pr} r_k(n)$, where $k\in E\backslash\{0\}=I$, we distinguish two main cases:

* $m_k=0$: Then $k\le i$, and we have two subcases:
  - $k=i$: Then $r_k(m)=m=n-e_i=r_k(n)$.
  - $k<i$: Then $m_j=n_j=0$ for all $j\le k$ and we find $r_k(m)=m\le_{pr} n=r_k(n)$.
* $m_k>0$: Then $r_k(m)=m-e_{j_1}$ with $j_1=\min\{ j \mid j\in I \text{ and } m_j=m_k \}$ and $r_k(n)=n-e_{j_2}$ with $j_2=\min\{ j \mid j\in I \text{ and } n_j=n_k \}$, and we distinguish three subcases:
  - $k<i$: Then $m_j=n_j$ for all $j\le k$, $j_1=j_2$ and $r_k(m)=m-e_{j_1}\le_{pr} n-e_{j_1}=r_k(n)$.
  - $k=i$: Then $j_2=k=i$, and we find $r_k(m)=m-e_{j_1}\le_{pr} m=n-e_i=r_k(n)$.
  - $k>i$: Then $m_j=n_j$ for $j=i+1,\ldots,k$. If $m_{i+1}=\cdots=m_k$ and $m_i=m_{i+1}-1$, then $j_1=i+1$, $j_2=i$ and $r_k(m)=m-e_{j_1}\le_{pr} m=n-e_i=r_k(n)$, else $j_1=j_2$ and $r_k(m)=m-e_{j_1}\le_{pr} n-e_{j_1}=r_k(n)$.

This completes the proof of (6.5) for the pairs $(m,n)\in P$ of the first type. For the pairs $(m,n)\in P$ of the second type, the proof of (6.5) may be given along the same lines and is left to the reader. □

The second step of the precedence relation method consists of the definition of flexible bound models. By using the precedence pairs given by P^*, we can define the following two flexible bounds models:

* *Threshold Jockeying (TJ):*
Since for the N-dimensional SSQS, the probability mass is concentrated around the states corresponding to situations with equal queue lengths, we let the truncated state space M' consist of all states for which the difference between the longest and the shortest queue length is at most equal to some threshold $T \geq 1$:

$$M' = \{ m \mid m=(m_1,\ldots,m_N) \text{ with } 0 \leq m_1 \leq \cdots \leq m_N \text{ and } m_N \leq m_1+T \}. \tag{6.6}$$

Then the only transitions from states of M' to states outside M' are the transitions from the states $m=(m_1,\ldots,m_N) \in M'$ with $m_1 > 0$, $m_N = m_1+T$, to $n=(m_1-1,m_2,\ldots,m_N)$, which are due to a service completion at one of the shortest queues at a moment that the difference between the longest and the shortest queue length has already reached its maximum value. Such a transition occurs with probability $j\mu$, where j denotes the number of shortest queues: $j=\max\{ i \mid i \in I \text{ and } m_i = m_1 \}$. Suppose that we want to derive a lower bound model, then these transitions must be redirected to more attractive states inside M'. To obtain a truncation model that approximates the original model as accurately as possible, we want each of these transitions to be redirected to a more attractive state $n' \in M'$ which is as close to n as possible, i.e. we want each of these transitions to be redirected to a state n' such that there is no state $r \in M'$ with $(n',r) \in P^*$ and $(r,n) \in P^*$ (otherwise redirecting to r would be better). Let the indices k_1 and k_2 be equal to $k_1 = \max\{ i \mid i \in I \text{ and } n_i \leq n_N-2 \}$ and $k_2 = \min\{ i \mid i \in I \text{ and } n_i = n_N \}$, then we can start with redirecting to the more attractive state $n' = n+e_{k_1}-e_{k_2}$, which means that a job of one of the longest queues is allowed to jockey to the closest queue where at least two jobs less are present. For the larger part of the states n as characterized above, it will hold that $n_{N-1} \leq n_N-2$, in which case $k_1 = N-1$, $k_2 = N$, and a job jockeys from the (unique) longest queue to a second longest queue. If $n' \in M'$, then we may stop, else we have to perform one or more extra steps; the procedure for the determination of n' is formally described in Figure 6.1. The resulting truncation model is called the Threshold Jockeying (TJ) model, and it is obvious that this model leads to a lower bound for the average costs g in the original Markov cost model if the direct costs $c(m)$ are chosen according to (6.2), i.e. it leads to a lower bound for the mean L_w of the total number of waiting jobs in the original SSQS.

* *Threshold Blocking (TB):*
Consider the TJ model, but let the transitions from the states $(m_1,\ldots,m_N)$ with $m_1 > 0$, $m_N = m_1+T$, to $(m_1-1,m_2,\ldots,m_N)$, be redirected to $(m_1,\ldots,m_N)$ itself. This means that if the difference m_N-m_1 between the longest and the shortest queue has already reached its maximum value T, then a service completion at a non-empty, shortest queue is not accompanied by a departure, and the job in service has to be served once more. It is easily seen that (because of the memory-less property of the exponential service times), this is equivalent to letting the servers at the shortest queues be blocked as long as $m_N-m_1 = T$, and therefore this modified model is called the Threshold Blocking (TB) model. Since the states $(m_1,\ldots,m_N)$ are less attractive than $(m_1-1,m_2,\ldots,m_N)$, the TB model produces upper bounds for L_w.

$\{m=(m_1,\ldots,m_N)\in M'$ with $m_1>0$, $m_N=m_1+T$, and $n=m-e_1\}$
$r:=n$;
repeat
  $k_1:=\max\{\, i \mid i\in I$ and $r_i\le r_N-2\,\}$;
  $k_2:=\min\{\, i \mid i\in I$ and $r_i=r_N\,\}$;
  $r:=r+e_{k_1}-e_{k_2}$
until $r_N-r_1\le T$;
$n':=r$

Figure 6.1. The procedure for determining the state n' to which a transition from a state m inside the truncated state space M' to a state n outside M' is redirected in the TJ model.

The TJ and TB model produce lower and upper bounds for the mean L_w of the total number of waiting jobs in the original SSQS. Let these bounds be denoted by $L_w^{TJ}(T)$ and $L_w^{TB}(T)$, and let the corresponding bounds for the mean normalized waiting time W be defined by (cf. (6.1))

$$W_{TJ}(T)=\frac{L_w^{TJ}(T)}{N\rho} \quad \text{and} \quad W_{TB}(T)=\frac{L_w^{TB}(T)}{N\rho}. \tag{6.7}$$

Since these formulae in fact represent Little's formula for the number of waiting jobs in the TJ model and the TB model, respectively, $W_{TJ}(T)$ and $W_{TB}(T)$ are precisely equal to the mean normalized waiting times in the TJ and TB model (where also for the TJ and TB model the normalized waiting time is defined as the waiting time divided by $1/\mu$, i.e. divided by the mean of one single service time).

In the previous chapter, we have mentioned a whole variety of monotonicity results that can be derived for the two-dimensional SSQS and the corresponding bound models. The TJ and TB model, as described here for the N-dimensional case, have been defined such that all monotonicity results concerning the two-dimensional SSQS and the TJ and TB model of the previous chapter, can be generalized to the N-dimensional case. We repeat some results which concern the mean normalized waiting time W and the corresponding bounds $W_{TJ}(T)$ and $W_{TB}(T)$.

By deriving precedence pairs of the types $(m,m+e_i)$ and $(m,m+e_i-e_j)$, $i>j$, for the TJ model, and considering the TJ model with threshold parameter T as a truncation model of the TJ model with threshold parameter $T+1$, it can be shown that $W_{TJ}(T)\le W_{TJ}(T+1)$ for all $T\ge 1$. Similarly, by deriving precedence pairs of the type $(m,m+e_i)$ for the TB model, and considering the TB model with threshold parameter T as a truncation model of the TB model with threshold parameter $T+1$, it can be shown that $W_{TB}(T)\ge W_{TB}(T+1)$ for all $T\ge 1$. Further, both $W_{TJ}(T)$ and $W_{TB}(T)$ will tend to W, as $T\to\infty$, since for $T=\infty$ both truncation models are identical to the original SSQS. So, we find that

$$W_{TJ}(T)\uparrow W \quad \text{and} \quad W_{TB}(T)\downarrow W, \quad \text{as } T\to\infty. \tag{6.8}$$

This result shows that the mean normalized waiting time W can be determined as accurately as desired by computing the bounds $W_{TJ}(T)$ and $W_{TB}(T)$ for increasing values of T, where for each T, the mean $(W_{TJ}(T)+W_{TB}(T))/2$ of the bounds is used as an approximation for W and

half of the difference between both bounds serves as an upper bound for the corresponding inaccuracy. The next section will be devoted to the only remaining problem: the computation of the bounds $W_{TJ}(T)$ and $W_{TB}(T)$ itself.

Remark 6.1. *(on bounds which are obtained by using alternative formulae for W)*

For formula (6.1), which we use for the determination of the mean normalized waiting time W of the SSQS, there are two alternative formulae. By applying Little's formula, it may be shown that the mean normalized response time is equal to $L/(N\rho)$, where L denotes the mean of the total number of jobs in the system (including the jobs in service). This leads to the alternative formula

$$W = \frac{L}{N\rho} - 1 . \tag{6.9}$$

Further, it is obvious that

$$W = L_{sq} , \tag{6.10}$$

where L_{sq} denotes the mean of the length of the shortest queue. The TB model can be proved to produce also upper bounds for L and L_{sq}, and therefore also the formulae (6.9) and (6.10) could be used to define the corresponding upper bound for W. It may be verified that, due to the fact that in the TB model some jobs are served more than once, using formula (6.9) would lead to a slightly worse upper bound for W and using formula (6.10) would lead to a slightly better upper bound for W (however, numerical results have pointed out that the differences become negligible as soon as the bounds are getting close to W). The TJ model can be proved to produce also an upper bound for L, but in this case using formula (6.9) would lead to the same bound for W as the bound $W_{TJ}(T)$ defined by (6.7).

Remark 6.2. *(on the bounds used in [2])*

In [2], also a TJ model and a TB model have been used to compute W within a given accuracy, however for both models the bounds obtained are slightly worse. The bounds in [2] stemming from the TB model are the bounds which we would obtain when using formula (6.9) for the definition of $W_{TB}(T)$ instead of formula (6.7). For the TJ model in [2], another type of jockeying has been used. Jockeying from the longest to the shortest queue has been used, when jumping to a state outside the truncated state space, instead of the jockeying as defined by the algorithm in Figure 6.1, which may be characterized as jockeying from the longest to the second longest queue (differences between both types of jockeying occur if $N \geq 3$ and $T \geq 2$). This means that there are transitions to states outside the truncated state space which are redirected to more attractive states than necessary. As a consequence, for this alternative TJ model the precedence pairs of the type $(m, m+e_i-e_j)$ with $i > j$ cannot be derived anymore, and hence it cannot be guaranteed that the bounds are monotonously non-decreasing for increasing values of T. Besides, for the TJ model presented in this chapter, one can prove that it leads to a better bound than the generalization of the Centralized Buffer (CB) model as presented in the previous chapter for the two-dimensional SSQS, whereas this does not hold for the alternative TJ model (instances have been found for which the CB model gives a better bound than the alternative TJ model).

6.4. Solving the flexible bound models by the matrix-geometric approach

For both the TJ model and the TB model, the structure of the state space and the transitions (with the corresponding transition probabilities) is such that they both can be solved by applying the *matrix-geometric approach*, as described by Neuts [58]. By using this method, we can check whether a bound model is positive recurrent (which is needed for the TB model), and if so, then we can compute the equilibrium distribution, which may be exploited to determine the bound for the mean normalized waiting time W. The main part of this section consists of showing that, due to the special structure of the transitions, for both bound models, the equilibrium distribution may be determined very efficiently by the matrix-geometric approach (the rate matrix R has only one row with nonnull elements, if the state space is appropriately partitioned into levels). After that, we shortly discuss the computation of the bounds for W and the order of magnitude of the computation time.

Determination of the equilibrium distribution for the TJ and TB model

Assume that the number of servers N, the workload ρ and the threshold parameter T are given, and that $N \geq 2$, $0 < \rho < 1$ and $T \geq 1$. Then for both bound models, we have an irreducible, discrete-time Markov process. The truncated state space M' is given by (6.6), and the transition probabilities and equilibrium probabilities are denoted by $q^{tr}_{m,n}$ and p^{tr}_m, respectively. Here, the indices tr may be replaced by TJ or TB, in case we are discussing a particular bound (truncation) model.

Application of the matrix-geometric approach requires a partitioning of the state space into subsets M'_l, $l \geq 0$, which are called *levels*. It appears to be appropriate to partition the state space on the basis of the number of jobs present at the longest queue. We define

$$M'_l = \{ m \in M' \mid m_N = l \} \quad \text{for all } l \geq 0.$$

Let the states be lexicographically ordered within each level M'_l, let the probability vector p^{tr}_l contain all equilibrium probabilities p^{tr}_m corresponding to the states $m \in M'_l$, and let the probability vector p^{tr} be equal to $(p^{tr}_0, p^{tr}_1, \cdots)$. Next, for simplicity we put together the levels $M'_0, \ldots, M'_{T-1}$, which have a less regular behavior, into one level $M'_{0\cdot}$. Then, for this partitioning, $p^{tr}_{0\cdot} = (p^{tr}_0, \ldots, p^{tr}_{T-1})$, $p^{tr} = (p^{tr}_{0\cdot}, p^{tr}_T, p^{tr}_{T+1}, \cdots)$, and the transition matrix P^{tr} is of the form

$$P^{tr} = \begin{bmatrix} B_{0\cdot 0\cdot} & B_{0\cdot 1} & 0 & 0 & 0 & \cdots \\ B_{10\cdot} & B_{11} & A_0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}. \tag{6.11}$$

Here, all submatrices are nonnegative and real valued; A_0, A_1, A_2 and B_{11} are $a \times a$ matrices (it is noted that for the TB model, B_{11} is equal to A_1), and $B_{0\cdot 0\cdot}$, $B_{0\cdot 1}$ and $B_{10\cdot}$ are of the size $b \times b$, $b \times a$ and $a \times b$, respectively, where a is the number of states for each of the levels M'_l with $l \geq T$ and b is the number of states for level $M'_{0\cdot}$. It is easily verified that a and b are equal to

$$a = \binom{N+T-1}{T}, \quad b = \sum_{l=0}^{T-1} \binom{N+l-1}{l}.$$

The matrix P^{tr} is said to be block tridiagonal, and the Markov process is called a *quasi-birth-and-death process.*

For the existence of the equilibrium distribution $\{p_m^{tr}\}$, it is required that we have a positive recurrent (i.e. ergodic) Markov process. Therefore, we first describe a simple, necessary and sufficient condition for the positive recurrence. Define the matrix A by $A := A_0 + A_1 + A_2$. Then A is obviously a stochastic matrix, and since two states of levels M_l' with $l > T$ can reach each other via paths not passing through levels M_l' with $l \leq T$, A is also irreducible. So, A is the transition matrix of a finite, irreducible, discrete-time Markov process, and the corresponding equilibrium vector $\pi = (\pi_1, \ldots, \pi_a)$ is the unique solution of the system of linear equations

$$\pi A = \pi, \quad \pi e = 1,$$

where e is an a-dimensional column vector consisting of all ones. The probability π_i denotes the probability to be in the i-th state, under the condition that we are at some level M_l', where l is very large. Now, the Markov process with transition matrix P^{tr} may be shown to be positive recurrent if and only if

$$\pi A_0 e < \pi A_2 e \tag{6.12}$$

(cf. Theorem 1.3.2 of [58]). This means that, for the levels far away from the origin, the drift into the direction of the higher levels must be smaller than the drift into the direction of the lower levels.

Suppose that condition (6.12) is satisfied. Then the equilibrium vector p^{tr} is characterized as the unique solution of the linear equations

$$p^{tr} P^{tr} = p^{tr}, \quad p^{tr} e = 1.$$

Writing out the first of these two equations suggests to look for a solution for which

$$p_l^{tr} = p_T^{tr} R^{l-T} \quad \text{for all } l \geq T, \tag{6.13}$$

where the $a \times a$ matrix R must be a solution of the quadratic matrix equation

$$R = A_0 + R A_1 + R^2 A_2. \tag{6.14}$$

To obtain an equilibrium distribution which can be normalized, it is required that the spectral radius $sp(R)$ of R is smaller than 1. Let R be defined as the minimal nonnegative solution of (6.14). Then, since we have a positive recurrent Markov process, it may be shown that $sp(R)$ indeed is smaller than 1, and that the equilibrium distribution has the *matrix-geometric structure* as denoted by (6.13) (see Theorem 1.5.1 of [58]). The matrix R is called the *rate matrix* and may be obtained by performing successive substitutions in (6.14), where $R = 0$ is used as starting matrix.

Due to our choice for the partitioning of the state space, for all $l \geq 0$, the only transition pointing from a state of level M_l' to a state of level M_{l+1}', is the transition from the state $(l, \ldots, l)$ to $(l, \ldots, l, l+1)$. This transition is due to an arrival of a new job in a situation that all queue lengths are equal; the corresponding transition rate is equal to λ. So, the matrix A_0 is equal to

$$A_0 = \begin{bmatrix} 0 \\ x \end{bmatrix},$$

where $x = (0, \ldots, 0, \lambda, 0, \ldots, 0)$ with λ on the $(a-(N-1))$-th position. This implies that the condition for the positive recurrence, as stated in (6.12), simplifies to

$$\lambda \pi_{a-(N-1)} < \pi A_2 e . \tag{6.15}$$

But, what is really important, is the simplification that is obtained for the rate matrix R. As one may derive from the successive substitutions scheme for the determination of R, each zero row of A_0 corresponds to a zero row of R (this also follows from the probabilistic interpretation of R as described in Section 1.2 of [58]). So, R also has the form

$$R = \begin{bmatrix} 0 \\ y \end{bmatrix}, \tag{6.16}$$

where $y = (y_1, \ldots, y_a)$, by which the formulae (6.13) and (6.14) simplify considerably. Since $R^2 = y_a R$ and $p_T^{tr} R = p_{(T,\ldots,T)}^{tr} y$, formula (6.13) reduces to

$$p_l^{tr} = p_{(T,\ldots,T)}^{tr} y_a^{l-(T+1)} y \quad \text{for all } l \geq T+1 . \tag{6.17}$$

Here, the factor y_a is smaller than 1, since y_a is equal to $sp(R)$. By insertion of the special forms of R and A_0, the quadratic matrix equation (6.14) simplifies to

$$y = x + y(A_1 + y_a A_2) . \tag{6.18}$$

Note that this equation still is quadratic. The vector y may be obtained by successive substitutions, where $y = 0$ is used as starting vector.

Now, the most natural way for exploiting the matrix-geometric structure as denoted by (6.17) in order to obtain the equilibrium distribution $\{p_m^{tr}\}$ seems to consist of substituting formula (6.17) into the equilibrium equations for the states of the levels $M'_0, \ldots, M'_T$ and into the normalization equation, and subsequently solving the resulting system of linear equations of order $a+b$ (note that the equilibrium equation for one of the states may be left out). However, due to the structure of the transitions in both bound models, we can determine the equilibrium probabilities p_m^{tr} for the states of the levels $M'_0, \ldots, M'_T$ in a much more efficient way. This is described below.

