A COMPENSATION APPROACH FOR QUEUEING PROBLEMS

I.J.B.F. ADAN

A COMPENSATION APPROACH FOR QUEUEING PROBLEMS

A COMPENSATION APPROACH FOR QUEUEING PROBLEMS

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 19 november 1991 om 16.00 uur

door

Ivo Jean-Baptiste François Adan

geboren te Roosendaal

druk: wibro dissertatiedrukkerij, helmond.

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

Table of Contents

Chapter 1

Introduction

In this monograph we study the equilibrium behaviour of two-dimensional Markov processes. Such processes are frequently used for the modelling of queueing problems. At present several techniques for the mathematical analysis of two-dimensional Markov processes are available. Most of these techniques are based on generating functions. A classical example is the analysis of the symmetric shortest queue problem. Kingman [44] and Flatto and McKean [23] use a uniformization technique to determine the generating function of the equilibrium distribution of the lengths of the two queues. From this generating function they obtain valuable insights in the asymptotic behaviour as well as in the specific form of the equilibrium probabilities. A similar uniformization approach has been used by Hofri to analyse a multiprogramming computer system with two queues involved (see Hofri [37] and Adan, Wessels and Zijm [1] for additional information) and by Flatto and Hahn [24] to analyse two $M|M|1$ queues with coupled arrivals. There are more general approaches regarding the analysis of generating functions of two-dimensional Markov processes. The work of Iasnogorodski and Fayolle [19,20,40] and Cohen and Boxma [14] shows that the study of the generating function of fairly general two-dimensional Markov processes can be reduced to that of a Riemann type boundary value problem. With some minor modifications this approach also proceeds for the time-dependent case. However, none of the approaches mentioned leads to an explicit characterization of the equilibrium probabilities, or can easily be used for numerical purposes.

A numerically-oriented method has been developed by Hooghiemstra, Kean and Van Ree [38]. This method is based on the calculation of power-series expansions for the equilibrium probabilities as functions of the traffic intensity and applies to fairly general exponential multi-dimensional queueing systems. For selected problems, the coefficients in these expansions may be found explicitly, see De Waard [58] who derives explicit relations for the coefficients in the power-series expansion for the equilibrium probabilities of the symmetric coupled processor problem. Blanc [10-12] reports that this approach works numerically satisfactory for several queueing problems. The theoretical foundation of this method, however, is still incomplete.

The main objective of the present monograph is to contribute to the development of techniques for the analysis of the equilibrium behaviour of Markov-processes with a two-dimensional state space. Our research was initiated with the analysis of the symmetric shortest queue problem. For this queueing problem we developed an approach to the characterization and calculation of the equilibrium probabilities. The essence of this approach is to characterize

the set of product form solutions satisfying the equations in the interior points and then to use the solutions in this set to construct a linear combination of product form solutions which also satisfies the boundary conditions. This construction is based on a compensation idea: after introducing the first term, terms are added so as to alternately compensate for errors on the two boundaries. This explains the name compensation approach. Keilson also developes a compensation method in his book [43]. Keilson's method, however, has not much affinity with our method. The compensation approach leads to an explicit characterization of the equilibrium probabilities, and therefore extends the work of Kingman [44] and Flatto and McKean [23]. Our results can easily be exploited for numerical analysis and lead to efficient algorithms with the advantage of tight error bounds.

As a first attempt to investigate the scope of the compensation approach, we apply these ideas to Markov processes on the lattice in the positive quadrant of $\mathbb{R}^2$. We consider processes for which the transition rates are constant in the interior of the state space and also constant on the two axes. To simplify the analysis, we assume that the transitions are restricted to neighbouring states. This class of processes is sufficiently rich in the sense that all queueing problems mentioned in the previous paragraphs can be modelled as Markov processes of this type. We derive conditions under which the compensation approach works. It appears that the essential condition is that transitions from interior states to the north, north-east and the east are not allowed. The symmetric shortest queue problem and the problem of multiprogramming queues can be formulated as Markov processes satisfying this condition. However, the other two queueing problems, mentioned in the previous paragraphs, violate this condition. Consequently, the compensation approach does not work for these two problems.

The compensation approach can be extended in various directions. Some of the possibilities are investigated in this monograph. The approach can easily be extended to the shortest queue model with a threshold-type jockeying. This means that one job jumps from the longest to the shortest queue if the difference between the lengths of the two queues exceeds some threshold value. For this model the main term already satisfies the boundary conditions. Thus no compensation arguments are required. Gertsbakh [29] studies this model by using the matrix-geometric approach developed by Neuts [51]. The relationship between these two approaches has been investigated in [8].

It appears that the compensation approach also works for the asymmetric shortest queue problem. This problem can be formulated as a Markov process on two adjacent quadrants of $\mathbb{R}^2$ with different stochastic properties in each quadrant. The compensation approach leads to an explicit characterization of the equilibrium probabilties. Although in this case the solution structures are rather complicated, our final results can easily be exploited for numerical purposes. Fayolle and Iasnogorodski [19,40] and Cohen and Boxma [14] show that the analysis of the generating function can be reduced to that of a *simultaneous* Riemann-Hilbert boundary

value problem. This type of boundary value problem, however, requires further research. The compensation approach further yields satisfactory results for the shortest delay problem with Erlang servers. This problem can be modelled as a Markov process for which transitions are not restricted to neighbouring states only. This process is not skipfree to the south, which is a basic assumption for the models studied in the book of Cohen and Boxma [14]. For the two problems mentioned no other analytical results are available in the literature.

In the following sections we give a short review of the different problems, which will be treated in the subsequent chapters, and a sketch of the solution approaches showing the kind of arguments that will be used. In the next section the compensation approach is outlined for the symmetric shortest queue problem. This section does not contain rigorous proofs, but is intended to sketch the basic ideas. Section 1.2 is devoted to an extension of the approach to a wider variety of problems. The next section briefly comments on several possibilities to further extend the approach. The compensation idea has an interesting analogue in the field of classical electrostatics, which is known as the *method of images*. This analogue is described in section 1.4. Finally, the contents of the subsequent chapters is summarized in section 1.5.

1.1. Introduction to the compensation approach

In this section we analyse the symmetric shortest queue problem. Our interest in this problem arose from problems in the design of flexible assembly systems. The final section in chapter 6 will be devoted to a short description of these problems (see also [2, 7]).

The symmetric shortest queue problem is characterized as follows. Consider a system with two identical servers (see figure 1.1). Jobs arrive according to a Poisson stream with rate 2ρ where $0 < \rho < 1$ On arrival a job joins the shortest queue, and, if queues have equal length, joins either queue with probability 1/2. The jobs require exponentially distributed service times with unit mean, the service times are supposed to be independent.

This problem has been addressed by many authors. Kingman [44] and Flatto and McKean [23] analyse the problem by using generating functions. They show that the generating function of the lengths of the two queues is a meromorphic function. By partial fraction decomposition of the generating function they can express the equilibrium probabilities as an infinite sum of products of powers. However, the decomposition leads to cumbersome expressions. An alternative approach can be found in Cohen and Boxma [14] and Fayolle and Iasnogorodski [19, 20, 40]. They show that the analysis of the functional equation for the generating function can be reduced to that of a Riemann-Hilbert boundary value problem. None of these approaches however, leads to an explicit characterization of the equilibrium probabilities or closes the matter from a numerical point of view.

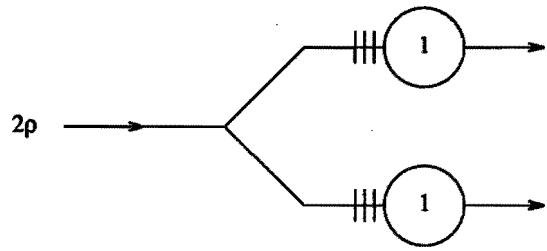


Figure 1.1.
The symmetric shortest queue model. Arriving jobs join the shortest queue and in case of equal queues, join either queue with probability 1/2.

In this section we show how the empirical finding that the asymptotic product form for the equilibrium probabilities (see theorem 5 in Kingman [44]) is already a good approximation at a short distance from the boundaries, can be exploited to develop a technique by which the probabilities can be found efficiently. This technique leads to an explicit characterization of the probabilities and therefore extends the results of Kingman [44] and Flatto and McKean [23]. Moreover, our results can be easily exploited for numerical calculations. The purpose of this section is not to provide rigorous proofs, but to illustrate the basic ideas.

The queueing system can be represented by a continuous-time Markov process, whose natural state space consists of the pairs (i, j) where i and j are the lengths of the two queues. Instead of i and j we use the state variables m and n where $m = \min(i, j)$ and $n = j - i$. So m is the length of the shortest queue and n is the difference between the queue lengths. Let $\{p_{m,n}\}$ be the equilibrium distribution. The transition-rate diagram is depicted in figure 1.2. The rates in the region $n \leq 0$ can be obtained by reflection in the m-axis. By symmetry $p_{m,n} = p_{m,-n}$. Hence the analysis can be restricted to the probabilities $p_{m,n}$ in the region $n \geq 0$.

The equilibrium equations for $\{p_{m,n}\}$ can be found by equating for each state the rate into and the rate out of that state. In the equations below we have eliminated the probabilities $p_{m,0}$ from (1.2) and (1.4) by substituting (1.5) and (1.6). This is done to simplify the presentation. The analysis can now be restricted to the probabilities $p_{m,n}$ with $n > 0$. These probabilities satisfy equations (1.1)-(1.4). The equations (1.5)-(1.6) are further treated as definition for $p_{m,0}$.

$$p_{m,n}2(\rho+1) = p_{m-1,n+1}2\rho + p_{m,n+1} + p_{m+1,n-1}\,, \qquad m>0,\, n>1 \quad (1.1)$$

$$p_{m,1}2(\rho+1) = p_{m-1,2}2\rho + p_{m,2} + (p_{m,1}2\rho + p_{m+1,1})\frac{1}{\rho+1} + (p_{m-1,1}2\rho + p_{m,1})\frac{\rho}{\rho+1}\,, \qquad m>0 \quad (1.2)$$

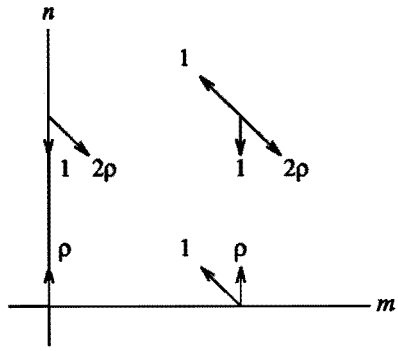


Figure 1.2.
Transition-rate diagram for the symmetric shortest queue model in figure 1.1.

$$p_{0,n}(2\rho+1) = p_{0,n+1} + p_{1,n-1}\,, \qquad n>1 \quad (1.3)$$

$$p_{0,1}(2\rho+1) = p_{0,2} + (p_{0,1}2\rho + p_{1,1})\frac{1}{\rho+1} + p_{0,1}\,, \qquad (1.4)$$

$$p_{m,0}(\rho+1) = p_{m-1,1}2\rho + p_{m,1}\,, \qquad m>0 \quad (1.5)$$

$$p_{0,0}\rho = p_{0,1}\,. \qquad (1.6)$$

Numerical experiments show that the probabilities $p_{m,n}$ behave as a product $K\alpha^m\beta^n$ already at a short distance from the boundaries. This feature is illustrated in table 1.1 for the case $\rho = 0.5$ by displaying the ratios $p_{m+1,n}/p_{m,n}$ and $p_{m,n+1}/p_{m,n}$. Actually we calculated *approximations* for $p_{m,n}$ by solving a *finite capacity* system exactly, that is, by means of a Markov chain analysis. In the example we calculated the probabilities for a system where each queue has a capacity of 15 jobs, which, for $\rho = 0.5$ approximates the infinite capacity system quite well. In table 1.1 we see that for almost all m and n the ratios $p_{m+1,n}/p_{m,n}$ and $p_{m,n+1}/p_{m,n}$ are constant, which suggests that for some constant K,

$$p_{m,n} \sim K\,\alpha^m\beta^n\,, \quad (n>0,\, m\to\infty)\,, \qquad (1.7)$$

where $\alpha = 0.25$ and $\beta = 0.1$ for the case $\rho = 0.5$. The first question that arises is: what are α and β in general? To obtain α, consider the process on the aggregate states k where k is the total number of jobs in the system. An approximation of the transition-rate diagram is depicted in figure 1.3.

n	0	1	2	3	4	5
6	0.19	0.24	0.25	0.25	0.25	0.25
5	0.19	0.24	0.25	0.25	0.25	0.25
4	0.19	0.24	0.25	0.25	0.25	0.25
3	0.19	0.24	0.25	0.25	0.25	0.25
2	0.20	0.24	0.25	0.25	0.25	0.25
1	0.28	0.25	0.25	0.25	0.25	0.25
0						

$m \rightarrow$

The ratios $p_{m+1,n} / p_{m,n}$

n	0	1	2	3	4	5
6	0.10	0.10	0.10	0.10	0.10	0.10
5	0.10	0.10	0.10	0.10	0.10	0.10
4	0.10	0.10	0.10	0.10	0.10	0.10
3	0.10	0.10	0.10	0.10	0.10	0.10
2	0.11	0.10	0.10	0.10	0.10	0.10
1	0.15	0.11	0.10	0.10	0.10	0.10
0						

$m \rightarrow$

The ratios $p_{m,n+1} / p_{m,n}$

Table 1.1.
The ratios $p_{m+1,n} / p_{m,n}$ and $p_{m,n+1} / p_{m,n}$ for the case $\rho = 0.5$.

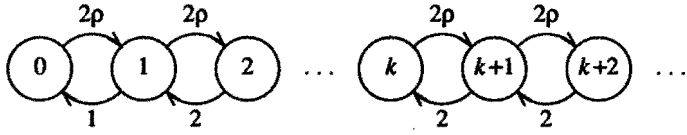


Figure 1.3.
Approximation of the transition-rate diagram for aggregated states k where k is the total number of jobs in the system.

In fact, this is the transition-rate diagram for the $M \mid M \mid 2$ system with arrival rate 2ρ and service rate 1 for both servers. It is an approximation for the shortest queue, since the average service rate in state $k > 1$ is less than 2 due to the fact that one of the servers can be idle. Intuitively it will be obvious that the average service rate tends to 2 as k tends to infinity, so the approximation is better for large k. From the transition-rate diagram we obtain for the equilibrium probabilities P_k that for large k

$$P_k \approx C\rho^k , \tag{1.8}$$

for some constant C. On the other hand, from (1.7) and using that empirically $\beta < \alpha$, for large k we get

$$P_{2k-1} = 2\sum_{l=1}^{k} p_{k-l,2l-1} \approx 2K\alpha^k \sum_{l=1}^{k} \frac{\beta^{2l-1}}{\alpha^l} \approx 2K\alpha^k \sum_{l=1}^{\infty} \frac{\beta^{2l-1}}{\alpha^l} . \tag{1.9}$$

Relations (1.8) and (1.9) suggest that

$$\alpha = \rho^2 ,$$

which agrees with the empirical value $\alpha = 0.25$ for $\rho = 0.5$. The parameter β can be found by observing that if the product (1.7) describes the asymptotic behaviour as $m \to \infty$, then it has to satisfy equations (1.1) and (1.2) for $m > 0$. Inserting $K\alpha^m\beta^n$ into (1.1) and then dividing by the common factor $K\alpha^{m-1}\beta^{n-1}$ leads to a *quadratic equation* for α and β.

Lemma 1.1.

The product $K\alpha^m\beta^n$ satisfies (1.1) if and only if

$$\alpha\beta 2(\rho + 1) = \beta^2 2\rho + \alpha\beta^2 + \alpha^2 . \tag{1.10}$$

Substituting $\alpha = \rho^2$ in (1.10) leads to a quadratic equation in β with roots $\beta = \rho$ and $\beta = \rho^2/(2+\rho)$. The first root yields the asymptotic solution $p_{m,n} \sim K\,\rho^{2m}\,\rho^n$ for some K, corresponding to the equilibrium distribution of two independent $M\,|M\,|\,1$ queues, each with workload ρ. The queues of the shortest queue model, however, are strongly dependent. Therefore, the only sensible choice is

$$\beta = \frac{\rho^2}{2+\rho} ,$$

which agrees with the empirical value $\beta = 0.1$ for $\rho = 0.5$. It can easily be verified that for these values of α and β equation (1.2) is also satisfied. Hence, we find that for some K

$$p_{m,n} \sim K\alpha_0^m\beta_0^n , \quad (n > 0, m \to \infty) \tag{1.11}$$

with

$$\alpha_0 = \rho^2 , \quad \beta_0 = \frac{\rho^2}{2+\rho} .$$

Actually, Kingman ([44], Theorem 5) and Flatto and McKean ([23], Section 3) gave rigorous proofs for this asymptotic result. We now come to the important question of how to exploit this asymptotic result to obtain better approximations. The product $\alpha_0^m\beta_0^n$ does not describe the behaviour near the vertical axis $m = 0$, as can be seen in table 1.1 for $\rho = 0.5$. Indeed $\alpha_0^m\beta_0^n$ violates equation (1.3) for $m = 0$. The idea to improve the initial approximation $\alpha_0^m\beta_0^n$ is:

> Try to find c_1, α, β with α, β satisfying (1.10) such that $\alpha_0^m\beta_0^n + c_1\alpha^m\beta^n$ satisfies (1.3).

Inserting this linear combination into (1.3) yields the condition:

$$A(\alpha_0, \beta_0)\beta_0^{n-1} + c_1 A(\alpha, \beta)\beta^{n-1} = 0\,, \quad n > 1 \tag{1.12}$$

with

$$A(x, y) = y(2\rho + 1) - y^2 - x\,.$$

Since $A(\alpha_0, \beta_0) \neq 0$ and condition (1.12) must hold for *all* $n > 1$, we have to take

$$\beta = \beta_0$$

and thus

$$\alpha = \alpha_1\,,$$

where α_1 is the second, smaller root of (1.10) with $\beta = \beta_0$ (the other root being α_0). The coefficient c_1 can now be solved from (1.12) with $\beta = \beta_0$, yielding

$$c_1 = -\frac{A(\alpha_0, \beta_0)}{A(\alpha_1, \beta_0)} = -\frac{\beta_0(2\rho + 1) - \beta_0^2 - \alpha_0}{\beta_0(2\rho + 1) - \beta_0^2 - \alpha_1}\,. \tag{1.13}$$

Since α_0 and α_1 are roots of (1.10) for $\beta = \beta_0$, we have

$$\alpha_0 + \alpha_1 = \beta_0 2(\rho + 1) - \beta_0^2\,,$$

so (1.13) can be simplified to

$$c_1 = -\frac{\alpha_1 - \beta_0}{\alpha_0 - \beta_0}\,.$$

For this choice of c_1 the sum $\alpha_0^m\beta_0^n + c_1\alpha^m\beta^n$ satisfies (1.3) and, of course, also (1.1). This procedure can be generalized as follows.

Lemma 1.2.

Let x_1 and x_2 be the roots of the quadratic equation (1.10) for fixed β. Then the linear combination $x_1^m\beta^n + cx_2^m\beta^n$ satisfies equations (1.1) and (1.3) if c is given by

$$c = -\frac{x_2 - \beta}{x_1 - \beta}\,. \tag{1.14}$$

For the case $\rho = 0.5$ we display in table 1.2 the same ratios as in table 1.1 for the approximation $\alpha_0^m\beta_0^n + c_1\alpha_1^m\beta_0^n$. Comparing tables 1.1 and 1.2 we see that $\alpha_0^m\beta_0^n + c_1\alpha_1^m\beta_0^n$ also describes the behaviour of the probabilities near the boundary $m = 0$. Hence, we find

$$p_{m,n} \sim K(\alpha_0^m\beta_0^n + c_1\alpha_1^m\beta_0^n)\,, \quad (n > 0,\ m + n \to \infty)\,,$$

n ↑	$m \rightarrow$ 0	1	2	3	4	5
6	0.19	0.24	0.25	0.25	0.25	0.25
5	0.19	0.24	0.25	0.25	0.25	0.25
4	0.19	0.24	0.25	0.25	0.25	0.25
3	0.19	0.24	0.25	0.25	0.25	0.25
2	0.19	0.24	0.25	0.25	0.25	0.25
1	0.19	0.24	0.25	0.25	0.25	0.25
0						

The ratios $\dfrac{(\alpha_0^{m+1} + c_1\alpha_1^{m+1})\beta_0^n}{(\alpha_0^m + c_1\alpha_1^m)\beta_0^n}$

n ↑	$m \rightarrow$ 0	1	2	3	4	5
6	0.10	0.10	0.10	0.10	0.10	0.10
5	0.10	0.10	0.10	0.10	0.10	0.10
4	0.10	0.10	0.10	0.10	0.10	0.10
3	0.10	0.10	0.10	0.10	0.10	0.10
2	0.10	0.10	0.10	0.10	0.10	0.10
1	0.10	0.10	0.10	0.10	0.10	0.10
0						

The ratios $\dfrac{(\alpha_0^m + c_1\alpha_1^m)\beta_0^{n+1}}{(\alpha_0^m + c_1\alpha_1^m)\beta_0^n}$

Table 1.2.
The ratios $p_{m+1,n} / p_{m,n}$ *and* $p_{m,n+1} / p_{m,n}$ *for the approximation* $p_{m,n} = \alpha_0^m\beta_0^n + c_1\alpha_1^m\beta_0^n$ *and the case* $\rho = 0.5$.

for some K. In fact, Flatto and McKean ([23], section 3) proved this statement, which is stronger than (1.11). We added $c_1\alpha_1^m\beta_0^n$ to compensate for the error on the vertical boundary $m = 0$ and by doing so introduced a new error on the horizontal boundary $n = 1$, since this term violates condition (1.2). Since $\alpha_1 < \alpha_0$, the term $c_1\alpha_1^m\beta_0^n$ is very small compared to $\alpha_0^m\beta_0^n$ even for small m. Therefore its disturbing effect near the horizontal boundary is neglegible. However, we can compensate for the error of $c_1\alpha_1^m\beta_0^n$ on the horizontal boundary by again adding a term:

> Try to find α, β, d_1 with α, β satisfying (1.10) such that $c_1\alpha_1^m\beta_0^n + d_1c_1\alpha^m\beta^n$ satisfies (1.2).

If we succeed, then the total sum $\alpha_0^m\beta_0^n + c_1\alpha_1^m\beta_0^n + d_1c_1\alpha_1^m\beta^n$ satisfies (1.2) by linearity. The procedure to find α, β, d_1 is analogous to the one used for the vertical boundary. To satisfy (1.2) for *all* $m > 0$ we are forced to take

$$\alpha = \alpha_1$$

and thus

$$\beta = \beta_1 ,$$

where β_1 is the second, smaller root of (1.10) with $\alpha = \alpha_1$ (the other root being β_0). Inserting $c_1\alpha_1^m\beta_0^n + c_1 d_1 \alpha_1^m \beta_1^n$ into (1.2) and dividing by $c_1\alpha_1^{m-1}$ leads to an equation for d_1 which is solved by

$$d_1 = -\frac{\dfrac{\beta_0(2\rho+\alpha_1)(\alpha_1+\rho)}{\rho+1} - \alpha_1^2}{\dfrac{\beta_1(2\rho+\alpha_1)(\alpha_1+\rho)}{\rho+1} - \alpha_1^2}. \tag{1.15}$$

Since β_0 and β_1 are the roots of the quadratic equation (1.10) for fixed $\alpha = \alpha_1$ we have

$$\beta_0\beta_1(2\rho+\alpha_1) = \alpha_1^2 .$$

This equality reduces (1.15) to

$$d_1 = -\frac{\dfrac{\alpha_1+\rho}{\beta_1} - (\rho+1)}{\dfrac{\alpha_1+\rho}{\beta_0} - (\rho+1)}.$$

This procedure can be generalized as follows.

Lemma 1.3.

Let y_1 and y_2 be the roots of the quadratic equation (1.10) for fixed α. Then the linear combination $\alpha^m y_1^n + d\alpha^m y_2^n$ satisfies (1.1) and (1.2) if d is given by

$$d = -\frac{\dfrac{\alpha+\rho}{y_2} - (\rho+1)}{\dfrac{\alpha+\rho}{y_1} - (\rho+1)}. \tag{1.16}$$

We added $d_1 c_1 \alpha_1^m \beta_1^n$ to compensate for $c_1\alpha_1^m\beta_0^n$ on the horizontal boundary and in doing so introduced a new error, since $d_1 c_1 \alpha_1^m \beta_1^n$ violates the vertical boundary conditions (1.2), so we have to add again a term, and so on. It is clear how to continue: the compensation procedure consists of adding on terms so as to compensate alternately for the error on the vertical boundary, according to lemma 1.2, and for the error on the horizontal boundary, according to lemma 1.3. This results in the infinite sum depicted in figure 1.4. Each term in the sum in figure 1.4 satisfies (1.1), each sum of two terms with the same β-factor satisfies (1.2) and each sum of two terms with the same α-factor satisfies (1.3). Since the equilibrium equations are linear, we can conclude that the sum in figure 1.4 formally satisfies the equations (1.1)-(1.3). Let us define $x_{m,n}$ as the infinite sum of compensation terms, so for $m \geq 0$, $n > 0$,

$$\overbrace{}^{H} \quad \overbrace{}^{H} \quad \overbrace{}^{H}$$

$$d_0c_0\alpha_0^m\beta_0^n + d_0c_1\alpha_1^m\beta_0^n + d_1c_1\alpha_1^m\beta_1^n + d_1c_2\alpha_2^m\beta_1^n + d_2c_2\alpha_2^m\beta_2^n + \cdots$$

$$\underbrace{}_{V} \quad \underbrace{}_{V}$$

Figure 1.4.
The final sum of compensation terms. By definition $c_0 = d_0 = 1$. Sums of two terms with the same β-factor satisfy the vertical boundary conditions (V) and sums of two terms with the same α-factor satisfy the horizontal boundary conditions (H).

$$x_{m,n} = \sum_{i=0}^{\infty} d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n \qquad \text{(pairs with the same β-factor)}, \quad (1.17)$$

$$= c_0d_0\beta_0^n\alpha_0^m + \sum_{i=0}^{\infty} c_{i+1}(d_i\beta_i^n + d_{i+1}\beta_{i+1}^n)\alpha_{i+1}^m \qquad \text{(pairs with the same α-factor)}. \quad (1.18)$$

Below we formulate the recursion relations for α_i, β_i, c_i and d_i.

For the initial values $\alpha_0 = \rho^2$ and $\beta_0 = \rho^2/(2+\rho)$, the sequence

$$\alpha_0 \nearrow \beta_0 \searrow \alpha_1 \nearrow \beta_1 \searrow \alpha_2 \nearrow \beta_2 \searrow \ldots$$

is generated such that for all $i \geq 0$ the numbers α_i and α_{i+1} are the roots of (1.10) for fixed $\beta = \beta_i$ and the numbers β_i and β_{i+1} are the roots of (1.10) for fixed $\alpha = \alpha_{i+1}$. The generation of α_i and β_i is graphically illustrated in figure 1.5.