The result stated in (6.17) is exploited for the determination of an unnormalized equilibrium distribution $\{\bar{p}_m^{tr}\}$, which afterwards only has to be normalized in order to obtain $\{p_m^{tr}\}$. Let $\bar{p}_{(T,\ldots,T)}^{tr} = 1$ and let the vectors $\bar{p}_l^{tr}$ for all $l \geq T+1$ be given by (6.17). Then the remaining unnormalized probabilities at the levels $M'_T, \ldots, M'_0$ may be obtained recursively. In the next paragraph we discuss how the equilibrium probabilities for level M'_l may be obtained from the equilibrium probabilities for level M'_{l+1}, where $0 \leq l \leq T$.

Let $1 \leq l \leq T$ and suppose that the unnormalized probability vector $\bar{p}_{l+1}^{tr}$ is known. Up to some multiplicative constant, p_l^{tr} is equal to the equilibrium probability vector of the Markov process restricted to level M'_l (i.e. excursions to other levels are not considered). Let the transition probabilities for the restricted Markov process be denoted by $\hat{q}_{m,n}$, with $m,n \in M'_l$. The rate $\hat{q}_{m,n}$ is given by $q_{m,n}^{tr}$, plus the rate due to excursions to other levels starting in m and ending in n. An excursion to the lower levels $M'_0, \ldots, M'_{l-1}$ always ends in $(l-1, \ldots, l-1, l)$ and an excursion to the higher levels $M'_{l+1}, M'_{l+2}, \cdots$ always starts at $(l, \ldots, l)$ and ends with probability z_n in state n of level M'_l, where z_n is given by (note that the unnormalized

equilibrium probabilities $\bar{p}_m^{tr}$ for the states of level M'_{l+1} denote relative frequencies with which the states at that level are visited)

$$z_n = \frac{\sum_{m \in M'_{l+1}} \bar{p}_m^{tr} q_{m,n}^{tr}}{\sum_{m \in M'_{l+1}} \bar{p}_m^{tr} \sum_{r \in M'_l} q_{m,r}^{tr}} .$$

Hence, for two states m and n of level M'_l, we find

$$\hat{q}_{m,n} = q_{m,n}^{tr} + \left\{ \begin{array}{ll} \sum_{r \in M'_{l-1}} q_{m,r}^{tr} & \text{if } n = (l-1, \ldots, l-1, l) \\ 0 & \text{otherwise} \end{array} \right\} + \left\{ \begin{array}{ll} \lambda z_n & \text{if } m = (l, \ldots, l) \\ 0 & \text{otherwise} \end{array} \right\} .$$

Thus, an unnormalized probability vector $\bar{p}_l^{tr}$ for the states of level M'_l can be obtained, once we have the unnormalized probability vector $\bar{p}_{l+1}^{tr}$ for the states of level M'_{l+1} available. Since we want to derive an unnormalized solution which satisfies all equilibrium equations, we scale the unnormalized vector $\bar{p}_l^{tr}$ such that we satisfy the equation stating that the flow from level M'_l to M'_{l+1} is equal to the flow from level M'_{l+1} to M'_l, i.e.

$$\lambda p_{(l,\ldots,l)}^{tr} = \sum_{m \in M'_{l+1}} p_m^{tr} \sum_{n \in M'_l} q_{m,n}^{tr}$$

(if $l = T$, then it suffices to scale p_l^{tr} such that $p_{(T,\ldots,T)}^{tr} = 1$).

Starting with $\bar{p}_{(T,\ldots,T)}^{tr} = 1$ and $\bar{p}_{T+1}^{tr}$ given by (6.17), we can use the approach sketched above to recursively compute the vectors $\bar{p}_T^{tr}, \ldots, \bar{p}_1^{tr}$. The unnormalized equilibrium distribution is completed by taking the only equilibrium probability $\bar{p}_{(0,\ldots,0)}^{tr}$ for level M'_0 such that the equilibrium equation for the origin $(0, \ldots, 0)$ is satisfied, i.e. by taking $\bar{p}_{(0,\ldots,0)}^{tr} = (\mu/\lambda)\bar{p}_{(0,\ldots,0,1)}^{tr}$. Finally, the desired equilibrium distribution $\{p_m^{tr}\}$ with total probability equal to 1 is obtained by dividing all probabilities $\bar{p}_m^{tr}$ by the normalizing constant

$$\begin{aligned} C &= \sum_{m \in M'} \bar{p}_m^{tr} = \sum_{l=0}^{T} \sum_{m \in M'_l} \bar{p}_m^{tr} + \sum_{l=T+1}^{\infty} y_a^{l-(T+1)} ye \\ &= \sum_{l=0}^{T} \sum_{m \in M'_l} \bar{p}_m^{tr} + \frac{1}{1-y_a} ye . \end{aligned} \tag{6.19}$$

This completes the description of the procedure for the determination of the equilibrium distribution, which works for both the TJ and TB model. We finally show that the procedure may be slightly simplified for the TJ model.

For the TJ model, the condition $\rho < 1$ may be shown to be necessary and sufficient for the positive recurrence, by which the check of condition (6.12) can be skipped and the component y_a of the vector y can be found explicitly. This is shown by using the following balance argument (see also [13]). Let V_l be the set of states with $m_1 + \cdots + m_N = l$ and let $P(V_l)$ be the equilibrium probability for the set V_l. By balancing the flow between the sets V_l and V_{l+1} it follows that for all $l \geq (N-1)T$

$$\lambda P(V_l) = N\mu P(V_{l+1}) ,$$

and by applying this relation N times, we obtain

$$P(V_{l+N}) = \rho^N P(V_l) \quad \text{for all } l \geq (N-1)T, \tag{6.20}$$

which proves that the TJ model is positive recurrent if and only if $\rho < 1$. A similar formula may be found by using (6.17). By (6.17), for all states $m \in M'$ with $m_N \geq T+1$

$$p_{(m_1+1,\ldots,m_N+1)} = y_a p_{(m_1,\ldots,m_N)} ,$$

from which it follows that

$$P(V_{l+N}) = y_a P(V_l) \quad \text{for all } l > NT. \tag{6.21}$$

Combining (6.20) and (6.21) yields

$$y_a = \rho^N . \tag{6.22}$$

As a result, the quadratic equation (6.18) for y simplifies to the linear equation

$$y = x + y(A_1 + \rho^N A_2), \tag{6.23}$$

from which the vector y can be obtained more efficiently, since this equation forms a contraction scheme. A convenient method to solve such schemes is the iteration method presented in Van der Wal and Schweitzer [73]. This method has the advantage that it provides upper and lower bounds for the vector y.

Determination of the bounds for the mean normalized waiting time

For both bound models, the bounds $W_{TJ}(T)$ and $W_{TB}(T)$ for W can be obtained from the equilibrium distribution $\{p_m^{tr}\}$ by first computing the corresponding bounds $L_w^{TJ}(T)$ and $L_w^{TB}(T)$ for the mean L_w of the total number of waiting jobs, and subsequently using the formulae stated in (6.7). For both the TJ and TB model, the bound $L_w^{tr}(T)$ itself may be determined as follows. Let for all $m \in M'$ the direct costs $c(m)$ be defined by (6.2), then the bound $L_w^{tr}(T)$ is precisely equal to the corresponding average costs. Next, let the column vector c_l, $l \geq 0$, contain the direct costs $c(m)$ for all states m of level M_l'. For these vectors c_l, we have the property that

$$c_l = c_{T+1} + (l-(T+1))Ne \quad \text{for all } l \geq T+1,$$

by which we find that

$$\begin{aligned} L_w^{tr}(T) &= \sum_{l=0}^{T} p_l^{tr} c_l + \sum_{l=T+1}^{\infty} p_{(T,\ldots,T)}^{tr} y_a^{l-(T+1)} y (c_{T+1} + (l-(T+1))Ne) \\ &= \sum_{l=0}^{T} p_l^{tr} c_l + p_{(T,\ldots,T)}^{tr} \left[\frac{1}{1-y_a} y c_{T+1} + \frac{N y_a}{(1-y_a)^2} ye \right] \\ &= \frac{1}{C} \left[\sum_{l=0}^{T} \bar{p}_l^{tr} c_l + \left[\frac{1}{1-y_a} y c_{T+1} + \frac{N y_a}{(1-y_a)^2} ye \right] \right] . \end{aligned} \tag{6.24}$$

Note that both the normalizing constant C and the part in brackets in formula (6.24) may be computed simultaneously with the computation of the equilibrium distribution; so, by using formula (6.24) for the determination of $L_w^{tr}(T)$, it is provided that during the computation of the vector y and the unnormalized probability vectors $\bar{p}_T^{tr}, \ldots, \bar{p}_0^{tr}$, in each step only the last computed vector has to be kept in memory.

Order of magnitude of the computation time

Let us finally discuss the order of magnitude of the computation time for solving the TJ or TB model. For both models, the main computational effort consists of solving systems of equations of order a. Systems of this order have to be solved to determine the equilibrium distribution π on behalf of the verification of condition (6.15) for the positive recurrence (only for the TB model), the vector y, and the unnormalized probability vector $\bar{p}_T^{tr}$ for the states of level M_T'. For the computation of the unnormalized equilibrium probabilities at the lower levels, one has to solve a number of smaller systems of equations, for which the computational effort may be ignored (especially when N is large). For the computation of the vector y, when solving the TB model, we will use successive substitutions. Here, Gauss-Seidel iterations are used to decrease the number of iterations required to determine the vector y within a given accuracy. In all other cases, we use the iteration method presented in Van der Wal and Schweitzer [73].

Let us consider the computation of the vector y. Independently of the iteration method used, the computation time is proportional to the number of iterations multiplied by the amount of work in each iteration. To decrease the amount of work in each iteration, and also the required memory space, only the nonnull elements of the sparse matrices A_1 and A_2 are stored. Then the amount of work in each iteration is mainly determined by the number of multiplications, and therefore the work per iteration is proportional to the total number of nonnull elements in A_1 and A_2. Since each nonzero element corresponds to a transition possibility from one state to another, for each row the number of nonnull elements in A_1 and A_2 is given by 1 (an arrival) plus $\min\{N,T+1\}$ (service completions), and the total number of nonnull elements is obtained by multiplying this expression by a. Hence, we may conclude that the order of magnitude O_{ct} of the computation time for the determination of y is given by

$$O_{ct} = (1+\min\{N,T+1\})\,a \cdot (\text{number of iterations}). \tag{6.25}$$

For all other systems of a equations which we have to solve, we also find this order of magnitude for the computation time. An indication of the number of iterations, an expression for the order of magnitude of the total computation time for solving a bound model and the computation times itself, are discussed at the end of the next section.

6.5. Numerical results

This section is devoted to some numerical results, which are obtained by a numerical procedure that is developed for the determination of the mean normalized waiting time W in the N-dimensional SSQS and that is based on the flexible bounds $W_{TJ}(T)$ and $W_{TB}(T)$ produced by the TJ and TB model.

The results of the previous two sections lead to the following procedure, with which, for given values of the number of servers N and the workload ρ, the mean normalized waiting time W is determined within a given, absolute accuracy ε_{abs}. By exploiting the results of Section 6.4, we compute both the lower bound $W_{TJ}(T)$ and the upper bound $W_{TB}(T)$ for $T=1,2,\cdots$ (both bounds may be assumed to be determined exactly, i.e. they are computed within a sufficiently high numerical accuracy, viz. an absolute numerical accuracy of at least

$10^{-2}\,\varepsilon_{abs}$). For each value of the threshold parameter T, the mean $(W_{TJ}(T)+W_{TB}(T))/2$ of the lower and upper bound is used as an approximation for W, and half of the difference between both bounds, i.e. $\Delta(T)=(W_{TB}(T)-W_{TJ}(T))/2$, is used as an upper bound for the absolute inaccuracy of this approximation. The computation process is stopped as soon as $\Delta(T)\leq\varepsilon_{abs}$ for some T. Note that a similar procedure can be used for the determination of W within a given, relative accuracy.

In Table 6.1, we have listed some numerical results which have been obtained by the above procedure. For varying values of ρ and N, which are given in the first two columns, we have determined the mean normalized waiting time W within absolute accuracy $\varepsilon_{abs}=0.005$. The third column denotes the smallest value of T for which this accuracy is reached, and the fourth column denotes the value of the size a of the largest systems of equations that have to be solved for this T. In the fifth and sixth column, the bounds $W_{TJ}(T)$ and $W_{TB}(T)$ are listed, and in the last two columns the corresponding approximation for W and the corresponding value for $\Delta(T)$ are given (it is noted that for $\Delta(T)$ the values obtained by rounding off upwards are given).

The results of Table 6.1 show that already small values of T are sufficient to approximate W within the desired accuracy; even for high workloads ρ, this appears to hold. Besides, the required value for T to reach the desired accuracy appears to be decreasing as a function of the number of servers N (an equivalent property is that for a fixed value of T the absolute accuracy, and also the relative accuracy, are decreasing as a function of N; see the values in the last column of Table 6.1). From these observations, we may conclude that, at least with respect to the mean normalized waiting time, the original SSQS is accurately approximated by the TJ and TB model for already small values of the threshold parameter T, and that this especially holds for SSQS-s consisting of many servers. We note that further it may be verified that for a fixed value of T, the original SSQS is more accurately approximated by the TJ model than by the TB model (see also the results presented in Table 1.8 and Figure 1.9 for the SSQS with $N=2$ servers).

The results in the last but one column of Table 6.1 show that the mean normalized waiting time W is strongly increasing for large values of the workload ρ, and that W is decreasing as a function of the number of servers N. It is noted that the behavior of the mean normalized waiting time W in the SSQS can be shown to be quite similar to the behavior of the mean normalized waiting time $W_{M\,|\,M\,|\,N}$ in the corresponding $M\,|\,M\,|\,N$ queueing system (recall that $W_{M\,|\,M\,|\,N}$ is equal to the mean normalized waiting time $W_{TJ}(T)$ in the TJ model with $T=1$, and study the values of $W-W_{TJ}(1)$; for the case with $N=2$ servers, some results for this difference are depicted in the second column of Table 1.8). This similarity can be exploited to obtain a simple, closed-form approximation formula for W (and, possibly, also to further improve the approximation formula for the purely exponential case, which has been presented by Nelson and Philips [57]).

There are two reasons for the fact that we have been able to determine the mean normalized waiting time W within the desired accuracy for systems with up to $N=50$ servers and workloads up to $\rho=0.95$. The first reason is that the bounds $W_{TJ}(T)$ and $W_{TB}(T)$ accurately approximate W for already small sizes of T, especially for large N, which provides that the TJ and TB model only have to be solved for relatively small sizes of the truncated state space. The second reason is that, as we already noticed in the previous section, the TJ and TB model can be solved very efficiently. This latter property is further discussed in the next paragraph.

ρ	N	T	a	$W_{TJ}(T)$	$W_{TB}(T)$	W	$\Delta(T)$
0.8	2	6	7	1.9552	1.9587	1.9570	0.0018
	5	4	70	0.7541	0.7581	0.7561	0.0021
	10	3	220	0.3557	0.3589	0.3573	0.0017
	15	3	680	0.2203	0.2206	0.2204	0.0002
	20	3	1540	0.1513	0.1514	0.1514	0.0001
	25	2	325	0.1098	0.1150	0.1124	0.0027
	30	2	465	0.0826	0.0843	0.0834	0.0009
	35	2	630	0.0637	0.0642	0.0640	0.0003
	40	2	820	0.0501	0.0502	0.0502	0.0001
	45	2	1035	0.0399	0.0400	0.0400	0.0001
	50	2	1275	0.0322	0.0322	0.0322	0.0001
0.9	2	7	8	4.4744	4.4831	4.4787	0.0044
	5	5	126	1.7974	1.8049	1.8012	0.0038
	10	4	715	0.9130	0.9171	0.9151	0.0021
	15	4	3060	0.6155	0.6160	0.6158	0.0003
	20	3	1540	0.4636	0.4686	0.4661	0.0025
	25	3	2925	0.3702	0.3717	0.3710	0.0008
	30	3	4960	0.3063	0.3068	0.3066	0.0003
	35	3	7770	0.2595	0.2597	0.2596	0.0001
	40	3	11480	0.2237	0.2237	0.2237	0.0001
	45	3	16215	0.1953	0.1953	0.1953	0.0001
	50	3	22100	0.1722	0.1722	0.1722	0.0001
0.95	2	9	10	9.4865	9.4914	9.4890	0.0025
	5	6	210	3.8269	3.8358	3.8314	0.0045
	10	5	2002	1.9571	1.9605	1.9588	0.0017
	15	4	3060	1.3349	1.3435	1.3392	0.0044
	20	4	8855	1.0223	1.0245	1.0234	0.0012
	25	4	20475	0.8330	0.8336	0.8333	0.0004
	30	4	40920	0.7053	0.7055	0.7054	0.0002
	35	3	7770	0.6127	0.6212	0.6169	0.0043
	40	3	11480	0.5422	0.5467	0.5445	0.0023
	45	3	16215	0.4864	0.4889	0.4877	0.0013
	50	3	22100	0.4411	0.4425	0.4418	0.0007

Table 6.1. The results which are obtained for the determination of the mean normalized waiting time W within an absolute accuracy of $\varepsilon_{abs}=0.005$ for varying values of the workload ρ and the number of servers N.

For both the TJ and the TB model, the computational effort mainly consists of the determination of the vectors π, y and $\bar{p}_T^{tr}$ (π only has to be determined for the TB model), for which systems of equations of order a have to be solved. At the end of the previous section, we have established that the order of magnitude of the computation time O_{ct} for solving these systems is equal to the expression given by (6.25). By studying the numbers of iterations for

ρ	N	T	a	Comp. Time (h:min:sec)
0.8	10	3	220	0:00:08
0.8	30	2	465	0:00:41
0.8	50	2	1275	0:03:17
0.9	10	4	715	0:00:55
0.9	30	3	4960	0:14:35
0.9	50	3	22100	2:04:13
0.95	10	5	2002	0:03:56
0.95	30	4	40920	4:26:37
0.95	50	3	22100	3:00:51

Table 6.2. Computation times on a SUN workstation for some instances of Table 6.1.

the different systems of a equations that we had to solve for the results listed in Table 6.1, we empirically find that (for a fixed value of the absolute numerical accuracy with which the systems of equations are solved) for all these systems the number of iterations is proportional to $N/(1-\rho)$, and thus we find that

$$O_{ct} = (1+\min\{N,T+1\})\, a\, \frac{N}{1-\rho} . \tag{6.26}$$

For sufficiently large values of N, the computation time required to solve the other systems of equations may be neglected. As a result, we find that for sufficiently large N, the expression for O_{ct} as stated in (6.26), also denotes the order of magnitude of the computation time for solving the whole TJ or TB model, and thus for the determination of $W_{TJ}(T)$ and/or $W_{TB}(T)$, and it also denotes the order of magnitude of the computation time for the numerical procedure for the determination of the mean normalized waiting time W within a given fixed accuracy. To give an impression of the computation times itself of the numerical procedure for the determination of W, for some of the instances of Table 6.1 we have listed in Table 6.2 the computation times which we obtained on a standard SUN workstation. It is finally noted that this machine had a memory space of 24 Megabyte, which has appeared to be about two or three times as much as needed for the instances of Table 6.1, but which will not be sufficient in case bound models with much larger values for N and T have to be solved.

6.6. Conclusions

In this chapter, we have applied the precedence relation method to the Symmetric Shortest Queue System (SSQS) with $N \geq 2$ parallel servers, in order to obtain flexible truncation models which produce lower and upper bounds for the mean normalized waiting time. Due to the shortest queue routing, which causes a strong drift to the states with equal queue lengths, we have been able to construct a flexible lower bound model, called the Threshold Jockeying (TJ) model, and a flexible upper bound model, called the Threshold Blocking (TB)

model, which produce tight bounds for already small sizes of the truncated state space. Since besides, as we have seen, both flexible bound models can be solved very efficiently by the matrix-geometric approach, the TJ and TB model are very appropriate for being exploited in a numerical procedure for the determination of the mean normalized waiting time of the original SSQS within an arbitrary, given, absolute or relative accuracy. By developing such a procedure, we have been able to determine this performance measure quite accurately for systems with many servers and high workloads (up to $N = 50$ servers and workloads up to 0.95). It is noted that also for all other relevant performance measures it will be possible to determine them efficiently by exploiting flexible bound models derived by the precedence relation method.

Chapter 7

Flexible Bound Models for the Shortest Queue System with a Job-Dependent Parallelism

7.1. Introduction

In Section 1.2 of the introductory chapter of this monograph, we have presented a multi-dimensional queueing model stemming from the production of Printed Circuit Boards by a flexible assembly system. We have called this model the *Shortest Queue System with a Job-Dependent Parallelism (SQS-JDP)*, and we established that this model represented a generalization of the Symmetric Shortest Queue System (SSQS). The latter model was known to be already a hard problem, and therefore we focused on the SSQS first. Since the analysis of the previous chapter has pointed out that the SSQS can successfully be treated by exploiting flexible bound models derived by the precedence relation method, we may now return to the SQS-JDP to investigate whether this system also can be analyzed successfully by using flexible bound models. This constitutes the main objective of this chapter.

It is noted that we would like to exploit flexible bound models for the SQS-JDP in a numerical procedure to determine the response times for a given assignment of components, which have to be mounted on the Printed Circuit Boards, to the parallel insertion machines of the flexible assembly system. Such a procedure could be a useful tool for the selection of good assignments of the components; see also Section 1.2.

As already noted in Section 1.2, the SQS-JDP or similar systems have hardly been studied in the literature, despite the fact that they occur in several practical situations. To our knowledge, the SQS-JDP itself has only been studied in Adan et al. [7], in which approximations are given for the mean waiting times for all different job types. Further, models similar to the SQS-JDP have been studied by Schwartz [63] (see also Roque [61]), who also gives approximations for the mean waiting times, and by Green [37], who derives flexible truncation models which can efficiently be solved by the matrix-geometric approach; see also Section 1.2, where we have discussed these papers more extensively. Finally, Hassin and Haviv [40] have studied a two-dimensional SSQS (with Threshold Jockeying), where an arriving job has to pay a fixed amount of money if he wants to have information on which queue is the shortest. This leads to a model, where an arriving jobs joins the shortest queue with

probability p and it joins a randomly chosen queue with probability $1-p$. This model in fact is a generalization of the Threshold Jockeying model which we obtained as a bound model for the SSQS, and, in case the threshold parameter T is equal to infinity, it constitutes a special case of the SQS-JDP with $N=2$ servers.

In this chapter, we apply the precedence relation method to the SQS-JDP. We shall derive a flexible lower bound model, called the Threshold Killing and Rejection (TKR) model, and an upper bound model, called the Threshold Blocking and Addition (TBA) model, which produce bounds for many performance measures, among which the distributions, and thus also the means, of the normalized waiting times for all different job types and for all job types together. The TBA model is a generalization of the Threshold Blocking model, which in the previous chapter has been used as an upper bound model for the SSQS, and the TKR model is a generalization of the Threshold Killing model, which in Chapter 5 has been presented as one of the four lower bound models for the SSQS with $N=2$ servers. Both the TKR and TBA model can be solved by the matrix-geometric approach again (but, in this case, we obtain full rate matrices R, and therefore they cannot be solved as efficiently as the TJ and TB model of the previous chapter). The TKR and TBA model are exploited in a numerical procedure for the determination of the mean normalized waiting times in a SQS-JDP with $N=2$ servers, and numerical results will be presented to show how well the original SQS-JDP can be approximated by its bound models.

The organization of this chapter is as follows. In Section 7.2, we describe the model for the SQS-JDP. The flexible bound models are derived in Section 7.3, and after that the solution of these models is discussed in Section 7.4. In Section 7.5, some numerical results are presented, and finally the conclusions are given in Section 7.6.

7.2. Model

The objective of this section is to describe a discrete-time, irreducible Markov cost model for the SQS-JDP, where direct cost functions are defined such that the corresponding average costs are equal to the mean normalized waiting times for the different job types. A side result that is obtained, is a condition for the ergodicity of a SQS-JDP. By a simple argument, it is shown that this condition is necessary for the ergodicity; after that, by studying a system which is identical to the SQS-JDP, but which has a static routing instead of the dynamic shortest queue routing, it is made plausible that this condition is also sufficient. Further, we shall define in which case a SQS-JDP is said to be *balanced* and/or *symmetric*.

Description of the Markov cost model for the SQS-JDP

The SQS-JDP consists of $N \geq 2$ servers, which all have their own queue, and there are several types of jobs which must be served by the SQS-JDP. Because of technical reasons, for example, each server can only serve a restricted set of job types. It is assumed that all service times are exponentially distributed with the same parameter $\mu > 0$; note that this implies that the service times are independent of the job type and that all servers work equally fast. Furthermore, it is assumed that for each job type, the jobs arrive according to a Poisson stream, and that each arriving job joins the shortest queue of all queues where the job can be served (ties are broken with equal probabilities). In Figure 7.1, we have depicted a SQS-JDP with $N=2$

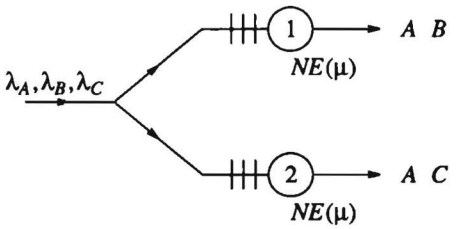


Figure 7.1. A SQS-JDP with $N=2$ servers and 3 job types.

servers and three types of jobs: jobs of type A, which arrive with intensity λ_A and which can be served by both servers, jobs of type B, which arrive with intensity λ_B and which can only be served by server 1, and jobs of type C, which arrive with intensity λ_C and which must be served by server 2.

We introduce the following notations. The servers are numbered $1,\ldots,N$ and the set I is defined by $I:=\{1,\ldots,N\}$. Let the set J contain the job types. For each $j\in J$, we let $I(j)$ denote the set of servers which can serve the jobs of type j. Assume that each job type can be served by at least one server and that each server can handle at least one job type; so, $I(j)\neq\emptyset$ for all $j\in J$, and $\cup_{j\in J} I(j)=I$. Further, for each $j\in J$, the arrival intensity of the Poisson stream of the jobs of type j is given by $\lambda_j>0$, and $\lambda:=\Sigma_{j\in J}\lambda_j$ denotes the total arrival intensity. For simplicity, we assume that time is scaled such that $\lambda+N\mu=1$. Finally, the average workload per server is given by $\rho=\lambda/(N\mu)$; it is obvious that we at least must satisfy the condition that $\rho<1$ in order to have an ergodic system.

For the application of the precedence relation method, it is required that the SQS-JDP is modeled as a discrete-time, irreducible Markov process. This is done as follows. Assume that the servers always work, but that a service completion is only accompanied by a departure of a job if there is a job present in the corresponding queue. Then the behavior of the SQS-JDP may be described by the discrete-time, irreducible Markov process on the time instants right after job arrivals and service completions, and with states $(m_1,\ldots,m_N)$, where m_i denotes the length of the queue at server i, $i\in I$ (jobs in service are included). So, the state space is equal to

$$M = \{\, m \mid m=(m_1,\ldots,m_N) \text{ with } m_i\in\mathbb{N}_0 \text{ for all } i\in I \,\}. \tag{7.1}$$

Let the transition probabilities be denoted by $q_{m,n}$. In Figure 7.2, we have depicted the transition probabilities for the SQS-JDP of Figure 7.1.

The performance measures we are interested in are the mean *normalized* waiting times $W^{(j)}$ for all job types $j\in J$, and the mean normalized waiting time for all job types together, which is equal to

$$W = \sum_{j\in J}\frac{\lambda_j}{\lambda}W^{(j)}; \tag{7.2}$$

here, the normalized waiting time is defined as the waiting time divided by the mean service time. It is obvious that for each $j\in J$,

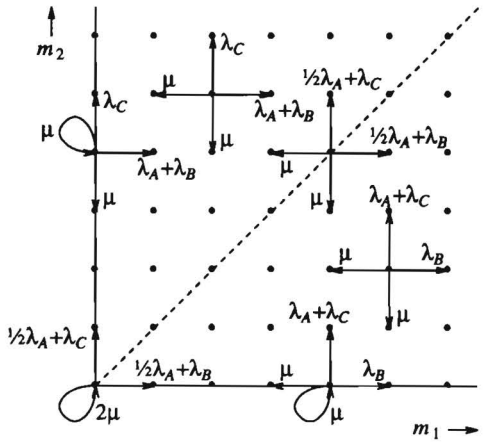


Figure 7.2. The transition probabilities for the discrete-time, irreducible Markov process for the SQS-JDP depicted in Figure 7.1.

$$W^{(j)} = L_{sq}^{(j)}, \tag{7.3}$$

where $L_{sq}^{(j)}$ denotes the mean of the shortest queue length of the queues at the servers $i \in I(j)$. We thus find that $W^{(j)}$ is equal to the average costs corresponding to the direct cost function $c^{(j)}(m)$ defined by

$$c^{(j)}(m) = \min_{i \in I(j)} m_i \quad \text{for all } m=(m_1, \ldots, m_N) \in M,\ j \in J. \tag{7.4}$$

This completes the description of the Markov cost model for the SQS-JDP.

A simple, necessary condition for the ergodicity of the SQS-JDP

By studying the routing of the jobs, we obtain a simple, *necessary* condition for the ergodicity of the SQS-JDP. For each subset $J' \subset J$, $J \neq \emptyset$, jobs of the types $j \in J'$ arrive with an intensity equal to $\Sigma_{j \in J'} \lambda_j$ and they must be served by the servers $\cup_{j \in J'} I(j)$. This shows that the SQS-JDP can only be ergodic if the following condition is satisfied:

$$\sum_{j \in J'} \lambda_j < | \bigcup_{j \in J'} I(j) | \mu \quad \text{for all } J' \subset J,\ J' \neq \emptyset. \tag{7.5}$$

Note that for $J'=J$, this inequality is equivalent to $\rho < 1$. For the SQS-JDP of Figure 7.1, condition (7.5) states that for the ergodicity at least the inequalities $\lambda_B < \mu$, $\lambda_C < \mu$ and $\lambda < 2\mu$ (or $\rho < 1$) must be satisfied.

Condition (7.5) is expected to be also *sufficient* for the ergodicity of the SQS-JDP. This is argued by considering so-called corresponding static systems.

A *corresponding static system* is a system that is identical to the SQS-JDP, but which uses a *static routing* instead of the dynamic shortest queue routing to route the arriving jobs to the servers. Such a static routing is described by discrete distributions $\{x_i^{(j)}\}_{i \in I(j)}$, $j \in J$, where for each $j \in J$ and $i \in I(j)$, the variable $x_i^{(j)}$ denotes the probability with which an

arriving job of type j is sent to server i. Under a static routing, for each $j \in J$ the Poisson stream of arriving jobs of type j may be divided into Poisson streams with intensities $x_{j,i} = \lambda_j x_i^{(j)}$, $i \in I(j)$, for arrivals of jobs of type j which join server i, and we obtain that the queues at the servers $i \in I$ constitute independent $M \mid M \mid 1$ queues with identical mean service times equal to $1/\mu$ and with arrival intensities $\Sigma_{j \in J,\,(j,i) \in A} x_{j,i}$, where

$$A := \{ (j,i) \mid j \in J,\ i \in I \text{ and } i \in I(j) \}.$$

As a result, we obtain a simple necessary and sufficient condition for the ergodicity of a corresponding static system, viz.

$$\sum_{\substack{j \in J \\ (j,i) \in A}} x_{j,i} < \mu \quad \text{for all } i \in I.$$

The observation that in situations with many jobs in the system the shortest queue routing will balance the queue lengths better (or at least equally good) than any static routing (more specifically, the shortest queue routing will better control the maximum queue length), leads to the *conjecture* that if there is a corresponding static system which is ergodic, then the SQS-JDP itself will also be ergodic. Further, from Lemma 7.1 stated below it follows that condition (7.5) guarantees that there exists a corresponding static system which is ergodic. By combining these two results, we find that it is reasonable to expect that condition (7.5) is not only necessary but also sufficient for the ergodicity of the SQS-JDP.

Lemma 7.1.

Condition (7.5) is necessary and sufficient for the existence of a corresponding static system which is ergodic.

Proof.

The necessity follows by the same arguments as used when deriving condition (7.5) for the SQS-JDP. For the sufficiency, we prove that condition (7.5) implies that there exists a nonnegative solution $\{x_{j,i}\}_{(j,i) \in A}$, with the set A defined as above, of the following equations and inequalities:

$$\sum_{\substack{i \in I \\ (j,i) \in A}} x_{j,i} = \lambda_j \quad \text{for all } j \in J, \qquad \sum_{\substack{j \in J \\ (j,i) \in A}} x_{j,i} < \mu \quad \text{for all } i \in I; \tag{7.6}$$

the equalities in (7.6) guarantee that the solution $\{x_{j,i}\}_{(j,i) \in A}$ corresponds to discrete distributions $\{x_i^{(j)}\}_{i \in I(j)}$ which describe a static routing, and the inequalities in (7.6) must be satisfied for the ergodicity.

Assume that condition (7.5) is satisfied. To prove that there exists a nonnegative solution $\{x_{j,i}\}_{(j,i) \in A}$ of (7.6), we consider a *transportation problem* with supply nodes $\hat{V}_1 = J \cup \{0\}$, demand nodes $\hat{V}_2 = I$, and arcs $\hat{A} = A \cup \{ (0,i) \mid i \in I \}$ (supply node 0 denotes an extra type of jobs, which may be served by all servers). Define the supplies $\hat{a}_j$ by $\hat{a}_j = \lambda_j$ for all $j \in \hat{V}_1 \setminus \{0\}$ and $\hat{a}_0 = N\mu - \lambda - N\varepsilon$, where

$$\varepsilon := \min_{\substack{J' \subset J \\ J' \neq \emptyset}} \frac{|\cup_{j \in J'} I(j)|\, \mu - \Sigma_{j \in J'} \lambda_j}{|\cup_{j \in J'} I(j)|}$$

(from (7.5), it follows that $\varepsilon > 0$, and $\hat{a}_0 \geq 0$ since by taking $J' = J$ we obtain the inequality

$\varepsilon \leq (N\mu-\lambda)/N$). Further, we define the demands $\hat{b}_i$ by $\hat{b}_i = \mu-\varepsilon$ for all $i \in \hat{V}_2$; note that $\Sigma_{j\in\hat{V}_1}\hat{a}_j = \Sigma_{i\in\hat{V}_2}\hat{b}_i$. For this transportation problem, we satisfy condition (5.25) stated in Lemma 5.4 (distinguish between the two cases $0 \in U$ and $0 \notin U$, to verify this). So, we find that there exists a feasible flow for the transportation problem, i.e. there exists a nonnegative solution $\{\hat{x}_{j,i}\}_{(j,i)\in\hat{A}}$ of the equations

$$\sum_{\substack{i\in\hat{V}_2\\(j,i)\in\hat{A}}} \hat{x}_{j,i} = \hat{a}_j \quad \text{for all } j \in \hat{V}_1, \qquad \sum_{\substack{j\in\hat{V}_1\\(j,i)\in\hat{A}}} \hat{x}_{j,i} = \hat{b}_i \quad \text{for all } i \in \hat{V}_2.$$

It is easily seen that then the solution $\{x_{j,i}\}_{(j,i)\in A}$ defined by $x_{j,i} = \hat{x}_{j,i}$ for all $(j,i) \in A$, is a nonnegative solution of (7.6), which completes the proof. □

Balanced and symmetric systems

From a practical point of view, for a SQS-JDP it is desirable that the job-dependent structure is such that the workloads for the different servers can be balanced. Formally, we say that a SQS-JDP is *balanced*, if there exists a corresponding static system for which the queues at all servers constitute $M \mid M \mid 1$ queues with equal workloads. This means that there must exist discrete distributions $\{x_i^{(j)}\}_{i\in I(j)}$ such that for each server $i \in I$, the arrival intensity $\Sigma_{j\in J,\,(j,i)\in A} x_{j,i}$ is equal to $\lambda/N = \rho\mu$, where $x_{j,i} = \lambda_j x_i^{(j)}$ for all $j \in J$ and $i \in I(j)$ and $A = \{(j,i) \mid j \in J,\ i \in I \text{ and } i \in I(j)\}$. Such discrete distributions exist if and only if there exists a nonnegative solution $\{x_{j,i}\}_{(j,i)\in A}$ of the equations

$$\sum_{\substack{i\in I\\(j,i)\in A}} x_{j,i} = \lambda_j \quad \text{for all } j \in J, \qquad \sum_{\substack{j\in J\\(j,i)\in A}} x_{j,i} = \frac{\lambda}{N} \quad \text{for all } i \in I. \tag{7.7}$$

These equations are the equations which must be satisfied by a feasible flow for the transportation problem with supply nodes $V_1 = J$, demand nodes $V_2 = I$, arcs A, supplies $a_j = \lambda_j$ for all $j \in V_1$ and demands $b_i = \lambda/N$ for all $i \in V_2$. Applying Lemma 5.4 shows that a necessary and sufficient condition for the existence of a nonnegative solution of (7.7), and thus also for a SQS-JDP to be balanced, is given by:

$$\sum_{j\in J'} \lambda_j \leq \Big|\bigcup_{j\in J'} I(j)\Big| \frac{\lambda}{N} \quad \text{for all } J' \subset J. \tag{7.8}$$

Note that for $J' = \emptyset$ and $J' = J$, this condition is satisfied by definition. Further, it follows that a balanced SQS-JDP satisfies condition (7.5) if and only if $\rho < 1$. So, for a balanced SQS-JDP, the simple condition $\rho < 1$ is not only necessary for the ergodicity, but it may be expected to be also sufficient.