$\{c_i\}$ is generated such that for all i the term $(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n$ satisfies (1.3). Application of lemma 1.2 yields that c_{i+1} can be obtained from c_i by

$$c_{i+1} = -\frac{\alpha_{i+1} - \beta_i}{\alpha_i - \beta_i}\, c_i\,, \quad i \geq 0\,,$$

where initially

$$c_0 = 1\,.$$

$\{d_i\}$ is generated such that for all i the term $(d_i\beta_i^n + d_{i+1}\beta_{i+1}^n)\alpha_{i+1}^m$ satisfies (1.2). Application of lemma 1.3 yields that d_{i+1} can be obtained from d_i by

$$d_{i+1} = -\frac{(\alpha_{i+1}+\rho)/\beta_{i+1} - (\rho+1)}{(\alpha_{i+1}+\rho)/\beta_i - (\rho+1)}\, d_i\,, \quad i \geq 0\,,$$

where initially

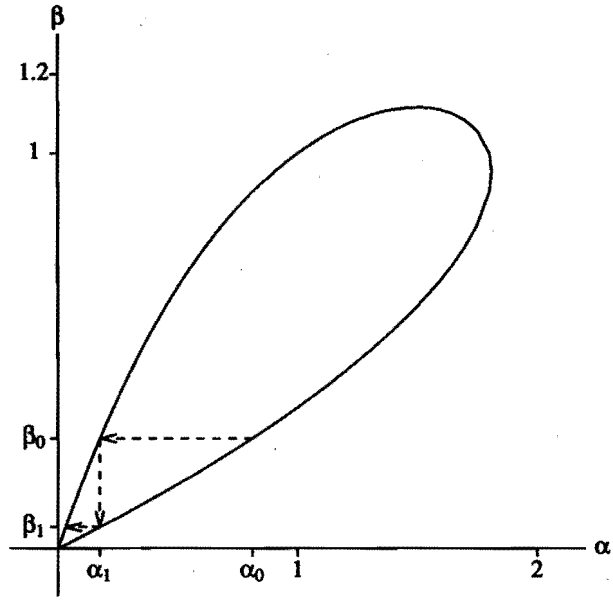


Figure 1.5.
The curve: $\beta^2 2\rho + \alpha\beta^2 + \alpha^2 - \alpha\beta 2(\rho + 1) = 0$ *in the positive quadrant for the case* $\rho = 0.9$. *This curve generates* $\{\alpha_i\}$ *and* $\{\beta_i\}$ *for the symmetric shortest queue problem.*

$d_0 = 1$.

$\{x_{m,n}\}$ is a *formal* solution of (1.1)-(1.3), including (1.4) due to the dependence of the equilibrium equations. The final problem is to prove the convergence of $\{x_{m,n}\}$. It can be shown that $x_{m,n}$ converges absolutely for fixed m and n (absolute convergence guarantees equality of (1.17) and (1.18)), that $x_{m,n} > 0$ for all m and n and that the sum of $x_{m,n}$ over all m and n converges (so normalization is possible). Now it can be concluded from a result of Foster (see appendix A) that the shortest queue problem is ergodic. Since the equilibrium distribution $\{p_{m,n}\}$ of an ergodic system is unique, the normalization of $\{x_{m,n}\}$ produces $\{p_{m,n}\}$.

The parameters α_i and β_i and the normalizing constant can be solved explicitly, so $\{x_{m,n}\}$ provides an explicit characterization of $\{p_{m,n}\}$. It can also be shown that the terms $d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n$ in (1.17) are alternating and monotonically decreasing in absolute value. Moreover, the convergence is exponentially fast. Therefore the series $x_{m,n}$ is suitable from a numerical point of view, since the error of each partial sum can be bounded by the

absolute value of its final term and a few terms suffice due to the exponential convergence. All details of this approach have been worked out in [3].

This concludes the treatment of the symmetric shortest queue problem. We exploited the feature that $p_{m,n}$ behaves asymptotically as a product $\alpha^m\beta^n$ to develop a technique to determine $p_{m,n}$ efficiently. We now give two queueing models for which $p_{m,n}$ has a more complex asymptotic behaviour involving factors $m^{-1/2}$ or $n^{-1/2}$. This suggests that the compensation approach does not work for these problems.

The first queueing model is characterized as follows. Consider a system with two identical parallel servers. The service times are exponentially distributed with mean μ^{-1}. Customers arrive according to a Poisson stream with rate 1. On arrival a customer generates two jobs served independently by the two servers. This model has been studied by Flatto and Hahn [24] (actually, they analyse the model with nonidentical servers). By using a uniformization technique they determine the generating function of the stationary queue length distribution $\{p_{m,n}\}$. From the generating function they are able to show that (see theorem 7.2 in [24])

$$p_{m,n} \sim \frac{K}{m^{1/2}\mu^m}, \quad (m \to \infty, \text{ fixed } n \geq 0),$$

for some constant K. This suggests that the compensation does not work for this problem. Moreover, the analogue of the quadratic equation (1.10) is

$$1 + \alpha^2\beta + \alpha\beta^2 - \alpha\beta(1 + 2\mu) = 0.$$

The curve in the $\alpha\beta$-plane with this equation generates the sequences $\{\alpha_i\}$ and $\{\beta_i\}$. Since this curve does not pass through the origin, these sequences cannot converge to zero. Hence, application of the compensation approach leads, most likely, to a divergent series in case infinitely many terms would be required.

The model above with general service time distributions has been studied by Klein [45]. He considers the workload process and shows that the functional equation for the Laplace-Stieltjes transform of the stationary distribution of this process can be reduced to a Fredholm integral equation.

The second model is the symmetric coupled processor. This model is characterized as follows. Consider a system with two identical parallel servers. At each queue jobs arrive according to a Poisson stream with rate ρ. An arriving job generates an exponentially distributed workload with unit mean. If both servers are busy, the service rate of each server is 1. If one of the servers is idle, the service rate of the busy one is 2. This model has been studied by Konheim, Meilijson and Melkman [47] and more general versions by Fayolle and Iasnogorodski [20] and Cohen and Boxma [14]. For the stationary queue length probabilities $p_{m,0}$ it has been shown that (see e.g. De Waard [58])

$$p_{m,0} \sim \frac{K\rho^m}{m^{1/2}}, \quad (m \to \infty),$$

for some constant K. This suggests that the compensation does not work for this problem. Moreover, the analogue of the quadratic equation (1.10) is

$$\beta\rho + \alpha\rho + \alpha^2\beta + \alpha\beta^2 - \alpha\beta 2(\rho + 1) = 0. \tag{1.19}$$

The curve with this equation passes through the origin. However, it does not enter the positive quadrant at this point. The part of this curve lying in the positive quadrant, is depicted in figure 1.6.

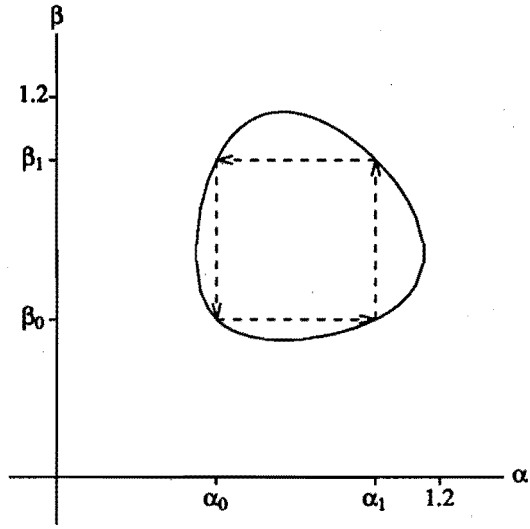


Figure 1.6.
The curve: $\beta\rho + \alpha\rho + \alpha^2\beta + \alpha\beta^2 - \alpha\beta 2(\rho + 1) = 0$ *in the positive quadrant for the case* $\rho = 0.5$. *For each initial pair* α_0, β_0 *the sequences* $\{\alpha_i\}$ *and* $\{\beta_i\}$ *generated by this curve are cycling. This is illustrated for* $\alpha_0 = \beta_0 = \rho$.

By using $\alpha_i\alpha_{i+1} = \beta_i\beta_{i+1} = \rho$ for all i, it follows that for each initial pair α_0, β_0 the curve with equation (1.19) generates the cycle

$$\alpha_0 \to \beta_0 \to \alpha_1 = \frac{\rho}{\alpha_0} \to \beta_1 = \frac{\rho}{\beta_0} \to \alpha_2 = \alpha_0 \to \beta_2 = \beta_0 \to \cdots$$

The compensation approach works if the four terms $\alpha_0^m\beta_0^n$, $\alpha_1^m\beta_0^n$, $\alpha_1^m\beta_1^n$ and $\alpha_0^m\beta_1^n$ would suffice. Indeed, the compensation approach constructs a linear combination of these four terms

satisfying all equilibrium equations. This construction only fails in the two cases $\alpha_0 = \beta_0 = \rho$ and $\alpha_0 = \beta_0 = 1$. However, for each initial pair α_0, β_0 at least one of the α_0, β_0, α_1 or β_1 has absolute value larger than or equal to one. So the solutions found by the compensation approach are not useful, since they cannot be normalized.

The compensation approach works for the shortest queue problem, but fails for the two problems mentioned above. Now the question arises: what is the scope of this approach? This is the subject of the following section.

1.2. The compensation approach applied to two-dimensional Markov processes

To investigate the scope of the compensation approach we study in chapter 2 a class of Markov processes on the lattice in the positive quadrant of $\mathbb{R}^2$ and explore under which conditions the approach works. We consider processes for which the transition rates are constant in the interior points and also constant on each of the axes. To simplify the analysis, we assume that the transitions are restricted to neighbouring states. The transition rates are depicted in figure 1.7.

This model can be analyzed by the fairly general approach developed by Fayolle and Iasnogorodski [19,20,40] and Cohen and Boxma [14]. They show that the analysis of the functional equation for the generating function can be reduced to that of a Riemann type boundary value problem. However, this approach does not lead to the explicit determination of the equilibrium probabilities and requires non-trivial algorithms for numerical calculations. It appears that the compensation approach works for a subset of these models only. On the other hand, our results lead to a fairly explicit characterization of the equilibrium probabilities and can be easily exploited for numerical purposes.

For the Markov process in figure 1.7 we obtain the quadratic equation (cf. (1.10))

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

The curve in the $\alpha\beta$-plane with this equation generates the sequences $\{\alpha_i\}$ and $\{\beta_i\}$. To aid convergence of the series of product form solutions obtained by application of the compensation approach, we require that these sequences converge to zero. This requirement directly has consequences for the transition possibilities. By considering the relations for $\alpha_i\alpha_{i+1}$ and $\alpha_i + \alpha_{i+1}$ and the similar ones for $\beta_i\beta_{i+1}$ and $\beta_i + \beta_{i+1}$ it is easy to show that the condition

$$q_{0,1} = q_{1,1} = q_{1,0} = 0$$

is necessary for convergence to zero of α_i and β_i. It appears that this condition is the crucial one to be imposed in order to successfully apply the compensation approach. Other conditions in chapter 2 are either not relevant (but imposed for convenience only) or imposed to ensure

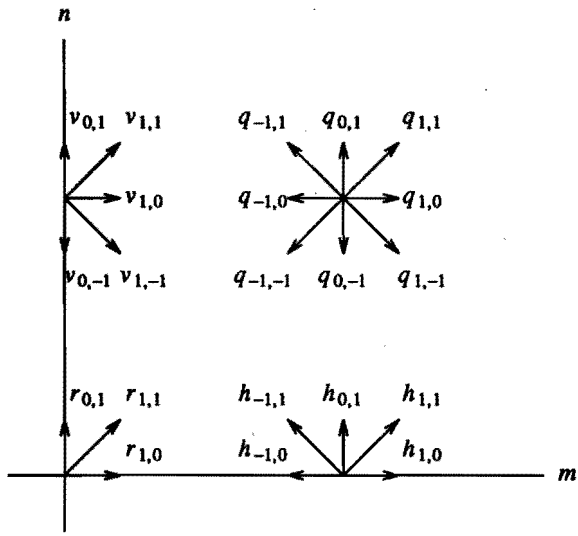


Figure 1.7.
Transition-rate diagram for a Markov process with constant rates and transitions restricted to neighbouring states. $q_{i,j}$ is the transition rate from (m, n) to $(m+i, n+j)$ with $m, n > 0$ and a similar notation is used for the transition rates on each of the axes.

ergodicity. The compensation approach now has the following new features. Depending on the boundary conditions, it is possible that more than one initial product form solution exists, each generating a series of compensation terms. Hence, the probabilities $p_{m,n}$ are represented by a *linear combination* of series of product form solutions. Moreover, it is possible that this series of product form solutions diverges for small m and n.

In chapter 3 we give a complete treatment of the symmetric shortest queue problem as an application of the general theory developed in chapter 2. In this treatment special attention is devoted to extra properties, which are exploited for numerical purposes. In chapter 4 we apply the general theory of chapter 2 to a queueing model for a multiprogramming computer system involving two queues. This model has originally been studied by Hofri [37]. He analyses this model by using generating functions.

1.3. Extensions

The compensation approach can be extended in several directions. Below we comment on several possibilities.

a. The shortest queue problem with threshold jockeying

In chapter 3 we shall also consider the shortest queue problem with a threshold-type jockeying. This means that a job jumps from the longest to the shortest queue as soon as the difference between the lengths of the two queues exceeds some threshold value T. Due to the jockeying, the state space is restricted to the pairs (m, n) satisfying $|n| \leq T$. It appears that the compensation approach also works for this model. In fact, the main term already satisfies the boundary conditions, so no compensation arguments are required.

There are several other techniques to analyse this model. The form of the state space suggests to apply the matrix-geometric approach developed by Neuts [51]. Actually, Gertsbakh [29] studies the threshold jockeying model by using this approach. In [8] the relationship between our approach and the matrix-geometric approach has been investigated. It appears that our approach suggests a state space partitioning which is definitely more useful than the one used by Gertsbakh [29]. In [4] it is shown that the matrix-geometric approach can also be used to analyse the threshold jockeying model with c parallel servers. The results in this paper emphasize the importance of a suitable choice of the state space partitioning. Another approach to the jockeying model with c parallel servers can be found in Grassmann and Zhao [63]. They use the concept of modified lumpability for continuous-time Markov processes. It is finally mentioned that the instantaneous jockeying model ($T = 1$) has been addressed by Haight [34] for $c = 2$ and by Disney and Mitchell [17], Elsayed and Bastani [18], Kao and Lin [42] and Zhao and Grassmann [31] for arbitrary c.

b. The asymmetric shortest queue problem

The Markov process in figure 1.7 is restricted to the first quadrant. In chapter 5 it will be shown that extensions with respect to this form of state space are possible. The subject of chapter 5 is the analysis of the shortest queue problem with *non-identical servers*. This problem is called the asymmetric shortest queue problem and can be modelled as a Markov process on the pairs of integers (m, n) with $m \geq 0$ and n free, which has different properties in the regions $n > 0$ and $n < 0$. It appears that the compensation approach works for this problem and leads to a series of product form solutions for the equilibrium probabilties $p_{m,n}$ in the region $n > 0$ and a similar series for $p_{m,n}$ in the region $n < 0$. The construction of these series, however, is more complicated than for the symmetric case. The interaction between the regions $n > 0$ and $n < 0$ gives rise to a *binary tree structure* of the sequences $\{\alpha_i\}$ and $\{\beta_i\}$ and a related structure of the series for the probabilities $p_{m,n}$. The binary tree structure of the sequences $\{\alpha_i\}$ and $\{\beta_i\}$ is

depicted in figure 1.8. These sequences are generated by using the different quadratic equations in the two regions $n > 0$ and $n < 0$.

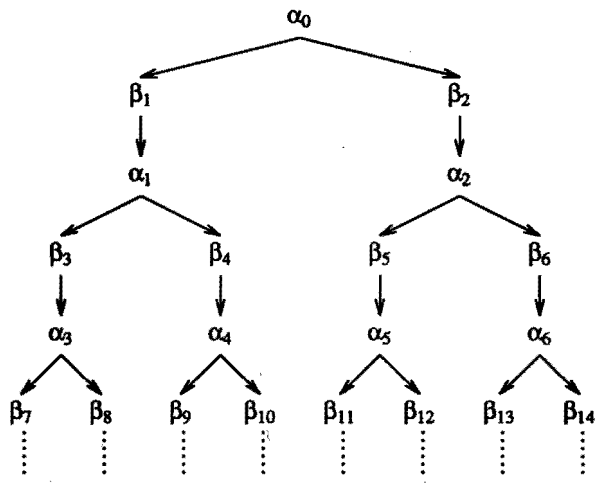


Figure 1.8.
The binary tree structure of $\{\alpha_i\}$ and $\{\beta_i\}$ for the asymmetric shortest queue problem. These sequences are generated by using the different quadratic equations in the two regions $n > 0$ and $n < 0$.

Although the solution structures are more complicated, our final results can easily be exploited for numerical purposes and lead to efficient algorithms for the calculation of the probabilities $p_{m,n}$ or other quantities of interest, with the advantage of tight error bounds.

Fayolle and Iasnogorodski [19, 40] and Cohen and Boxma showed that the analysis of the generating function can be reduced to that of a *simultaneous* Riemann-Hilbert boundary value problem. This type of boundary value problem stems from the coupling between the regions $n > 0$ and $n < 0$ and requires further research. For the asymmetric shortest queue problem no further analytical results are available in the literature.

c. The symmetric shortest delay problem for Erlang servers

To simplify the analysis in chapter 2 we considered Markov processes with transitions restricted to neighbouring states. This restriction does not seem to be essential. However, the boundary conditions for processes with more transition possibilities become definitely more complicated and therefore a general treatment of this type of processes may rise severe complications.

For some special cases the approach is tractable. The first case is the symmetric shortest delay problem with Erlang servers. This problem is characterized as follows.

Consider a system with two identical parallel servers. The service times are Erlang-l distributed with mean l. Jobs arrive according to a Poisson stream with rate 2λ. To ensure that the system can handle the offered load, we assume that $\lambda l < 1$. An arriving job can be thought of as a batch of l identical subjobs, where each subjob requires an exponentially distributed service time with unit mean. Arriving jobs join the queue with the smallest number of subjobs, where ties are broken with probability 1/2. This routing policy is called *shortest delay routing*.

The problem can be modelled as a Markov process on the pairs of integers (m, n) where m is the number of subjobs in the shortest queue and n is the difference between the number of subjobs in the two queues. For this model formulation transitions are not restricted to neighbouring states, as, for instance, can be seen for state (m, n) with $n > 0$ for which a transition is possible to state $(m+l, n-l)$ with rate 2λ (since an arriving job generates a batch of l subjobs). Moreover, the process is not skipfree to the south, which is a basic assumption for the models studied in the book of Cohen and Boxma [14]. It appears that application of the compensation approach leads to a series of product form solutions for the equilibrium probabilities $p_{m,n}$. In fact, the probabilities $p_{m,n}$ can be expressed as a linear combination of series of product form solutions, each with the structure of an l-fold tree. A paper on the detailed analysis of this problem is forthcoming. Some of the features of the approach will be sketched in chapter 6. To our knowledge, no further analytical results for the shortest delay problem are available in the literature.

The next comment is devoted to a second case for which the compensation approach is tractable.

d. The $M|E_r|c$ queue

The $M|E_r|c$ queue can be formulated as a Markov process on the states $(n_0, n_1, ..., n_c)$ where n_0 is the number of waiting jobs and n_i is the number of remaining service phases for server i, $i = 1, ..., c$. For this model formulation transitions are not restricted to neighbouring states, as, for instance, can be seen for state $(n_0, 1, n_2, ..., n_c)$ with $n_0 > 0$ from which a transition to $(n_0-1, r, n_2, ..., n_c)$ is possible due to a service completion of server 1. Moreover, the form of the state space is special. It is bounded in each direction, except in the n_0-direction.

The $M|E_r|c$ queue has been extensively studied in the literature. We mention the work of Mayhugh and McCormick [49] and Heffer [36]. They use generating functions to analyse this problem. Their analysis, however, does not lead to an explicit determination of the equilibrium probabilities. Shapiro [57] studied the $M|E_2|c$ queue for which a simpler formulation of the state space is possible. His analysis has some affinity with our approach.

Our approach first tries to characterize the set of solutions of the form $\alpha^{n_0}\beta_1^{n_1}\cdots\beta_c^{n_c}$ satisfying the equilibrium equations in the interior points, that is, the points with $n_0 > 0$. It appears that this set consists of *finitely* many solutions. However, the set is sufficiently rich, since it is possible to construct a (non-trivial) linear combination of the product form solutions in this set also satisfying the boundary conditions. Similar to the solution of the model under point **a**, this construction is not of a compensation-type. A detailed description of the results can be found in [59]. The analysis can be extended to the $E_k|E_r|c$ queue. A paper on the analysis of the $E_k|E_r|c$ queue is forthcoming.

In the next section we briefly outline an interesting analogue of the compensation approach in the field of electrostatics. This analogue was communicated to us by Prof. P. J. Schweitzer.

1.4. Method of images

Many problems in electrostatics concern the determination of the potential in an arbitrary point P in a region involving boundary surfaces on which the potential or surface charge density is specified. A special approach to these problems is the method of images (see e.g. Maxwell [48] and Jackson [41]). The method of images deals with the problem of a number of point charges in the presence of boundary surfaces, such as, for example conductors held at fixed potentials. Usually, the sum of the potentials of the point charges does not satisfy the boundary conditions. Under favourable conditions it is possible to place a number of additional point charges *outside* the region of interest, such that the sum of the potentials of the point charges inside and outside the region satisfies the boundary conditions. The charges placed outside the region are called the *image charges* and the replacement of the original problem with boundaries by an enlarged region with image charges and no boundaries is called *the method of images*. The image charges must be external to the region of interest, since their potentials must be solutions to the Laplace equation inside the region. The particular solution to the Poisson equation inside the region is provided by the sum of the potentials of the charges inside the region.

Figure 1.9 shows a simple example where a point charge is located in front of an infinite plane conductor which is held at fixed potential $\Phi = 0$. It is clear that this problem is equivalent to the problem of the point charge together with an equal but opposite charge which is located at the mirror image point on the other side of the plane. Let P be any point in the space at the right side of the infinite plane conductor, whose distance from the charges q and $-q$ is r_1 and r_2 respectively. Then the value of the potential at P is given by

$$\Phi = \frac{q}{r_1} - \frac{q}{r_2}.$$

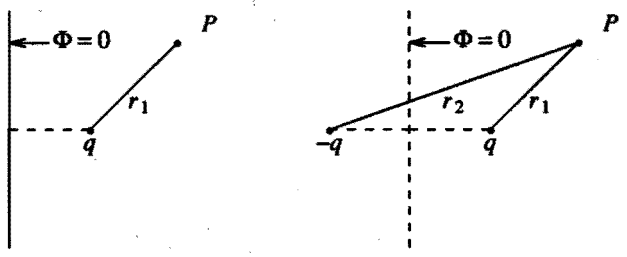


Figure 1.9.
Solution by the method of images. The original potential problem on the left is equivalent with the image problem on the right.

The next example clearly illustrates the analogue with the compensation approach demonstrated in section 1.1. This example is worked out in more detail in the appendix to chapter XI in Maxwell's book [48].

Consider two non intersecting conducting spheres, whose centers are A and B, their radii a and b and their potentials Φ_a and 0 respectively. Suppose that their distance of centers is c (see figure 1.10). Below it is shown that the potential Φ at any point P can be found by producing an infinite sequence of image charges.

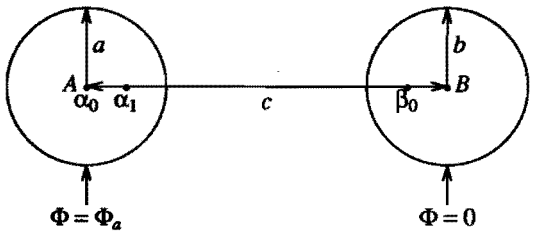


Figure 1.10.
Problem of two conducting spheres A and B held at fixed potentials Φ_a and 0 respectively.

If the spheres did not influence each other ($c = \infty$), then the potential Φ is that of the image charge $\alpha_0 = a\Phi_a$ located at A. However, since c is finite, that potential does not vanish on the sphere B. To compensate for that error we place inside the sphere B a new image charge β_0 at

distance c_0 from B on the ray AB, and try to choose β_0 and c_0 such that the sum of the potentials of the charges α_0 and β_0 vanishes on the sphere B. Indeed, the sum of these potentials vanishes by taking

$$\beta_0 = -\frac{b}{c}\alpha_0 , \quad c_0 = \frac{b^2}{c} .$$

We now added a charge β_0 inside the sphere B to compensate for the error of the potential of charge α_0 on the sphere B. At the same time, the charge β_0 alters the potential on the sphere A. To keep that potential unaltered we place inside the sphere A an image charge α_1 at distance d_1 from A on the ray AB, and try to choose α_1 and d_1 such that the potential of α_1 and β_0 vanishes on the sphere A (so the sum of the potentials of α_0, α_1 and β_0 equals Φ_a on A). That leads to

$$\alpha_1 = -\frac{a}{c-c_0}\beta_0 , \quad d_1 = \frac{a^2}{c-c_0} .$$

We now compensated for the error on the sphere A, but in doing so, we introduced a new error on the sphere B, since the potential of the new image charge α_1 does not vanish on the sphere B. It is clear that we can continue by adding on image charges inside the spheres A and B so as to alternately satisfy the boundary conditions on these spheres. This results in an infinite sequence of image charges. Let α_i and d_i be the charge and distance from A of the ith image charge inside the sphere A on the ray AB, and let β_i and c_i be the charge and distance from B of the ith image charge inside the sphere B on the ray AB. Then we obtain for all $i \geq 0$ the following recursion relations for α_i, β_i, c_i and d_i.

$$\beta_i = -\frac{b}{c-d_i}\alpha_i , \qquad c_i = \frac{b^2}{c-d_i} ,$$

$$\alpha_{i+1} = -\frac{a}{c-c_i}\beta_i , \quad d_{i+1} = \frac{a^2}{c-c_i} ,$$

where initially

$$\alpha_0 = a\Phi_a , \qquad d_0 = 0 .$$

These recursion relations can easily be solved explicitly. Once the image charges and their distances are known, the value of the potential at any point P in the space outside the two spheres is given by the sum of the potentials of the image charges.

The analogue with the approach in section 1.1 will be clear: in the example above point charges are subsequently added so as to alternately satisfy the boundary conditions on the two spheres, whereas in section 1.1 product form solutions are subsequently added so as to alternately satisfy the boundary conditions on the two axes.

1.5. Summary of the subsequent chapters

In chapter 2 we consider a fairly general class of two-dimensional Markov processes. The object in that chapter is to investigate under what conditions these processes have a solution in the form of a series of products of powers which can be found by a compensation approach. In chapter 3 the symmetric shortest queue problem is treated as an application of the theory in chapter 2. For this problem some special properties are worked out in detail and used for numerical purposes. In chapter 4 the general theory is applied to a queueing model for multiprogramming queues. In chapter 5 the compensation approach is further extended to the asymmetric shortest queue problem. Chapter 6 is devoted to conclusions and comments. In particular, the approach will be sketched for the symmetric shortest delay problem with Erlang servers and the $M \mid E_r \mid c$ queue.

Chapter 2

The compensation approach applied to two-dimensional Markov processes

In section 1.1 we have seen for the symmetric shortest queue problem how the feature that the equilibrium probabilities $p_{m,n}$ behave asymptotically as a product of powers can be exploited to develop an approach to find the probabilities $p_{m,n}$ explicitly and efficiently. We further mentioned two other queueing problems, that is, the coupled processor problem and the problem of two $M \mid M \mid 1$ queues with coupled arrivals, for which the probabilities $p_{m,n}$ have a more complicated asymptotic behaviour involving extra factors $m^{-1/2}$ or $n^{-1/2}$. This suggests that the approach does not work for these problems. Now the question arises: what is the scope of the compensation approach? In this chapter first attempts are made to answer this question.

To investigate the scope of the compensation approach we apply this approach to a class of Markov processes on the lattice in the positive quadrant of $\mathbb{R}^2$ and investigate under which conditions the approach works. We consider processes for which the transition rates are constant in the interior points and also constant on each of the axes. To simplify the analysis, we assume that the transitions are restricted to neighbouring states. The class of processes is sufficiently rich in the sense that all problems mentioned in the previous paragraph can be formulated as Markov processes of this type. The class of models fits into the general framework developed by Fayolle and Iasnogorodski [19,40] and Cohen and Boxma [14]. They show that the analysis of the functional equation for the generating function can be reduced to that of a Riemann type boundary value problem. Moreover, with some minor modifications the approach also proceeds for the time-dependent case. However, this approach does not lead to an explicit determination of the probabilities and requires non-trivial algorithms for numerical calculations. In this chapter it will be investigated under what conditions the compensation approach works. The essence of the approach is to characterize the set of product form solutions satisfying the equations in the interior points and then to use the solutions in this set to construct a linear combination of product form solutions which also satisfies the boundary conditions. This construction is based on a compensation idea: after introducing the first term, terms are added so as to alternately compensate for errors on the two boundaries. It is pointed out that the compensation approach first tries to satisfy the conditions in the interior and then tries to satisfy the boundary conditions, whereas generating function approaches combine these conditions into a functional equation for the generating function. The compensation approach leads to formal, possibly divergent solutions of the equilibrium equations. Therefore we shall

explore under which conditions the approach leads to convergent solutions. The essential condition appears to be that transitions from the interior points to the north, north-east and east are not allowed. Hence, the compensation approach works for a subclass of models only. On the other hand, our results lead to a fairly explicit characterization of the equilibrium probabilities and can be easily exploited for numerical purposes, due to the algorithmic nature of the approach. A new feature of the approach is that, depending on the boundary conditions, possibly more than one initial product form solution exists, each generating a series of compensation terms. Hence, the equilibrium probabilities can be expressed as a linear combination of series of product form solutions. Furthermore, it is possible that this series diverges near the origin of the state space.

The organization in this chapter is as follows. In the next section we formulate the model and the equilibrium equations. In section 2.2 the compensation method is outlined and the resulting formal solutions $x_{m,n}(\alpha_0, \beta_0)$ are defined. Section 2.3 introduces the convergence requirements and analyzes its consequences with respect to the transition structure in the interior of the state space. The next three sections are devoted to the derivation and interpretation of conditions for the existence of feasible initial pairs α_0, β_0. In section 2.7 it is shown that the formal solutions $x_{m,n}(\alpha_0, \beta_0)$ simplify for feasible pairs α_0, β_0. It is investigated in section 2.8 whether the construction of these solutions can fail. In the next two sections the absolute convergence of these solutions is treated. In section 2.11 we prove that the solutions $x_{m,n}(\alpha_0, \beta_0)$, with feasible α_0 and β_0, are linearly independent. In section 2.12 we prove our main result, stating that on a subset of the state space the equilibrium probabilities can be expressed as a linear combination of the solutions $x_{m,n}(\alpha_0, \beta_0)$ with feasible α_0 and β_0. In section 2.13 we comment on a condition, which arose out of the analysis in section 2.8. Section 2.14 treats some pathological cases, which are initially excluded in section 2.1. The final section is devoted to conclusions.

2.1. Model and equilibrium equations

We shall consider a Markov process on the pairs (m, n) of nonnegative integers, which is characterized by the property that transitions are restricted to neighbouring states and that the transition rates are constant on the set of all pairs (m, n) of positive integers and also constant on each of the axes. The transition rates are depicted in figure 2.1. Let $\{p_{m,n}\}$ be the equilibrium distribution, which we suppose to exist. Furthermore, we assume that the Markov process is irreducible. The following assumption is made to initially exclude some pathological cases. In section 2.17 we comment on these cases.

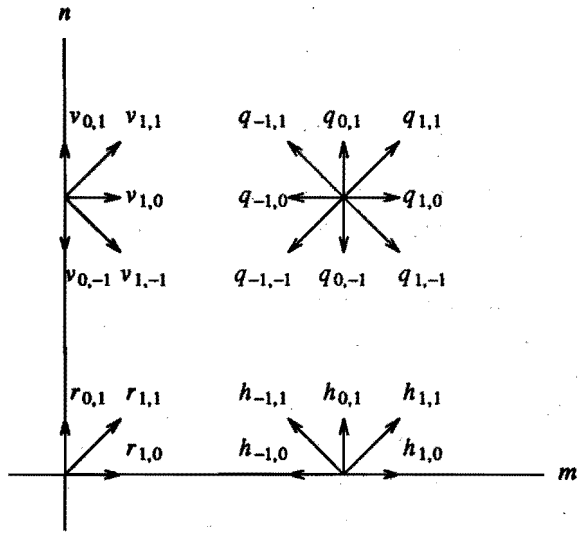


Figure 2.1.
Transition-rate diagram for a Markov process with constant rates and transitions restricted to neighbouring states. $q_{i,j}$ is the transition rate from (m, n) to $(m+i, n+j)$ with $m, n > 0$ and a similar notation is used for the transition rates on each of the axes.

Assumption 2.1.

(i)	$q_{1,1} + q_{1,0} + q_{1,-1} > 0$	*(there is a rate component to the east);*
(ii)	$q_{-1,1} + q_{-1,0} + q_{-1,-1} > 0$	*(there is a rate component to the west);*
(iii)	$q_{-1,1} + q_{0,1} + q_{1,1} > 0$	*(there is a rate component to the north);*
(iv)	$q_{-1,-1} + q_{0,-1} + q_{1,-1} > 0$	*(there is a rate component to the south);*
(v)	$\nu_{1,1} + \nu_{1,0} + \nu_{1,-1} > 0$	*(reflecting n-axis);*
(vi)	$h_{-1,1} + h_{0,1} + h_{1,1} > 0$	*(reflecting m-axis).*

The equilibrium equations for $\{p_{m,n}\}$ can be found by equating for each state the rate into and the rate out of that state. These equations are formulated below. The equations in (1, 1), (0, 1), (1, 0) and (0, 0) are left out, since they will appear to be of minor importance to the

analysis.

$$p_{m,n}q = p_{m+1,n-1}q_{-1,1} + p_{m,n-1}q_{0,1} + p_{m-1,n-1}q_{1,1} + p_{m-1,n}q_{1,0} + p_{m-1,n+1}q_{1,-1}$$
$$+ p_{m,n+1}q_{0,-1} + p_{m+1,n+1}q_{-1,-1} + p_{m+1,n}q_{-1,0}, \qquad m>1, n>1 \quad (2.1)$$

$$p_{1,n}q = p_{2,n-1}q_{-1,1} + p_{1,n-1}q_{0,1} + p_{0,n-1}v_{1,1} + p_{0,n}v_{1,0} + p_{0,n+1}v_{1,-1}$$
$$+ p_{1,n+1}q_{0,-1} + p_{2,n+1}q_{-1,-1} + p_{2,n}q_{-1,0}, \qquad n>1 \quad (2.2)$$

$$p_{0,n}v = p_{1,n-1}q_{-1,1} + p_{0,n-1}v_{0,1} + p_{0,n+1}v_{0,-1} + p_{1,n+1}q_{-1,-1} + p_{1,n}q_{-1,0}, \qquad n>1 \quad (2.3)$$

$$p_{m,1}q = p_{m+1,0}h_{-1,1} + p_{m,0}h_{0,1} + p_{m-1,0}h_{1,1} + p_{m-1,1}q_{1,0} + p_{m-1,2}q_{1,-1}$$
$$+ p_{m,2}q_{0,-1} + p_{m+1,2}q_{-1,-1} + p_{m+1,1}q_{-1,0}, \qquad m>1 \quad (2.4)$$

$$p_{m,0}h = p_{m-1,0}h_{1,0} + p_{m-1,1}q_{1,-1} + p_{m,1}q_{0,-1} + p_{m+1,1}q_{-1,-1} + p_{m+1,0}h_{-1,0}, \qquad m>1 \quad (2.5)$$

where

$$q = q_{-1,1} + q_{0,1} + q_{1,1} + q_{1,0} + q_{1,-1} + q_{0,-1} + q_{-1,-1} + q_{-1,0}, \quad (2.6)$$

$$v = v_{0,1} + v_{1,1} + v_{1,0} + v_{1,-1} + v_{0,-1}, \quad (2.7)$$

$$h = h_{-1,1} + h_{0,1} + h_{1,1} + h_{1,0} + h_{-1,0}. \quad (2.8)$$

2.2. The compensation approach

In this section we develop the compensation approach. This approach constructs a *formal solution* to the equilibrium equations (2.1)-(2.5) by using linear combinations of products $\alpha^m\beta^n$ satisfying equation (2.1) in the interior of the state space. Inserting $\alpha^m\beta^n$ into (2.1) and then dividing both sides of that equation by the common factor $\alpha^{m-1}\beta^{n-1}$ leads to the following characterization (cf. lemma 1.1 in section 1.1).

Lemma 2.2.

The product $\alpha^m\beta^n$ *is a solution of equation (2.1) if and only if* α *and* β *satisfy*

$$\alpha\beta q = \alpha^2 q_{-1,1} + \alpha q_{0,1} + q_{1,1} + \beta q_{1,0} + \beta^2 q_{1,-1}$$
$$+ \alpha\beta^2 q_{0,-1} + \alpha^2\beta^2 q_{-1,-1} + \alpha^2\beta q_{-1,0}. \quad (2.9)$$

Any linear combination of products $\alpha^m\beta^n$ with α, β satisfying the quadratic equation (2.9), is a solution of (2.1). We now have to find a linear combination satisfying the boundary

conditions (2.2)-(2.5). Consider an arbitrary product $\alpha_0^m \beta_0^n$ with complex α_0, β_0 satisfying (2.9) and suppose that $\alpha_0^m \beta_0^n$ violates the vertical boundary conditions (2.2)-(2.3). The idea to satisfy these conditions is:

Try to find α, β, c_1 with α, β satisfying (2.9) such that
$\alpha_0^m \beta_0^n + c_1 \alpha^m \beta^n$ satisfies the boundary conditions (2.2)-(2.3).

Inserting this linear combination into (2.2)-(2.3) yields two conditions of the form:

$$A(\alpha_0, \beta_0)\beta_0^{n-1} + c_1 A(\alpha, \beta)\beta^{n-1} = 0\,, \quad n > 1\,, \tag{2.10}$$

$$B(\alpha_0, \beta_0)\beta_0^{n-1} + c_1 B(\alpha, \beta)\beta^{n-1} = 0\,, \quad n > 1\,, \tag{2.11}$$

where at least one of the $A(\alpha_0, \beta_0)$ and $B(\alpha_0, \beta_0)$ is nonzero. To satisfy (2.10) and (2.11) for *all* $n > 1$ we are forced to take

$$\beta = \beta_0$$

and thus

$$\alpha = \alpha_1\,,$$

where α_1 is the other root of the quadratic equation (2.9) with $\beta = \beta_0$. Dividing (2.10) and (2.11) by the common factor β_0^{n-1} leads to two linear equations for c_1, which have, in general, no solution. Therefore, we introduce an extra coeffient by considering

$$\begin{aligned} &\alpha_0^m \beta_0^n + c_1 \alpha_1^m \beta_0^n \quad \text{for} \quad m > 0\,,\, n > 0\,, \\ &e_0 \beta_0^n \qquad\qquad\qquad\quad \text{for} \quad m = 0\,,\, n > 0\,. \end{aligned}$$

Inserting this form into the boundary conditions (2.2)-(2.3) and then dividing by the common factor β_0^{n-1} leads to two linear equations for c_1 and e_0, which can readily be solved using Cramer's rule. The resulting expressions for c_1 and e_0 can be simplified by using (2.9). This procedure is generalized in the following lemma (cf. lemmas 1.2 and 1.3). Part (ii) formulates the analogue for the horizontal boundary.

Lemma 2.3.

(i) *Let x_1 and x_2 be the roots of the quadratic equation (2.9) for fixed β and let*

$$z_{m,n} = \begin{cases} x_1^m \beta^n + c x_2^m \beta^n & \text{for} \quad m > 0,\, n > 0, \\ e\beta^n & \text{for} \quad m = 0,\, n > 0. \end{cases}$$

Then $z_{m,n}$ satisfies (2.1), (2.2) and (2.3) if c and e are given by

$$c = -\frac{\dfrac{\beta^2 v_{1,-1} + \beta v_{1,0} + v_{1,1}}{x_2} + v_{0,1} + \beta^2 v_{0,-1} - \beta v}{\dfrac{\beta^2 v_{1,-1} + \beta v_{1,0} + v_{1,1}}{x_1} + v_{0,1} + \beta^2 v_{0,-1} - \beta v}\,, \tag{2.12}$$

$$e = -\frac{\left[\beta^2 q_{1,-1} + \beta q_{1,0} + q_{1,1}\right]\left[\dfrac{1}{x_2} - \dfrac{1}{x_1}\right]}{\dfrac{\beta^2 v_{1,-1} + \beta v_{1,0} + v_{1,1}}{x_1} + v_{0,1} + \beta^2 v_{0,-1} - \beta v}. \tag{2.13}$$

(ii) *Let y_1 and y_2 be the roots of the quadratic equation (2.9) for fixed α and let*

$$w_{m,n} = \begin{cases} \alpha^m y_1^n + d\alpha^m y_2^n & \text{for } m > 0,\ n > 0, \\ f\alpha^m & \text{for } m > 0,\ n = 0. \end{cases}$$

Then $w_{m,n}$ satisfies (2.1), (2.4) and (2.5) if d and f are given by

$$d = -\frac{\dfrac{\alpha^2 h_{-1,1} + \alpha h_{0,1} + h_{1,1}}{y_2} + h_{1,0} + \alpha^2 h_{-1,0} - \alpha h}{\dfrac{\alpha^2 h_{-1,1} + \alpha h_{0,1} + h_{1,1}}{y_1} + h_{1,0} + \alpha^2 h_{-1,0} - \alpha h}, \tag{2.14}$$

$$f = -\frac{\left[\alpha^2 q_{-1,1} + \alpha q_{0,1} + q_{1,1}\right]\left[\dfrac{1}{y_2} - \dfrac{1}{y_1}\right]}{\dfrac{\alpha^2 h_{-1,1} + \alpha h_{0,1} + h_{1,1}}{y_1} + h_{1,0} + \alpha^2 h_{-1,0} - \alpha h}. \tag{2.15}$$

We added $c_1\alpha_1^m\beta_0^n$ to compensate for the error of $\alpha_0^m\beta_0^n$ on the vertical boundary and by doing so introduced a new error on the horizontal boundary, since $c_1\alpha_1^m\beta_0^n$ violates these boundary conditions. To compensate for this error we add $c_1 d_1 \alpha_1^m \beta_1^n$ where β_1 is the other root of (2.9) with $\alpha = \alpha_1$. The coefficient d_1 follows from lemma 2.3(ii). However, this term violates the vertical boundary conditions, so we have to add again a term, and so on. Thus the compensation of $\alpha_0^m\beta_0^n$ on the vertical boundary generates an infinite sequence of compensation terms. An analogous sequence is generated by starting the compensation of $\alpha_0^m\beta_0^n$ on the horizontal boundary. This results in the sum of terms depicted in figure 2.2.

$$\underbrace{\cdots + d_{-1}c_{-1}\alpha_{-1}^m\beta_{-1}^n + d_{-1}c_0\alpha_0^m\beta_{-1}^n}_{V} + \underbrace{d_0c_0\alpha_0^m\beta_0^n + d_0c_1\alpha_1^m\beta_0^n}_{V} + d_1c_1\alpha_1^m\beta_1^n + \cdots$$

$$\overbrace{\qquad\qquad}^{H} \qquad \overbrace{\qquad\qquad}^{H}$$

Figure 2.2.

The final sum of compensation terms. By definition $c_0 = d_0 = 1$. Sums of two terms with the same β-factor satisfy the vertical boundary conditions (V) and sums of two terms with the same α-factor satisfy the horizontal boundary conditions (H).

Each term in the sum in figure 2.2 satisfies (2.1), each sum of two terms with the same β-factor satisfies the vertical boundary conditions (2.2)-(2.3) and each sum of two terms with the same α-factor satisfies the horizontal boundary conditions (2.4)-(2.5). Since the equilibrium equations are linear, we can conclude that the sum in figure 2.2 formally satisfies the equations (2.1)-(2.5). Let us define $x_{m,n}(\alpha_0, \beta_0)$ as the infinite sum of compensation terms. For all $m > 0$, $n > 0$ set

$$x_{m,n}(\alpha_0, \beta_0) = \sum_{i=-\infty}^{\infty} d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n \qquad \text{(pairs with same β-factor)}, \quad (2.16)$$

$$= \sum_{i=-\infty}^{\infty} c_{i+1}(d_i\beta_i^n + d_{i+1}\beta_{i+1}^n)\alpha_{i+1}^m \qquad \text{(pairs with same α-factor)}. \quad (2.17)$$

The pairs in (2.18) and (2.17) reflect the compensation on the vertical, respectively horizontal boundary. The compensation on these boundaries requires the introduction of new coefficients for the terms in $x_{0,n}(\alpha_0, \beta_0)$ and $x_{m,0}(\alpha_0, \beta_0)$. For all $m = 0$, $n > 0$ set

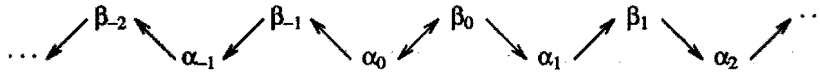
$$x_{0,n}(\alpha_0, \beta_0) = \sum_{i=-\infty}^{\infty} d_i e_i \beta_i^n , \qquad (2.18)$$

and for all $m > 0$, $n = 0$

$$x_{m,0}(\alpha_0, \beta_0) = \sum_{i=-\infty}^{\infty} c_{i+1} f_{i+1} \alpha_{i+1}^m . \qquad (2.19)$$

Note that we do not define $x_{0,0}(\alpha_0, \beta_0)$, since in the first place the equations in (1, 1), (0, 1), (1, 0) and (0, 0) are not considered and in the second place it is not clear whether $x_{0,0}(\alpha_0, \beta_0)$ should be defined by the series (2.18) with $n = 0$ or by the series (2.19) with $m = 0$. Below we formulate the recursion relations for α_i, β_i, c_i, d_i, e_i and f_i.

For the initial roots α_0 and β_0 of the quadratic equation (2.9) the sequence



is generated such that for all i the numbers α_i and α_{i+1} are the roots of (2.9) with fixed $\beta = \beta_i$ and β_i and β_{i+1} are the roots of (2.9) with fixed $\alpha = \alpha_{i+1}$.

$\{c_i\}$ and $\{e_i\}$ are generated such that for all i the terms $(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n$ and $e_i\beta_i^n$ satisfy the vertical boundary conditions (2.2)-(2.3). Initially set

$$c_0 = 1 .$$

Application of lemma 2.3(i) yields that c_{i+1} and e_i for $i \geq 0$ can be obtained from c_i by

$$c_{i+1} = -\frac{\dfrac{\beta_i^2 v_{1,-1} + \beta_i v_{1,0} + v_{1,1}}{\alpha_{i+1}} + v_{0,1} + \beta_i^2 v_{0,-1} - \beta_i v}{\dfrac{\beta_i^2 v_{1,-1} + \beta_i v_{1,0} + v_{1,1}}{\alpha_i} + v_{0,1} + \beta_i^2 v_{0,-1} - \beta_i v} c_i , \quad i \geq 0 , \qquad (2.20)$$

$$e_i = -\frac{\left(\beta_i^2 q_{1,-1} + \beta_i q_{1,0} + q_{1,1}\right)\left(\dfrac{1}{\alpha_{i+1}} - \dfrac{1}{\alpha_i}\right)}{\dfrac{\beta_i^2 v_{1,-1} + \beta_i v_{1,0} + v_{1,1}}{\alpha_i} + v_{0,1} + \beta_i^2 v_{0,-1} - \beta_i v} c_i, \qquad i \geq 0,$$

and analogously c_i and e_i for $i < 0$ can be obtained from c_{i+1} by

$$c_i = -\frac{\dfrac{\beta_i^2 v_{1,-1} + \beta_i v_{1,0} + v_{1,1}}{\alpha_i} + v_{0,1} + \beta_i^2 v_{0,-1} - \beta_i v}{\dfrac{\beta_i^2 v_{1,-1} + \beta_i v_{1,0} + v_{1,1}}{\alpha_{i+1}} + v_{0,1} + \beta_i^2 v_{0,-1} - \beta_i v} c_{i+1}, \quad i < 0,$$

$$e_i = -\frac{\left(\beta_i^2 q_{1,-1} + \beta_i q_{1,0} + q_{1,1}\right)\left(\dfrac{1}{\alpha_i} - \dfrac{1}{\alpha_{i+1}}\right)}{\dfrac{\beta_i^2 v_{1,-1} + \beta_i v_{1,0} + v_{1,1}}{\alpha_{i+1}} + v_{0,1} + \beta_i^2 v_{0,-1} - \beta_i v} c_{i+1}, \quad i < 0.$$

$\{d_i\}$ and $\{f_i\}$ are generated such that for all i the terms $(d_i\beta_i^n + d_{i+1}\beta_{i+1}^n)\alpha_{i+1}^m$ and $f_{i+1}\alpha_{i+1}^m$ satisfy the horizontal boundary conditions (2.4)-(2.5). Initially set

$$d_0 = 1.$$

Application of lemma 2.3(ii) yields that d_{i+1} and f_{i+1} for $i \geq 0$ can be obtained from d_i by

$$d_{i+1} = -\frac{\dfrac{\alpha_{i+1}^2 h_{-1,1} + \alpha_{i+1} h_{0,1} + h_{1,1}}{\beta_{i+1}} + h_{1,0} + \alpha_{i+1}^2 h_{-1,0} - \alpha_{i+1} h}{\dfrac{\alpha_{i+1}^2 h_{-1,1} + \alpha_{i+1} h_{0,1} + h_{1,1}}{\beta_i} + h_{1,0} + \alpha_{i+1}^2 h_{-1,0} - \alpha_{i+1} h} d_i, \quad i \geq 0, \tag{2.21}$$

$$f_{i+1} = -\frac{\left(\alpha_{i+1}^2 q_{-1,1} + \alpha_{i+1} q_{0,1} + q_{1,1}\right)\left(\dfrac{1}{\beta_{i+1}} - \dfrac{1}{\beta_i}\right)}{\dfrac{\alpha_{i+1}^2 h_{-1,1} + \alpha_{i+1} h_{0,1} + h_{1,1}}{\beta_i} + h_{1,0} + \alpha_{i+1}^2 h_{-1,0} - \alpha_{i+1} h} d_i, \quad i \geq 0,$$

and analogously d_i and f_{i+1} for $i < 0$ can be obtained from d_{i+1} by

$$d_i = -\frac{\dfrac{\alpha_{i+1}^2 h_{-1,1} + \alpha_{i+1} h_{0,1} + h_{1,1}}{\beta_i} + h_{1,0} + \alpha_{i+1}^2 h_{-1,0} - \alpha_{i+1} h}{\dfrac{\alpha_{i+1}^2 h_{-1,1} + \alpha_{i+1} h_{0,1} + h_{1,1}}{\beta_{i+1}} + h_{1,0} + \alpha_{i+1}^2 h_{-1,0} - \alpha_{i+1} h} d_{i+1}, \quad i < 0, \tag{2.22}$$

$$f_{i+1} = -\frac{\left(\alpha_{i+1}^2 q_{-1,1} + \alpha_{i+1} q_{0,1} + q_{1,1}\right)\left(\dfrac{1}{\beta_i} - \dfrac{1}{\beta_{i+1}}\right)}{\dfrac{\alpha_{i+1}^2 h_{-1,1} + \alpha_{i+1} h_{0,1} + h_{1,1}}{\beta_{i+1}} + h_{1,0} + \alpha_{i+1}^2 h_{-1,0} - \alpha_{i+1} h} d_{i+1}, \; i < 0.$$

This concludes the definition of $x_{m,n}(\alpha_0, \beta_0)$. Each solution $x_{m,n}(\alpha_0, \beta_0)$ has its *own sequence* $\{\alpha_i, \beta_i\}$ depending on the initial values α_0 and β_0, and its associated sequence of coefficients $\{c_i, d_i, e_i, f_i\}$. For any pair α_0, β_0 satisfying equation (2.9) the series $x_{m,n}(\alpha_0, \beta_0)$ formally satisfies the equations (2.1)-(2.5). In the next section it will be investigated for what α_0, β_0 the series $x_{m,n}(\alpha_0, \beta_0)$ converges.

Remark 2.4.

If the rates on the vertical boundary are the truncation of the rates in the interior points, that is, $v_{1j} = q_{1j}$, then $e_i = c_i + c_{i+1}$ for all i and thus the series (2.18) is identical to (2.16) with $m = 0$. An analogous remark holds if $h_{j1} = q_{j1}$.

2.3. Analysis of the sequence of α_i and β_i

Under favourable conditions $x_{m,n}(\alpha_0, \beta_0)$ reduces to a *finite sum*. This happens if c_i or d_i vanishes for some $i \geq 0$ and for some $i \leq 0$, which means that from there no more compensation is needed (all subsequent coefficients vanish; cf. (2.20)-(2.22)). A simple example is that of two independent $M|M|1$ queues, each with workload ρ. For $\alpha_0 = \beta_0 = \rho$ no compensation is needed at all, so $x_{m,n}(\rho, \rho)$ reduces to

$$x_{m,n}(\rho, \rho) = \rho^m \rho^n .$$

Conditions for getting such product form solutions are well-known (see e.g. [9]).

Under unfortunate conditions compensation *fails*. This happens if for some value of i the equation (2.9) with fixed $\beta = \beta_i$ or fixed $\alpha = \alpha_{i+1}$ reduces to a linear equation, so the necessary second root does not exist. If the second root is equal to the first one, then it can be verified that the compensation procedure constructs the null solution. Furthermore, compensation fails if for some value of i the denominator in the definition of the coefficients vanishes (cf. (2.20)-(2.22)).

Let us suppose in this section that for the initial α_0, β_0 in at least one direction infinitely many compensation terms are needed and that compensation is always possible. We want to know under what conditions the infinite sum $x_{m,n}(\alpha_0, \beta_0)$ converges. To aid convergence of $x_{m,n}(\alpha_0, \beta_0)$ for *fixed* m and n we require that α_i and β_i tend to zero as $|i|$ tends to infinity. To aid convergence of the sum of $x_{m,n}(\alpha_0, \beta_0)$ over *all values* m and n (necessary for normalization) we require that $|\alpha_i| < 1$ and $|\beta_i| < 1$ for all i.