If a SQS-JDP is balanced, i.e. if there exists a static routing under which the workloads for the servers of the system are balanced, then this does not necessarily mean that these workloads are also balanced under the shortest queue routing. This can be seen by considering the SQS-JDP depicted in Figure 7.1. According to condition (7.8), this SQS-JDP is balanced if and only if $\lambda_B \leq \frac{1}{2}\lambda$ and $\lambda_C \leq \frac{1}{2}\lambda$, i.e. if and only if $\lambda_B \leq \lambda_A + \lambda_C$ and $\lambda_C \leq \lambda_A + \lambda_B$. This condition is obviously satisfied if we take $\lambda_C = \lambda_A + \lambda_B$. For this case, equal workloads for both servers can only be obtained if all jobs of type A are sent to server 1. But, under the shortest queue routing, there always will occur situations in which jobs of type A are sent to server 2, and therefore server 2 will have a higher workload than server 1. Nevertheless, it is expected that for a balanced SQS-JDP, the shortest queue routing at least ensures that the

workloads for all servers will not differ too much.

A subclass of balanced systems is constituted by the symmetric systems. A SQS-JDP is said to be *symmetric*, if

$$\lambda(I_1) = \lambda(I_2) \quad \text{for all } I_1, I_2 \subset I \text{ with } |I_1| = |I_2|, \tag{7.9}$$

where

$$\lambda(I') := \sum_{\substack{j \in J \\ I(j) = I'}} \lambda_j \ , \quad I' \subset I.$$

So, a SQS-JDP is symmetric, if for all subsets $I' \subset I$ with the same number of servers $|I'|$, the arrival intensity $\lambda(I')$ for the jobs which can be served by precisely the servers of I', is the same. The SQS-JDP of Figure 7.1 is symmetric if $\lambda_B = \lambda_C$.

For a symmetric SQS-JDP, all queue lengths are equally distributed which implies that all servers have equal workloads. The behavior of a symmetric SQS-JDP may be described by a Markov process with states $(m_1, \ldots, m_N)$, where m_i denotes the length of the i-th shortest queue, $i \in I$. In this case, we can derive precedence pairs of the type $(m, m+e_i-e_j)$, $i > j$. As a consequence, by using the precedence relation method, it can be shown that a symmetric SQS-JDP has a stochastically smaller number of jobs, and thus also a smaller mean normalized waiting time for all job types together, than the corresponding system consisting of N independent $M|M|1$ queues with workload ρ (see also Remark 5.1). So, it may be concluded that $\rho < 1$ is a necessary and sufficient condition for the ergodicity of a symmetric SQS-JDP. Further, it can be shown that for a symmetric SQS-JDP, the shortest queue routing minimizes the total number of jobs in the system and thus also the mean normalized waiting time (this may be done by the technique used by Hordijk and Koole [43, 44]).

7.3. Application of the precedence relation method

In this section, we apply the precedence relation method to the Markov cost model for a general SQS-JDP. We shall derive two flexible truncation models, which lead to lower and upper bounds for the mean normalized waiting times $W^{(j)}$ for the different job types $j \in J$, and, by (7.2), also to lower and upper bounds for the mean normalized waiting time W for all job types together.

The first step of the precedence relation method consists of the derivation of precedence pairs (m,n), $m,n \in M$. For the cost functions $c^{(j)}(m)$ given by (7.4), which all are nondecreasing in each component, we can derive precedence pairs of the type $(m, m+e_i)$ (it is noted that for a general SQS-JDP, we *cannot* derive precedence pairs of the type $(m, m+e_i-e_j)$, with $m_i \geq m_j$, also not in case the cost functions $c^{(j)}(m)$ would be defined such that $c^{(j)}(m) \leq c^{(j)}(n)$ for pairs (m,n) of this type). We define

$$P = \{ (m,n) \mid m,n \in M \text{ and } n = m+e_i \text{ for some } i \in I \}, \tag{7.10}$$

and the corresponding completed set P^* is given by

$$P^* = \{ (m,n) \mid m,n \in M \text{ and } m \leq n \}, \tag{7.11}$$

where the inequality $m \leq n$ for two vectors m and n means that m must be smaller than or

equal to n in each component. That the sets P and P^* consist of precedence pairs is proved by applying Lemma 5.1.

Lemma 7.2.

The set P defined by (7.10) and the corresponding completed set P^* given by (7.11) consist of precedence pairs for all cost functions $c(m)$ which are nondecreasing in each component.

Proof.

Consider the Markov cost model with direct cost function $c(m)$, and assume that $c(m)$ is nondecreasing in each component. Further, let the functions $v_t(m)$ denote the t-period costs Applying Lemma 5.1 shows that we must prove that

$$\sum_{\substack{r\in M \\ q_{m,r}>0}} q_{m,r}\, v_t(r) \le \sum_{\substack{r\in M \\ q_{n,r}>0}} q_{n,r}\, v_t(r) \quad \text{for all } (m,n)\in P, \tag{7.12}$$

where it is known that $v_t(m)\le v_t(n)$ for all $(m,n)\in P^*$ and t is a fixed nonnegative integer.

In the Markov cost model for the SQS-JDP, for each state $m\in M$ all outgoing transitions are due to service completions and arrivals of jobs, and it holds that

$$\sum_{\substack{r\in M \\ q_{m,r}>0}} q_{m,r}\, v_t(r) = \sum_{i\in I} \mu\, v_t(r_i(m)) + \sum_{j\in J} \lambda_j \sum_{i\in I(j;m)} \frac{1}{|I(j;m)|}\, v_t(m+e_i)\,, \tag{7.13}$$

where $r_i(m)$ denotes the unique state to which a transition is made if a Markov process is in state m and a service completion at server i occurs, and $I(j;m)$ denotes the set of states to which transitions may be made if the Markov process is in state m and a job of type j arrives at the system, i.e. $I(j;m)$ denotes the servers of $I(j)$ for which the corresponding queues are the shortest:

$$I(j;m) = \{\, i\in I(j) \mid m_i = \min_{k\in I(j)} m_k \,\}\,. \tag{7.14}$$

Exploiting (7.13) shows that for the proof of (7.12) it suffices to prove that for all $(m,n)\in P$,

$$v_t(r_i(m)) \le v_t(r_i(n)) \quad \text{for all } i\in I; \tag{7.15}$$

$$\sum_{i\in I(j;m)} \frac{1}{|I(j;m)|}\, v_t(m+e_i) \le \sum_{i\in I(j;n)} \frac{1}{|I(j;n)|}\, v_t(n+e_i) \quad \text{for all } j\in J. \tag{7.16}$$

Let $(m,n)\in P$ with $n=m+e_l$ for some $l\in I$. We first prove (7.15) for a given $i\in I$. It suffices to show that $r_i(m)\le r_i(n)$, since, according to the set P^*, this implies that $v_t(r_i(m))\le v_t(r_i(n))$. We distinguish three cases:

- $m_i=0$ and $i=l$: Then $r_i(m)=m=n-e_i=r_i(n)$.
- $m_i=0$ and $i\ne l$: Then also $n_i=0$, and $r_i(m)=m\le n=r_i(n)$.
- $m_i>0$: In this case, we find $r_i(m)=m-e_i\le n-e_i=r_i(n)$.

The proof of (7.16) for a given $j\in J$ is slightly more complicated. We again distinguish three cases:

- $l\notin I(j;m)$: Then $I(j;n)=I(j;m)$, and the inequality stated in (7.15) immediately follows from the property that $m+e_i\le n+e_i$ for all $i\in I(j;m)$.

- $l \in I(j;m)$ and $|I(j;m)| = 1$: In this case, we find that

$$\sum_{i \in I(j;m)} \frac{1}{|I(j;m)|} v_t(m+e_i) = v_t(m+e_l) = v_t(n) \leq \sum_{i \in I(j;n)} \frac{1}{|I(j;n)|} v_t(n+e_i) .$$

- $l \in I(j;m)$ and $|I(j;m)| \geq 2$: Then $I(j;n) = I(j;m) \setminus \{l\}$, and

$$\begin{aligned} \sum_{i \in I(j;m)} \frac{1}{|I(j;m)|} v_t(m+e_i) &= \frac{1}{|I(j;m)|} v_t(n) + \sum_{i \in I(j;n)} \frac{1}{|I(j;m)|} v_t(m+e_i) \\ &\leq \sum_{i \in I(j;n)} \frac{1}{|I(j;m)|\,|I(j;n)|} v_t(n+e_i) + \sum_{i \in I(j;n)} \frac{1}{|I(j;m)|} v_t(n+e_i) \\ &= \sum_{i \in I(j;n)} \frac{1}{|I(j;n)|} v_t(n+e_i) . \end{aligned}$$

This completes the proof of Lemma 7.2. □

The second step of the precedence relation method consists of the definition of flexible bound models. By using the precedence pairs given by P^*, we derive the following two flexible bounds models:

* *Threshold Killing and Rejection (TKR):*
Because of the shortest queue routing, in general there will be a drift to the states with equal queue lengths. Therefore, we define the truncated state space M' by

$$M' = \{ m \in M \mid m = (m_1, \ldots, m_N) \text{ and } m_i \leq \min(m) + T_i \text{ for all } i \in I \} , \tag{7.17}$$

where $\min(m) := \min_{i \in I} m_i$ and $T_1, \ldots, T_N$ are positive integers. The definition of M' shows that a state $m \in M$ is contained in the truncated state space M' if and only if for each $i \in I$ the length m_i of the queue at server i is at most T_i larger than the length of any other queue. The variables T_i are called the threshold parameters and they are contained in the threshold vector $\hat{T} = (T_1, \ldots, T_N)$.
For the truncated state space M' defined by (7.17), there are two types of transitions pointing from states inside M' to states outside M': transitions due to service completions and transitions due to arrivals of new jobs. Let $m = (m_1, \ldots, m_N) \in M'$ be a state for which the set $I' = \{ i \in I \mid m_i = \min(m) + T_i \}$ is not empty. If $\min(m) > 0$, then a service completion at one of the servers or queues $k \in I$ with $m_k = \min(m)$ leads to a transition from m to state $n = m - e_k \notin M'$. This transition occurs with probability μ and is redirected to the state $n' = m - e_k - \sum_{i \in I'} e_i$, which is more attractive than state n according to the set P^*. Further, an arrival of a new job at one of the queues $i \in I'$ leads to a transition from m to the state $n = m + e_i \notin M'$. This transition occurs with probability $\sum_{j \in J} |I(j;m)|^{-1} \lambda_j 1_{\{i \in I(j;m)\}}$, where $I(j;m)$ is defined by (7.14) (note that this probability may be equal to 0), and this transition is redirected to the more attractive state $n' = m$ itself.
The physical interpretation of the redirections of the first type is that a departure of a job at a non-empty shortest queue is accompanied by a destruction or killing of one job at each of the queues $i \in I'$, for which the difference with respect to the shortest queues has already reached its maximum value T_i. The physical interpretation of the redirections of the second type is that a new job arriving at one of the servers $i \in I'$ is rejected. Therefore, the truncation model obtained by these redirections, is called the Threshold Killing

and Rejection (TKR) model. Since all transitions ending in states outside M' have been redirected to more attractive states inside M', the TKR model leads to lower bounds for all average costs corresponding to direct cost functions $c(m)$ which are nondecreasing in each component.

* *Threshold Blocking and Addition (TBA):*
To obtain an upper bound model, the same truncated state space M' is taken as for the TKR model, but the transitions ending in states outside M' are redirected to *less* attractive states. For each $m=(m_1,\ldots,m_N)\in M'$ with $I'=\{\, i\in I \mid m_i=\min(m)+T_i \,\}\neq\emptyset$, the transitions from state m to states n outside of M' are redirected as follows. If $\min(m)>0$, then the transition from m to $n=m-e_k\notin M'$, which is due to a service completion at a server $k\in I$ with $m_k=\min(m)$, is redirected to the less attractive state $n'=m$ itself. This means that if for some queues the difference with respect to the shortest queues has already reached its maximum value, then a service completion at a non-empty shortest queue is not accompanied by a departure, and the job in service has to be served once more; (because of the memory-less property of the exponential service times) this is equivalent to saying that then the servers at the shortest queues are blocked. Further, for each $i\in I'$, the transition from state m to state $n=m+e_i\notin M'$, which is due to an arrival of a new job at queue i, is redirected to the less attractive state $n'=m+e_i+\Sigma_{k\in I_{sq}} e_k$ with $I_{sq}=\{\, k\in I \mid m_k=\min(m) \,\}$. This means that an arrival of a new job at one of the queues for which the difference with respect to the shortest queues has already reached its maximum value, is accompanied by the addition of one extra job at each of the shortest queues. The upper bound model that is obtained by these redirections, is called the Threshold Blocking and Addition (TBA) model, and it leads to upper bounds for all average costs corresponding to direct cost functions $c(m)$ which are non-decreasing in each component.

In Figure 7.3, we have depicted the lower and upper bound model which are obtained for the SQS-FDP of Figure 7.1.

Since the TKR and TBA model are bound models for all cost functions $c(m)$ which are nondecreasing in each component, they lead to lower and upper bounds for the distribution and all moments of the shortest queue length of the queues at the servers $i\in I(j)$, where $j\in J$. As a result, it may be shown that they also lead to lower and upper bounds for the distributions and all moments of the normalized waiting times for the different job types $j\in J$ and for all job types together. However, in this chapter we only focus on the bounds for the means of the normalized waiting times.

For the cost function $c^{(j)}(m)$ given by (7.4), the TKR and TBA model produce bounds for the mean normalized waiting time $W^{(j)}$ for job type $j\in J$, which are denoted by $W^{(j)}_{TKR}(\hat{T})$ and $W^{(j)}_{TBA}(\hat{T})$, where $\hat{T}=(T_1,\ldots,T_N)$ is the threshold vector which determines the size of the truncated state space M'. The corresponding bounds for the mean normalized waiting time W for all job types together are given by (cf. (7.2))

$$W_{TKR}(\hat{T}) = \sum_{j\in J}\frac{\lambda_j}{\lambda} W^{(j)}_{TKR}(\hat{T}) \quad \text{and} \quad W_{TBA}(\hat{T}) = \sum_{j\in J}\frac{\lambda_j}{\lambda} W^{(j)}_{TBA}(\hat{T}). \tag{7.18}$$

It is noted that the lower bounds $W^{(j)}_{TKR}(\hat{T})$ and $W_{TKR}(\hat{T})$ are somewhat larger than the normalized waiting times (i.e. the waiting times divided by the mean $1/\mu$ of one regular service) in the TKR model itself, and that the upper bounds $W^{(j)}_{TBA}(\hat{T})$ and $W_{TBA}(\hat{T})$ are somewhat smaller

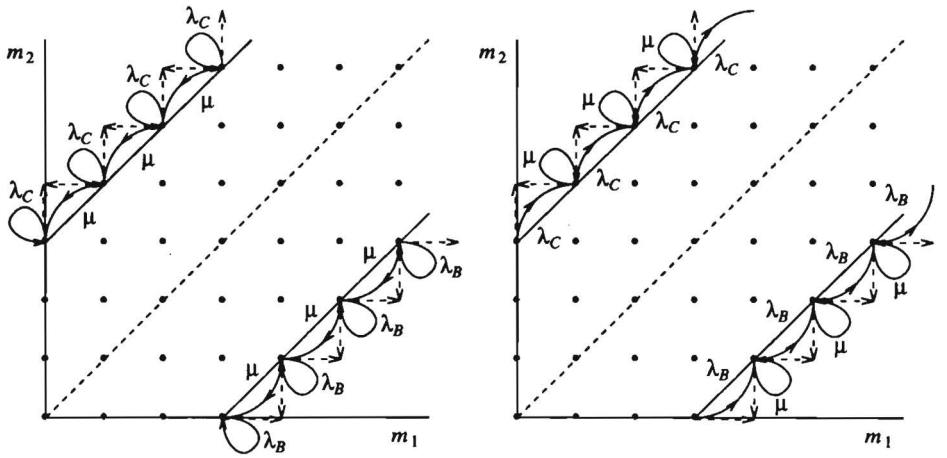


Figure 7.3. The redirections for the two flexible bound models which are obtained for the SQS-JDP depicted in Figure 7.1. For both models, the threshold vector has been taken equal to $\hat{T}=(3,3)$.

than the normalized waiting times in the TBA model itself; see also Remark 6.1.

By deriving precedence pairs of the type $(m,m+e_i)$ for the TKR and TBA model itself and applying the precedence relation method, it may be shown that the lower bounds $W^{(j)}_{TKR}(\hat{T})$ and $W_{TKR}(\hat{T})$ are monotonously non-decreasing and the upper bounds $W^{(j)}_{TBA}(\hat{T})$ and $W_{TBA}(\hat{T})$ are monotonously non-increasing for increasing values of one or more threshold parameters T_i. Since besides the bounds $W^{(j)}_{TKR}(\hat{T})$ and $W^{(j)}_{TBA}(\hat{T})$ will tend to $W^{(j)}$ for all $j \in J$ and the bounds $W_{TKR}(\hat{T})$ and $W_{TBA}(\hat{T})$ will tend to W as $T_i \to \infty$ for each $i \in I$, i.e. as $\hat{T}=(T_1,\ldots,T_N)\to(\infty,\ldots,\infty)$, we find that

$$W^{(j)}_{TKR}(\hat{T}) \uparrow W^{(j)} \quad \text{and} \quad W^{(j)}_{TBA}(\hat{T}) \downarrow W^{(j)} \quad \text{for all } j \in J, \tag{7.19}$$

$$W_{TKR}(\hat{T}) \uparrow W \quad \text{and} \quad W_{TBA}(\hat{T}) \downarrow W, \tag{7.20}$$

as $\hat{T} \to (\infty,\ldots,\infty)$. This shows that the mean normalized waiting times $W^{(j)}$ and W can be determined as accurately as desired by computing the corresponding lower and upper bounds for increasing values of the parameters T_i of the threshold vector $\hat{T}$. In Section 7.5, it will be investigated whether this leads to an efficient procedure; the efficiency is expected to depend mainly on how well the TKR and TBA model approximate the original model for the SQS-JSP, i.e. on the strength of the drift to the states with equal queue lengths (note that, to obtain a sufficiently strong drift to these states, a SQS-JDP must at least be balanced). But first, in Section 7.4, we discuss the computation of the bounds itself.

7.4. Solving the flexible bound models by the matrix-geometric approach

Just as the two flexible bound models which we derived for the SSQS in the previous chapter, the TKR and TBA model can be solved by the matrix-geometric approach. Therefore, the bounds produced by the TKR and TBA model can be computed quite efficiently. However, since in this case we obtain full rate matrices R, unfortunately, they cannot be computed as efficiently as the bounds for the SSQS.

Assume that the model parameters for the SQS-JDP and the threshold vector $\hat{T} = (T_1, \ldots, T_N)$ are given. Then we have a discrete-time, irreducible Markov process for the original model, and similarly for the TKR and TBA model. Next, assume that condition (7.5) is satisfied, by which it may be expected that the Markov process for the original model is positive recurrent (= ergodic), and thus also that the Markov process for the TKR model is positive recurrent.