Convergence requirements 2.5.

(i) α_i *and* β_i *tend to zero as* $|i|$ *tends to infinity;*

(ii) $|\alpha_i| < 1$ *and* $|\beta_i| < 1$ *for all* i.

Below we investigate the implications of these convergence requirements for the transition possibilities in the interior of the state space. Numerical evidence suggests that the behaviour of α_i and β_i when $q_{1,1} + q_{0,1} + q_{1,0} > 0$ is *essentially different* from the behaviour of α_i and β_i when $q_{1,1} + q_{0,1} + q_{1,0} = 0$. This is illustrated in the figures 2.3 and 2.4.

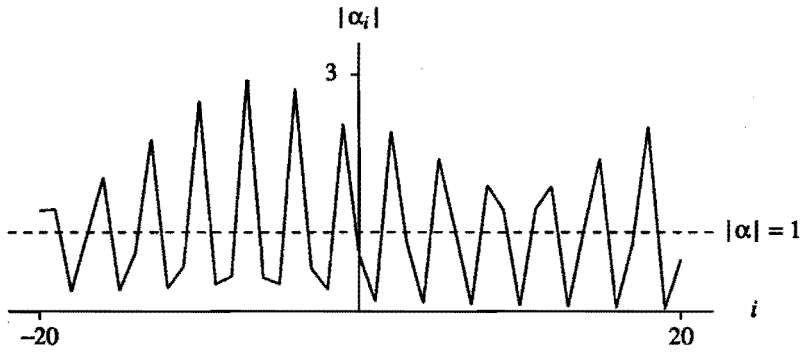


Figure 2.3.
The behaviour of $|\alpha_i|$ *for* $\alpha_0 = (1+\sqrt{-1})/2$ *and rates* $q_{1,-1} = 2$, $q_{-1,-1} = q_{-1,1} = q_{0,-1} = 1$, $q_{-1,0} = 0$ *and* $q_{1,1} = 0$, $q_{0,1} = q_{1,0} = 1/2$.

The behaviour of $|\alpha_i|$ for an example with $q_{1,1} + q_{0,1} + q_{1,0} > 0$ is depicted in figure 2.3. For that example α_i does *not* converge to zero as $|i|$ tends to infinity, and in fact demonstrates an oscillating behaviour. Numerical experiments suggest that this is a typical feature of α_i when $q_{1,1} + q_{0,1} + q_{1,0} > 0$.

The behaviour of $|\alpha_i|$ for an example with $q_{1,1} + q_{0,1} + q_{1,0} = 0$ is depicted in figure 2.4. For that example α_i converges to zero very fast as $|i|$ tends to infinity, with a single trip *outside* the open unit disk. Numerical evidence suggests that this is a typical feature of α_i when $q_{1,1} + q_{0,1} + q_{1,0} = 0$.

Below we try to prove these features. Since α_i and α_{i+1} are the roots of the quadratic equation (2.9) with $\beta = \beta_i$ we have

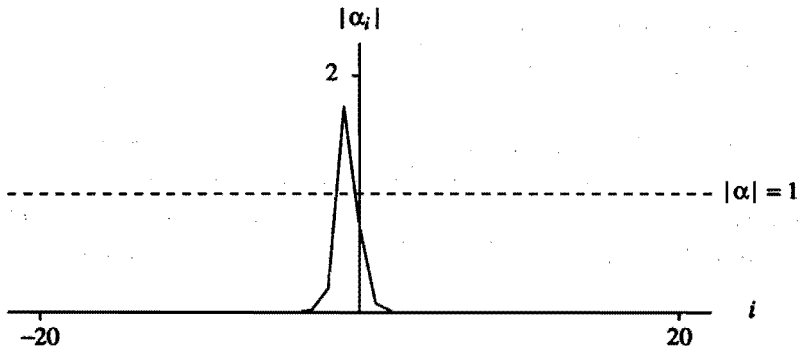


Figure 2.4.
The behaviour of $|\alpha_i|$ *for* $\alpha_0 = (1+\sqrt{-1})/2$ *and rates* $q_{1,-1} = 2$, $q_{-1,-1} = q_{-1,1} = q_{0,-1} = 1$, $q_{-1,0} = 0$ *and* $q_{1,1} = q_{0,1} = q_{1,0} = 0$.

$$\alpha_i \alpha_{i+1} = \frac{\beta_i^2 q_{1,-1} + \beta_i q_{1,0} + q_{1,1}}{\beta_i^2 q_{-1,-1} + \beta_i q_{-1,0} + q_{-1,1}}, \tag{2.23}$$

$$\alpha_i + \alpha_{i+1} = \frac{\beta_i q - q_{0,1} - \beta_i^2 q_{0,-1}}{\beta_i^2 q_{-1,-1} + \beta_i q_{-1,0} + q_{-1,1}}, \tag{2.24}$$

and accordingly β_i and β_{i+1} are the roots of (2.9) with $\alpha = \alpha_{i+1}$ and therefore satisfy

$$\beta_i \beta_{i+1} = \frac{\alpha_{i+1}^2 q_{-1,1} + \alpha_{i+1} q_{0,1} + q_{1,1}}{\alpha_{i+1}^2 q_{-1,-1} + \alpha_{i+1} q_{0,-1} + q_{1,-1}}, \tag{2.25}$$

$$\beta_i + \beta_{i+1} = \frac{\alpha_{i+1} q - q_{1,0} - \alpha_{i+1}^2 q_{-1,0}}{\alpha_{i+1}^2 q_{-1,-1} + \alpha_{i+1} q_{0,-1} + q_{1,-1}}. \tag{2.26}$$

If α_i and β_i converge to zero as $|i|$ tends to infinity, then we deduce from (2.23)-(2.26) that

$$q_{1,1} = q_{0,1} = q_{1,0} = 0.$$

This condition is necessary for convergence to zero of α_i and β_i. The case of two independent $M|M|1$ queues mentioned at the beginning of this section violates this condition and therefore the compensation approach would not work. Fortunately, in this case no compensation is needed at all. In other cases violating this condition, like two $M|M|1$ queues with coupled arrivals (cf.[24,45]) and the coupled processor problem (cf. [14,20,47]), compensation is needed, but it would not work. Indeed, the solutions are essentially more complicated in these cases. We suppose from now on that the condition above is satisfied and moreover, to exclude

pathological cases, that there is a rate component to the south-west (cf. assumption 2.1).

Assumption 2.6.

(i) $q_{1,1} = q_{0,1} = q_{1,0} = 0$ *(there is no rate component to the north, north-east and east);*

(ii) $q_{-1,0} + q_{-1,-1} + q_{0,-1} > 0$ *(there is a rate component to the south-west).*

Note that assumption 2.6(i) and assumptions 2.1(i) and 2.1(iii) imply that

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

We now investigate whether assumption 2.6(i) guarantees convergence for any starting pair of roots α_0, β_0 of equation (2.9) satisfying $|\alpha_0| < 1$ and $|\beta_0| < 1$. By assumption 2.6(i) the equation (2.9) simplifies to

$$\alpha\beta q = \alpha^2 q_{-1,1} + \beta^2 q_{1,-1} + \alpha\beta^2 q_{0,-1} + \alpha^2\beta^2 q_{-1,-1} + \alpha^2\beta q_{-1,0}. \tag{2.27}$$

For each fixed α or fixed β equation (2.27) has two roots, except when (2.27) reduces to a linear equation, but then it is sensible to define ∞ as second root. We state the following lemma for the roots of equation (2.27). To prove the lemma we use *Rouché's theorem*, rather then the explicit formulas for the roots of equation (2.27). Rouché's theorem reads as follows. Let the bounded region D have as its boundary a closed Jordan curve C. Let the function $f(z)$ and $g(z)$ be analytic both in D and on C, and assume that $|f(z)| < |g(z)|$ on C (so automatically $f(z) + g(z) \neq 0$ on C). Then $f(z) + g(z)$ has in D the same number of zeros as $g(z)$, all zeros counted according to their multiplicity.

Lemma 2.7.

For each fixed α satisfying $0 < |\alpha| < 1$, equation (2.27) has exactly one root β with $0 < |\beta| < |\alpha|$ and one root β with $|\beta| > |\alpha|$.
The same holds with α and β interchanged.

Proof.

Dividing equation (2.27) by α^2 and using the new variable $z = \beta/\alpha$, we obtain the following equation for z,

$$z^2(\alpha^2 q_{-1,-1} + \alpha q_{0,-1} + q_{1,-1}) - z(q - \alpha q_{-1,0}) + q_{-1,1} = 0. \tag{2.28}$$

Let

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

$$g(z) = -z(q - \alpha q_{-1,0}) + q_{-1,1}.$$

Recall that q is the sum of the q_{ij}'s (see (2.6)). Then for all $|z| = 1$,

$$|f(z)| \le q_{-1,-1} + q_{0,-1} + q_{1,-1},$$

$$|g(z)| \ge q_{-1,-1} + q_{0,-1} + q_{1,-1},$$

where, by assumption 2.6, at least one of the inequalities is strict. Now the lemma follows by applying Rouché's theorem to $f(z)$ and $g(z)$ above and the unit circle for C. □

Lemma 2.7 leads to the following properties of the sequences $\{\alpha_i\}$ and $\{\beta_0\}$. Note that in the formulation of corollary 2.8 the case $1 > |\beta_0| = |\alpha_0| > 0$ is not possible by lemma 2.7.

Corollary 2.8.

Let α_0 and β_0 be roots of equation (2.27) satisfying $1 > |\alpha_0| > |\beta_0| > 0$.
Then there exists a negative value of i for which $|\alpha_i| \ge 1$ or $|\beta_i| \ge 1$ and

$$1 > |\alpha_{i+1}| > |\beta_{i+1}| > \cdots > |\alpha_0| > |\beta_0| > |\alpha_1| > |\beta_1| > \cdots \downarrow 0.$$

A similar result holds if α_0 and β_0 satisfy $1 > |\beta_0| > |\alpha_0| > 0$.

Proof.

The monotonicity follows directly from lemma 2.7. To prove that there exists a negative value of i for which $|\alpha_i| \ge 1$ or $|\beta_i| \ge 1$, we also need information about the β-roots of equation (2.27) for fixed α with $|\alpha| = 1$.

For fixed α with $|\alpha| = 1$ and $\alpha \ne 1, -1$ it follows by applying Rouché's theorem, similarly as in the proof of lemma 2.7, that equation (2.28) has one root z with $|z| < 1$ and one root z with $|z| > 1$. The same result holds for $\alpha = -1$ if at least one of the rates $q_{-1,0}$ and $q_{0,-1}$ is positive. For fixed $\alpha = 1$ and, if $q_{-1,0} = q_{0,-1} = 0$, also for $\alpha = -1$ equation (2.28) is solved by $z = 1$ and $z = q_{-1,1} / (q_{-1,-1} + q_{0,-1} + q_{1,-1})$, respectively.

Hence, if $q_{-1,1} < q_{-1,-1} + q_{0,-1} + q_{1,-1}$, we can define $z(\alpha)$ as the root of (2.28) for fixed α with $|\alpha| \le 1$, which satisfies $z(\alpha) < 1$. Since $z(\alpha)$ is continuous, the maximum of $|z(\alpha)|$ for $|\alpha| \le 1$ exists and is less than one. So $|\beta_i / \alpha_i| = |z(\alpha_i)| \le \max_{|\alpha| \le 1} |z(\alpha)| < 1$ as long as $|\alpha_i| < 1$. This proves that $|\alpha_i|$ and $|\beta_i|$ decrease exponentially fast to zero as i tends to infinity, and that $|\alpha_i| \ge 1$ or $|\beta_i| \ge 1$ for some negative value of i.

If $q_{-1,1} \ge q_{-1,-1} + q_{0,-1} + q_{1,-1}$, then from assumption 2.6(ii) we obtain the inequality $q_{1,-1} < q_{-1,-1} + q_{-1,0} + q_{-1,1}$. Hence, we can repeat the arguments above by considering the roots of equation (2.27) for fixed β instead of fixed α. □

When the sequence of α_i and β_i is started with roots α_0 and β_0 of (2.27) satisfying $0 < |\alpha_0| < 1$ and $0 < |\beta_0| < 1$, then $|\alpha_0| < |\beta_0|$ or $|\alpha_0| > |\beta_0|$ by lemma 2.7 (equality is not possible). Hence, by corollary 2.8, $|\alpha_i|$ and $|\beta_i|$ decrease to zero in at least one direction. In the opposite direction $|\alpha_i|$ and $|\beta_i|$ increase and eventually $|\alpha_i| \geq 1$ or $|\beta_i| \geq 1$ for some i. Therefore we cannot meet the convergence requirements in that direction, unless in that direction c_i or d_i vanishes for some i before $|\alpha_i| \geq 1$ or $|\beta_i| \geq 1$. After renumbering the terms this amounts to the requirement that the initial product $\alpha_0^m \beta_0^n$ fits the horizontal boundary conditions ($d_{-1} = 0$) if $|\alpha_0| > |\beta_0|$ or otherwise the vertical boundary conditions ($c_1 = 0$). In such a case we have to generate compensation terms in the decreasing direction only. Pairs α_0, β_0 satisfying these requirements will be called *feasible pairs*.

Definition 2.9.

A pair α_0, β_0 will be called feasible if:

(i) *α_0 and β_0 are roots of (2.27) with $0 < |\alpha_0| < 1$ and $0 < |\beta_0| < 1$;*

(ii) $|\alpha_0| > |\beta_0| \Rightarrow d_{-1} = 0$;

(iii) $|\alpha_0| < |\beta_0| \Rightarrow c_1 = 0$.

Conclusion 2.10.

For convergence the following two conditions are crucial:

(i) The Markov process has to satisfy $q_{0,1} = q_{1,1} = q_{1,0} = 0$;

(ii) We have to initialize $x_{m,n}(\alpha_0, \beta_0)$ with a feasible pair α_0, β_0.

We end this section with a theorem stating that α_0, β_0, α_1, β_1, ... can be solved explicitly if $1 > |\alpha_0| > |\beta_0|$. The same result holds for β_0, α_0, β_{-1}, α_{-1}, ... if $1 > |\beta_0| > |\alpha_0|$ (cf. lemma 3 in Kingman [44]).

Theorem 2.11 (Explicit solution of α_i and β_i).

Let α_0 and β_0 be roots of equation (2.31) with $1 > |\alpha_0| > |\beta_0|$.
Then there exist complex numbers a and b, depending on α_0 and β_0, such that for $i \geq 0$

$$\frac{1}{\alpha_i} = A + \sqrt{\gamma}\left(a\lambda^i + \frac{1}{a\lambda^i}\right), \tag{2.29}$$

$$\frac{1}{\beta_i} = B + \sqrt{\delta}\left(b\lambda^i + \frac{1}{b\lambda^i}\right),$$

where

$$A = \frac{qq_{-1,0} + 2q_{0,-1}q_{-1,1}}{q^2 - 4q_{1,-1}q_{-1,1}}, \tag{2.30}$$

$$B = \frac{qq_{0,-1} + 2q_{-1,0}q_{1,-1}}{q^2 - 4q_{1,-1}q_{-1,1}},$$

$$\gamma = \frac{A^2 q_{1,-1}q_{-1,1}}{q^2} + \frac{q_{-1,1}q_{-1,-1}q^2 + q_{-1,1}q_{0,-1}(qq_{-1,0} + q_{-1,1}q_{0,-1})}{q^2(q^2 - 4q_{1,-1}q_{-1,1})}, \tag{2.31}$$

$$\delta = \frac{B^2 q_{1,-1}q_{-1,1}}{q^2} + \frac{q_{1,-1}q_{-1,-1}q^2 + q_{1,-1}q_{-1,0}(qq_{0,-1} + q_{1,-1}q_{-1,0})}{q^2(q^2 - 4q_{1,-1}q_{-1,1})},$$

$$\lambda = \frac{q - \sqrt{q^2 - 4q_{1,-1}q_{-1,1}}}{q + \sqrt{q^2 - 4q_{1,-1}q_{-1,1}}}. \tag{2.32}$$

Proof.

We prove the expressions for $1/\alpha_i$. The expressions for $1/\beta_i$ can be obtained similarly (replace q_{ij} by q_{ji}). By corollary 2.8 the numbers α_i and β_i are nonzero for $i \geq 0$, and equation (2.27) does not reduce to a linear equation for fixed $\beta = \beta_i$ or $\alpha = \alpha_{i+1}$, so the denominator in (2.23)-(2.26) does not vanish for $i \geq 0$. Then, from (2.23) and (2.24) we obtain for $i \geq 0$ that (recall that $q_{1,1} = q_{1,0} = 0$),

$$\frac{1}{\alpha_i} + \frac{1}{\alpha_{i+1}} = \frac{1}{\beta_i}\frac{q}{q_{1,-1}} - \frac{q_{0,-1}}{q_{1,-1}}. \tag{2.33}$$

Adding this relation to the one with i replaced by $i+1$ yields for $i \geq 0$,

$$\frac{1}{\alpha_i} + \frac{2}{\alpha_{i+1}} + \frac{1}{\alpha_{i+2}} = \left[\frac{1}{\beta_i} + \frac{1}{\beta_{i+1}}\right]\frac{q}{q_{1,-1}} - 2\frac{q_{0,-1}}{q_{1,-1}}. \tag{2.34}$$

From (2.25) and (2.26) we obtain for $i \geq 0$ the analogue of (2.30) for $1/\beta_i$,

$$\frac{1}{\beta_i} + \frac{1}{\beta_{i+1}} = \frac{1}{\alpha_{i+1}}\frac{q}{q_{-1,1}} - \frac{q_{-1,0}}{q_{-1,1}}.$$

Inserting this equality into (2.34) gives for $i \geq 0$

$$\frac{1}{\alpha_i} + \frac{2}{\alpha_{i+1}} + \frac{1}{\alpha_{i+2}} = \left[\frac{1}{\alpha_{i+1}}\frac{q}{q_{-1,1}} - \frac{q_{-1,0}}{q_{-1,1}}\right]\frac{q}{q_{1,-1}} - 2\frac{q_{0,-1}}{q_{1,-1}}.$$

This is a second order inhomogeneous recursion relation for $1/\alpha_i$, the solution of which is given by

$$\frac{1}{\alpha_i} = A + a_1\lambda^i + a_2\lambda^{-i}, \tag{2.35}$$

where A and λ are given by (2.30) and (2.32) (note that by assumption 2.6(ii) the denominator $q^2 - 4q_{1,-1}q_{-1,1}$ is positive), and a_1 and a_2 are complex constants, which follow from the initial values $1/\alpha_0$ and $1/\alpha_1$. To establish (2.29), with $a = a_1/\sqrt{\gamma}$, it remains to prove that a_1 and a_2 satisfy

$$a_1 a_2 = \gamma.$$

First note that (2.23) is equivalent to

$$\frac{1}{\alpha_i}\frac{1}{\alpha_{i+1}} = \frac{q_{-1,1}}{q_{1,-1}}\frac{1}{\beta_i^2} + \frac{q_{-1,0}}{q_{1,-1}}\frac{1}{\beta_i} + \frac{q_{-1,-1}}{q_{1,-1}}. \tag{2.36}$$

Eliminating $1/\beta_i$ from (2.33) and (2.36) leads to

$$\begin{aligned}(q_{1,-1}q^2 - 2q_{1,-1}^2 q_{-1,1})&\frac{1}{\alpha_i}\frac{1}{\alpha_{i+1}} \\ &= q_{-1,-1}q^2 + q_{0,-1}(qq_{-1,0} + q_{-1,1}q_{0,-1}) \\ &+ \left[q_{1,-1}(qq_{-1,0} + q_{-1,1}q_{0,-1}) + q_{0,-1}q_{-1,1}q_{1,-1}\right]\left[\frac{1}{\alpha_i} + \frac{1}{\alpha_{i+1}}\right] \\ &+ q_{1,-1}^2 q_{-1,1}\left[\frac{1}{\alpha_i^2} + \frac{1}{\alpha_{i+1}^2}\right].\end{aligned} \tag{2.37}$$

By inserting the expression (2.35) for $1/\alpha_0$ and $1/\alpha_1$, equation (2.37) reduces to the identity

$$a_1 a_2 = \gamma.$$

□

Remark 2.12.

Assumption 2.6(ii) excludes the special case that all q_{ij}, except $q_{-1,1}$ and $q_{1,-1}$, are zero. The results in this section and in the following sections are essentially still valid for this special case (except when $q_{-1,1} = q_{1,-1}$), and often simplify. In particular, in this case equation (2.27) further simplifies to

$$(\alpha - \beta)(\alpha q_{-1,1} - \beta q_{1,-1}) = 0,$$

for which it is easy to prove, if $\alpha_0 = \beta_0$, that for all i

$$\alpha_i = \left[\frac{q_{1,-1}}{q_{-1,1}}\right]^i \alpha_0, \qquad \beta_i = \left[\frac{q_{1,-1}}{q_{-1,1}}\right]^i \beta_0.$$

The generation of α_i and β_i, and consequently the compensation method, fails if $q_{-1,1} = q_{1,-1}$.

2.4. On the existence of feasible pairs

One of the questions that now arise is how many feasible pairs of α_0, β_0 exist and whether these pairs are real or complex. Another question concerns conditions for the existence of feasible pairs. In this section we show, by considering a Markov process closely related to the original process, that the number of feasible pairs directly follows from the transition structure at the boundaries and that all there pairs are real. In the next section we derive conditions for the existence of feasible pairs. The analysis is restricted to the feasible pairs with respect to the *horizontal boundary*. This means that we only consider roots α_0, β_0 of equation (2.27) with $1 > |\alpha_0| > |\beta_0| > 0$ satisfying $d_{-1} = 0$. Feasibility with respect to the vertical boundary can be treated similarly.

We first treat the case that $h_{1,1} > 0$ and consider an irreducible Markov process on the set $\{ (m, n) | m + n > 0, n \geq 0 \} \cup \{ (-1, 0), (0, 0) \}$. The transition rates from states with $m + n > 0$ are given by h_{ij} if $n = 0$, and by q_{ij} if $n > 0$. We assume that from states with $m + n = 1$ and $n > 0$ transitions are possible to $(-1, 0)$ and $(0, 0)$ with rate $(q_{-1,0} + q_{-1,-1} + q_{0,-1})/2$ to each of the two states. From $(-1, 0)$ and $(0, 0)$ the Markov process can *reenter* the set of states with $m + n > 0$ with rates h_{ij}. The transition-rate diagram is depicted in figure 2.5.

The equilibrium equation in state (m, n) with $m + n > 0$ is given by (2.1) for $n > 1$, by (2.4) for $n = 1$ and finally, by (2.5) for $n = 0$. The equations in $(-1, 0)$ and $(0, 0)$ are different from (2.5), which is mainly due to the incoming rates from states with $m + n = 1$. Now define for each α_0 and β_0,

$$z_{m,n}(\alpha_0, \beta_0) = \begin{cases} \alpha_0^m \beta_0^n & \text{for } m+n > 0,\ n > 0; \\ f_0 \alpha_0^m & \text{for } m \geq -1,\ n = 0. \end{cases}$$

Then, for roots α_0, β_0 of equation (2.27) with $1 > |\alpha_0| > |\beta_0| > 0$ satisfying $d_{-1} = 0$, the product $z_{m,n}(\alpha_0, \beta_0)$ satisfies the equilibrium equations in *all states with* $m+n > 0$, and

$$\sum_{\substack{m+n>0 \\ n \geq 0}} |z_{m,n}(\alpha_0, \beta_0)| < \infty . \tag{2.38}$$

For each finite set of pairs α_0, β_0 the corresponding products $z_{m,n}(\alpha_0, \beta_0)$ are *linearly independent on the set of states with* $m+n > 0$, which follows from the next lemma. Lemma 2.13 can easily be proved by using properties of the *Vandermonde* matrix and therefore is omitted. In fact, a generalization of this lemma to infinite sums will be proved later on.

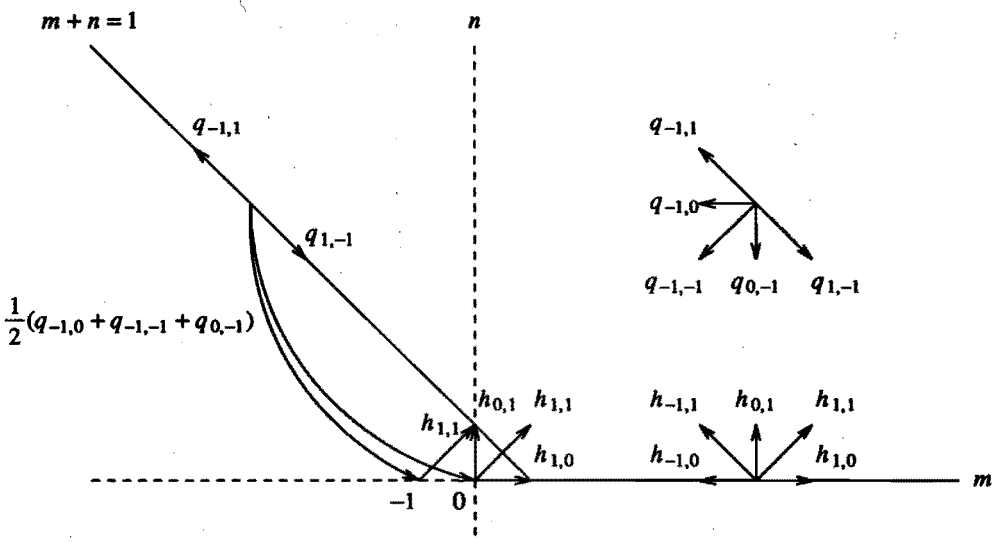


Figure 2.5.
Transition rates for the Markov process in the proof of conclusion 2.13 on the set $\{ (m, n) \mid m + n > 0, n \geq 0 \} \cup \{ (-1, 0), (0, 0) \}$. *It is assumed that* $h_{1,1} > 0$.

Lemma 2.13.

Let $(a_0, b_0), \ldots, (a_l, b_l)$ *be distinct. Then* $\sum_{i=0}^{l} k_i a_i^m b_i^n = 0 \ (m \geq 0, n \geq 0) \Leftrightarrow k_0 = \ldots = k_l = 0$.

The first question is how many feasible pairs there are. Suppose that $(\hat{\alpha}_0, \hat{\beta}_0)$ and $(\tilde{\alpha}_0, \tilde{\beta}_0)$ are feasible pairs. Then there exist nonnull coefficients $\hat{k}$ and $\tilde{k}$ such that the linear combination $\hat{k} z_{m,n}(\hat{\alpha}_0, \hat{\beta}_0) + \tilde{k} z_{m,n}(\tilde{\alpha}_0, \tilde{\beta}_0)$ satisfies the (homogeneous) equilibrium equation in state (0, 0). The remaining equation in state (−1, 0) is also satisfied, since inserting this linear combination of $z_{m,n}(\hat{\alpha}_0, \hat{\beta}_0)$ and $z_{m,n}(\tilde{\alpha}_0, \tilde{\beta}_0)$ into the equations on the set of states with $m + n \geq 0$ and $n \geq 0$ and then summing these equations and changing summations, exactly yields the equilibrium equation in state (−1, 0). Changing summations is justified by the absolute convergence in (2.38). So $\hat{k} z_{m,n}(\hat{\alpha}_0, \hat{\beta}_0) + \tilde{k} z_{m,n}(\tilde{\alpha}_0, \tilde{\beta}_0)$ is an absolutely convergent solution of the equilibrium equations of the Markov process in figure 2.5, and further, this linear combination is nonnull, since $z_{m,n}(\hat{\alpha}_0, \hat{\beta}_0)$ and $z_{m,n}(\tilde{\alpha}_0, \tilde{\beta}_0)$ are linearly independent. Hence, by a result of Foster (see appendix A), the Markov process in figure 2.5 is ergodic and normalization of $\hat{k} z_{m,n}(\hat{\alpha}_0, \hat{\beta}_0) + \tilde{k} z_{m,n}(\tilde{\alpha}_0, \tilde{\beta}_0)$ produces the equilibrium distribution. Since the equilibrium

distribution of an ergodic Markov process is unique and for different pairs of α_0 and β_0 the products $z_{m,n}(\alpha_0, \beta_0)$ are linearly independent, we can conclude that there exist *at most two feasible pairs.*

The second question is whether there exist complex feasible pairs α_0, β_0. Suppose α_0 is complex. Then there exists a nonnull coefficient k such that the sum of $kz_{m,n}(\alpha_0, \beta_0)$ and the complex conjugate of this term satisfy all equations of the Markov process in figure 2.5. However, for fixed $n \geq 0$ this sum is of the form

$$K\alpha_0^m + \bar{K}\bar{\alpha}_0^m ,$$

which has positive and negative values for $m > 0$ and therefore cannot produce probabilities. Hence, α_0 must be *real.* By a similar argument, it follows that β_0 must be real and moreover, if there exist two feasible pairs, then at least one of the α_0 (and one of the β_0) must be *positive.*

It remains to consider the two cases $h_{1,1} = 0$, $h_{0,1} + h_{1,0} > 0$ and $h_{1,1} = h_{0,1} = h_{1,0} = 0$. In the former case there exists at most one feasible pair of (positive) α_0 and β_0. This can be proved analogously to the case $h_{1,1} > 0$ by considering the irreducible Markov process on the set $\{ (m, n) | m + n > 0, n \geq 0 \} \cup \{ (0, 0) \}$ where the rate $q_{-1,0} + q_{-1,-1} + q_{0,-1}$ from states with $m + n = 1$ and $n > 0$, which is split up between $(-1, 0)$ and $(0, 0)$ in figure 2.5, is now completely directed to $(0, 0)$. In case $h_{1,1} = h_{0,1} = h_{1,0} = 0$, it can easily be derived from the definition of d_{-1} that there exist no feasible pairs.

The conclusion is that we can see directly from the transition structure at the horizontal boundary how many feasible pairs α_0, β_0 there are at most and further, that all feasible pairs are real.

Conclusion 2.14.

There are at most two feasible pairs α_0, β_0 with respect to the horizontal boundary. These pairs are always real. The maximum number of feasible pairs depends as follows on the transition structure at the horizontal boundary:

(i) If $h_{1,1} > 0$, then there are at most two pairs. If there are indeed two pairs, then at least one of the α_0 (and one of the β_0) must be positive;

(ii) If $h_{1,1} = 0$ and $h_{0,1} + h_{1,0} > 0$, then there is at most one pair. These roots are positive;

(iii) If $h_{1,1} = h_{0,1} = h_{1,0} = 0$, then there are no pairs.

2.5. Conditions for the existence of feasible pairs

We now proceed to derive conditions for the existence of roots α_0, β_0 of equation (2.27) with $1 > |\alpha_0| > \beta_0| > 0$ satisfying $d_{-1} = 0$, that is, by definition (2.22),

$$\frac{\alpha_0^2 h_{-1,1} + \alpha_0 h_{0,1} + h_{1,1}}{\beta_{-1}} + h_{1,0} + \alpha_0^2 h_{-1,0} - \alpha_0 h = 0\,, \tag{2.39}$$

$$\frac{\alpha_0^2 h_{-1,1} + \alpha_0 h_{0,1} + h_{1,1}}{\beta_0} + h_{1,0} + \alpha_0^2 h_{-1,0} - \alpha_0 h \neq 0\,. \tag{2.40}$$

The roots β_0 and β_{-1} will be regarded as *functions of* α_0. By conclusion 2.13 the analysis can be restricted to *real* α_0. It is readily verified that inequality (2.40) is always valid for nonzero $\alpha_0 \in (-1, 1)$ satisfying (2.39). To analyse (2.39) we insert the explicit formula for the root β_{-1}, for which we first derive some useful properties.

For fixed α equation (2.27) is solved by

$$X_{\pm}(\alpha) = \alpha\, \frac{q - \alpha q_{-1,0} \pm \sqrt{(q - \alpha q_{-1,0})^2 - 4(\alpha^2 q_{-1,-1} + \alpha q_{0,-1} + q_{1,-1}) q_{-1,1}}}{2(\alpha^2 q_{-1,-1} + \alpha q_{0,-1} + q_{1,-1})}\,. \tag{2.41}$$

The denominator in (2.41) may vanish for some $\alpha < 0$. For such an α, $X_-(\alpha)$ and $X_+^{-1}(\alpha)$ can be extended by taking $X_-(\alpha) = \alpha q_{-1,1} / (q - \alpha q_{-1,0})$ and $X_+^{-1}(\alpha) = 0$, thus $X_+(\alpha) = \infty$. Let $Y_{\pm}(\beta)$ be the roots of (2.27) for fixed β. We now prove the following monotonicity properties.

Lemma 2.15.

For all $0 < \alpha < 1$

(i) *the ratio* $X_+(\alpha) / \alpha$ *is decreasing and* $X_-(\alpha) / \alpha$ *is increasing;*

(ii) $|X_+(-\alpha)| \geq X_+(\alpha) > \alpha > X_-(\alpha) \geq -X_-(-\alpha) > 0\,.$

The same properties hold for $Y_{\pm}(\beta)$.

Proof.

For all $-1 < \alpha < 1$ the discriminant $D(\alpha)$ in (2.41), defined by

$$D(\alpha) = (q - \alpha q_{-1,0})^2 - 4(\alpha^2 q_{-1,-1} + \alpha q_{0,-1} + q_{1,-1}) q_{-1,1}\,,$$

is positive, which follows by using the fact that $D(\alpha)$ is decreasing for $\alpha \geq 0$, so for $0 \leq \alpha < 1$

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

Hence $X_-(\alpha)$ and $X_+(\alpha)$ are real for $-1 < \alpha < 1$.
Since $X_+(\alpha) / \alpha$ is decreasing for $0 < \alpha < 1$ we obtain for $0 < \alpha < 1$

$$\frac{|X_+(-\alpha)|}{|-\alpha|} \geq \frac{X_+(\alpha)}{\alpha} > X_+(1)\,. \tag{2.42}$$

From (2.41) follows

$$X_{\pm}(1) = \frac{q_{-1,-1}+q_{0,-1}+q_{1,-1}+q_{-1,1} \pm |q_{-1,-1}+q_{0,-1}+q_{1,-1}-q_{-1,1}|}{2(q_{-1,-1}+q_{0,-1}+q_{1,-1})}\,,$$

i.e.,

$$X_+(1) = 1 \quad \text{if} \quad q_{-1,-1}+q_{0,-1}+q_{1,-1} \geq q_{-1,1}\,; \tag{2.43}$$

$$> 1 \quad \text{otherwise.}$$

Hence, from (2.42) we can conclude that for $0 < \alpha < 1$

$$\frac{|X_+(-\alpha)|}{|-\alpha|} \geq \frac{X_+(\alpha)}{\alpha} > 1\,.$$

The root $X_-(\alpha)/\alpha$ can be rewritten as

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

Hence, for all $0 < \alpha < 1$ the ratio $X_-(\alpha)/\alpha$ is increasing, so

$$0 < \frac{X_-(-\alpha)}{-\alpha} \leq \frac{X_-(\alpha)}{\alpha} < X_-(1) \leq 1\,.$$

This completes the proof of lemma 2.15. □

Using lemma 2.15 we can refine corollary 2.8 as follows.

Lemma 2.16.

Let α_0 and β_0 be roots of equation (2.27) satisfying $1 > |\alpha_0| > |\beta_0| > 0$.
If $0 < \alpha_0 < 1$, then

$$\alpha_0 > X_-(\alpha_0) = \beta_0 > Y_-(\beta_0) = \alpha_1 > X_-(\alpha_1) = \beta_1 > \cdots > 0\,,$$

and if $-1 < \alpha_0 < 0$, then

$$\alpha_0 < X_-(\alpha_0) = \beta_0 < Y_-(\beta_0) = \alpha_1 < X_-(\alpha_1) = \beta_1 < \cdots < 0\,.$$

Since $|\alpha_0| > |\beta_0|$ it follows that $|\alpha_0| < |\beta_{-1}|$, so by lemma 2.15 we can set

$$\beta_{-1} = X_+(\alpha_0)$$

(and $\beta_0 = X_-(\alpha_0)$). Substituting this identity into equation (2.39) and rearranging terms we obtain that α_0 has to be a root of the equation (h is the sum of the h_{ij}'s, see (2.8))

$$\alpha h = \frac{\alpha^2 h_{-1,1} + \alpha h_{0,1} + h_{1,1}}{X_+(\alpha)} + \alpha^2 h_{-1,0} + h_{1,0} \,. \tag{2.44}$$

Denote by $LH(\alpha)$ the left-hand side of (2.44) and by $RH(\alpha)$ the right-hand side. Figure 2.6 shows $LH(\alpha)$ and $RH(\alpha)$ for the case $q_{-1,1} = q_{0,-1} = h_{-1,0} = 2$, $q_{1,-1} = h_{1,1} = 1$ and all other q_{ij} and h_{ij} are zero.

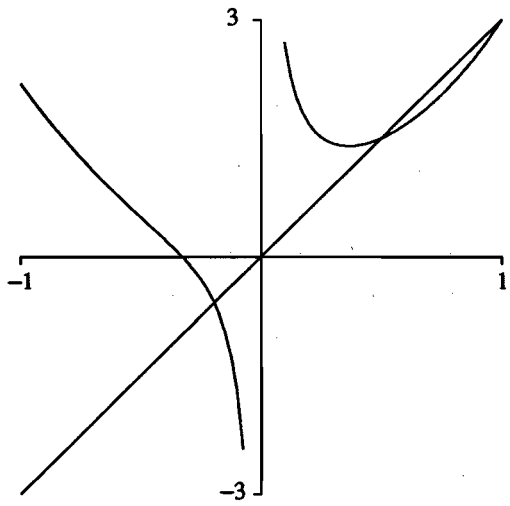


Figure 2.6.
The left-hand side $LH(\alpha)$ and the right-hand side $RH(\alpha)$ of equation (2.48) for the case $q_{-1,1} = q_{0,-1} = h_{-1,0} = 2$, $q_{1,-1} = h_{1,1} = 1$ and all other q_{ij} and h_{ij} are zero.

Figure 2.6 suggests the following lemma.

Lemma 2.17.

The right-hand side $RH(\alpha)$ of equation (2.44) is strictly convex for $0 < \alpha < 1$.

To prove lemma 2.17 we need the following elementary convexity properties.

Convexity properties:

(i) *If $f(x)$ and $g(x)$ are positive, nondecreasing and convex, then $f(x)g(x)$ is convex, and strictly convex if $f(x)$ or $g(x)$ is strictly convex;*

(ii) *If $f(x)$ and $g(x)$ are convex, then $f(x)+g(x)$ is convex, and strictly convex if $f(x)$ or $g(x)$ is strictly convex;*

(iii) *If $f(x)$ is positive and strictly concave, then $1/f(x)$ is strictly convex;*

(iv) *If $f(x)$ is positive, decreasing and strictly concave on $(0, 1)$, then $xf(x)$ is strictly concave on $(0, 1)$.*

Proof of lemma 2.17.

By inserting the formula (2.41) for $X_+(\alpha)$ into $RH(\alpha)$ we obtain

$$RH(\alpha)=\frac{2(\alpha^2 q_{-1,-1}+\alpha q_{0,-1}+q_{1,-1})(\alpha^2 h_{-1,1}+\alpha h_{0,1}+h_{1,1})}{\alpha(q-\alpha q_{-1,0}+\sqrt{D(\alpha)})}+\alpha^2 h_{-1,0}+h_{1,0}\,, \quad (2.45)$$

where $D(\alpha)$ is the discriminant in (2.41). We first prove that $q-\alpha q_{-1,0}+\sqrt{D(\alpha)}$ is positive, decreasing and strictly concave for $0<\alpha<1$. The first and second property are easily verified. To establish the third property it suffices to show that for $0<\alpha<1$

$$\left[\sqrt{D(\alpha)}\right]''<0\,. \quad (2.46)$$

This second derivative is given by

$$\left[\sqrt{D(\alpha)}\right]''=\frac{2D''(\alpha)D(\alpha)-D'(\alpha)^2}{4D(\alpha)^{3/2}}$$

where

$$D'(\alpha)=-2\left[q_{-1,0}(q-\alpha q_{-1,0})+2q_{-1,1}(2\alpha q_{-1,-1}+q_{0,-1})\right],$$

$$D''(\alpha)=2q_{-1,0}^2-8q_{-1,1}q_{-1,-1}\,.$$

Hence, if $D''(\alpha)\le 0$, then (2.46) follows directly. If $D''(\alpha)>0$, then (2.46) follows by using

$$0<D(\alpha)<(q-\alpha q_{-1,0})^2\,,\quad 0<D''(\alpha)\le 2q_{-1,0}^2\,,\quad D'(\alpha)^2\ge 4q_{-1,0}^2(q-\alpha q_{-1,0})^2\,,$$

valid for all $0<\alpha<1$. Hence, we can conclude that $q-\alpha q_{-1,0}+\sqrt{D(\alpha)}$ is positive, decreasing and strictly concave for $0<\alpha<1$. Then follows from convexity property (iv) that the numerator in (2.45) is strictly concave for $0<\alpha<1$ and so, by the properties (i) and (iii), that the quotient in (2.45) is strictly convex. Finally, by convexity property (ii), the sum in (2.45) is strictly convex for $0<\alpha<1$. This completes the proof of lemma 2.17. □

It is easily verified that at the boundary $\alpha = 0$

$$RH(0^+) > LH(0) \quad \text{if} \quad h_{1,1} + h_{0,1} + h_{1,0} > 0 ;$$

$$= LH(0) \quad \text{otherwise} ;$$

and, by using (2.43), that at the boundary $\alpha = 1$

$$RH(1) < LH(1) \quad \text{if} \quad q_{-1,-1} + q_{0,-1} + q_{1,-1} > q_{-1,1} ;$$

$$= LH(1) \quad \text{otherwise} .$$

Hence, by lemma 2.17, in case $h_{1,1} + h_{0,1} + h_{1,0} > 0$ there is a root of equation (2.44) in the interval (0, 1) if $q_{-1,-1} + q_{0,-1} + q_{1,-1} > q_{-1,1}$ and otherwise the following extra condition is required:

Condition: $RH'(1) > LH'(1)$.

Remark that the square root in (2.41) vanishes at $\alpha = 1$ if $q_{-1,-1} + q_{0,-1} + q_{1,-1} = q_{-1,1}$ and thus the (left) derivative of this square root at $\alpha = 1$ is $-\infty$. Consequently, in this case the derivative of RH at $\alpha = 1$ is $+\infty$, so the condition above trivially holds.

Lemma 2.18.

If the following two conditions hold:

(i) $h_{1,1} + h_{0,1} + h_{1,0} > 0$;

(ii) $q_{-1,-1} + q_{0,-1} + q_{1,-1} > q_{-1,1} \Rightarrow RH'(1) > LH'(1)$,

then equation (2.44) has a unique solution in (0, 1).
If condition (i) or (ii) does not hold, then equation (2.44) has no solution in (0, 1).

We know from section 2.6 that equation (2.44) may have a second solution in (–1, 1) if $h_{1,1} > 0$. We show that condition (ii) in lemma 2.18 also guarantees the existence of a second solution in (–1, 0) if $h_{1,1} > 0$. Since $RH(0^-) = -\infty$ and $RH(\alpha)$ and $LH(\alpha)$ are continuous on [–1,0), there exists a root in (–1, 0) if for some $\alpha \in [-1, 0)$

$$RH(\alpha) > LH(\alpha) .$$

This inequality trivially holds if $\alpha^2 h_{-1,1} + \alpha h_{0,1} + h_{1,1} = 0$. Now suppose there is no such α. Then, by taking $\alpha = -1$, we find

$$h_{-1,1} - h_{0,1} + h_{1,1} > 0 . \qquad (2.47)$$

By lemma 2.5 and (2.43),

$$\frac{1}{|X_+(-1)|} \leq \frac{1}{X_+(1)} \leq 1 .$$

So we also find

$$\frac{1}{X_+(-1)} \geq -1 , \tag{2.48}$$

with equality if and only if $q_{-1,-1} + q_{1,-1} \geq q_{-1,1}$ and $q_{-1,0} = q_{0,-1} = 0$. Combining the inequalities (2.47) and (2.48) yields

$$RH(-1) \geq LH(-1) , \tag{2.49}$$

with equality if and only if $q_{-1,-1} + q_{1,-1} \geq q_{-1,1}$ and $q_{-1,0} = q_{0,-1} = 0$ and all $h_{ij} = 0$ with the exception of $h_{-1,1}$ and $h_{1,1}$. In case of equality in (2.49) we have $RH(-\alpha) = -RH(\alpha)$ for $0 < \alpha < 1$ and thus condition (ii) in lemma 2.18 guarantees that equation (2.44) has a solution in $(-1, 0)$. Combining these results we can formulate the following theorem.

Theorem 2.19.

If $h_{1,1} + h_{0,1} + h_{1,0} > 0$, then the maximum number of feasible pairs with respect to the horizontal boundary is found if and only if the following condition is satisfied:

$$q_{-1,-1} + q_{0,-1} + q_{1,-1} > q_{-1,1} \Rightarrow RH'(1) > LH'(1) ; \tag{2.50}$$

Depending on the boundary behaviour the feasible pairs have the following properties:

(i) *If $h_{1,1} > 0$, then there are two feasible pairs. One α_0 is the solution of equation (2.44) in $(0, 1)$ and the other α_0 is its solution in $(-1, 0)$;*

(ii) *If $h_{1,1} = 0$ and $h_{0,1} + h_{1,0} > 0$, there is one feasible pair. The α_0 is the solution of equation (2.44) in $(0, 1)$.*

If $h_{1,1} + h_{0,1} + h_{1,0} = 0$, then there are no feasible pairs with respect to the horizontal boundary.

In the next section it is shown that condition (2.50) can be interpreted as a drift condition.

2.6. Neuts' mean drift condition

Condition (2.50) in theorem 2.19 states that if the rate downwards exceeds the rate upwards, then inequality $RH'(1) > LH'(1)$ must hold. This inequality can be interpreted as a mean drift condition. In fact, we will show that inequality $RH'(1) > LH'(1)$ corresponds to *Neuts' mean drift condition* ([51], Theorem 1.7.1.).

Consider a Markov process with generator Q of the form

$$Q = \begin{bmatrix} B_1 & B_0 & 0 & 0 & 0 & \cdots \\ A_2 & A_1 & A_0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}, \tag{2.51}$$

where all elements are (finite) $(k+1)\times(k+1)$ matrices. The states are denoted by (m, n), $m = 0, 1, 2, \ldots, n = 0, 1, \ldots, k$, and are lexicographically ordered, that is, $(0, 0), \ldots (0, k), (1, 0), \ldots, (1, k), (2, 0), \ldots$. The set of states $(m, 0), (m, 1), \ldots, (m, k)$ is called level m. Suppose that $A_0 + A_1 + A_2$ is irreducible and let π be the solution of

$$\pi(A_0 + A_1 + A_2) = 0, \quad \pi e = 1,$$

where e is the column vector with all its elements equal to one. Then by Theorem 1.7.1. in Neuts' book [51] the Markov process Q is ergodic if and only if

$$\pi A_0 e < \pi A_2 e. \tag{2.52}$$

The left-hand side of (2.52) can be defined as the mean drift from level m to $m+1$, and the right-hand side as the mean drift from level $m+1$ to m, $m > 0$, where the mean is taken with respect to the distribution π. Then (2.52) states that the mean drift to the higher level should be less than the mean drift to the lower level, and therefore is called the mean drift condition.

In our case, the generator Q is also of the form (2.51), but all elements are infinite matrices and level m is the infinite set of states $(m, 0), (m, 1), (m, 2), \ldots$. Hence, we cannot conclude that the mean drift condition (2.52) is necessary and sufficient for ergodicity of Q. However, if

$$q_{1,-1} + q_{0,-1} + q_{-1,-1} > q_{-1,1}$$

(so $X_-(1) < 1$), then the row vector $\pi = (\pi_0, \pi_1, \ldots)$ is given by

$$\pi_n = C\frac{q_{-1,1}}{h_{-1,1} + h_{0,1} + h_{1,1}} \quad \text{if} \quad n = 0,$$

$$= CX_-^n(1) \quad \text{if} \quad n > 0,$$

where C is the normalizing constant. The mean drift condition (2.52) then becomes

$$\frac{X_-(1)}{1-X_-(1)}q_{1,-1}+\frac{q_{-1,1}}{h_{-1,1}+h_{0,1}+h_{1,1}}(h_{1,1}+h_{1,0})$$

$$<\frac{X_-(1)}{1-X_-(1)}(q_{-1,1}+q_{-1,0}+q_{-1,-1})+\frac{q_{-1,1}}{h_{-1,1}+h_{0,1}+h_{1,1}}(h_{-1,0}+h_{-1,1}) \qquad (2.53)$$

The interesting point is that inequality $RH'(1) > LH'(1)$ can be rewritten as (2.53): First insert the identity (see (2.25))

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

into equation (2.44), then differentiate equation (2.44) and insert the identity

$$X_-'(1)=X_-(1)+\frac{X_-(1)q_{-1,0}}{q_{1,-1}+q_{0,-1}+q_{-1,-1}-q_{-1,1}}+\frac{X_-^2(1)(q_{0,-1}+2q_{-1,-1})}{q_{1,-1}+q_{0,-1}+q_{-1,-1}-q_{-1,1}},$$

which can be derived by straightforward calculation.

Conclusion 2.20.

If $q_{1,-1}+q_{0,-1}+q_{-1,-1} > q_{-1,1}$, then condition (2.50) in theorem 2.19 is equivalent to Neuts' mean drift condition (2.53).

2.7. Simplifications of the formal solutions with feasible pairs

In the previous sections we derived necessary and sufficient conditions for the existence of feasible initial pairs (α_0, β_0) with respect to the horizontal boundary. There are at most two such pairs. We denote these pairs by

$$(\alpha_+, X_-(\alpha_+)) \quad \text{and} \quad (\alpha_-, X_-(\alpha_-)),$$

where α_+ is the solution of equation (2.44) in (0, 1) and α_- its solution on (−1, 0). Analogous conditions can be derived on the vertical boundary. The feasible pairs on the vertical boundary are denoted by

$$(\beta_+, X_-(\beta_+)) \quad \text{and} \quad (\beta_-, X_-(\beta_-)),$$

where β_+ is the solution of the β-equivalent of equation (2.44) in (0, 1) and β_- its solution on (−1, 0). For $\alpha_0 = \alpha_+$ and $\beta_0 = X_-(\alpha_+)$ we abbreviate the notation $x_{m,n}(\alpha_0, \beta_0)$ to $x_{m,n}(\alpha_+)$. Similar abbreviations are used for the other feasible pairs.

The formal solutions $x_{m,n}(\alpha_0, \beta_0)$ with feasible initial pairs simplify with respect to the general definition in section 2.2. If we take $\alpha_0 = \alpha_+$ and $\beta_0 = X_-(\alpha_+)$, then $d_{-1} = 0$ and so

$d_i = f_i = 0$ for all $i < 0$. Then for $m > 0$ and $n > 0$ the series $x_{m,n}(\alpha_+)$ simplifies to (see (2.16)-(2.17))

$$x_{m,n}(\alpha_+) = \sum_{i=0}^{\infty} d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n \tag{2.54}$$

$$= d_0c_0\beta_0^n\alpha_0^m + \sum_{i=0}^{\infty} c_{i+1}(d_i\beta_i^n + d_{i+1}\beta_{i+1}^n)\alpha_{i+1}^m \; ; \tag{2.55}$$

for $m = 0$ and $n > 0$ and for $m > 0$ and $n = 0$ to (see (2.18)-(2.19))

$$x_{0,n}(\alpha_+) = \sum_{i=0}^{\infty} d_ie_i\beta_i^n \; , \tag{2.56}$$

$$x_{m,0}(\alpha_+) = c_0f_0\alpha_0^m + \sum_{i=0}^{\infty} c_{i+1}f_{i+1}\alpha_{i+1}^m \; , \tag{2.57}$$

respectively, where the sequence $\{\alpha_i, \beta_i\}$ is initialized by

$$\alpha_0 = \alpha_+ \; , \quad \beta_0 = X_-(\alpha_+) \; .$$

The solution $x_{m,n}(\alpha_-)$ simplifies accordingly. If we take $\alpha_0 = Y_-(\beta_+)$ and $\beta_0 = \beta_+$, then $c_1 = 0$ and so $c_i = e_i = 0$ for $i > 0$. Then for $m > 0$ and $n > 0$ the series $x_{m,n}(\beta_+)$ simplifies to

$$x_{m,n}(\beta_+) = d_0c_0\alpha_0^m\beta_0^n + \sum_{i=-\infty}^{-1} d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n$$

$$= \sum_{i=-\infty}^{-1} c_{i+1}(d_i\beta_i^n + d_{i+1}\beta_{i+1}^n)\alpha_{i+1}^m \; ;$$

for $m = 0$ and $n > 0$ and for $m > 0$ and $n = 0$ to

$$x_{0,n}(\beta_+) = d_0e_0\beta_0^n + \sum_{i=-\infty}^{-1} d_ie_i\beta_i^n \; , \tag{2.58}$$

$$x_{m,0}(\beta_+) = \sum_{i=-\infty}^{-1} c_{i+1}f_{i+1}\alpha_{i+1}^n \; , \tag{2.59}$$

where the sequence $\{\alpha_i, \beta_i\}$ is initialized by

$$\beta_0 = \beta_+ \; , \quad \alpha_0 = Y_-(\alpha_+) \; .$$

The formal solution $x_{m,n}(\beta_-)$ simplifies accordingly.

In the next section we investigate whether for feasible initial pairs the construction of $x_{m,n}(\alpha_0, \beta_0)$ may fail because of a vanishing denominator in the definition of the coefficients c_i, d_i, e_i or f_i (cf. (2.20)-(2.22)).

2.8. On the construction of the formal solutions

In this section we investigate whether for $\alpha_0 = \alpha_+$ and $\beta_0 = X_-(\alpha_+)$ the construction of $x_{m,n}(\alpha_0, \beta_0)$ can fail. The construction of $x_{m,n}(\alpha_0, \beta_0)$ for the other three potential feasible pairs can be investigated accordingly.