For both the TKR and TBA model, the truncated state space M' is given by (7.17). The transition probabilities and equilibrium probabilities are denoted by $q^{tr}_{m,n}$ and p^{tr}_m, respectively, where the indices tr may be replaced by *TKR* or *TBA*.

On behalf of the application of the matrix-geometric approach, the state space M' is partioned into levels M'_l:

$$M'_l = \{ m \in M' \mid \min(m) = l \} \quad \text{for all } l \geq 0.$$

It is easily verified that each level M'_l contains

$$a = \sum_{\substack{I' \subset I \\ I' \neq \emptyset}} \prod_{i \notin I'} T_i = \prod_{i \in I} (T_i + 1) - \prod_{i \in I} T_i$$

states; note that $a = 1 + T_1 + T_2$ if $N = 2$. Within each level M'_l the states are lexicographically ordered, the vector p^{tr}_l denotes the equilibrium probabilities p^{tr}_m for the states m of level M'_l, and p^{tr} is equal to $(p^{tr}_0, p^{tr}_1, \cdots)$. The transition matrix P^{tr} is of the form

$$P^{tr} = \begin{bmatrix} B_1 & A_0 & 0 & 0 & 0 & \cdots \\ A_2 & A_1 & A_0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}, \tag{7.21}$$

where A_0, A_1, A_2 and B_1 are squared, real-valued, nonnegative matrices of order a.

For the existence of the equilibrium distribution $\{p^{tr}_m\}$, it is required that we have a positive recurrent Markov process. The Markov process with transition matrix P^{tr} may be shown to be positive recurrent if and only if

$$\pi A_0 e < \pi A_2 e, \tag{7.22}$$

where $\pi = (\pi_1, \ldots, \pi_a)$ is the unique equilibrium distribution of the finite, discrete-time, irreducible Markov process with transition matrix $A := A_0 + A_1 + A_2$ (cf. Theorem 1.3.2 of [58]). Condition (7.22) provides us with a simple, necessary and sufficient condition for the positive recurrence. Although, under the assumption that condition (7.5) is satisfied, the TKR

model is expected to be positive recurrent, we nevertheless advise to check whether condition (7.22) is satisfied for this model. If condition (7.22) is not satisfied, then the bounds $W_{ir}^{(j)}(\hat{T})$ and $W_{tr}(\hat{T})$ may be taken equal to ∞, else they can be obtained from the equilibrium distribution $\{p_m^{tr}\}$, as described below.

Now, suppose that condition (7.22) is satisfied. Then, by applying Theorem 1.3.2 of [58], we find that

$$p_l^{tr} = p_0^{tr} R^l \quad \text{for all } l \geq 0, \tag{7.23}$$

where the $a \times a$ matrix R is the minimal nonnegative solution of the quadratic matrix equation

$$R = A_0 + RA_1 + R^2 A_2 . \tag{7.24}$$

The matrix R is called the *rate matrix* and has a spectral radius $sp(R)$ smaller than 1; it is noted that R may be obtained by performing successive substitutions in (7.24), where $R = 0$ is used as starting matrix. By the matrix-geometric result stated in (7.23), all equilibrium probabilities are known, once the probability vector p_0^{tr} is determined. The vector p_0^{tr} is characterized as the unique solution of the linear equations

$$p_0^{tr} = p_0^{tr} (B_1 + RA_2), \quad p_0^{tr} (I-R)^{-1} e = 1 , \tag{7.25}$$

which are obtained by substituting (7.23) into the equilibrium equations for the states of level M_0' and into the normalization equation.

After having determined the rate matrix R and the probability vector p_0^{tr}, we can compute the bounds $W_{ir}^{(j)}(\hat{T})$ and $W_{tr}(\hat{T})$ as follows. For each $j \in J$, the bound $W_{ir}^{(j)}(\hat{T})$ is equal to the average costs corresponding to the direct cost function $c^{(j)}(m)$ given by (7.4). Let the column vector $c_l^{(j)}$ contain the direct costs $c^{(j)}(m)$ for all states m of level M_l'. Then

$$c_l^{(j)} = c_0^{(j)} + l\, e \quad \text{for all } l \geq 0,$$

and, by using this result and the matrix-geometric result stated in (7.23), we find that

$$\begin{aligned} W_{ir}^{(j)}(\hat{T}) &= \sum_{l=0}^{\infty} p_l^{tr} c_l^{(j)} = \sum_{l=0}^{\infty} p_0^{tr} R^l (c_0^{(j)} + l\, e) \\ &= p_0^{tr} (I-R)^{-1} c_0^{(j)} + p_0^{tr} R (I-R)^{-1} e . \end{aligned} \tag{7.26}$$

The bound $W_{tr}(\hat{T})$ follows from its definition given in (7.18).

Contrary to the rate matrices R for the bound models which we derived for the SSQS in the previous chapter, the rate matrices R for the TKR and TBA model have almost no zero rows. As a consequence, for both the TKR and TBA model, the determination of the rate matrix R requires a large computational effort, and it constitutes the main part of the computation time required for the determination of the bounds $W_{ir}^{(j)}(\hat{T})$ and $W_{tr}(\hat{T})$. For that reason, it is important that R is determined as efficiently as possible, and therefore we shall use the algorithm 'Ex', as developed by Latouche and Ramaswami [52], instead of the method of successive substitutions. This algorithm provides that, for a given threshold vector $\hat{T}$, the order of magnitude of the computation time O_{ct} is restricted to $O_{ct} = a^3$; note that O_{ct} does not depend on the workload ρ.

7.5. Numerical results

In this section, we present some numerical results obtained by a numerical procedure which we have developed for the determination of the mean normalized waiting times for the SQS-JDP depicted in Figure 7.1, and which has been based on the bounds produced by the TKR and TBA model. The main purpose of this section is to investigate how the values for the threshold parameters T_1 and T_2 which are needed to approximate the mean normalized waiting times within the desired accuracy, depend on the model parameters. In particular, we are interested in the dependence on the model parameters which determine the drift to the states with equal queue lengths.

The procedure that we have developed for the SQS-JDP of Figure 7.1, determines for a given instance satisfying condition (7.5), all mean normalized waiting times $W^{(j)}$ for the job types $j \in J = \{A,B,C\}$ and the mean normalized waiting time W for all job types together within a given, absolute accuracy ε_{abs}. Condition (7.5) is a necessary condition for the ergodicity of a SQS-JDP. Up to now, the numerical procedure has always found a threshold vector $\hat{T}$ for which the TBA model is positive recurrent (= ergodic), which supports the conjecture that condition (7.5) is not only necessary but also sufficient for the ergodicity of a SQS-JDP. Under the assumption that this conjecture indeed is true, the TKR model is positive recurrent for all threshold vectors $\hat{T}$, and thus it always produces finite bounds $W^{(j)}_{TKR}(\hat{T})$ and $W_{TKR}(\hat{T})$. In the case that we have this latter property, the numerical procedure works as follows.

By exploiting the results of the previous section, the lower bounds $W^{(j)}_{TKR}(\hat{T})$ and $W_{TKR}(\hat{T})$ and the upper bounds $W^{(j)}_{TBA}(\hat{T})$ and $W_{TBA}(\hat{T})$ are computed for increasing values of the threshold parameters T_1 and T_2 of the vector $\hat{T}=(T_1,T_2)$, where initially $\hat{T}=(1,1)$ is taken. For each vector $\hat{T}$, the mean $(W^{(j)}_{TKR}(\hat{T})+W^{(j)}_{TBA}(\hat{T}))/2$ and half of the difference, i.e. $\Delta^{(j)}(\hat{T})=(W^{(j)}_{TBA}(\hat{T})-W^{(j)}_{TKR}(\hat{T}))/2$, are used as an approximation for $W^{(j)}$ and an upper bound for the corresponding absolute inaccuracy, where $j \in J$, and similarly $(W_{TKR}(\hat{T})+W_{TBA}(\hat{T}))/2$ and $\Delta(\hat{T})=(W_{TBA}(\hat{T})-W_{TKR}(\hat{T}))/2$ are used as an approximation and an upper bound for the corresponding absolute inaccuracy for W. Here, all means and upper bounds $\Delta^{(j)}(\hat{T})$ and $\Delta(\hat{T})$ are taken equal to ∞, if the TBA model is not positive recurrent. If for some $\hat{T}$ each of the upper bounds $\Delta^{(j)}(\hat{T})$ and $\Delta(\hat{T})$ is smaller than or equal to ε_{abs}, then the computation process is stopped, otherwise the procedure continues with the computation of improved bounds and approximations for a next threshold vector $\hat{T}$ with increased values for at least one of the threshold parameters T_1 and T_2. In the latter case, the decision on how to increase the threshold parameters is based on the fractions of redirections $p_{rd}(1)$ and $p_{rd}(2)$; this decision is described in the next paragraph.

The variable $p_{rd}(1)$ is used to denote the fraction of periods in which a redirection occurs on the boundary $m_1=m_2+T_1$ of the truncated state space. If for the present threshold vector $\hat{T}$ only the TKR model is positive recurrent, then we let $p_{tr}(1)$ denote the fraction for this TKR model, else we let $p_{tr}(1)$ denote the mean of the fraction found for the TKR model and the fraction found for the TBA model. Similarly, $p_{tr}(2)$ is used to denote the fraction of periods in which a redirection occurs on the boundary $m_2=m_1+T_2$ of the truncated state space. Both fractions can be computed in a similar way as the bounds for the mean normalized waiting times $W^{(j)}$. The higher the values of $p_{rd}(1)$ and $p_{rd}(2)$, the more important it is

to increase the values of T_1 and T_2, respectively. If it is established that for the present threshold vector $\hat{T}$ the desired accuracy for the approximations for the mean normalized waiting times $W^{(j)}$ and W is not reached yet, then in our numerical procedure, the value of T_1 is increased by 1 if $p_{tr}(1) > p_{rd}(2)$, the value of T_2 is increased by 1 if $p_{tr}(2) > p_{rd}(1)$, and both T_1 and T_2 are increased by 1 unit if $p_{tr}(1) = p_{rd}(2)$. Note that for a symmetric SQS-JDP, the fractions $p_{tr}(1)$ and $p_{tr}(2)$ will always be equal, by which the numerical procedure will compute the bounds and approximations for $W^{(j)}$ and W for the successive threshold vectors $\hat{T} = (1,1),(2,2), \cdots$.

The numerical procedure described above has been applied to three series of instances for the SQS-JDP of Figure 7.1. For all instances, the absolute accuracy ε_{abs} has been taken equal to $\varepsilon_{abs} = 0.005$. Further, the instance with

$$\rho = 0.9, \quad \lambda = 2\rho\mu, \quad \lambda_A = p\lambda \text{ with } p = \tfrac{1}{2}, \quad \lambda_B = \lambda_C = \tfrac{1}{2}(1-p)\lambda \tag{7.27}$$

(note that μ is uniquely determined by the assumption that $\lambda + 2\mu = 1$) has been taken as a basic instance. Since $\lambda_B = \lambda_C$, this instance represents a symmetric SQS-JDP.

The following three series of instances have been considered. In the first series, we have varied the workload ρ of the basic instance. All instances of this series represent a symmetric SQS-JDP, and the corresponding results obtained by the numerical procedure are presented in Table 7.1. The first column of this table denotes the chosen values for ρ, while the second column depicts the values of the parameters T_1 and T_2 of the first threshold vector $\hat{T} = (T_1, T_2)$ for which the desired absolute accuracy ε_{abs} was reached; note that $T_1 = T_2$ because of the symmetry. In the third, fifth and seventh column, we have listed the approximations which for this threshold vector $\hat{T}$ have been obtained for $W^{(A)}$, $W^{(B)} = W^{(C)}$ (because of the symmetry, also the waiting times for the types B and C are equal) and W; and, in the fourth, sixth and eighth column, we have listed the upper bounds $\Delta^{(A)}(\hat{T})$, $\Delta^{(B)}(\hat{T}) = \Delta^{(C)}(\hat{T})$ and $\Delta(\hat{T})$ for the corresponding absolute inaccuracies (it is noted that for these variables the values obtained by rounding off upwards are given). In the second series of instances, we have varied the fraction p of the arriving jobs which can be served by both servers. Also the instances of this series represent symmetric systems, and the numerical results for this series are presented in Table 7.2. The third series concerns instances for which the SQS-JDP is not symmetric, but still balanced. In this series, we have varied the values of λ_B and λ_C, under the restriction that $\lambda_B + \lambda_C = \tfrac{1}{2}\lambda$. We have taken $\lambda_B = \tfrac{1}{2}\hat{p}\lambda$ and $\lambda_C = \tfrac{1}{2}(1-\hat{p})\lambda$, where $\hat{p}$ is varied from 0.0 up to 0.5; cases with $\hat{p} > 0.5$ have not been considered, since they lead to the same results as the cases with $\hat{p} < 0.5$, but with the roles of the types B and C interchanged. The results obtained for this third series have been presented in Table 7.3. The Tables 7.2 and 7.3 contain about the same information as Table 7.1, and therefore these tables need no further explanation.

The results in Table 7.1 show that, as expected, the threshold parameters T_1 and T_2 which are needed to approximate the mean normalized waiting times within the desired absolute accuracy, are increasing as a function of the workload ρ. Further, the results in the Tables 7.2 and 7.3 show how the required values for T_1 and T_2 depend on the strength of the drift to the states with equal queue lengths, i.e. to the states on the diagonal. In Table 7.2, a smaller value for p corresponds to a weaker drift to the states on the diagonal. From this table, it follows that the weaker the drift to the diagonal, the larger the required values for T_1 and T_2. We also see that very large values for T_1 and T_2 are needed, if the drift to the diagonal is very small. In the extreme case with $p = 0.0$, in which the corresponding SQS-JDP

ρ	$T_1 = T_2$	$W^{(A)}$	$\Delta^{(A)}(\hat{T})$	$W^{(B)} = W^{(C)}$	$\Delta^{(\cdot)}(\hat{T})$	W	$\Delta(\hat{T})$
0.1	2	0.0146	0.0006	0.1059	0.0007	0.0603	0.0006
0.2	3	0.0558	0.0006	0.2282	0.0008	0.1420	0.0007
0.3	3	0.1281	0.0034	0.3746	0.0043	0.2514	0.0038
0.4	4	0.2351	0.0030	0.5577	0.0038	0.3964	0.0034
0.5	5	0.3966	0.0034	0.7977	0.0042	0.5971	0.0038
0.6	7	0.6468	0.0018	1.1337	0.0021	0.8902	0.0019
0.7	8	1.0723	0.0039	1.6532	0.0044	1.3628	0.0041
0.8	11	1.9222	0.0027	2.6142	0.0029	2.2682	0.0028
0.9	15	4.4516	0.0032	5.2782	0.0033	4.8649	0.0033
0.95	18	9.4729	0.0048	10.3800	0.0048	9.9265	0.0048
0.98	23	24.4883	0.0032	25.4495	0.0032	24.9689	0.0032
0.99	26	49.4939	0.0031	50.4742	0.0031	49.9841	0.0031

Table 7.1. The mean normalized waiting times $W^{(j)}$ and W determined within an absolute accuracy of $\varepsilon_{abs} = 0.005$ for the SQS-JDP of Figure 7.1 for increasing values of ρ and with $\lambda = 2\rho\mu$, $\lambda_A = \frac{1}{2}\lambda$, $\lambda_B = \lambda_C = \frac{1}{4}\lambda$.

consists of 2 independent $M \mid M \mid 1$ queues, T_1 and T_2 have to be equal to 85 in order to reach the desired accuracy, while in the other extreme case with $p = 1.0$, in which we have a pure SSQS as studied in Chapter 6, T_1 and T_2 only have to be equal to 8. In Table 7.3, a smaller value for $\hat{p}$ corresponds to a stronger drift to the diagonal in the region $m_1 \leq m_2$, but to a smaller drift to the diagonal in the region $m_2 \leq m_1$ (see Figure 7.1). As a result, for decreasing values of $\hat{p}$, we find a decreasing behavior for the parameter T_1 of the first threshold vector $\hat{T} = (T_1, T_2)$ for which the desired accuracy is reached by the numerical procedure, while an increasing behavior is found for the parameter T_2. However, for small values of $\hat{p}$, the increasing effect for T_2 is much larger than the decreasing effect for T_1. In the extreme case with $\hat{p} = 0.0$, in which we have the behavior of a pure SSQS in the region $m_1 \leq m_2$ and the behavior of two independent $M \mid M \mid 1$ queues in the region $m_2 \leq m_1$, we find a value for T_1 that is almost equal to the values obtained for T_1 and T_2 in Table 7.2 for $p = 1.0$ and we find a value for T_2 that is almost equal to the values obtained for T_1 and T_2 in Table 7.2 for $p = 0.0$. In a similar way, also for the other cases depicted in Table 7.3, the values found for T_1 and T_2 can be explained on the basis of the values obtained for T_1 and T_2 in Table 7.2.

From the results for the threshold parameters T_1 and T_2 which are needed to approximate the mean normalized waiting times within the desired accuracy, it may be concluded that the TKR and TBA model only lead to tight bounds, if the drift to the states with equal queue lengths is sufficiently strong. This will also hold for a general SQS-JDP with $N \geq 2$ servers. It is noted that the existence of a certain drift to the states with equal queue lengths has been a point of departure when we constructed the TKR and TBA model. So, if there is only a weak drift to the states with equal queue lengths, then the probability mass will not be concentrated around these states, and one should focus on bound models with alternative truncated state spaces.

p	$T_1=T_2$	$W^{(A)}$	$\Delta^{(A)}(\hat{T})$	$W^{(B)}=W^{(C)}$	$\Delta^{(\cdot)}(\hat{T})$	W	$\Delta(\hat{T})$
0.0	85	4.2648	0.0024	8.9976	0.0046	8.9976	0.0046
0.1	43	4.3594	0.0038	6.8002	0.0046	6.5561	0.0045
0.2	29	4.4027	0.0041	6.0435	0.0045	5.7154	0.0044
0.3	22	4.4266	0.0040	5.6619	0.0042	5.2913	0.0041
0.4	18	4.4414	0.0033	5.4320	0.0034	5.0357	0.0034
0.5	15	4.4516	0.0032	5.2782	0.0033	4.8649	0.0033
0.6	13	4.4589	0.0027	5.1682	0.0028	4.7426	0.0028
0.7	11	4.4645	0.0034	5.0856	0.0035	4.6509	0.0034
0.8	10	4.4688	0.0025	5.0212	0.0025	4.5793	0.0025
0.9	9	4.4722	0.0021	4.9697	0.0021	4.5220	0.0021
1.0	8	4.4751	0.0022	4.9275	0.0022	4.4751	0.0022

Table 7.2. The mean normalized waiting times $W^{(j)}$ and W determined within an absolute accuracy of $\varepsilon_{abs}=0.005$ for the SQS-JDP of Figure 7.1 with $\rho=0.9$, $\lambda=2\rho\mu$, $\lambda_A=p\lambda$, $\lambda_B=\lambda_C=\frac{1}{2}(1-p)\lambda$, and varying p.