The construction of $x_{m,n}(\alpha_+)$ fails if for some nonnegative value of i the denominator in the definition of the coefficients c_{i+1}, e_i, d_{i+1} or f_{i+1} vanishes (see (2.20)-(2.25)). Since $0 < \alpha_0 = \alpha_+ < 1$ and $0 < \beta_0 = X_-(\alpha_+) < \alpha_0$, from lemma 2.16 we obtain

$$\alpha_0 > \beta_0 > \alpha_1 > \beta_1 > \cdots > 0\,, \tag{2.60}$$

where for all $i \geq 0$,

$$\beta_i = X_-(\alpha_i)\,, \quad \alpha_{i+1} = Y_-(\beta_i)\,, \tag{2.61}$$

and therefore also

$$\alpha_i = Y_+(\beta_i)\,, \quad \beta_i = X_+(\alpha_{i+1})\,. \tag{2.62}$$

Inserting (2.62) into the common denominator of the definitions of d_{i+1} and f_{i+1} (cf. (2.21)) we see that this denominator is equal to $RH(\alpha_{i+1}) - LH(\alpha_{i+1})$. Hence, since $\alpha_{i+1} < \alpha_+$ by (2.60), we conclude that the denominator in the definitions of d_{i+1} and f_{i+1} is positive for all $i \geq 0$. Similarly, it can be seen that the denominator in the definitions of c_{i+1} and e_i (cf. (2.20)) vanishes if β_i solves the β-equivalent of equation (2.44):

$$\beta v = \frac{\beta^2 v_{1,-1} + \beta v_{1,0} + v_{1,1}}{Y_+(\beta)} + v_{0,1} + \beta^2 v_{0,-1}\,,$$

i.e., if $\beta_i = \beta_+$. The following example shows that this possibility really may occur.

Example 2.21.

Consider the process for which $q_{0,-1} = q_{1,-1} = v_{1,-1} = h_{1,0} = 1$, $q_{-1,1} = h_{-1,1} = 2$, $v_{0,-1} = 1$, $v_{0,1} = 1/2$ and all other rates $q_{i,j}$, $h_{i,j}$ and $v_{i,j}$ are zero (see figure 2.7). By theorem 2.19(ii) and its β-analogue, the roots α_+ and β_+ exist and it is easily verified that $\alpha_+ = 1/2$ and $\beta_+ = 1/3$ (the value for α_+ is suggested by the property that the marginal distribution $\{p_m\}$ is that of an $M|M|1$ queue with load 1/2). The construction of $x_{m,n}(\alpha_+)$ fails, since $\beta_0 = X_-(\alpha_+) = 1/3 = \beta_+$, so the denominator in c_1 vanishes.

This leads to the following condition:

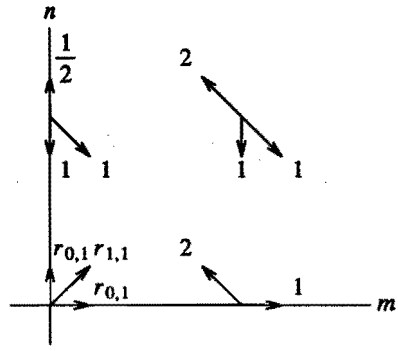


Figure 2.7.
An example for which the construction of $x_{m,n}(\alpha_+)$ fails, due to the fact that the denominator in the definition of c_1 vanishes.

Condition 2.22.

If the roots α_+ and β_+ exist, then none of the β_i of the sequence $\{\alpha_i, \beta_i\}_{i=0}^{\infty}$ with initial values $\alpha_0 = \alpha_+$ and $\beta_0 = X_-(\alpha_+)$ may be equal to β_+.

By considering small perturbations of the boundary conditions it can be shown that violation of this condition is exceptional (see the remark at the end of this section). In fact, in section 2.13 we shall demonstrate that the probabilities of processes violating condition 2.22 can be obtained from a limiting argument. At the end of this section we formulate the analogous conditions for the sequences $\{\alpha_i, \beta_i\}$ associated with $x_{m,n}(\alpha_-)$, $x_{m,n}(\beta_+)$ and $x_{m,n}(\beta_-)$.

Above we investigated whether the denominator in the definitions of c_{i+1} and d_{i+1} vanishes for some nonnegative value of i. Alternatively we may investigate whether the numerator in these definitions vanishes. In this case all subsequent c_j or d_j vanish and $x_{m,n}(\alpha_+)$ reduces to a finite sum. Since $\beta_{i+1} < \beta_i$ by (2.60), it follows that the numerator in the definition of d_{i+1} is larger than the denominator, which is positive for all $i \geq 0$. This proves that for all $i \geq 0$,

$$\frac{d_{i+1}}{d_i} < 0 .$$

Inserting (2.61) into the numerator of the definition of c_{i+1} we see that this numerator vanishes if β_i solves the equation:

$$\beta v = \frac{\beta^2 v_{1,-1} + \beta v_{1,0} + v_{1,1}}{Y_-(\beta)} + v_{0,1} + \beta^2 v_{0,-1} .$$

In the following example the numerator in c_1 vanishes, so $c_i = e_i = 0$ for all $i > 0$.

Example 2.23.

Consider the process with the same rates as in example 2.21, except that $\nu_{0,-1} = 3$ and $\nu_{0,1} = 0$. By theorem 2.9(ii) the root α_+ exists and it is easily verified that $\alpha_+ = 1/2$. Hence, by taking $\alpha_0 = \alpha_+$ and $\beta_0 = X_-(\alpha_+) = 1/3$, it follows that $d_{-1} = 0$ and $e_0 = f_0 = 1$, but more importantly, also $c_1 = 0$, so for all m and n the solution $x_{m,n}(\alpha_+)$ reduces to

$$x_{m,n}(\alpha_+) = \frac{1}{2}^m \frac{1}{3}^n .$$

We now formulate the analogues of condition 2.22 for the sequences $\{\alpha_i, \beta_i\}$ associated with $x_{m,n}(\alpha_-)$, $x_{m,n}(\beta_+)$ and $x_{m,n}(\beta_-)$. Remark that from lemma 2.16 it follows that the sequence $\{\alpha_i, \beta_i\}_{i=0}^{\infty}$ is *positive and decreasing* if we take $\alpha_0 = \alpha_+$ and $\beta_0 = X_-(\alpha_+)$, and this sequence is *negative and increasing* if we take $\alpha_0 = \alpha_-$ and $\beta_0 = X_-(\alpha_-)$. The similar remark holds for the sequence $\{\alpha_i, \beta_i\}_{i=0}^{-\infty}$ with $\beta_0 = \beta_+$ and $\alpha_0 = Y_-(\beta_+)$ and with $\beta_0 = \beta_-$ and $\alpha_0 = Y_-(\beta_-)$ respectively. Hence, to guarantee that the construction of the formal solutions with feasible initial pairs succeeds, we have to impose the following condition.

Condition 2.24.

If the roots α_+ and β_+ exist, then:

(i) *None of the β_i of $\{\alpha_i, \beta_i\}_{i=0}^{\infty}$ with $\alpha_0 = \alpha_+$ and $\beta_0 = X_-(\alpha_+)$ may be equal to β_+;*

(ii) *None of the α_i of $\{\alpha_i, \beta_i\}_{i=0}^{-\infty}$ with $\beta_0 = \beta_+$ and $\alpha_0 = Y_-(\beta_+)$ may be equal to α_+.*

The same property should be satisfied with α_+, β_+ replaced by α_-, β_-.

Remark 2.25.

Let us investigate the effect of small perturbations of the horizontal boundary behaviour on the solutions α_+ and α_-. If we replace $h_{1,0}$ by $h_{1,0} + \varepsilon$ with $\varepsilon \geq 0$, then equation (2.44) becomes

$$LH(\alpha) + \alpha\varepsilon = RH(\alpha) + \varepsilon . \tag{2.63}$$

Let $\alpha_+(\varepsilon)$ be the solution of (2.63) in (0, 1) and $\alpha_-(\varepsilon)$ its solution in (–1, 0). Then it is readily verified that $\alpha_+(\varepsilon)$ and $\alpha_-(\varepsilon)$ are continuous and increasing in ε. Specifically, to prove that $\alpha_+(\varepsilon)$ is increasing in ε, its derivative has to be evaluated. From (2.63) follows

$$\alpha_+'(\varepsilon) = \frac{1-\alpha_+(\varepsilon)}{LH'(\alpha_+(\varepsilon)) - RH'(\alpha_+(\varepsilon)) + \varepsilon}. \tag{2.64}$$

This derivative is positive for $\varepsilon \geq 0$, since $LH'(\alpha_+(\varepsilon)) > RH'(\alpha_+(\varepsilon))$ by the strict convexity of $RH(\alpha)$ (cf. lemma 2.17). If $\nu_{0,1}$ is replaced by $\nu_{0,1} + \varepsilon$ with $\varepsilon \geq 0$, then the same properties for the solutions $\beta_+(\varepsilon)$ and $\beta_-(\varepsilon)$ are readily derived.

It can now be concluded that if condition 2.24 is not satisfied by the process at hand, it is satisfied by some ε-perturbed process. Hence, violation of this condition is exceptional. In fact, in section 2.13 it will be argued by using a sequence of ε-perturbed processes, that in case of violation of condition 2.22 the probabilities can be obtained from a limiting argument.

2.9. Absolute convergence of the formal solutions

We now try to prove that for all feasible initial pairs α_0, β_0 the series $x_{m,n}(\alpha_0, \beta_0)$ converges absolutely. We need absolute convergence to guarantee equality of (2.16) and (2.17). It appears however, that $x_{m,n}(\alpha_0, \beta_0)$ possibly *diverges* in states near the origin of the state space, but we will prove:

Theorem 2.25 (Absolute convergence).

There exists an integer N such that for all feasible pairs (α_0, β_0)*:*

(i) *The series* $\sum_{i=-\infty}^{\infty} d_i c_i \alpha_i^m \beta_i^n$ *and* $\sum_{i=-\infty}^{\infty} d_i c_{i+1} \alpha_{i+1}^m \beta_i^n$ *the sum of which defines* $x_{m,n}(\alpha_0, \beta_0)$ *for* $m > 0$ *and* $n > 0$*, both converge absolutely for all* $m \geq 0$, $n \geq 0$ *with* $m + n > N$*;*

(ii) *The series* $\sum_{i=-\infty}^{\infty} d_i e_i \beta_i^n$ *the sum of which defines* $x_{0,n}(\alpha_0, \beta_0)$ *for* $n > 0$*,*

$$\text{converges absolutely for all} \begin{cases} n > N-2 & \text{if } \nu_{1,1} > 0; \\ n > N-1 & \text{if } \nu_{1,1} = 0,\ \nu_{1,0}+\nu_{0,1} > 0; \\ n > N & \text{if } \nu_{1,1} = \nu_{1,0} = \nu_{0,1} = 0; \end{cases}$$

(iii) *The series* $\sum_{i=-\infty}^{\infty} c_{i+1} f_{i+1} \alpha_{i+1}^m$ *the sum of which defines* $x_{m,0}(\alpha_0, \beta_0)$ *for* $m > 0$*,*

$$\text{converges absolutely for all} \begin{cases} m > N-2 & \text{if } h_{1,1} > 0; \\ m > N-1 & \text{if } h_{1,1} = 0,\ h_{0,1}+h_{1,0} > 0; \\ m > N & \text{if } h_{1,1} = h_{0,1} = h_{1,0} = 0; \end{cases}$$

(iv) $\sum_{\substack{m\geq 0, n\geq 0 \\ m+n>N}} |x_{m,n}(\alpha_0, \beta_0)| < \infty$.

In the next section we will define N as *the smallest integer* that guarantees the absolute convergence, stated in theorem 2.25.

2.10. Proof of theorem 2.25

We only prove theorem 2.25 for $\alpha_0 = \alpha_+$ and $\beta_0 = X_-(\alpha_+)$. The proof is similar for the other three potential feasible initial pairs. We first derive the limiting behaviour of the sequence $\{\alpha_i, \beta_i\}_{i=0}^{\infty}$ and the associated sequence of coefficients $\{c_i, d_i, e_i, f_i\}_{i=0}^{\infty}$. Using these results the proof of theorem 2.25 appears to be simple.

Lemma 2.26.

(i) *Let* $|x_1| > |x_2|$ *be the roots of (2.27) for fixed* β *with* $0 < |\beta| < 1$ *and let*

$$z_{m,n} = \begin{cases} x_1^m \beta^n + c x_2^m \beta^n, & m > 0,\ n > 0; \\ e\beta^n, & m = 0,\ n > 0, \end{cases}$$

where c and e are given by (2.12) and (2.13) respectively. Then, as $\beta \to 0$,

$$\frac{\beta}{x_1} \to \frac{1}{A_2} \quad \textit{where} \quad A_2 = \frac{q + \sqrt{q^2 - 4q_{1,-1}q_{-1,1}}}{2q_{-1,1}};$$

$$\frac{x_2}{\beta} \to A_1 \quad \textit{where} \quad A_1 = \frac{q - \sqrt{q^2 - 4q_{1,-1}q_{-1,1}}}{2q_{-1,1}};$$

$$c \to -V \quad \textit{where} \quad V = \begin{cases} \dfrac{A_2}{A_1} & \textit{if} \quad v_{1,1} > 0; \\ \dfrac{A_1^{-1}v_{1,0} + v_{0,1}}{A_2^{-1}v_{1,0} + v_{0,1}} & \textit{if} \quad v_{1,1} = 0,\ v_{1,0} + v_{0,1} > 0; \\ \dfrac{A_1^{-1}v_{1,-1} - v}{A_2^{-1}v_{1,-1} - v} & \textit{if} \quad v_{1,1} = v_{1,0} = v_{0,1} = 0. \end{cases} \tag{2.65}$$

$$e\beta^{-2} \to -\frac{q_{1,-1}(A_1^{-1} - A_2^{-1})}{A_2^{-1}v_{1,1}} \quad \textit{if} \quad v_{1,1} > 0;$$

$$e\beta^{-1} \to -\frac{q_{1,-1}(A_1^{-1} - A_2^{-1})}{A_2^{-1}v_{1,0} + v_{0,1}} \qquad \text{if} \quad v_{1,1} = 0,\ v_{1,0} + v_{0,1} > 0;$$

$$e \to -\frac{q_{1,-1}(A_1^{-1} - A_2^{-1})}{A_2^{-1}v_{1,-1} - v} \qquad \text{if} \quad v_{1,1} = v_{1,0} = v_{0,1} = 0.$$

(ii) *Let* $|y_1| > |y_2|$ *be the roots of (2.27) for fixed* α *with* $0 < |\alpha| < 1$ *and let*

$$w_{m,n} = \begin{cases} \alpha^m y_1^n + d\alpha^m y_2^n, & m > 0,\ n > 0; \\ f\alpha^m, & m > 0,\ n = 0, \end{cases}$$

where d and f are given by (2.14) and (2.15) respectively. Then, as $\alpha \to 0$,

$$\frac{\alpha}{y_1} \to A_1;$$

$$\frac{y_2}{\alpha} \to \frac{1}{A_2};$$

$$d \to -H \quad \text{where} \quad H = \begin{cases} \dfrac{A_2}{A_1} & \text{if} \quad h_{1,1} > 0; \\ \dfrac{A_2 h_{0,1} + h_{1,0}}{A_1 h_{0,1} + h_{1,0}} & \text{if} \quad h_{1,1} = 0,\ h_{0,1} + h_{1,0} > 0; \\ \dfrac{A_2 h_{-1,1} - h}{A_1 h_{-1,1} - h} & \text{if} \quad h_{1,1} = h_{0,1} = h_{1,0} = 0; \end{cases}$$

$$f\alpha^{-2} \to -\frac{q_{-1,1}(A_2 - A_1)}{A_1 h_{1,1}} \qquad \text{if} \quad h_{1,1} > 0;$$

$$f\alpha^{-1} \to -\frac{q_{-1,1}(A_2 - A_1)}{A_1 h_{0,1} + h_{1,0}} \qquad \text{if} \quad h_{1,1} = 0,\ h_{0,1} + h_{1,0} > 0;$$

$$f \to -\frac{q_{-1,1}(A_2 - A_1)}{A_1 h_{-1,1} - h} \qquad \text{if} \quad h_{1,1} = h_{0,1} = h_{1,0} = 0.$$

Proof.

We prove part (ii). Part (i) can be proved similarly. Let $z(\alpha)$ be the smaller root of (2.28), so $y_2/\alpha = z(\alpha)$. Hence, since $z(\alpha)$ is continuous, as $\alpha \to 0$,

$$\frac{y_2}{\alpha} = z(\alpha) \to z(0) = \frac{1}{A_2}, \tag{2.66}$$

and from (2.25),

$$\frac{\alpha}{y_1} = \frac{\alpha^2 q_{-1,-1} + \alpha q_{0,-1} + q_{1,-1}}{q_{-1,1}} \frac{y_2}{\alpha} \to \frac{q_{1,-1}}{q_{-1,1}} \frac{1}{A_2} = A_1. \tag{2.67}$$

The limits of d and f can directly be obtained by letting $\alpha \to 0$ in (2.14) and (2.15) and then inserting (2.66) and (2.67). □

For $\alpha_0 = \alpha_+$ and $\beta_0 = X_-(\alpha_+)$ it follows from corollary 2.8 that α_i and β_i tend to zero as i tends to plus infinity. Then we directly obtain the desired limiting behaviour from lemma 2.26 and the definitions of α_i, β_i, c_i, d_i, e_i and f_i in section 2.2. Note that the ratios β_i / α_i and α_{i+1} / β_i are monotonically decreasing, which follows from $1 > \alpha_0 > \beta_0 > 0$ and the lemmas 2.15 and 2.16. Furthermore, lemma 2.27 is formulated for $\alpha_0 = \alpha_+$. If α_+ exists, then $h_{1,1} + h_{1,0} + h_{0,1} > 0$ by lemma 2.18 and thus the case $h_{1,1} = h_{1,0} = h_{0,1} = 0$ is not relevant for the limit of f_{i+1} (cf. lemma 2.26(ii)).

Lemma 2.27.

Consider the feasible initial pair $\alpha_0 = \alpha_+$, $\beta_0 = X_-(\alpha_+)$ and let i tends to infinity. Then we have:

$$\frac{\beta_i}{\alpha_i} \downarrow \frac{1}{A_2};$$

$$\frac{\alpha_{i+1}}{\beta_i} \downarrow A_1;$$

If $c_i = 0$ for some $i > 0$, then $c_j = e_j = 0$ for all $j > i$. Otherwise, as $i \to \infty$, then

$$\frac{c_{i+1}}{c_i} \to -V;$$

$$\frac{d_{i+1}}{d_i} \to -H;$$

$$\frac{e_i}{c_i \beta_i^2} \to -\frac{q_{1,-1}(A_1^{-1} - A_2^{-1})}{v_{1,1} A_2^{-1}} \quad \text{if} \quad v_{1,1} > 0;$$

$$\frac{e_i}{c_i \beta_i} \to -\frac{q_{1,-1}(A_1^{-1} - A_2^{-1})}{v_{1,0} A_2^{-1} + v_{0,1}} \quad \text{if} \quad v_{1,1} = 0,\ v_{1,0} + v_{0,1} > 0;$$

$$\frac{e_i}{c_i} \to -\frac{q_{1,-1}(A_1^{-1} - A_2^{-1})}{v_{1,-1} A_2^{-1} - v} \quad \text{if} \quad v_{1,1} = v_{1,0} = v_{0,1} = 0;$$

$$\frac{f_{i+1}}{d_i \alpha_{i+1}^2} \to -\frac{q_{-1,1}(A_2 - A_1)}{A_1 h_{1,1}} \quad \text{if} \quad h_{1,1} > 0;$$

$$\frac{f_{i+1}}{d_i \alpha_{i+1}} \to -\frac{q_{-1,1}(A_2 - A_1)}{A_1 h_{0,1} + h_{1,0}} \quad \text{if} \quad h_{1,1} = 0,\ h_{0,1} + h_{1,0} > 0.$$

Theorem 2.25 is trivial if $c_i = 0$ for some $i > 0$, since then $c_j = e_j = 0$ for all $j > i$, so $x_{m,n}(\alpha_+)$ simplifies to a finite sum. Now suppose that c_i never vanishes. To prove theorem 2.25 in this case, consider a *fixed* $m > 0$ and $n > 0$. Then by lemma 2.27, as $i \to \infty$,

$$\frac{|d_{i+1}c_{i+1}\alpha_{i+1}^m\beta_{i+1}^n|}{|d_ic_i\alpha_i^m\beta_i^n|} \text{ and } \frac{|d_{i+1}c_{i+1}\alpha_{i+2}^m\beta_{i+1}^n|}{|d_ic_{i+1}\alpha_{i+1}^m\beta_i^n|} \to |HV|(A_1/A_2)^{m+n}. \tag{2.68}$$

Hence, if $|HV|(A_1/A_2)^{m+n} < 1$, then the series $x_{m,n}(\alpha_+)$ *converges absolutely*, and if the limit $|HV|(A_1/A_2)^{m+n} > 1$, then the series $x_{m,n}(\alpha_+)$ *diverges*. Finally, nothing can be said in general if $|HV|(A_1/A_2)^{m+n} = 1$.

Similarly, by lemma 2.27, for fixed values $m > 0$ and $n > 0$,

$$\frac{|d_{i+1}e_{i+1}\beta_{i+1}^n|}{|d_ie_i\beta_i^n|} \to \begin{cases} |HV|(A_1/A_2)^{n+2} & \text{if } \nu_{1,1} > 0; \\ |HV|(A_1/A_2)^{n+1} & \text{if } \nu_{1,1} = 0,\ \nu_{1,0} + \nu_{0,1} > 0; \\ |HV|(A_1/A_2)^{n} & \text{if } \nu_{1,1} = \nu_{1,0} = \nu_{0,1} = 0; \end{cases} \tag{2.69}$$

$$\frac{|c_{i+2}f_{i+2}\alpha_{i+2}^m|}{|c_{i+1}f_{i+1}\alpha_{i+1}^m|} \to \begin{cases} |HV|(A_1/A_2)^{m+2} & \text{if } h_{1,1} > 0; \\ |HV|(A_1/A_2)^{m+1} & \text{if } h_{1,1} = 0,\ h_{0,1} + h_{1,0} > 0, \end{cases}$$

as $i \to \infty$. Because $0 < A_1 < 1 < A_2$ we can define N, mentioned in theorem 2.25, as follows.

Definition 2.28.

Let N be the smallest nonnegative integer such that $|HV|(A_1/A_2)^{N+1} < 1$.

From this definition and the limits (2.68) and (2.69) it follows that N is the smallest integer which guarantees the absolute convergence stated in theorem 2.25(i)-(iii). We finally prove theorem 2.25(iv) stating that the sum

$$\sum_{\substack{m\geq 0, n\geq 0 \\ m+n>N}} |x_{m,n}(\alpha_+)|$$

converges. Inserting the definitions (2.54), (2.56) and (2.57) for $x_{m,n}(\alpha_+)$ into this sum yields

$$\sum_{\substack{m\geq 0, n\geq 0 \\ m+n>N}} |x_{m,n}(\alpha_+)|$$

$$= \sum_{m=1}^{N} \sum_{n=N+1-m}^{\infty} |x_{m,n}(\alpha_+)| + \sum_{m=N+1}^{\infty} \sum_{n=1}^{\infty} |x_{m,n}(\alpha_+)|$$

$$+ \sum_{n=N+2}^{\infty} |x_{0,n}(\alpha_+)| + \sum_{m=N+2}^{\infty} |x_{m,0}(\alpha_+)|$$

$$\le \sum_{m=1}^{N} \sum_{n=N+1-m}^{\infty} \sum_{i=0}^{\infty} |d_i|(|c_i|\alpha_i^m + |c_{i+1}|\alpha_{i+1}^m)\beta_i^n + \sum_{m=N+1}^{\infty} \sum_{n=1}^{\infty} \sum_{i=0}^{\infty} |d_i|(|c_i|\alpha_i^m + |c_{i+1}|\alpha_{i+1}^m)\beta_i^n$$

$$+ \sum_{n=N+2}^{\infty} \sum_{i=0}^{\infty} |d_i e_i|\beta_i^n + \sum_{m=N+2}^{\infty} \sum_{i=-1}^{\infty} |c_{i+1} f_{i+1}|\alpha_{i+1}^m$$

$$= \sum_{m=1}^{N} \sum_{i=0}^{\infty} |d_i|(|c_i|\alpha_i^m + |c_{i+1}|\alpha_{i+1}^m)\frac{\beta_i^{N+1-m}}{1-\beta_i} + \sum_{i=0}^{\infty} |d_i|(|c_i|\frac{\alpha_i^{N+1}}{1-\alpha_i} + |c_{i+1}|\frac{\alpha_{i+1}^{N+1}}{1-\alpha_{i+1}})\frac{\beta_i}{1-\beta_i}$$

$$+ \sum_{i=0}^{\infty} |d_i e_i|\frac{\beta_i^{N+2}}{1-\beta_i} + \sum_{i=-1}^{\infty} |c_{i+1} f_{i+1}|\frac{\alpha_{i+1}^{N+2}}{1-\alpha_{i+1}} < \infty ,$$

since the ratio of successive terms in each of these series tends to $|HV|(A_1/A_2)^{N+1} < 1$ as i tends to infinity (for the last two series the limit of this ratio may be smaller). This completes the proof of theorem 2.25. □

Remark 2.29.

If $h_{0,1} + h_{1,1} + h_{1,0} > 0$ and $v_{1,0} + v_{1,1} + v_{0,1} > 0$, then

$$1 \le H \le \frac{A_2}{A_1}, \quad 1 \le V \le \frac{A_2}{A_1},$$

so $N \le 2$. In particular, $N = 2$ if $h_{1,1} > 0$ and $v_{1,1} > 0$. However, if $h_{0,1} + h_{1,1} + h_{1,0} = 0$ or $v_{1,0} + v_{1,1} + v_{0,1} = 0$, then N can be arbitrary large. This is illustrated by the following example. Consider the process for which $q_{-1,1} = h_{-1,1} = 2$, $q_{1,-1} = v_{1,-1} = h_{1,0} = 1 - \delta$, $q_{0,-1} = v_{0,-1} = \delta$, where $0 < \delta < 1$ and all other rates $q_{i,j}$, $h_{i,j}$ and $v_{i,j}$ are zero (see figure 2.8). For this example it can readily be verified that

$$A_1 = \frac{3}{4} - \frac{1}{4}\sqrt{1+8\delta}, \quad A_2 = \frac{3}{4} + \frac{1}{4}\sqrt{1+8\delta},$$

$$H = 1, \quad V = \frac{(1-\delta)A_1^{-1} - 1}{(1-\delta)A_2^{-1} - 1} = \frac{1+\sqrt{1+8\delta}}{1-\sqrt{1+8\delta}}.$$

Hence, the integer N is the smallest nonnegative integer for which

$$\frac{\sqrt{1+8\delta}+1}{\sqrt{1+8\delta}-1}\left[\frac{3-\sqrt{1+8\delta}}{3+\sqrt{1+8\delta}}\right]^{N+1} < 1.$$

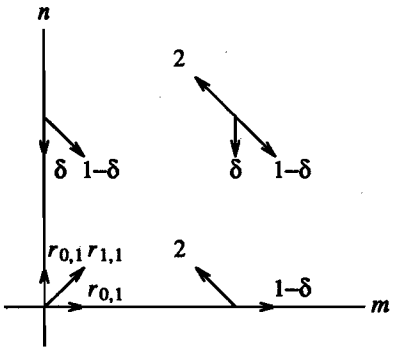


Figure 2.8.
An example for which $N \to \infty$ *as* $\delta \downarrow 0$.