$\hat{p}$	(T_1,T_2)	$W^{(A)}$	$\Delta^{(A)}(\hat{T})$	$W^{(B)}$	$\Delta^{(B)}(\hat{T})$	$W^{(C)}$	$\Delta^{(C)}(\hat{T})$	W	$\Delta(\hat{T})$
0.0	(7,90)	4.2756	0.0017	4.3427	0.0017	13.0498	0.0047	8.6627	0.0032
0.1	(8,44)	4.3744	0.0034	4.5244	0.0034	8.5221	0.0048	6.2484	0.0041
0.2	(9,30)	4.4175	0.0037	4.6754	0.0037	6.9337	0.0043	5.4497	0.0040
0.3	(11,22)	4.4386	0.0038	4.8365	0.0038	6.1181	0.0041	5.0861	0.0039
0.4	(12,18)	4.4487	0.0042	5.0294	0.0043	5.6185	0.0044	4.9158	0.0043
0.5	(15,15)	4.4516	0.0032	5.2782	0.0033	5.2782	0.0033	4.8649	0.0033

Table 7.3. The mean normalized waiting times $W^{(j)}$ and W determined within an absolute accuracy of $\varepsilon_{abs}=0.005$ for the SQS-JDP of Figure 7.1 with $\rho=0.9$, $\lambda=2\rho\mu$, $\lambda_A=\frac{1}{2}\lambda$, $\lambda_B=\frac{1}{2}\hat{p}\lambda$, $\lambda_C=\frac{1}{2}(1-\hat{p})\lambda$, and varying $\hat{p}$.

The values presented in the Tables 7.1-7.3 for the mean normalized waiting times itself, also deserve some attention. The results in Table 7.1 show that the behavior of the waiting times as a function of the workload ρ is similar to the behavior found for the pure SSQS (see Table 1.8). Further, we observe only a small difference between the waiting times for the types B and C and the waiting time for type A, even for high workloads ρ. From the results in Table 7.2, it follows that the mean normalized waiting time W for all job types together is more than proportionally decreasing as a function of the fraction p of jobs which can be served by both servers. A similar behavior is observed in Table 7.3 for the mean normalized waiting time W as a function of $\hat{p}$. In this table, small values of $\hat{p}$ correspond to situations in which, under the shortest queue routing, the workloads for the servers 1 and 2 are badly balanced; for small $\hat{p}$, there will be several jobs of type A which join server 2 while they better could join server 1 in order to repair the asymmetry in the job-dependent structure. Due to

this behavior, for server 2 a significantly higher workload is obtained than for server 1 and the mean normalized waiting time W becomes relatively large.

From the results of the Tables 7.2 and 7.3, we may conclude the following for the production situation of Printed Circuit Boards by a flexible assembly system, as described in Section 1.2. The results in these tables point out that, in order to obtain small mean waiting times for the given total workload, the assignment of the components to the insertion machines should be such that the resulting job-dependent structure for the corresponding SQS-JDP leads to a strong drift to the states with equal queue lengths in all regions.

The results presented in the Tables 7.1-7.3 have been computed on a standard SUN workstation. For the instances of Table 7.1, the computation times consumed by the numerical procedure as described in this section, varied from a negligibly small time for $\rho = 0.1$ to about 3 minutes for $\rho = 0.99$; for Table 7.2, they varied from 2 seconds for $p = 1.0$ to about 5 hours for $p = 0.0$, and for Table 7.3, they varied from 18 seconds for $\hat{p} = 0.5$ to about 1 hour for $\hat{p} = 0.0$. The largest computation time, viz. about 5 hours for the case $p = 0.0$ in Table 7.2, concerns the total time needed for the computation of the lower and upper bounds for all threshold vectors $\hat{T} = (1,1),(2,2),\ldots,(85,85)$, for which we had to determine rate matrices R of the orders $a = 3,5,\ldots,171$. Since the computation for the last vector $\hat{T} = (85,85)$ costed only about 14 minutes, it is obvious that we could have obtained a much smaller total computation time for this case by increasing the threshold parameters T_1 and T_2 by say 5 or 10 units instead of 1 unit after each step; and, similarly for all other cases for which we obtained large computation times. We repeat that for a given threshold vector $\hat{T}$, the computation time for solving the TKR and TBA model has an order of magnitude equal to $O_{ct} = a^3$; this order has been confirmed by the values which we found for the computation times.

The following may be concluded on whether the TKR and TBA model are appropriate for being used for the analysis of a general SQS-JDP with $N \geq 2$ servers. Let us focus on the question whether they can be used to obtain sufficiently accurate approximations for the mean normalized waiting times for a given instance of a SQS-JDP. We can say that the TKR and TBA model are at least appropriate for all systems with $N = 2$ servers. For larger systems, with say up to $N = 5$ servers, they will also be appropriate, but then, because of the limitations stemming from the computational efforts, it will only be possible to obtain accurate approximations for instances with a sufficiently strong drift to the states with equal queue lengths.

Finally, we comment on whether the TKR and TBA model are appropriate for analyzing the production situation of the Printed Circuit Boards by a flexible assembly system; see Section 1.2. As we observed, the condition that there is a sufficiently strong drift to the states with equal queue lengths, must be satisfied by each sensible assignment of the components to the insertion machines. Therefore, the TKR and TBA model will be appropriate for evaluating each SQS-JDP corresponding to a sensible assignments. For a bad (or non-sensible) assignment, the TKR and TBA model will not be appropriate to determine accurate approximations for the mean normalized waiting times in the corresponding SQS-JDP, but then it may be expected that still sufficiently tight lower bounds are obtained in order to classify the given assignment as a bad assignment. So, the TBA and TKR model seem to be appropriate for selecting a restricted set of assignments which lead to the lowest waiting times. For completeness, we recall that these assignments represent the best assignments for the simplified model of the SQS-JDP as studied in this chapter, and that a simulation study for a more general model could be used to determine the 'real' quality for each of these assignments.

7.6. Conclusions

This chapter has been devoted to the so-called Shortest Queue System with a Job-Dependent Parallelism (SQS-JDP), which consists of $N \geq 2$ parallel servers and which constitutes a generalization of the Symmetric Shortest Queue System (SSQS) studied in the previous chapter. The SQS-JDP has been encountered in Chapter 1, when studying the mounting of components of Printed Circuit Boards by a flexible assembly system consisting of a number of parallel insertion machines.

By applying the precedence relation method, we have derived a flexible lower bound model, called the Threshold Killing and Rejection (TKR) model, and a flexible upper bound model, called the Threshold Blocking and Addition (TBA) model, which produce lower and upper bounds for the mean normalized waiting times for all different job types and for all job types together. Both the TKR and the TBA model have been constructed such that they can be solved by the matrix-geometric approach. Numerical results for a SQS-JDP with $N=2$ servers have shown that the TKR and TBA model lead to tight bounds for not too large sizes of the truncated state space if and only if in the original model there is a sufficiently strong drift to the states corresponding to situations with equal queue lengths. If this latter condition is satisfied, then the TKR and TBA model will be appropriate for the determination of the mean normalized waiting times for systems with up to about 5 servers (and workloads up to 0.95).

The numerical results for the SQS-JDP with $N=2$ servers also have pointed out that for the production problem of the Printed Circuit Boards, in order to prevent too large waiting times, each sensible assignment of components to the insertion machines must be such that the resulting job-dependent structure for the corresponding SQS-JDP leads to the required strong drift to the states with equal queue lengths. As a result, the TKR and TBA model seem to be appropriate for selecting a restricted number of assignments which lead to the best performance characteristics.

Chapter 8

Conclusions and Suggestions for Future Research

Many queueing systems can be modeled as Markov processes on multi-dimensional state spaces, which are discrete and infinite in each component. The relevant performance measures of such a queueing system usually may be obtained from the equilibrium distribution of the corresponding Markov process. However, the determination of the equilibrium distribution of a multi-dimensional Markov process for which the state space is infinite in each component, is a hard problem in general. In this monograph, we have described two new approaches for this problem: the *compensation approach*, which is a direct approach for the determination of the equilibrium distribution, and the *precedence relation method*, which can be used for the derivation of flexible truncation models which produce approximations for the equilibrium distribution of the original model or Markov process such that bounds for the relevant performance measures are obtained. For both approaches, we describe below the main conclusions and some suggestions for future research.

Conclusions and suggestions for future research for the compensation approach

The components m_i of the states $(m_1, \ldots, m_N)$ of a multi-dimensional Markov process describing the behavior of a queueing system, usually represent quantities such as queue lengths, which often leads to a certain *homogeneity* in the transition probabilities/rates. In that case, a Markov process is also referred to as a random walk. A well-known class of homogeneous random walks is the class of *product-form networks*, for which the equilibrium distribution is a product-form distribution, i.e. the equilibrium distribution can be written as a product of powers of fixed factors (see Baskett et al. [15]). These fixed factors are obtained by substituting a product-form solution in the equilibrium equations and solving the remaining system of non-linear equations. The class of product-form networks seems to be the only class of multi-dimensional Markov processes for which the equilibrium distribution can be determined this easily and this explicitly.

Recent research by Adan et al. [12] (see also [3]) has suggested that, beside the product-form networks, there exists another class of homogeneous random walks for which the equilibrium distribution can be determined explicitly. Adan et al. [12] have developed the so-called *compensation approach*, which has enabled them to determine the equilibrium distribution for a class of two-dimensional, homogeneous random walks on the states (m_1, m_2) with $m_1, m_2 \in \mathbb{N}_0$. The *main idea* of the compensation approach is to characterize a set of product-form solutions which satisfy the equilibrium equations for all states in the interior of

the state space and next to construct a linear combination of these product-form solutions which also satisfies the equilibrium equations for the states at the boundaries and the origin. If this approach works for a given problem, then it is proved that the equilibrium distribution may be written as a linear combination of possibly infinitely many product-form solutions, and explicit expressions are obtained for all coefficients and product factors of this linear combination. Obviously, the main idea of this approach can also be applied to higher-dimensional problems, and therefore the analysis of [12], has led to the conjecture that there exists a class of $N(\geq 2)$-dimensional problems for which the equilibrium distribution can be determined explicitly by the compensation approach, or, better, by an extended version of it.

In the Chapters 2-4 of this monograph, we have applied the main idea of the compensation approach as developed in [12], and extended the method itself, to the class of N-dimensional, irreducible, positive recurrent, homogeneous, nearest-neighboring random walks with the projection property on the states $(m_1, \ldots, m_N)$, where $N \geq 2$ and $m_i \in \mathbb{N}_0$ for all $i \in I := \{1, \ldots, N\}$. The main results have been stated in Theorem 3.4 and were proved by induction with respect to N in the Chapters 2 and 3. We have shown that for a random walk of the considered class, the equilibrium distribution can be determined by using the compensation approach if and only if

$$q_{t_1,\ldots,t_N} = 0 \quad \text{if } t_i + t_j > 0 \text{ for some } i,j \in I,\ i \neq j, \tag{8.1}$$

where for each direction $(t_1, \ldots, t_N)$, the variable $q_{t_1,\ldots,t_N}$ denotes the transition probability/rate for a transition from a state $(m_1, \ldots, m_N)$ in the interior of the state space to the state $(m_1+t_1, \ldots, m_N+t_N)$. This condition stems from convergence requirements for the linear combinations of product-form solutions constructed by the compensation approach, and it states that no transitions can be made from the states in the interior into directions which lead to a larger sum of the components m_i and m_j of the state for some indices $i,j \in I$, $i \neq j$. Further, we have proved that if condition (8.1) is satisfied, then the equilibrium distribution may be written as *an alternating sum of infinitely many, pure product-form distributions*, which constitute solutions for the equilibrium equations for the states in the interior of the state space; and, similarly for all marginal distributions.

For the determination of the product factors of all product-form distributions required for the equilibrium distribution and all marginal distributions of a random walk satisfying condition (8.1), we have presented simple, recursive formulae. These product-form distributions are in fact obtained from a certain tree, which has appeared to have an interesting, geometric structure; see Chapter 4. A detailed analysis of this geometric structure has led to explicit expressions for upper bounds for the absolute errors of the approximations of the equilibrium probabilities by finite, alternating sums of product-form solutions. The error bounds have been exploited in efficient numerical procedures for the computation of the equilibrium distribution and related quantities within some desired accuracy.

Condition (8.1) implies that for each state $(m_1, \ldots, m_N)$ in the interior, an outgoing transition which leads to a positive step for one component m_i, $i \in I$, may only have a positive probability/rate, if it leads to negative steps for all other components m_j, $j \in I \backslash \{i\}$. It is obvious that this condition restricts the applicability of the compensation approach, especially for $N \geq 3$. A queueing system which satisfies condition (8.1), also for $N \geq 3$, is the $2 \times N$ buffered switch; numerical results for this system have been presented in the Chapters 2 and 4.

Condition (8.1) constitutes an extension of the condition obtained by Adan et al. [12] for the class of two-dimensional, irreducible, positive recurrent, homogeneous, nearest-neighboring random walks, which do not necessarily satisfy the projection property. They found that for this class, the necessary and sufficient condition under which the compensation approach works, is given by

$$q_{0,1} = q_{1,0} = q_{1,1} = 0\,, \tag{8.2}$$

which means that for the states in the interior no transitions to the North, East and North-East are allowed. Problems which are in this class and satisfy this condition, are the Symmetric Shortest Queue System (SSQS) with two parallel servers, a multiprogramming queues system, and, of course, the 2×2 buffered switch.

A generalization of both condition (8.1) and condition (8.2) is obtained when considering the class of N-dimensional, irreducible, positive recurrent, homogeneous, nearest-neighboring random walks, which do not necessarily the projection property. We conjecture that the equilibrium distribution of a random walk of this class may be determined by using the compensation approach if and only if for all $J \subset I$,

$$q^J_{t_1,\ldots,t_N} = 0 \quad \text{if } t_i + t_j > 0 \text{ for some } i,j \in J,\ i \neq j, \tag{8.3}$$

where the variables $q^J_{t_1,\ldots,t_N}$ denote the probabilities/rates for the outgoing transitions from the states $(m_1,\ldots,m_N)$ with $m_i > 0$ for all $i \in J$ and $m_i = 0$ for all $i \notin J$ (note that the variables $q^I_{t_1,\ldots,t_N}$ correspond to the variables $q_{t_1,\ldots,t_N}$, which occur in the conditions (8.1) and (8.2)). Condition (8.3) is satisfied by definition for all $J \subset I$ with $|J| \leq 1$, by which (8.3) reduces to (8.2) for the case $N = 2$, and it follows from the definition of the projection property that (8.3) is equivalent to (8.2) for a random walk with the projection property. A second conjecture is that condition (8.3) also represents the necessary and sufficient condition under which the compensation approach works, if, beside the projection property, also the nearest-neighboring property, i.e. the property that only transitions to nearest neighbors occur, is omitted. Both conjectures should be confirmed by future research.

Another interesting topic for future research is constituted by homogeneous random walks on multi-dimensional state spaces which are infinite in exactly two components. A problem in this class is the SSQS with two parallel servers and a Markovian arrival process. The behavior of this system is described by a Markov process with states (m_1, m_2, m_3), where m_1 denotes the length of the shortest queue, m_2 denotes the difference between both queue lengths and m_3 denotes the state of the Markovian arrival process; so, $m_1, m_2 \in \mathbb{N}_0$ and $m_3 \in \{1,\ldots,K\}$, where K is the number of states for the Markovian arrival process. A numerical experiment has indicated that the equilibrium distribution for this system consists of a linear combination of solutions of the form

$$p_{m_1,m_2} = (p_{m_1,m_2,1},\ldots,p_{m_1,m_2,K}) = a\,\alpha^{m_1}\beta^{m_2}\,, \quad m_1, m_2 \in \mathbb{N}_0,$$

where $a = (a_1,\ldots,a_K)$ is a K-dimensional row vector with components a_i that do not depend on m_1 and m_2. Future research has to reveal whether the equilibrium distribution indeed consists of solutions of this form, and whether linear combinations of solutions of this form also lead to the equilibrium distribution for similar problems.

Another interesting topic for future research is to return to the class of two-dimensional random walks as studied by Adan et. al. [12], and to investigate what kind of structure the equilibrium distribution has if condition (8.2) is not satisfied. Here, assume that the random

walks also have the projection property, which implies that each of the two marginal distributions is equal to a simple product-form distribution. Numerical results, which we obtained from truncated random walks, have indicated that for random walks which violate condition (8.2), the equilibrium distribution seems to have a more complicated structure than a linear combination of product-form distributions (except for the case in which we precisely have a product-form network). More information on the structure may be gained by further investigating why both the compensation approach and the uniformization technique fail if condition (8.2) is violated, and by studying the expressions which are obtained for the generating function of the equilibrium distribution by applying the boundary value method.

Conclusions and suggestions for future research for the precedence relation method

For many multi-dimensional Markov processes for which the state space is infinite in each component, it is not possible to determine the equilibrium distribution in an analytical way, and then alternative methods are needed. One alternative method is the power-series algorithm (see Blanc [18]), with which the equilibrium distribution and the corresponding relevant performance measures can be determined within a given accuracy, provided that the corresponding requirements with respect to the computational efforts are met. Another alternative is constituted by the use of solvable truncation models which can approximate the original model or Markov process as accurately as desired. This latter property may be obtained by defining the truncated state space such that its size depends on one or more truncation parameters. Truncation models with this property are called flexible truncation models. Truncation models in fact lead to approximations for the equilibrium distribution of the original model, and therefore also to approximations for the relevant performance measures. Of a particular interest are flexible truncation models which produce bounds for the relevant performance measures. Such models are also called *flexible bound models*, and they may be derived by the so-called *precedence relation method*, which we have developed in Chapter 5 of this thesis.

In principle, the precedence relation method is a method for proving a monotonicity result between the average costs of two, discrete-time, irreducible Markov cost models, of which the state space of one model is a subset of the state space of the other model. The *main idea* of the precedence relation method is that the proof of such a monotonicity result may be simplified by first deriving so-called precedence pairs of states of the model with the larger state space; such pairs must satisfy a certain *precedence relation* which denotes that the first state of a pair is more attractive with respect to the costs than the second state. We have mainly focused on how the precedence relation method can be used for deriving flexible truncation models which lead to bounds for one or more of the relevant performance measures of a given, original Markovian (queueing) system. Such models can be derived by performing the following two steps. In the first, preliminary, step, one must derive precedence pairs for the original model. After that, in the second step, one can easily define one or more flexible bound models: a flexible lower (upper) bound model is obtained by first defining a truncated state space M' with a flexible size and next redirecting all transitions ending in states outside M' to states inside M', which are more (less) attractive according to the precedence pairs obtained for the original model. The derivation of the precedence pairs in the preliminary step is essential for obtaining flexible bound models which accurately approximate the original model for already small sizes of the truncated state space.

The precedence relation method can be applied to any Markov process, but it depends on the structure of the state space and the transition probabilities/rates of a particular model whether it is possible to derive flexible bound models of a good *quality*. The quality of a flexible bound model is determined by the ratio of the tightness of the bounds and the efficiency with which the truncation model itself can be solved.