From this inequality it follows that $N \to \infty$ as $\delta \downarrow 0$.

2.11. Linear independence of the formal solutions

In this section we prove the following property.

Lemma 2.30.

The solutions $x_{m,n}(\alpha_0, \beta_0)$ *for different feasible pairs are linearly independent on the set of states* $\{(m, n) \mid m \geq 0, n \geq 0, m + n > N\}$.

To prove the lemma we first need the following result.

Lemma 2.31.

Let the numbers $a_0, a_1, a_2, \ldots$ *satisfy* $0 < |a_i| < 1$ *for* $i \geq 0$, $a_i \neq a_j$ *for* $i \neq j$ *and* $a_i \to 0$ *as* $i \to \infty$; *define for* $m \geq 0$

$$x_m = \sum_{i=0}^{\infty} k_i a_i^m$$

with $\sum_{i=0}^{\infty} |k_i| < \infty$. *Then* $x_m \equiv 0 \iff k_i \equiv 0$.

Proof.

Define

$$f(z) = \sum_{m=0}^{\infty} x_m z^m = \sum_{i=0}^{\infty} \frac{k_i}{1 - a_i z}, \quad z \in \mathbb{C} \setminus \{\frac{1}{a_0}, \frac{1}{a_1}, \cdots\};$$

$$\delta_j = \inf_{i \neq j} |\frac{1}{a_i} - \frac{1}{a_j}|, \quad j \geq 0.$$

δ_j is the distance from the pole $1/a_j$ of $f(z)$ to its other poles. $\delta_j > 0$, since all a_i are distinct and $a_i \to 0$ as $i \to \infty$. Using Cauchy's theorem of residues yields

$$x_m \equiv 0 \Rightarrow f(z) \equiv 0 \Rightarrow 0 = \frac{1}{2\pi i} \int_{|z - \frac{1}{a_j}| = \frac{\delta_j}{2}} f(z)dz = \frac{k_j}{a_j} \Rightarrow k_j = 0 \quad (j \geq 0).$$

The implication $k_i \equiv 0 \Rightarrow x_m \equiv 0$ is trivial. □

From this lemma we can derive an extension of lemma 2.13 to infinite sums.

Corollary 2.32.

Let the pairs $(a_0, b_0), (a_1, b_1), (a_2, b_2), \ldots$ *satisfy* $0 < |a_i| < 1, 0 < |b_i| < 1$ *for* $i \geq 0$, $(a_i, b_i) \neq (a_j, b_j)$ *for* $i \neq j$ *and* $(a_i, b_i) \to (0, 0)$ *as* $i \to \infty$; *define for* $m \geq 0, n \geq 0$

$$x_{m,n} = \sum_{i=0}^{\infty} k_i a_i^m b_i^n$$

with $\sum_{i=0}^{\infty} |k_i| < \infty$. *Then* $x_{m,n} \equiv 0 \Longleftrightarrow k_i \equiv 0$.

Proof.

By first considering $x_{m,n}$ for fixed m we obtain from lemma 2.31

$$x_{m,n} \equiv 0 \Rightarrow \sum_{b_i = b_j} k_i a_i^m = 0 \quad (m \geq 0, j \geq 0) \Rightarrow k_i = 0 \quad (i \geq 0).$$

The implication $k_i \equiv 0 \Rightarrow x_{m,n} \equiv 0$ is trivial. □

Condition 2.24 implies that the solutions $x_{m,n}(\alpha_0, \beta_0)$ for feasible initial pairs have no products in common. Then lemma 2.30 easily follows from corollary 2.32.

2.12. Main result

We now have all ingredients to prove our main result, stating that under certain *drift* conditions, the probabilities $p_{m,n}$ can be expressed as a linear combination of the series $x_{m,n}(\alpha_0, \beta_0)$, with (α_0, β_0) running through the set of at most four feasible pairs, on a *subset of the state space*. Essentially, this subset is the set on which the series $x_{m,n}(\alpha_0, \beta_0)$ converge absolutely. By theorem 2.25, this set is given by

$$\mathcal{A}(N) = \{(m, n) \mid m \geq 0,\ n \geq 0,\ m + n > N\} \cup \mathcal{B}(N),$$

where the set $\mathcal{B}(N)$ depends on the transition structure on the axes, that is,

$$\mathcal{B}(N) = \begin{cases} \{(N-1, 0), (N, 0)\} & \text{if } h_{1,1} > 0; \\ \{(N, 0)\} & \text{if } h_{1,1} = 0,\ h_{0,1} + h_{1,0} > 0; \\ \varnothing & \text{if } h_{1,1} = h_{0,1} = h_{1,0} = 0; \end{cases}$$

$$\cup \begin{cases} \{(0, N-1), (0, N)\} & \text{if } v_{1,1} > 0; \\ \{(0, N)\} & \text{if } v_{1,1} = 0,\ v_{1,0} + v_{0,1} > 0; \\ \varnothing & \text{if } v_{1,1} = v_{1,0} = v_{0,1} = 0. \end{cases}$$

In figure 2.9 the set $\mathcal{A}(N)$ is depicted for the case $h_{1,1} > 0$ and $v_{1,1} > 0$.

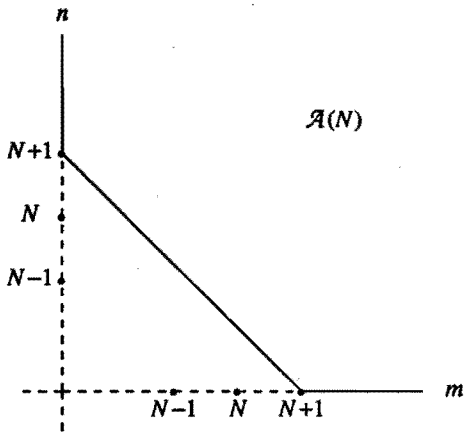


Figure 2.9.
The set $\mathcal{A}(N)$ on which the series $x_{m,n}(\alpha_0, \beta_0)$ converge absolutely, for the case $h_{1,1} > 0$ and $v_{1,1} > 0$. For states outside $\mathcal{A}(N)$ the series $x_{m,n}(\alpha_0, \beta_0)$ may diverge.

Theorem 2.33 (main result).

Let the following conditions be satisfied:

(i) *If* $h_{0,1} + h_{1,1} + h_{1,0} > 0$, *then condition (2.50) should be satisfied;*

(ii) *If* $v_{1,0} + v_{1,1} + v_{0,1} > 0$, *then the analogous drift condition should be satisfied on the vertical boundary.*

Then there exists an integer M such that on the set $\mathcal{A}(M)$

$$p_{m,n} = \sum_{(\alpha_0, \beta_0)} k(\alpha_0, \beta_0) x_{m,n}(\alpha_0, \beta_0),$$

where (α_0, β_0) *runs through the set of at most four feasible pairs and* $k(\alpha_0, \beta_0)$ *is an appropriately chosen coefficient.*

Proof.

Take $M \geq N$ and $M > 1$. The latter inequality is required to preclude possible complications in the states (0, 0), (0, 1), (1, 1) and (1, 0) due to the rates r_{ij} in the origin. Then, to prove the main theorem, we shall consider *the Markov process restricted to the set* $\mathcal{A}(M)$, that is, visits to states outside $\mathcal{A}(M)$ are not considered.

In all states with $m + n > M$ the equilibrium equations, associated with the restricted process, are identical to the ones of the original process, that is, the equations (2.1)-(2.5)). Hence, for each feasible pair (α_0, β_0) the series $x_{m,n}(\alpha_0, \beta_0)$, which converges absolutely on the set $\mathcal{A}(M)$, satisfies the equilibrium equations associated with the restricted process in all states with $m + n > M$. The boundary equations on $\mathcal{B}(M)$ are not given by the equations (2.3) and (2.5), but have an extra incoming rate. This is due to excursions of the original process to states *outside* $\mathcal{A}(M)$, which, because of the special transition structure in the interior of the state space, always end at one of the states in $\mathcal{B}(M)$. To satisfy the equations on $\mathcal{B}(M)$, we will try to fit a linear combination of series $x_{m,n}(\alpha_0, \beta_0)$ with different feasible pairs (α_0, β_0) on these equations.

Since the original Markov process is supposed to be irreducible, $h_{0,1} + h_{1,1} + h_{1,0} > 0$ or $v_{1,0} + v_{1,1} + v_{0,1} > 0$. The conditions (i) and (ii) in theorem 2.33 are necessary and sufficient for the existence of a number of feasible pairs at least equal to the number of states in $\mathcal{B}(M)$. Hence, by first omitting one arbitrarily chosen equation on $\mathcal{B}(M)$, there exist nonnull coefficients $k(\alpha_0, \beta_0)$ such that the linear combination

$$\sum_{(\alpha_0, \beta_0)} k(\alpha_0, \beta_0) x_{m,n}(\alpha_0, \beta_0), \tag{2.70}$$

where (α_0, β_0) runs through the set of feasible pairs, satisfies the remaining (*homogeneous*) equilibrium equations on $\mathcal{B}(M)$. The equation on $\mathcal{B}(M)$, which is initially omitted, is also

satisfied, since inserting the linear combination (2.70) into the other equations on the set $\mathcal{A}(M)$ and then summing these equations and changing summations exactly yields the desired equation. Changing summations is allowed by the absolute convergence stated in theorem 2.25(iv). The linear combination (2.70) is *nonnull*, because, by lemma 2.30, the series $x_{m,n}(\alpha_0, \beta_0)$ for different feasible pairs are linearly independent on the set of states with $m + n > M$. By a result of Foster (see appendix A), this proves that the process restricted to $\mathcal{A}(M)$ is ergodic and normalization of the linear combination (2.70) produces the equilibrium distribution $\{p_{m,n}(M)\}$ of the process restricted to $\mathcal{A}(M)$. Since the complement of $\mathcal{A}(M)$ is finite, it follows that the original process is also ergodic and the probabilities $p_{m,n}$ and $p_{m,n}(M)$ are related by

$$p_{m,n} = p_{m,n}(M)\, P(\mathcal{A}(M)), \quad (m, n) \in \mathcal{A}(M),$$

where $P(\mathcal{A}(M))$ is the probability that the original process is in the set $\mathcal{A}(M)$. Since $p_{m,n}(M)$ equals the sum (2.70) up to a normalizing constant, this finally proves theorem 2.33. □

2.13. Comment on condition 2.24

In this section it is shown by using a sequence of ε-perturbed processes how the probabilities $p_{m,n}$ of the process in figure 2.7 violating condition 2.22 may be obtained by a limiting procedure. In the resulting expression for $p_{m,n}$ products of the form $m\alpha^m\beta^n$ and $n\alpha^m\beta^n$ appear.

For the process in figure 2.7 the construction of $x_{m,n}(\alpha_+)$ fails due to a vanishing denominator in the definition of c_1. The construction of $x_{m,n}(\beta_+)$, however, succeeds. $N = 0$, since $h_{1,1} = h_{0,1} = \nu_{1,1} = \nu_{1,0} = 0$, so $x_{m,n}(\beta_+)$ converges for all $m + n > 0$ and satisfies all equilibrium equations, except in the states (0, 0), (0, 1), (1, 0) and (1, 1). It remains to define $x_{0,0}(\beta_+)$ for which we can use definition (2.58) or (2.59). If we specify $r_{1,1} = h_{1,1}$ and $r_{1,0} = h_{1,0}$, then equations (2.7) and (2.8) are identical to (2.4) and (2.5) for $m = 1$. Hence, if $x_{0,0}(\beta_+)$ is defined by (2.59), then $x_{m,n}(\beta_+)$ also satisfies the equilibrium equations in (1, 1) and (1, 0). The equation in (0, 0) is given by

$$p_{0,0}(1 + r_{0,1}) = p_{0,1}, \tag{2.71}$$

which may be satisfied by $x_{m,n}(\beta_+)$ for some special $r_{0,1}$. In that case, the equation in (0, 1) is also satisfied, due to the dependence of the equilibrium equations, so $x_{m,n}(\beta_+)$ can be normalized to produce the equilibrium distribution. However, we assume that $r_{0,1}$ is such that equation (2.71) is violated by $x_{m,n}(\beta_+)$, i.e.,

$$x_{0,0}(\beta_+)(1 + r_{0,1}) - x_{0,1}(\beta_+) \neq 0. \tag{2.72}$$

To find the equilibrium distribution though, we proceed as follows.

Perturb the vertical boundary behaviour of the process in figure 2.7 by adding some small $\varepsilon > 0$ to $\nu_{0,1} = 1/2$ (see figure 2.10). For this ε-perturbed process the coefficients in the

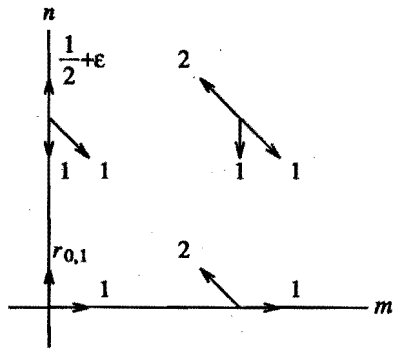


Figure 2.10.
Transition-rate diagram of the ε-perturbed Markov process.

solutions $x_{m,n}(\alpha_+)$ and $x_{m,n}(\beta_+)$ depend on ε. Moreover, since β_+ depends on ε, the parameters α_i and β_i in $x_{m,n}(\beta_+)$ also depend on ε. We write $\beta_+(\varepsilon)$, $x_{m,n}(\alpha_+, \varepsilon)$ and $x_{m,n}(\beta_+(\varepsilon), \varepsilon)$ to indicate the dependence of ε. Since $\beta_+(\varepsilon) > \beta_+$ (see remark 2.25), the construction of $x_{m,n}(\alpha_+, \varepsilon)$ now succeeds. However, as $\varepsilon \downarrow 0$,

$$c_1(\varepsilon) = -\frac{\varepsilon + 7/6}{\varepsilon} \downarrow -\infty .$$

Therefore it is sensible to *rescale* $x_{m,n}(\alpha_+, \varepsilon)$, by taking instead of $d_0 = 1$,

$$d_0(\varepsilon) = \frac{1}{c_1(\varepsilon)} = -\frac{\varepsilon}{\varepsilon + 7/6} , \tag{2.73}$$

from which it follows that

$$x_{m,n}(\alpha_+, 0) = x_{m,n}(\beta_+) . \tag{2.74}$$

Since $N = 0$ the series $x_{m,n}(\alpha_+, \varepsilon)$ and $x_{m,n}(\beta_+(\varepsilon), \varepsilon)$, which may be defined in the origin by (2.57) and (2.59) respectively, converge absolutely for all $m \geq 0$ and $n \geq 0$, and satisfy all equilibrium equations, except in (0, 0) and (0, 1). The equation in (0, 0) (and then also in (0, 1)) can be satisfied by

$$x_{m,n}(\alpha_+, \varepsilon) + k(\varepsilon) x_{m,n}(\beta_+(\varepsilon), \varepsilon)) , \tag{2.75}$$

where, from (2.71),

$$k(\varepsilon) = -\frac{x_{0,0}(\alpha_+, \varepsilon)(1 + r_{0,1}) - x_{0,1}(\alpha_+, \varepsilon)}{x_{0,0}(\beta_+(\varepsilon), \varepsilon)(1 + r_{0,1}) - x_{0,1}(\beta_+(\varepsilon), \varepsilon)} . \tag{2.76}$$

It is readily verified for the sequence $\{\alpha_i(\varepsilon), \beta_i(\varepsilon)\}$ associated with $x_{m,n}(\beta_+(\varepsilon), \varepsilon)$ that

$$\alpha_i(\varepsilon) \to 0\,,\ \beta_i(\varepsilon) \to 0 \quad (i \to -\infty\,,\ \text{uniformly for small } \varepsilon \geq 0)\,. \tag{2.77}$$

By use of (2.77) it can be proved that the series $x_{m,n}(\beta_+(\varepsilon), \varepsilon)$ converges uniformly for small $\varepsilon \geq 0$, and hence, is continuous in ε. The similar properties can be proved for $x_{m,n}(\alpha_+, \varepsilon)$. Then, by (2.72), the denominator in (2.76) does not vanish for small ε and by (2.74) the coefficient $k(\varepsilon) \to -1$ as $\varepsilon \downarrow 0$. Hence (2.75) tends to the null solution as $\varepsilon \downarrow 0$. Therefore we have to investigate higher order terms. Since $\beta_0(\varepsilon) = \beta_+(\varepsilon)$ is differentiable (cf. (2.64)) and for $i \leq 0$

$$\alpha_i(\varepsilon) = Y_-(\beta_i(\varepsilon))\,, \quad \beta_{i-1}(\varepsilon) = Y_-(\alpha_i(\varepsilon))\,,$$

it follows by induction that the parameters $\alpha_i(\varepsilon)$ and $\beta_i(\varepsilon)$ in $x_{m,n}(\beta_+(\varepsilon), \varepsilon)$ are differentiable for all $i \leq 0$. Hence, each term in $x_{m,n}(\beta_+(\varepsilon), \varepsilon)$ can be differentiated with respect to ε. Introduce

$$x_{m,n}'(\beta_+(\varepsilon), \varepsilon) = \text{term-by-term derivative of } x_{m,n}(\beta_+(\varepsilon), \varepsilon) \text{ with respect to } \varepsilon\,.$$

In the series for $x_{m,n}'(\beta_+(\varepsilon), \varepsilon)$ terms will appear of the form $m\alpha^m\beta^n$ and $n\alpha^m\beta^n$. It can be proved that for fixed ε the series $x_{m,n}'(\beta_+(\varepsilon), \varepsilon)$ converges absolutely for all $m \geq 0$ and $n \geq 0$ and that its sum over all $m \geq 0$ and $n \geq 0$ converges absolutely. Furthermore, by use of (2.77) it can be proved that for all $m \geq 0$ and $n \geq 0$ the series $x_{m,n}'(\beta_+(\varepsilon), \varepsilon)$ is uniformly convergent for small $\varepsilon \geq 0$, from which it follows that $x_{m,n}(\beta_+(\varepsilon), \varepsilon)$ is differentiable with respect to ε and its derivative can be obtained by differentiating term-by-term. The similar properties can be proved for the term-by-term derivative $x_{m,n}'(\alpha_+, \varepsilon)$. If we denote the derivative of $k(\varepsilon)$ by $k'(\varepsilon)$, then

$$x_{m,n}(\alpha_+, \varepsilon) + k(\varepsilon)x_{m,n}(\beta_+(\varepsilon), \varepsilon) = \Big[x_{m,n}'(\alpha_+,0) + k(0)x_{m,n}'(\beta_+(0),0) + k'(0)x_{m,n}(\beta_+(0),0)\Big]\varepsilon$$

$$+ o(\varepsilon), \quad (\varepsilon \downarrow 0)\,. \tag{2.78}$$

By letting $\varepsilon \downarrow 0$ in (2.78), we conclude that the sum

$$x_{m,n}'(\alpha_+, 0) + k(0)x_{m,n}'(\beta_+(0), 0) + k'(0)x_{m,n}(\beta_+(0), 0) \tag{2.79}$$

satisfies all equilibrium equations of the original process ($\varepsilon = 0$). For $m \to \infty$ and fixed $n > 0$ the dominating term in (2.79) is the first term in $x_{m,n}'(\alpha_+, 0)$ which is given by (see (2.73))

$$d_0'(0)\alpha_0^m\beta_0^n = \frac{6}{7}\alpha_+^m X_-^n(\alpha_+)\,.$$

This proves that (2.79) is nonnull. Hence, we can finally conclude that the sum (2.79) is an absolutely convergent and nonnull solution of all equilibrium equations of the original process, so normalization of this solution produces the desired equilibrium distribution.

2.14. Comment on assumption 2.1

In this section we comment on the cases that are initially excluded by assumption 2.1 in section 2.1. We first consider the case that part (i) of assumption 2.1 is violated, so

$q_{1,1} + q_{1,0} + q_{1,-1} = 0$ (there is no rate component to the east),

where, to avoid trivialities, we also assume that

$h_{1,1} + h_{1,0} > 0$.

Then it can be proved under certain drift conditions that for all $m > 1$ the probabilities $p_{m,n}$ can be expressed as a linear combination of the initial products $\alpha_0^m \beta_0^n$ which can be fitted to the horizontal boundary ($d_{-1} = 0$). This can be established by restricting the Markov process to the set of states $\{(m, n) | m > 1, n \geq 0\} \cup \{(1, 0)\}$, and then proceeding in the same way as in sections 2.4 and 2.5. To be able to restrict the process to the above set of states we need to know that as soon as the process enters the set of states with $m \leq 1$, then the expected time to return to (1, 0) is *finite*; a necessary and sufficient condition for this can be derived from Neuts' mean drift condition ([51], theorem 1.7.1.).

Now we consider the case that part (ii) of assumption 2.1 is violated, so

$q_{-1,1} + q_{-1,0} + q_{-1,-1} = 0$ (there is no rate component to the west),

where, to avoid trivialities, we also assume that

$h_{-1,1} + h_{-1,0} > 0$.

This case is illustrated by the following example.

Example 2.34 (the longer queue model).

Consider a system consisting of two queues that are served by one server. The service times are exponentially distributed with unit mean. The server always works on the longer queue and treats the jobs in the longer queue with preemptive priority with respect to the jobs in the shorter queue. In each queue jobs arrive according to a Poisson stream with intensity $\rho/2$. This model is known as the longer queue model. The state space consists of the pairs (m, n), $m, n = 0, 1, \ldots$ where m is the length of the shorter queue and $m+n$ the length of the longer queue. Jobs in service are also counted as being in queue. The transition rates are depicted in figure 2.11.

Now equation (2.9) reduces to a linear equation in α. Therefore, the generation of compensation terms fails and, due to the vertical boundary conditions, the probabilities $p_{m,n}$ cannot be expressed as a linear combination of the initial products $\alpha_0^m \beta_0^n$ which can be fitted to the horizontal boundary. Several alternatives to the compensation approach are available for

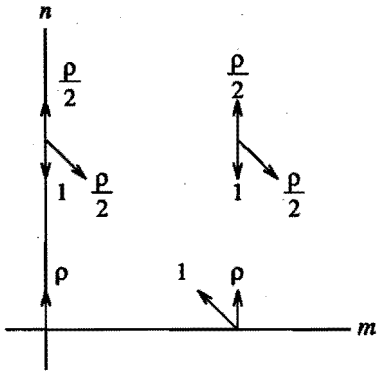


Figure 2.11.
Transition rates for the longer queue model.

solving the longer queue model. In fact, this problem has been extensively studied by Zheng and Zipkin [64] and by Flatto [25]. The method of Zheng and Zipkin consists of directly solving the equilibrium equations. First the probabilities $p_{0,n}$ are solved from the equilibrium equations for $m = 0$ which form a second order *homogeneous* recursion relation. For $n \geq 0$ this yields

$$p_{0,n} = \lambda^{n-1} \rho p_{0,0} , \tag{2.80}$$

where

$$\lambda = \frac{\rho + 1 - \sqrt{\rho^2 + 1}}{2}$$

and, from a balance argument, $p_{0,0} = 1 - \rho$. Inserting (2.80) into the equilibrium equations for $m = 1$ leads to an *inhomogeneous* recursion relation for $p_{1,n}$, which is solved by

$$p_{1,n} = (an + b)\lambda^n , \quad n > 0 , \tag{2.81}$$

where

$$a = \frac{\lambda\rho(1-\rho)}{\rho/2 - \lambda^2}$$

and b is an arbitrary constant. Next $p_{1,0}$ can be determined from the equilibrium equation in state (0, 1) and then, by inserting (2.81), the constant b follows from the equation in (1, 0). By repeating this procedure for $m = 2, 3, \ldots$ and using induction it can be proved that $p_{m,n}$ is of the form

$$p_{m,n} = \lambda^n P_m(n),$$

where $P_m(n)$ is a polynomial in n of degree m. So the probabilities do not have a geometric form. The method of Flatto consists of transforming the equilibrium equations to a functional equation for the probability generating function $F(x, y)$. The functional equation for this model can easily be solved explicitly (see sections 2 and 3 in [25]). We have assumed that jobs in the longer queue are treated with preemptive priority with respect to the jobs in the shorter queue. The longer queue model, where jobs in the longer queue are treated with nonpreemptive priority, has been studied by Cohen [15]. He treats the case of general service time distributions and reduces the relevant functional equation to a Riemann type boundary value problem.

Finally we point out that the parts (v) and (vi) of assumption 2.1 can be relaxed; it suffices that at least one of the axes is reflecting.

2.15. Conclusion

In this chapter we applied the compensation approach to two-dimensional Markov processes on the lattice in the positive quadrant of $\mathbb{R}^2$. We considered Markov processes for which the transition rates are constant in the interior points and also constant on the two axes. To simplify the analysis, we assumed that the transitions are restricted to neighbouring states. We characterized the structural properties of Markov processes for which the compensation method can be used: there may be no transitions possibilities from the interior points to the north, nort-east and east. The compensation approach does not work for processes which do not have this property. In some cases (like two independent $M \mid M \mid 1$ queues) there is no compensation needed. In other cases (like two $M \mid M \mid 1$ queues with coupled arrivals, cf. [24,45], and the coupled processor problem, cf. [14,20,47]) compensation is needed, but it would not work. Indeed, the solutions become essentially more complicated for these cases. In the next two chapters we give a complete treatment of two queueing problems as an application of the theory developed in this chapter. In doing so, special attention is devoted to extra properties of these problems.

The emphasis in this chapter was on the development of analytical results. The results are obtained by a numerically-oriented approach and therefore can easily be exploited for numerical purposes. Numerical procedures have not been worked out for the general model in this chapter. For the queueing problems in the two subsequent chapters it will be shown that the compensation approach indeed leads to efficient and accurate numerical algorithms for the calculation of the equilibrium probabilities or other quantities of interest, such as mean waiting times and mean queue lengths. Moreover, these algorithms have the advantage that tight error bounds can be given.

Chapter 3

The symmetric shortest queue problem

In the previous chapter we studied a class of two-dimensional Markov processes on the lattice in the positive quadrant of $\mathbb{R}^2$. We explored under which conditions the compensation approach works. It appeared that the processes for which this approach works are characterized by the property that transitions from state (m, n) with $m > 0$, $n > 0$ to any of the neighbouring states $(m, n+1)$, $(m+1, n+1)$ or $(m+1, n)$ are not allowed. In this chapter we treat the symmetric shortest queue problem as an application of the theory in chapter 2. This problem is characterized as follows. Jobs arrive according to a Poisson stream at a system consisting of two identical parallel servers. The jobs require exponentially distributed service times. On arrival a job joins the shortest queue and, if queues have equal length, joins either queue with probability 1/2. This problem can be formulated as a Markov process satisfying the condition on the transition possibilities just mentioned. Therefore, we may apply the compensation method leading to an explicit characterization of the equilibrium probabilities. In section 1.1 we have shown that this problem can also be treated as a direct application of the compensation approach, i.e., without use of the general theory of chapter 2. All details of this dircet applications are worked out in [3].