In the Chapters 6 and 7 of this monograph, we have applied the precedence relation method to the Symmetric Shortest Queue System (SSQS) and to a generalization of it, which we called the Shortest Queue System with a Job-Dependent Parallelism (SQS-JDP) and which stems from a production situation of Printed Circuit Boards by a flexible assembly system (see Chapter 1). For the SSQS, we have been able to derive a flexible lower bound model and a flexible upper bound model, which both produce tight bounds for the mean normalized waiting time for already small sizes of the truncated state space, and which both can be solved very efficiently by the matrix-geometric approach, as described by Neuts [58]. These bound models have been exploited in an efficient numerical procedure, with which the mean normalized waiting time has been determined quite accurately for systems with many servers and high workloads. For the SQS-JDP, the application of the precedence relation method has led to similar flexible bound models as for the SSQS. Also for this system, the obtained flexible bound models have appeared to be appropriate for numerically determining the mean waiting times, and possibly several other performance measures, within a given accuracy.

The research of the Chapters 5-7 gives rise to several interesting topics for future research. First of all, it should be investigated for which other queueing models than the SSQS and the SQS-JDP, the precedence relation method leads to flexible bound models of a good quality. It is noted that in [1], the precedence relation method has led to flexible bound models with which the symmetric longest queue system has been analyzed efficiently. Secondly, by using flexible bound models, one can gather numerical data for interesting performance measures, after which these data may be exploited for obtaining good approximation formulae for the corresponding performance measures. For example, by using the flexible bound models derived for the SSQS, one can gather numerical data for the mean normalized waiting time, after which comparing these data to the corresponding data for the $M \mid M \mid N$ queue may lead to a simple and new approximation formula for this most interesting performance measure for the SSQS. A third topic concerns how well the flexible bound models derived by the precedence relation method can be used for the determination of the transient behavior of a queueing model. Finally, the last topic that we mention, is the application of the main idea of the precedence relation method to other types of models than Markov cost models; for example, to Markov decision models.

References

1. ADAN, IVO, VAN HOUTUM, GEERT-JAN, AND VAN DER WAL, JAN, "The symmetric longest queue system," Memorandum COSOR 94-33, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., 1994. Submitted for publication.
2. ADAN, IVO, VAN HOUTUM, GEERT-JAN, AND VAN DER WAL, JAN, "Upper and lower bounds for the waiting time in the symmetric shortest queue system," *Ann. Oper. Res.*, vol. 48, pp. 197-217, 1994.
3. ADAN, I.J.B.F., *A Compensation Approach for Queueing Problems,* CWI Tract 104, Stichting Mathematisch Centrum, Amsterdam, 1994.
4. ADAN, I.J.B.F., VAN DE WAARSENBURG, W.A., AND WESSELS, J., "Analysing $E_k|E_r|c$ queues," *Eur. J. Oper. Res.*, 1995. To appear.
5. ADAN, I.J.B.F., VAN HOUTUM, G.J., WESSELS, J., AND ZIJM, W.H.M., "A compensation procedure for multiprogramming queues," *OR Spektrum*, vol. 15, pp. 95-106, 1993.
6. ADAN, I.J.B.F. AND WESSELS, J., "Analysing shortest expected delay routing for Erlang servers," Memorandum COSOR 93-38, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., 1993.
7. ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "Queuing analysis in a flexible assembly system with a job-dependent parallel structure," in *Operations Research Proceedings 1988*, pp. 551-558, Springer-Verlag, Berlin, 1989.
8. ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "Analysis of the symmetric shortest queue problem," *Stochastic Models*, vol. 6, pp. 691-713, 1990.
9. ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "Analysis of the asymmetric shortest queue problem with threshold jockeying," *Stochastic Models*, vol. 7, pp. 615-627, 1991.
10. ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "Analysis of the asymmetric shortest queue problem," *Queueing Systems*, vol. 8, pp. 1-58, 1991.
11. ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "Analysing multiprogramming queues by generating functions," *SIAM J. Appl. Math.*, vol. 53, pp. 1123-1131, 1993.
12. ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "A compensation approach for two-dimensional Markov processes," *Adv. Appl. Prob.*, vol. 25, pp. 783-817, 1993.
13. ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "Matrix-geometric analysis of the shortest queue problem with threshold jockeying," *Oper. Res. Lett.*, vol. 13, pp. 107-112, 1993.
14. AHUJA, RAVINDRA K., MAGNANTI, THOMAS L., AND ORLIN, JAMES B., *Network Flows: Theory, Algorithms, and Applications,* Prentice-Hall, Englewood Cliffs, New Jersey, 1993.
15. BASKETT, F., CHANDY, K.M., MUNTZ, R., AND PALACIOS-GOMEZ, F., "Open, closed and mixed networks of queues with different classes of customers," *Journal of the*

ACM, vol. 22, pp. 248-260, 1975.

16. BERTSIMAS, D., "An analytic approach to a general class of $G|G|s$ queueing systems," *Oper. Res.*, vol. 38 , pp. 139-155, 1990.
17. BLANC, J.P.C., "The power-series algorithm applied to the shortest-queue model," *Oper. Res.*, vol. 40, pp. 157-167, 1992.
18. BLANC, J.P.C., "Performance analysis and optimization with the power-series algorithm," in *Performance Evaluation of Computer and Communication Systems*, ed. R.D. Nelson, pp. 53-90, North-Holland, Amsterdam, 1993.
19. BOXMA, O.J. AND VAN HOUTUM, G.J., "The compensation approach applied to a 2x2 switch ," *Prob. Engineer. Inform. Sci.*, vol. 7, pp. 471-493, 1993.
20. COHEN, J.W., "A two-queue, one-server model with priority for the longer queue," *Queueing Systems*, vol. 2, pp. 261-283, 1987.
21. COHEN, J.W., "Boundary value problems in queueing theory," *Queueing Systems*, vol. 3, pp. 97-128, 1988.
22. COHEN, J.W., "On a class of two-dimensional nearest-neighbour random walks," in *Studies in Applied Probability (Journal of Applied Probability, Special Volume 31A)*, ed. J. Gani, pp. 207-237, 1994.
23. COHEN, J.W. AND BOXMA, O.J., *Boundary Value Problems in Queueing System Analysis*, North-Holland, Amsterdam, 1983.
24. CONOLLY, B.W., "The autostrada queueing problem," *J. Appl. Prob.*, vol. 21, pp. 394-403, 1984.
25. DE KLEIN, JAN, *Fredholm Integral Equations in Queueing Analysis*, Thesis, University of Utrecht, Utrecht, 1988.
26. DISNEY, R.L. AND MITCHELL, W.E., "A solution for queues with instantaneous jockeying and other customer selection rules," *Naval Res. Log.*, vol. 17, pp. 315-325, 1971.
27. DUNFORD, NELSON AND SCHWARTZ, JACOB T., *Linear Operators, Part I: General Theory*, Interscience, New York, 1958.
28. ELSAYED, E.A. AND BASTANI, A., "General solutions of the jockeying problem," *Eur. J. Oper. Res.*, vol. 22, pp. 387-396, 1985.
29. FAYOLLE, G., *Méthodes Analytiques pour les Files d'Attente Couplées*, Thesis, Univ. de Paris VI, Paris, 1979.
30. FAYOLLE, G. AND IASNOGORODSKI, R., "Two coupled processors: the reduction to a Riemann-Hilbert problem," *Z. Wahrsch. Verw. Gebiete*, vol. 47, pp. 325-351, 1979.
31. FLATTO, L., "The longer queue model," *Prob. Engineer. Inform. Sci.*, vol. 3, pp. 537-559, 1989.
32. FLATTO, L. AND HAHN, S., "Two parallel queues created by arrivals with two demands I," *SIAM J. Appl. Math.*, vol. 44, pp. 1041-1053, 1984.
33. FLATTO, L. AND MCKEAN, H.P., "Two queues in parallel," *Comm. Pure Appl. Math.*, vol. 30, pp. 255-263, 1977.
34. FOSCHINI, G.J. AND SALZ, J., "A basic dynamic routing problem and diffusion," *IEEE Trans. Commun. COM-26*, pp. 320-327, 1978.

35. GERTSBAKH, I., "The shorter queue problem: a numerical study using the matrix-geometric solution," *Eur. J. Oper. Res.*, vol. 15, pp. 374-381, 1984.

36. GRASSMANN, W.K., "Transient and steady state results for two parallel queues," *OMEGA Int. J. of Mgmt. Sci. 8*, pp. 105-112, 1980.

37. GREEN, L., "A queueing system with general-use and limited-use servers," *Oper. Res.*, vol. 33, pp. 168-182, 1985.

38. HAIGHT, F.A., "Two queues in parallel," *Biometrica*, vol. 45, pp. 401-410, 1958.

39. HALFIN, S., "The shortest queue problem," *J. Appl. Prob.*, vol. 22, pp. 865-878, 1985.

40. HASSIN, REFAEL AND HAVIV, MOSHE, "Equilibrium strategies and the value of information in a two line queueing system with threshold jockeying," *Stochastic Models*, vol. 10, pp. 415-435, 1994.

41. HOFRI, M., "A generating-function analysis of multiprogramming queues," *Int. J. Comp. Inform. Sci.*, vol. 7, pp. 121-155, 1978.

42. HOOGHIEMSTRA, G., KEANE, M., AND REE, S. VAN DE, "Power series for stationary distributions of coupled processor models," *SIAM J. Appl. Math.*, vol. 48, pp. 1159-1166, 1988.

43. HORDIJK, ARIE AND KOOLE, GER, "On the optimality of the generalized shortest queue policy," *Prob. Engineer. Inform. Sci.*, vol. 4, pp. 477-487, 1990.

44. HORDIJK, ARIE AND KOOLE, GER, "On the assignment of customers to parallel queues," *Prob. Engineer. Inform. Sci.*, vol. 6, pp. 495-511, 1992.

45. IASNOGORODSKI, R., *Problèmes-Frontières dans les Files d'Attente*, Thesis, Univ. de Paris VI, Paris, 1979.

46. JAFFE, S., "Equilibrium results for a pair of coupled discrete-time queues," Ultracomputer Note, NYA Ultracomputer Research Lab, Courant Institute of Mathematical Sciences, New York, 1989.

47. JAFFE, S., "The equilibrium distribution for a clocked buffered switch," *Prob. Engineer. Inform. Sci.*, vol. 6, pp. 425-438, 1992.

48. KAO, E.P.C. AND LIN, C., "A matrix-geometric solution of the jockeying problem," *Eur. J. Oper. Res.*, vol. 44, pp. 67-74, 1990.

49. KINGMAN, J.F.C., "Two similar queues in parallel," *Ann. Math. Statist.*, vol. 32, pp. 1314-1323, 1961.

50. KNESSL, C., MATKOWSKY, B.J., SCHUSS, Z., AND TIER, C., "Two Parallel Queues with Dynamic Routing," *IEEE Trans. Commun.*, vol. 34, pp. 1170-1175, 1986.

51. KONHEIM, A.G., MEILIJSON, J., AND MELKMAN, A., "Processor-sharing of two parallal lines," *J. Appl. Prob.*, vol. 18, pp. 952-956, 1981.

52. LATOUCHE, GUY AND RAMASWAMI, V., "A logarithmic reduction algorithm for quasi-birth-death processes," *J. Appl. Prob.*, vol. 30, pp. 650-674, 1993.

53. LUI, JOHN C.S. AND MUNTZ, RICHARD R., "Algorithmic approach to bounding the mean response time of a minimum expected delay routing system," *Performance Evaluation Review*, vol. 20, pp. 140-151, 1992.

54. LUI, J.C.S., MUNTZ, R.R., AND TOWSLEY, D., "Bounding the mean response time of a minimum expected delay routing system: an algoritmic approach," CMPSCI Technical

report 93-68, University of Massachusetts, 1993. Submitted for publication.

55. MALYSHEV, V.A., "Classification of two-dimensional positive random walks and almost linear semimartingales," *Dokl. Akad. Nauk. SSSR (Engl. transl. in Soviet Math. Dokl. 13)*, vol. 202, pp. 526-528, 1972.
56. NELSON, R. AND PHILIPS, T.K., "An approximation to the response time for the shortest queue routing," *Performance Evaluation Review*, vol. 7, pp. 181-189, 1989.
57. NELSON, RANDOLPH D. AND PHILIPS, THOMAS K., "An approximation for the mean response time for the shortest queue routing with general interarrival and service times," *Performance Evaluation*, vol. 17, pp. 123-139, 1993.
58. NEUTS, MARCEL F., *Matrix-Geometric Solutions in Stochastic Models*, Johns Hopkins University Press, Baltimore, 1981.
59. NEUTS, MARCEL F., *Structured Stochastic Matrices of M|G|1 Type and Their Applications*, Marcel Dekker, New York, 1989.
60. RAO, B.M. AND POSNER, M.J.M., "Algorithmic and approximation analysis of the shorter queue model," *Naval Res. Log.*, vol. 34, pp. 381-398, 1987.
61. ROQUE, D.R., "A note on "Queueing models with lane selection"," *Oper. Res.*, vol. 28, pp. 419-420, 1980.
62. SCHASSBERGER, R., "A service system with two parallel queues," *Computing*, vol. 4, pp. 24-29, 1969.
63. SCHWARTZ, B.L., "Queuing models with lane selection: a new class of problems," *Oper. Res.*, vol. 22, pp. 331-339, 1974.
64. SENETA, E., *Non-negative Matrices and Markov Chains*, Springer-Verlag, New York, 1981.
65. STEPANOV, S.N., "Increasing the efficiency of numerical methods for models with repeat calls," *Problems of Information Transmission*, vol. 22, pp. 313-326, 1986.
66. STEPANOV, S.N., "Numerical analysis of queueing models with repeated calls," Research Reports on Information Sciences B-274, Department of Information Sciences, Tokyo Institute of Technology, 1993.
67. STEPANOV, S.N. AND TSITOVICH, I.I., "Qualitative methods of analysis of systems with repeat calls," *Problems of Information Transmission*, vol. 23, pp. 156-173, 1987.
68. STEWART, WILLIAM J. (ED.), *Numerical Solution of Markov Chains*, Marcel Dekker, New York, 1991.
69. TAKAHASHI, Y., "Asymptotic exponentiality of the tail of the waiting time distribution in a *Ph|Ph|c* queue," *Adv. Appl. Prob.*, vol. 13, pp. 619-630, 1981.
70. TIJMS, H.C., *Stochastic Modelling and Analysis: A Computational Approach*, Wiley, New York, 1986.
71. TITCHMARSH, E.C., *The Theory of Functions*, Oxford University Press, Oxford, 1960.
72. VAN DER WAL, J., "Monotonicity of the throughput of a closed exponential queueing network in the number of jobs," *OR Spektrum*, vol. 11, pp. 97-100, 1989.
73. VAN DER WAL, J. AND SCHWEITZER, P.J., "Iterative bounds on the equilibrium distribution of a finite Markov chain," *Prob. Engineer. Inform. Sci.*, vol. 1, pp. 117-131, 1987.

74. VAN DIJK, N., *Queueing Networks and Product Forms: A Systems Approach*, Wiley, New York, 1993.

75. VAN DIJK, N. AND LAMOND, B.F., "Simple bounds for finite single-server exponential tandem queues," *Oper. Res.*, vol. 36, pp. 470-477, 1988.

76. VAN DIJK, N. AND VAN DER WAL, J., "Simple bounds and monotonicity results for finite multi-server exponential tandem queues," *Queueing Systems*, vol. 4, pp. 1-16, 1989.

77. VAN HOUTUM, G.J., ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "The compensation approach for three or more dimensional random walks," in *Operations Research Proceedings 1992*, pp. 342-349, Springer-Verlag, Berlin, 1993.

78. VAN HOUTUM, G.J., ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "The equilibrium distribution for a class of multi-dimensional random walks," Memorandum COSOR 94-01, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., 1994. Submitted for publication.

79. VAN HOUTUM, G.J., ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "On the precedence relation method for deriving flexible bound models for queueing systems," Memorandum COSOR 94-27, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., 1994. Submitted for publication.

80. VAN HOUTUM, G.J., ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "The equilibrium distribution for a class of multi-dimensional random walks: structure analysis," Memorandum COSOR, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., 1995. In preparation.

81. VAN HOUTUM, G.J., ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "The precedence relation method for deriving flexible bound models for queueing systems," Memorandum COSOR, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., 1995. In preparation.

82. WRIGHT, PAUL E., "Two parallel processors with coupled inputs," *Adv. Appl. Prob.*, vol. 24, pp. 986-1007, 1992.

83. ZHAO, Y. AND GRASSMANN, W.K., "The shortest queue model with jockeying," *Naval Res. Log.*, vol. 37, pp. 773-381, 1990.

84. ZHAO, Y. AND GRASSMANN, W.K., "A numerically stable algorithm for two server queue models," *Queueing Systems*, vol. 8, pp. 59-80, 1991.

85. ZHENG, Y.S. AND ZIPKIN, P., "A queuing model to analyze value of centralized inventory information," *Oper. Res.*, vol. 38, pp. 296-307, 1990.

86. ZIJM, W.H.M., "Operational control of automated PCB assembly lines," in *Modern production concepts: theory and applications*, ed. G. Fandel, G. Zaepfel, pp. 146-164, Springer-Verlag, Berlin, 1991.

Samenvatting

Dit proefschrift is gewijd aan de ontwikkeling van twee methoden voor het bepalen van evenwichtsverdelingen, en met behulp daarvan de relevante prestatiematen, van wachtrijsystemen waarvan het gedrag beschreven wordt door multi-dimensionale Markov modellen. Deze methoden hebben we respectievelijk de *compensatiemethode* en de *precedentierelatie methode* genoemd. Van beide methoden beschrijven we hieronder in het kort de belangrijkste aspecten.

Voor vele wachrijsystemen wordt het gedrag beschreven door $N(\geq 2)$-dimensionale Markov modellen met toestanden $(m_1, \ldots, m_N)$, waarbij de componenten m_i grootheden zoals lengten van wachtrijen representeren. Dit leidt tot een toestandsruimte die oneindig is in iedere richting, en vaak ook tot een zekere vorm van *homogeniteit* in de overgangskansen. Multi-dimensionale Markov processen waarvan de toestandsruimte oneindig is in iedere richting, zijn in het algemeen nauwelijks of niet expliciet oplosbaar. Echter, een bepaalde homogeniteit in de overgangskansen kan er voor zorgen dat het afleiden van expliciete formules voor de evenwichtsverdeling wel mogelijk is. Het is bekend dat voor de klasse van *produktvorm netwerken*, de evenwichtsverdeling gelijk is aan een produktvorm oplossing, en dat deze oplossing kan worden verkregen door het substitueren van een produktvorm oplossing in de evenwichtsvergelijkingen en vervolgens het resterende stelsel van niet-lineaire vergelijkingen op te lossen. In de hoofdstukken 2-4 van dit proefschrift zijn wij er in geslaagd om voor een tweede, en tot nu toe enige andere, klasse van multi-dimensionale Markov modellen waarvan de toestandsruimte oneindig is in iedere richting, expliciete formules voor de evenwichtsverdeling af te leiden. Deze expliciete formules konden worden afgeleid door gebruik te maken van de zogenaamde *compensatiemethode*.

De compensatiemethode is oorspronkelijk ontwikkeld door Adan, Wessels en Zijm voor een klasse van twee-dimensionale, homogene Markov modellen. Het *basisidee* achter de methode is dat men voor homogene Markov modellen de evenwichtsverdeling kan bepalen door eerst een klasse van produktvorm oplossingen te definiëren, die voldoen aan de evenwichtsvergelijkingen voor de toestanden in het inwendige van de toestandsruimte, en door vervolgens lineaire combinaties van deze produktvorm oplossingen te construeren, die ook voldoen aan de evenwichtsvergelijkingen voor de toestanden op de randen van de toestandsruimte.

In de hoofdstukken 2-4 hebben wij het basisidee van de compensatiemethode toegepast op, en de compensatiemethode zelf uitgebreid naar, de klasse van N-dimensionale, irreducibele, positief recurrente, homogene Markov modellen met de zogeheten projectie eigenschap en de eigenschap dat vanuit iedere toestand alleen overgangen naar buurtoestanden worden gemaakt. Met behulp van inductie naar de dimensie N, hebben we in de hoofdstukken 2 en 3 laten zien dat de evenwichtsverdeling voor een Markov model van deze klasse kan worden bepaald met behulp van de compensatie aanpak dan en slechts dan als er geen overgangen kunnen worden gemaakt vanuit toestanden $(m_1, \ldots, m_N)$ in het inwendige van de

toestandsruimte in richtingen die leiden tot een verhoging van de som van de componenten m_i en m_j voor zekere indices $i,j \in \{1,\ldots,N\}$, $i \neq j$. Deze conditie vormt een generalisatie van de conditie die door Adan e.a. werd verkregen voor de door hen onderzochte klasse van twee-dimensionale Markov modellen. Het is duidelijk dat deze conditie de toepasbaarheid van de compensatiemethode beperkt, zeker voor drie- en hoger-dimensionale Markov modellen. Een wachtrijsysteem dat ook voor hogere N aan de conditie voldoet, is het zogenaamde $2\times N$ switch systeem.

Voor een Markov model van de onderzochte klasse, waarvoor aan de genoemde conditie wordt voldaan, leidt de toepassing van de compensatiemethode tot zeer expliciete resultaten. Het leidt tot het bewijs dat de evenwichtsverdeling kan worden geschreven als een *alternerende som van aftelbaar veel pure produktvorm verdelingen*, en dat dit tevens geldt voor alle marginale verdelingen. Voor het bepalen van alle benodigde produktvorm verdelingen zijn eenvoudige, recursieve formules beschikbaar. Verder hebben we in hoofdstuk 4 door het bestuderen van de struktuur achter deze recursieve formules eenvoudige foutengrenzen afgeleid voor het benaderen van de evenwichtskansen door eindige sommen van produktvorm verdelingen. Deze foutengrenzen hebben geleid tot efficiënte procedures voor het berekenen van evenwichtskansen en van de bijbehorende prestatiematen. Deze procedures zijn gebruikt voor het berekenen van enige interessante, numerieke resultaten voor het $2\times N$ switch systeem.

Voor vele multi-dimensionale wachtrijsystemen en Markov modellen is het niet mogelijk om de evenwichtsverdeling, en de relevante prestatiematen, expliciet te bepalen met behulp van een analytische methode. Voor zulke modellen kan het gebruik van flexibele truncatie-modellen een geschikt alternatief vormen. Flexibele truncatie-modellen zijn truncatie-modellen die het oorspronkelijke systeem of model willekeurig dicht kunnen benaderen. Zulke modellen kunnen worden verkregen door de grootte van de afgeknotte toestandsruimte af te laten hangen van bepaalde truncatieparameters. Ze leiden tot benaderingen voor de evenwichtsverdeling van het originele model en ze zijn extra interessant indien ze leiden tot grenzen voor de relevante prestatiematen van het oorspronkelijke model. In dat laatste geval, noemen we ze ook wel *flexibele grensmodellen*. In hoofdstuk 5 van dit proefschrift hebben we de zogenaamde *precedentierelatie methode* ontwikkeld, die zeer geschikt blijkt te zijn voor het afleiden van flexibele grensmodellen.

De precedentierelatie methode is in principe een methode voor het vergelijken van de gemiddelde kosten in twee Markov kostenmodellen, waarbij de toestandsruimte van het ene model een deelverzameling is van de toestandsruimte van het andere model. Het *basisidee* van de precedentierelatie methode is dat deze vergelijking kan worden vereenvoudigd door eerst voor paren van toestanden van het model met de grootste toestandsruimte te laten zien dat ze voldoen aan een bepaalde *precedentierelatie* die aangeeft dat de eerste toestand van een paar aantrekkelijker is met betrekking tot de kosten dan de tweede toestand. Paren die aan de precedentierelatie voldoen heten precedentieparen. Wij hebben ons vooral gericht op het gebruik van de precedentierelatie methode voor het verkrijgen van flexibele truncatiemodellen die grenzen voortbrengen voor één of meerdere relevante prestatiematen van een gegeven, origineel Markov (wachtrij-)systeem. Zulke grensmodellen kunnen worden afgeleid via de volgende twee stappen. In de eerste, voorbereidende, stap dient men precedentieparen voor het originele model af te leiden. Daarna kunnen in de tweede stap eenvoudig flexibele

grensmodellen worden verkregen: een flexibel ondergrensmodel (resp. bovengrensmodel) verkrijgt men door eerst een flexibele afgeknotte toestandsruimte M' te definiëren, en vervolgens alle overgangen die eindigen in toestanden buiten M' terug te koppelen naar toestanden binnen M', die aantrekkelijker (resp. onaantrekkelijker) zijn volgens de afgeleide precedentieparen voor het originele model. Het zal duidelijk zijn dat het afleiden van de precedentieparen in de voorbereidende stap essentieel is voor het kunnen verkrijgen van flexibele grensmodellen die reeds voor kleine afgeknotte toestandsruimten leiden tot nauwkeurige grenzen voor de relevante prestatiematen van het originele model.

De precedentierelatie methode kan in principe worden toegepast op ieder multidimensionaal wachtrijsysteem of Markov model. Maar het zal met name van de struktuur van zowel de toestandsruimte als de mogelijke overgangen met de bijbehorende overgangskansen afhangen of deze methode kan leiden tot flexibele grensmodellen van een voldoende hoge *kwaliteit*. De kwaliteit van een flexibel grensmodel wordt bepaald door de verhouding tussen de nauwkeurigheid van de geproduceerde grenzen en de efficiëncy waarmee het grensmodel zelf kan worden opgelost.

In de hoofdstukken 6 en 7 van dit proefschrift hebben we de precedentierelatie methode toegepast op het als zeer lastig bekend staande symmetrische kortste rij systeem en op een generaliseerd systeem dat wordt verkregen indien er meerdere klanttypen zijn en elk klanttype door (slechts) een deelverzameling van alle parallelle servers bediend kan worden. Voor het symmetrische kortste rij systeem hebben we een flexibel ondergrensmodel en een flexibel bovengrensmodel afgeleid, die allebei al voor kleine afgeknotte toestandsruimten tot nauwkeurige grenzen voor de gemiddelde wachttijd leiden en die allebei efficiënt zijn op te lossen met behulp van de matrix-geometrische methode. Deze twee grensmodellen hebben dientengevolge geleid tot een efficiënte numerieke procedure voor het berekenen van de gemiddelde wachttijd binnen een gegeven, gewenste nauwkeurigheid. Met behulp van deze procedure zijn we in staat gebleken om de gemiddelde wachttijd tamelijk nauwkeurig te berekenen voor systemen bestaande uit vele parallelle loketten en hoge werklasten. Voor de generalisatie van het symmetrische kortste rij systeem heeft de precedentierelatie methode geleid tot soortgelijke grensmodellen als voor het symmetrische kortste rij systeem zelf. Ook voor dit systeem zijn de verkregen grensmodellen geschikt gebleken voor het berekenen van de relevantie prestatiematen (in dit geval, de gemiddelde wachttijden voor ieder soort klanten apart en voor alle soorten klanten tezamen). Een derde wachtrijsysteem waarvoor inmiddels gebleken is dat de precedentierelatie methode leidt tot bruikbare grensmodellen voor de prestatieanalyse, is het symmetrische *langste* rij systeem.

Curriculum Vitae

De auteur van dit proefschrift, Geert-Jan van Houtum, werd geboren op 21 april 1967 te Erp. Van 1979 to 1985 bezocht hij het Monseigneur Zwijsen College te Veghel. Na het behalen van het VWO-diploma aldaar, begon hij in september 1985 aan zijn studie Technische Wiskunde aan de Technische Universiteit Eindhoven. Hij studeerde af in de afstudeerrichting Besliskunde in oktober 1990 na het uitvoeren van een afstudeerproject onder begeleiding van prof.dr. W.H.M. Zijm. Het afstudeerproject betrof de analyse van multi-echelon voorraadsystemen.

Van november 1990 tot en met oktober 1994 is de auteur voor het verrichten van promotieonderzoek verbonden geweest als assistent in opleiding aan de faculteit Wiskunde en Informatica van de Technische Universiteit Eindhoven. Het resultaat van dit onderzoek is weergegeven in dit proefschrift. Het promotieonderzoek werd gefinancierd door het Landelijk Netwerk Mathematische Besliskunde en het werd verricht onder begeleiding van en in samenwerking met prof.dr. J. Wessels, prof.dr. W.H.M. Zijm en dr.ir. I.J.B.F. Adan. Verder heeft er samenwerking plaatsgevonden met prof.dr. O.J. Boxma en dr.ir. J. van der Wal. Gedurende twee korte perioden werd het onderzoek verricht tijdens werkverblijven aan het International Institute for Applied Systems Analysis, als gast van prof.dr. J. Wessels, en aan de University of Arizona, als gast van prof.dr. M.F. Neuts. Sinds november 1994 is de auteur werkzaam als Universitair Docent bij de faculteit Werktuigbouwkunde van de Universiteit Twente, in een overeenstemmingsrelatie met prof.dr. W.H.M. Zijm.

STELLINGEN

behorende bij het proefschrift

New Approaches for Multi-Dimensional Queueing Systems

van

Geert-Jan van Houtum

Stelling 1.

In de operations research kunnen numerieke experimenten zeer waardevol zijn voor het verschaffen van extra inzicht in een bepaald probleem en voor het vinden van de juiste weg om een probleem analytisch op te lossen; zie bijv. hoofdstuk 1 van dit proefschrift, waarin getoond wordt hoe een eenvoudig numeriek experiment de sleutel aanreikt voor de analytische oplossing van het befaamde symmetrische kortste rij systeem met twee parallelle servers. Numerieke experimenten zouden daarom meer aandacht moeten krijgen in de vakliteratuur.

Stelling 2.

In [1-3] worden drie verschillende flexibele truncatie-modellen beschreven voor het symmetrische kortste rij probleem met twee parallelle servers. Met behulp van de precedentierelatie methode, zoals beschreven in hoofdstuk 5 van dit proefschrift, kan men op een eenvoudige manier bewijzen dat de twee truncatie-modellen van Conolly [1] en Gertsbakh [2] leiden tot *ondergrenzen* voor de gemiddelde wachttijd in het oorspronkelijke kortste rij systeem en dat het truncatie-model van Rao en Posner [3] leidt tot *bovengrenzen* voor deze relevante prestatiemaat. Deze resultaten kunnen ook worden bewezen voor direkte generalisaties van de drie truncatie-modellen voor het geval dat men *twee of meer* parallelle servers heeft; ook voor dit algemenere geval kan het bewijs eenvoudig worden gegeven met behulp van de precedentierelatie methode.

1. CONOLLY, B.W., "The autostrada queueing problem," *J. Appl. Prob.*, vol. 21, pp. 394-403, 1984.
2. GERTSBAKH, I., "The shorter queue problem: a numerical study using the matrix-geometric solution," *Eur. J. Oper. Res.*, vol. 15, pp. 374-381, 1984.
3. RAO, B.M. AND POSNER, M.J.M., "Algorithmic and approximation analysis of the shorter queue model," *Naval Res. Log.*, vol. 34, pp. 381-398, 1987.

Stelling 3.

De optimaliteit van aanvulstrategieën (Base Stock policies) is bewezen voor bepaalde multi-echelon voorraadsystemen met een pure assemblage- of convergente struktuur (zie bijv. [1-3]) en, onder de zogenaamde balansaanname, voor bepaalde multi-echelon voorraadsystemen met een pure distributie- of divergente struktuur (zie bijv. [2,4]). Met behulp van de aanpak zoals beschreven in [4], en onder de balansaanname, kan de optimaliteit van aanvulstrategieën worden bewezen voor algemenere multi-echelon voorraadsystemen met een willekeurige convergente struktuur tussen de externe leveranciers en een bepaald centraal voorraadpunt en een willekeurige divergente struktuur tussen het centrale voorraadpunt en de eindvoorraadpunten.

1. CLARK, A.J., AND SCARF, H., "Optimal policies for a multi-echelon inventory problem," *Management Science*, vol. 6, pp. 475-490, 1960.
2. LANGENHOFF, L.J.G., AND ZIJM, W.H.M., "An analytical theory of multi-echelon production/distribution systems," *Statistica Neerlandica*, vol. 44, pp. 149-174, 1990.
3. VAN HOUTUM, G.J., AND ZIJM, W.H.M., "Computational procedures for stochastic multi-echelon production systems," *Int. J. Prod. Econom.*, vol. 23, pp. 223-237, 1991.
4. VAN HOUTUM, G.J., AND ZIJM, W.H.M., "Theoretische en numerieke analyse van multi-echelon voorraadsystemen met distributiestruktuur," Working paper, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., 1992.

Stelling 4.

Onder de balansaanname, wordt in [1] een aanvulstrategie afgeleid die de gemiddelde (voorraad)kosten minimaliseert in een distributiesysteem (divergent multi-echelon voorraadsysteem) bestaande uit *één* centraal depot en een willekeurig aantal lokale vemen. De karakterisering van deze optimale aanvulstrategie laat zien dat men in het centrale depot *geen tussenvoorraad* hoeft aan te houden, indien de toegevoegde voorraadkosten voor de lokale vemen gelijk aan 0 zijn (d.w.z. indien de voorraadkosten per produkt voor de lokale vemen even groot zijn als voor het centrale depot). Deze eigenschap zal *niet* worden verkregen, indien de balansaanname *niet* wordt gemaakt. In feite impliceert de balansaanname het kostenloos vervoer van produkten tussen de diverse lokale vemen.

1. LANGENHOFF, L.J.G., AND ZIJM, W.H.M., "An analytical theory of multi-echelon production/distribution systems," *Statistica Neerlandica*, vol. 44, pp. 149-174, 1990.

Stelling 5.

In de literatuur over cyclische codes beperkt men zich gewoonlijk tot codes over alfabet $\mathbb{F}_q$ en ter lengte n waarbij $\text{ggd}(q,n)=1$, d.w.z. waarbij n relatief priem is t.o.v. q. Deze gewoonte lijkt gerechtvaardigd te zijn op basis van het in [1-3] uitgevoerde onderzoek dat heeft aangetoond dat codes over alfabet $\mathbb{F}_q$ en met een lengte n waarvoor $\text{ggd}(q,n)>1$ in het algemeen slechter zijn dan codes met een lengte n waarvoor wel geldt dat $\text{ggd}(q,n)=1$.

1. BLOEMEN, A.A.F., VAN HOUTUM, G.J.J.A.N., AND VERHAEGH, W.F.J., "Over cyclische codes over alfabet $\mathbb{F}_q$ en met lengte $q^k n$," Working paper, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., 1989.
2. CASTAGNOLI, GUY, MASSEY, JAMES L., SCHOELLER, PHILIP A., AND VON SEEMANN, NIKLAUS, "On repeated-root cyclic codes," *IEEE Trans. Inform. Theory*, vol. 37, pp. 337-342, 1991.
3. VAN LINT, J.H., "Repeated-root cyclic codes," *IEEE Trans. Inform. Theory*, vol. 37, pp. 343-345, 1991.

Stelling 6.

Het bepalen van de relevante prestatiematen van een moeilijk oplosbaar wachtrijsysteem via relatief eenvoudig te analyseren flexibele onder- en bovengrensmodellen, zoals beschreven in de hoofdstukken 5-7 van dit proefschrift, is in feite equivalent aan de volgende aanpak voor het uit het hoofd bepalen van de waarde van een wortel. Bijvoorbeeld, voor $\sqrt{2}$ verkrijgt men onmiddelijk de grenzen $1 \le \sqrt{2} \le 2$ door te bedenken dat het niet-quadratische (niet-oplosbare) getal 2 tussen de kwadraten 1 en 4 ligt; vervolgens vindt men via de observatie dat het getal 200 tussen de kwadraten 196 en 225 ligt, de nauwkeurigere grenzen $\sqrt{2} \ge \sqrt{1,96} = 1,4$ en $\sqrt{2} \le \sqrt{2,25} = 1,5$; enz.

Stelling 7.

De doorsnee speler van het spel Blackjack, zoals dat gespeeld wordt in de Holland Casino's, speelt dit spel volgens een strategie waarbij geen rekening wordt gehouden met de reeds getrokken kaarten. Bekend is dat zo'n strategie een negatieve verwachte winst per eenheid inzet oplevert (zie [1]), en derhalve zal de geldvoorraad van de doorsnee speler zich gedragen volgens een één-dimensionale stochastische wandeling (random walk) met een negatieve drift (d.w.z. met een drift naar de positie 0).

1. VAN DER GENUGTEN, B.B., *Blackjack in Holland Casino's: hoe de dealer te verslaan!*, Tilburg University Press, Tilburg, 1993.

Stelling 8.

Huisvrouwen of -mannen die bij hun wekelijkse bezoek aan een supermarkt precies die spullen kopen die in de afgelopen week zijn opgegaan, hanteren in feite het uit de voorraadtheorie bekende Kanban-systeem.

Stelling 9.

Ter bevordering van de kwaliteit van de verschillende studierichtingen, en tegelijkertijd ter bestrijding van studierichtingen met goedklinkende namen maar een slechte inhoud, zou de Nederlands overheid moeten besluiten om in de toekomst de universiteiten niet af te rekenen op basis van de aantallen afgestudeerden, maar op basis van de aantallen afgestudeerden die in staat zijn om binnen een afzienbare tijd een baan te vinden met de genoten opleiding.

Stelling 10.

Omdat het bij een voetbalwedstrijd vooral gaat om de mate waarin een relatief klein aantal kansen wordt benut, is er sprake van een vrij sterke invloed van de stochastiek (het toeval) op de einduitslag. Met name voetbaljournalisten en andere pseudo-voetbalkenners willen die invloed nogal eens onderschatten en vertalen een uitslag liever met behulp van termen als vorm en lekker in de wedstrijd zitten.

Stelling 11.

Een bekende Eindhovense voetbalvereniging heeft de oudere en relatief duurdere werknemers aan de kant gezet en tegelijkertijd geïnvesteerd in jonge en relatief goedkope krachten om een ommekeer in de neergaande spiraal te bewerkstelligen. Een bekende Eindhovense onderzoeks- en onderwijsinstelling dreigt de tegenovergestelde weg te gaan bewandelen. Beide organisaties lijken daarmee een verre van optimale strategie te (gaan) volgen.