The symmetric shortest problem has been addressed by many authors. Haight [34] introduced the problem. Kingman [44] and Flatto and McKean [23] analyse the problem by generating functions. Using a uniformization approach they show that the generating function for the equilibrium distribution of the lengths of the two queues is a meromorphic function and they find explicit relations for the poles and residues. Then, by partial fraction decomposition of the generating function, it follows that the equilibrium probabilities can be expressed as an infinite linear combination of product forms. However, the decomposition leads to cumbersome formulae for the equilibrium probabilities. Another analytic approach is given in Cohen and Boxma [14] and Fayolle and Iasnogorodski [19,20,40]. They show that the analysis of the symmetric shortest queue problem can be reduced to that of a Riemann-Hilbert boundary value problem. The approaches mentioned, however, do not lead to an explicit characterization of the equilibrium probabilities. The main advantage of the compensation approach over the analytic results of Kingman [44] and Flatto and McKean [23] is that the compensation method yields explicit relations for the coefficients in the infinite linear combination of product forms and thereby an explicit characterization of the equilibrium probabilities.

So far, the available analytic results, though mathematically elegant, offered no practical means for computing the performance characteristics and therefore didn't close the matter in this respect. For this reason, many numerical studies have appeared on the present problem. Most studies, however, deal with the evaluation of *approximating* models. For instance, Gertsbakh [29], Grassmann [30], Rao and Posner [52] and Conolly [16] treat the shortest queue problem by truncating one or more state variables. Using linear programming, Halfin [35] obtains upper and lower bounds for the queue length distribution. Foschini and Salz [26] obtain heavy traffic diffusion approximations for the queue length distribution. Knessl, Matkowsky, Schuss and Tier [46] derive asymptotic expressions for the queue length distribution. Based on a formula, given by Flatto and McKean [22], for the probability that there are k jobs in each queue, where $k = 0, 1, ...$, Zhao and Grassmann [32] derive a numerically stable algorithm for computing these probabilities and then use these probabilities to calculate recursively the other queue length probabilities from the equilibrium equations. Schassberger [54,55] uses an iterative method to numerically obtain approximating values for the queue length probabilities. These studies are all restricted to systems with two parallel queues. Hooghiemstra, Keane and Van de Ree [38] develop a power series method to calculate the stationary queue length distribution for fairly general multidimensional exponential queueing systems. Their method is not restricted to systems with two queues, but applies equally well to systems with more queues. As far as the shortest queue problem is concerned, Blanc [10,11] reports that the power series method is numerically satisfactory for the shortest queue system with up to 25 parallel queues. The theoretical foundation of this method is, however, still incomplete. Nelson and Philips [50] derive an approximation for the mean response time for the shortest queue system with multiple queues. They report that their approximation has a relative error of less than 2 percent for systems with at most 16 queues and with service utilizations over the range from 0 to 0.99. Finally, a common disadvantage of the numerical methods mentioned is that in general no error bounds can be given.

Since the compensation method is constructive in nature, the analytical results can easily be exploited for numerical purposes. It appears that these results offer an efficient numerical procedure, with tight bounds on the error of each partial sum. Also, expressions are obtained for the first and second moment of the waiting time, which are suitable for numerical evaluation. *These algorithms apply to the exact model.*

Many authors addressed the problem of proving that the shortest queue policy is optimal. Winston [61] studies the model with c identical exponential servers, infinite buffers and Poisson arrivals. He proves that the shortest queue policy maximizes stochastically the number of jobs served by any time t. Hordijk and Koole [39] extend Winston's results to systems allowing finite buffers and batch arrivals. Moreover, they consider general arrival processes. Ephremides, Varaiya and Walrand study the model with $c = 2$ and prove that the shortest queue policy

minimizes the sum of the expected sojourn times of all jobs arriving before a certain time t.

This chapter is organized as follows. In section 3.1 we formulate the model and the equilibrium equations. In section 3.2 we apply the general theory of chapter 2 to this model. It appears that only the feasible pair $(\alpha_+, X_-(\alpha_+))$ plays a role. We prove that for all m, n the stationary queue length probabilities $p_{m,n}$ can be expressed as $x_{m,n}(\alpha_+)$ up to some normalizing constant C. The general theory however, does not explicitly yield α_+ and C. In sections 3.3 and 3.4 it is shown that for this problem α_+ and C can be found explicitly. Monotonicity properties of the terms in the series for $x_{m,n}(\alpha_+)$ are derived in section 3.5, leading to bounds on the error of each partial sum of $x_{m,n}(\alpha_+)$. An asymptotic expansion of $x_{m,n}(\alpha_+)$ as $m + n \to \infty$ is given in section 3.6. Product form expressions for global performance measures are presented in section 3.7, and section 3.8 presents some numerical results. In section 3.9 we develop a recursive algorithm to numerically compute the stationary queue length probabilities. Sections 3.10 and 3.11 deal with some simple variants of the symmetric shortest queue problem. The final section is devoted to conclusions.

3.1. Model and equilibrium equations

Consider a system with two identical servers (see figure 3.1). Jobs arrive according to a Poisson stream with rate 2ρ where $0 < \rho < 1$. On arrival a job joins the shortest queue. Ties are broken with equal probabilities. The jobs require exponentially distributed service times with unit mean, the service times are supposed to be independent. This model is known as the symmetric shortest queue model.

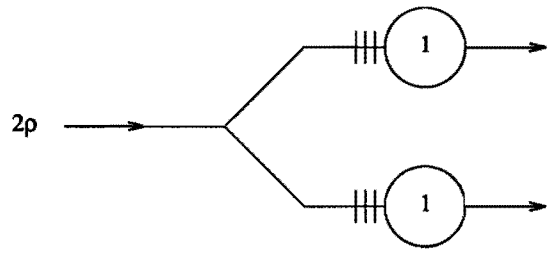


Figure 3.1.
The symmetric shortest queue model. Arriving jobs join the shortest queue. Ties are broken with equal probabilities. It is assumed that $0 < \rho < 1$.

This queueing system can be represented by a continuous-time Markov process whose natural state space consists of the pairs (i, j) where i and j are the lengths of the two queues. Instead of i and j we use the state variables m and n where $m = \min(i, j)$ and $n = j - i$. Let $\{p_{m,n}\}$ be the equilibrium distribution. The transition-rate diagram is depicted in figure 3.2. The rates in the region $n \leq 0$ can be obtained by reflection in the m-axis. By symmetry $p_{m,n} = p_{m,-n}$. Hence, the analysis can be restricted to the probabilities $p_{m,n}$ in the region $n \geq 0$.

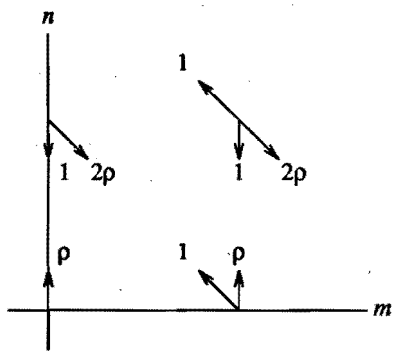


Figure 3.2.
Transition-rate diagram for the symmetric shortest queue model in figure 3.1.

The equilibrium equations for $\{p_{m,n}\}$ can be found by equating for each state the rate into and the rate out of that state. These equations are formulated below.

$$p_{m,n}2(\rho+1) = p_{m-1,n+1}2\rho + p_{m,n+1} + p_{m+1,n-1}, \qquad m > 1, n > 1 \quad (3.1)$$

$$p_{1,n}2(\rho+1) = p_{0,n+1}2\rho + p_{1,n+1} + p_{2,n-1}, \qquad n > 1 \quad (3.2)$$

$$p_{0,n}(2\rho+1) = p_{0,n+1} + p_{1,n-1}, \qquad n > 1 \quad (3.3)$$

$$p_{m,1}2(\rho+1) = p_{m-1,2}2\rho + p_{m,2} + p_{m+1,0} + p_{m,0}\rho, \qquad m > 1 \quad (3.4)$$

$$p_{m,0}(\rho+1) = p_{m-1,1}2\rho + p_{m,1}, \qquad m > 1 \quad (3.5)$$

$$p_{1,1}2(\rho+1) = p_{0,2}2\rho + p_{1,2} + p_{2,0} + p_{1,0}\rho, \qquad (3.6)$$

$$p_{1,0}(\rho+1) = p_{0,1}2\rho + p_{1,1}, \qquad (3.7)$$

$$p_{0,1}(2\rho+1) = p_{0,2} + p_{1,0} + p_{0,0}\rho, \qquad (3.8)$$

$$p_{0,0}\rho = p_{0,1}. \qquad (3.9)$$

This formulation can be simplified by observing that equation (3.2) is identical to (3.1) with $m = 1$. The reason for not doing so, is that the equations (3.1)-(3.5) correspond with the equations (2.1)-(2.5), and therefore are suited to application of the general theory of chapter 2. In the next section we investigate how the compensation approach works out here.

3.2. Application of the compensation approach

To facilitate application of the theory in chapter 2 we translate the transition rates q_{ij}, v_{ij}, h_{ij} and r_{ij} in terms of the rates in figure 3.2. From the transition-rate diagrams in figures 2.2 and 3.2 it follows that,

$$\begin{aligned}
&q_{-1,1} = 1\,, \quad q_{0,-1} = 1\,, \quad q_{1,-1} = 2\rho\,, \quad q_{-1,-1} = q_{-1,0} = q_{0,1} = q_{1,1} = q_{1,0} = 0\,; \\
&v_{0,-1} = 1\,, \quad v_{1,-1} = 2\rho\,, \qquad\qquad v_{1,0} = v_{1,1} = v_{0,1} = 0\,; \\
&h_{-1,1} = 1\,, \quad h_{0,1} = \rho\,, \qquad\qquad h_{-1,0} = h_{1,1} = h_{1,0} = 0\,; \\
&r_{0,1} = \rho\,, \qquad\qquad r_{1,1} = r_{1,0} = 0\,.
\end{aligned} \tag{3.10}$$

For the transition rates in the interior points we directly obtain

$$q_{0,1} = q_{1,1} = q_{1,0} = 0\,,$$

which is the essential condition for application of the compensation approach (cf. assumption 2.6). In general there are at most four feasible pairs. For this model it follows from the boundary behaviour that $(\alpha_+, X_-(\alpha_+))$ is the only feasible pair possible (cf. conclusion 2.14(ii)). By theorem 2.19 this pair exists if and only if condition (2.50) is satisfied. This condition is verified below. Since

$$q_{-1,-1} + q_{0,-1} + q_{1,-1} \;=\; 1 + 2\rho \;>\; 1 \;=\; q_{-1,1}\,,$$

we must have $RH'(1) > LH'(1)$. Inserting (3.10) in $RH(\alpha)$ and $LH(\alpha)$ (cf. (2.44)) yields

$$RH(\alpha) = \frac{(\alpha+\rho)(\alpha+2\rho)}{\rho + 1 + \sqrt{\rho^2 + 1 - \alpha}}\,, \tag{3.11}$$

$$LH(\alpha) = \alpha(\rho + 1)\,, \tag{3.12}$$

from which it easily follows that

$$RH'(1) = \frac{3\rho + 1}{2\rho}$$

$$LH'(1) = \rho + 1\,.$$

So $RH'(1) > LH'(1)$, since $0 < \rho < 1$. Hence, the feasible pair $(\alpha_+, X_-(\alpha_+))$ indeed exists. We now investigate convergence properties of $x_{m,n}(\alpha_+, X_-(\alpha_+))$, which may be abbreviated by

$x_{m,n}(\alpha_+)$. The integer N mentioned in theorem 2.25 is defined as the smallest nonnegative integer such that (see definition 2.28)

$$|HV|(A_1/A_2)^{N+1} < 1 . \tag{3.13}$$

Inserting (3.10) into the definitions of A_1, A_2, H and V (see lemma 2.26) yields

$$A_1 = \rho + 1 - \sqrt{\rho^2+1}\,, \quad A_2 = \rho + 1 + \sqrt{\rho^2+1}\,,$$

$$H = \frac{A_2}{A_1}\,, \qquad V = \frac{A_1^{-1}2\rho - (2\rho+1)}{A_2^{-1}2\rho - (2\rho+1)} = \frac{1-A_1}{1-A_2}\,. \tag{3.14}$$

To obtain the latter equality, we substituted the identities $2\rho = A_1A_2$ and $2(\rho+1) = A_1 + A_2$. Substitution of the expressions for A_1, A_2, H and V into (3.13) yields $N = 0$. It then follows from theorem 2.25 that the series (2.54) defining $x_{m,n}(\alpha_+)$ for $m > 0$ and $n > 0$ converges absolutely for all $m > 0$ and $n > 0$; the series (2.56) defining $x_{0,n}(\alpha_+)$ for $n > 0$ converges absolutely for all $n > 0$; the series (2.57) defining $x_{m,0}(\alpha_+)$ for $m > 0$ converges absolutely for all $m > 0$, but also for $m = 0$, so we may define $x_{0,0}(\alpha_+)$ by the series (2.57) with $m = 0$ (note that $x_{0,0}(\alpha_+)$ has not been defined in chapter 2); and finally it follows from theorem 2.25 that the sum of $x_{m,n}(\alpha_+)$ over all $m \geq 0$ and $n \geq 0$ converges absolutely. Further, by lemma 2.30, $\{x_{m,n}(\alpha_+)\}$ is nonnull.

$\{x_{m,n}(\alpha_+)\}$ satisfies the conditions (3.1)-(3.5). Since $v_{1j} = q_{1j}$, we have $e_i = c_i + c_{i+1}$ for all i, so the series (2.56) defining $x_{0,n}(\alpha_+)$ for $n > 0$ is identical to (2.54) with $m = 0$ (cf. remark 2.4). Hence, it easily follows that $\{x_{m,n}(\alpha_+)\}$ also satisfies (3.6)-(3.7). Below it is shown that $\{x_{m,n}(\alpha_+)\}$ also satisfies (3.8)-(3.9). First, we rewrite (3.8) as

$$p_{0,1}(2\rho+1) - p_{0,2} = p_{1,0} + p_{0,0}\rho\,. \tag{3.15}$$

Inserting the series (2.56) into the left-hand side yields

$$x_{0,1}(\alpha_+)(2\rho+1) - x_{0,2}(\alpha_+) = \sum_{i=0}^{\infty} d_i(e_i\beta_i(1+2\rho) - e_i\beta_i^2)\,. \tag{3.16}$$

$\{c_i\}$ and $\{e_i\}$ are such that for all i the terms $(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n$ and $e_i\beta_i^n$ satisfy (3.3). Hence, substituting these terms into (3.3) leads to

$$e_i\beta_i^n(1+2\rho) - e_i\beta_i^{n+1} = (c_i\alpha_i + c_{i+1}\alpha_{i+1})\beta_i^{n-1}\,.$$

Dividing both sides of this equality by β_i^{n-1} and then inserting into (3.16) yields

$$x_{0,1}(\alpha_+)(2\rho+1)-x_{0,2}(\alpha_+)=\sum_{i=0}^{\infty} d_i(c_i\alpha_i+c_{i+1}\alpha_{i+1})$$

$$=d_0c_0\alpha_0+\sum_{i=0}^{\infty} c_{i+1}(d_i+d_{i+1})\alpha_{i+1}\,, \tag{3.17}$$

where equality of the two series follows from theorem 2.25(i) with $N=0$. On the other hand, inserting the series (2.57) into the right-hand side of equation (3.15) yields

$$x_{1,0}(\alpha_+)+x_{0,0}(\alpha_+)\rho=f_0(\rho+\alpha_0)+\sum_{i=0}^{\infty} f_{i+1}(\rho+\alpha_{i+1})\,. \tag{3.18}$$

$\{f_i\}$ and $\{d_i\}$ are such that for all i the terms $(d_i\beta_i^n+d_{i+1}\beta_{i+1}^n)\alpha_{i+1}^m$ and $f_{i+1}\alpha_{i+1}^m$ satisfy (3.4). Hence, substituting these terms into (3.4) gives

$$f_{i+1}\alpha_{i+1}^m\rho+f_{i+1}\alpha_{i+1}^{m+1}=(d_i\beta_i+d_{i+1}\beta_{i+1})\alpha_{i+1}^m 2(\rho+1)$$

$$-(d_i\beta_i^2+d_{i+1}\beta_{i+1}^2)\alpha_{i+1}^{m-1}2\rho-(d_i\beta_i^2+d_{i+1}\beta_{i+1}^2)\alpha_{i+1}^m\,.$$

Dividing this equality by α_{i+1}^{m-1} and inserting into the right-hand side the quadratic equation (2.9), which by use of (3.10) simplifies to

$$\alpha\beta 2(\rho+1)=\alpha^2+\beta^2 2\rho+\alpha\beta^2\,,$$

we obtain

$$f_{i+1}(\rho+\alpha_{i+1})=(d_i+d_{i+1})\alpha_{i+1}\,. \tag{3.19}$$

This relation reduces the right-hand side of (3.18) to (3.17) (note that $d_{-1}=0$). So $x_{m,n}(\alpha_+)$ does indeed satisfy (3.8). The remaining equation (3.9) is also satisfied by $x_{m,n}(\alpha_+)$, since inserting $x_{m,n}(\alpha_+)$ into the equations for $m+n>0$ and then summing these equations and changing summations, exactly yields equation (3.9). Changing summations is allowed by the absolute convergence stated in theorem 2.25(iv).

Now we can finally conclude that $\{x_{m,n}(\alpha_+)\}$ is a nonnull absolutely convergent solution of all equilibrium equations. Hence, by a result of Foster (see appendix A), the Markov process is ergodic and normalization of $\{x_{m,n}(\alpha_+)\}$ produces $\{p_{m,n}\}$. Before summarizing the results we restate the definition of $x_{m,n}(\alpha_+)$, which for the present model simplifies considerably.

Since $v_{1j}=q_{1j}$, we have $e_i=c_i+c_{i+1}$ for all i, so the series (2.56) defining $x_{0,n}(\alpha_+)$ for $n>0$ is identical to (2.54) with $m=0$ (cf. remark 2.4). Hence,

$$x_{m,n}(\alpha_+)=\sum_{i=0}^{\infty} d_i(c_i\alpha_i^m+c_{i+1}\alpha_{i+1}^m)\beta_i^n\,,\quad m\geq 0\,,\ n>0\,; \tag{3.20}$$

$$x_{m,0}(\alpha_+)=c_0f_0\alpha_0^m+\sum_{i=0}^{\infty} c_{i+1}f_{i+1}\alpha_{i+1}^m\,,\quad m\geq 0\,. \tag{3.21}$$

Insertion of (3.10) into (2.20)-(2.21) leads to the following recursion relations for c_i and d_i. We simplified the recursion relation for c_i by using the relations for $\alpha_i\alpha_{i+1}$ and $\alpha_i + \alpha_{i+1}$ (cf. (2.23)-(2.24)). The equation for f_{i+1} directly follows from (3.19).

$$c_{i+1} = -\frac{\beta_i - \alpha_{i+1}}{\beta_i - \alpha_i} c_i \qquad (i = 0, 1, \cdots), \tag{3.22}$$

$$d_{i+1} = -\frac{\dfrac{\alpha_{i+1}^2 + \alpha_{i+1}\rho}{\beta_{i+1}} - \alpha_{i+1}(\rho + 1)}{\dfrac{\alpha_{i+1}^2 + \alpha_{i+1}\rho}{\beta_i} - \alpha_{i+1}(\rho + 1)} d_i \qquad (i = 0, 1, \cdots), \tag{3.23}$$

$$f_{i+1} = (d_i + d_{i+1})\frac{\alpha_{i+1}}{\rho + \alpha_{i+1}} \qquad (i = 0, 1, \cdots), \tag{3.24}$$

with $c_0 = d_0 = 1$ and, since $d_{-1} = 0$,

$$f_0 = d_0 \frac{\alpha_0}{\rho + \alpha_0}. \tag{3.25}$$

We summarize our findings in the following theorem.

Theorem 3.1.

For all $m \geq 0, n \geq 0$,

$$p_{m,n} = C^{-1} x_{m,n}(\alpha_+),$$

where C is the normalizing constant.

Theorem 3.1 is incomplete in the sense that the theory in chapter 2 does not yield *explicit expressions* for α_+ and C. However, for the present problem α_+ and C can be derived explicitly. This will be shown in the next two sections.

3.3. Explicit determination of α_+

To determine α_+ explicitly, consider the process on the *aggregate states* k, where k denotes the total number of jobs in the system (so $k = 2m + |n|$). Let P_k be the probability that there are k jobs in the system. The average rate from state k to $k+1$ is given by the arrival intensity 2ρ. The average rate from state $k+1$ to state k is obtained by observing that the service rate

in state $k+1$ is 2, except when all jobs are in one queue only, in which case the service rate is 1. Hence, the average rate from state $k+1$ to state k is given by

$$2 - \frac{p_{0,k+1} + p_{0,-k-1}}{P_{k+1}} = 2 - \frac{2p_{0,k+1}}{P_{k+1}} .$$

The transition-rate diagram is depicted in figure 3.3.

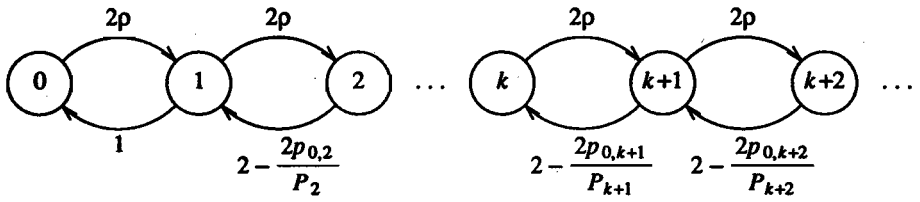


Figure 3.3.
The transition-rate diagram for the aggregate states k where k is the total number of jobs in the system and P_k is the probability that there are k jobs in the system.

Balancing the rates in figure 3.3 gives for all $k \geq 0$,

$$\frac{P_{k+1}}{P_k} = \frac{\rho}{1 - p_{0,k+1}/P_k} . \tag{3.26}$$

To obtain a relation for $\alpha_0 = \alpha_+$ we let $k \to \infty$ in the equality (3.26). Then we first need the asymptotic behaviour of P_k. The probabilities P_k can be expressed in terms of the detailed probabilities $p_{m,n}$. This yields for all $k \geq 0$,

$$P_{2k} = p_{k,0} + 2\sum_{l=1}^{k} p_{k-l,2l} ,$$

$$P_{2k+1} = 2\sum_{l=0}^{k} p_{k-l,2l+1} .$$

Hence, the asymptotics of P_k as $k \to \infty$ can be derived from that of $p_{m,n}$ as $m + n \to \infty$.

Below we prove that the asymptotic behaviour of $p_{m,n} = C^{-1} x_{m,n}(\alpha_+)$ is determined by the first term in the series $x_{m,n}(\alpha_+)$, viz.

$$\frac{x_{m,n}(\alpha_+)}{d_0(c_0\alpha_0^m + c_1\alpha_1^m)\beta_0^n} \to 1 , \quad \text{as} \quad m+n \to \infty , n > 0 , \tag{3.27}$$

$$\frac{x_{m,0}(\alpha_+)}{c_0 f_0 \alpha_0^m} \to 1 , \quad \text{as} \quad m \to \infty , \tag{3.28}$$

i.e., we have the following asymptotic equivalences

$$x_{m,n}(\alpha_+) \sim d_0(c_0\alpha_0^m + c_1\alpha_1^m)\beta_0^n\,, \quad (m+n \to \infty,\ n>0)\,, \tag{3.29}$$

$$x_{m,0}(\alpha_+) \sim c_0 f_0 \alpha_0^m\,, \qquad (m \to \infty)\,. \tag{3.30}$$

First, since (see (2.60) in section 2.8)

$$1 > \alpha_0 > \beta_0 > \alpha_1 > \beta_1 > \cdots > 0\,, \tag{3.31}$$

it follows from (3.22) that $c_1 > 0$. Hence, $d_0(c_0\alpha_0^m + c_1\alpha_1^m)\beta_0^n$ and $f_0\alpha_0^m$ are positive, so the quotients (3.27) and (3.28) are well defined. From (3.31) we obtain that for $m \geq 0$, $n > 0$,

$$\begin{aligned} |x_{m,n}(\alpha_+) - d_0(c_0\alpha_0^m + c_1\alpha_1^m)\beta_0^n| &\leq \sum_{i=1}^{\infty} |d_i|(|c_i|\alpha_i^m + |c_{i+1}|\alpha_{i+1}^m)\beta_i^n \\ &\leq \alpha_1^m\beta_1^{n-1}\sum_{i=1}^{\infty} |d_i|(|c_i| + |c_{i+1}|)\beta_i\,, \end{aligned}$$

where, since $N = 0$, the latter series converges by theorem 2.25(i) with $m = 0$ and $n = 1$. This inequality yields the asymptotic equivalence (3.29) by observing that $\alpha_1 < \alpha_0$ and $\beta_1 < \beta_0$. The asymptotic equivalence (3.30) can be established similarly. Inserting (3.29) and (3.30) in the expression for P_{2k} yields

$$P_{2k} \sim C^{-1}\left\{ c_0 f_0 \alpha_0^k + 2\sum_{l=1}^{k} d_0(c_0\alpha_0^{k-l} + c_1\alpha_1^{k-l})\beta_0^{2l} \right\}, \quad (k \to \infty)\,.$$

By inserting the identity

$$\sum_{l=1}^{k} x^{k-l}y^l = y\,\frac{x^k - y^k}{x - y}$$

in this expression and using that $1 > \alpha_0 > \beta_0 > \alpha_1 > 0$, the asymptotic formula for P_{2k} simplifies to

$$P_{2k} \sim L\alpha_0^k\,, \qquad (k \to \infty)\,,$$

where the constant L is given by

$$L = C^{-1}\left[c_0 f_0 + d_0 c_0 \frac{\beta_0^2}{\alpha_0 - \beta_0^2} \right],$$

which is independent of k. Similarly, we obtain

$$P_{2k+1} \sim M\alpha_0^k\,, \quad (k \to \infty)\,,$$

for some constant M independent of k. Now we have all ingredients to find $\alpha_0 = \alpha_+$ explicitly. First, since $0 < \beta_0 < \alpha_0 < 1$, it follows from the asymptotic formulas for $p_{0,k+1}$ and P_{k+1} that

$p_{0,k+1} / P_{k+1} \to 0$ as $k \to \infty$. Hence, from (3.26) we obtain

$$\frac{P_{k+1}}{P_k} \sim \rho , \qquad (k \to \infty) .$$

Combining this relation for $2k$ and $2k+1$, we find

$$\frac{P_{2k+2}}{P_{2k}} \sim \rho^2 , \quad (k \to \infty) . \tag{3.32}$$

Finally, substitution of the asymptotic formula for P_{2k} into (3.32) yields

$$\alpha_0 = \alpha_+ = \rho^2 .$$

It is easily verified that $\alpha = \rho^2$ does indeed satisfy the equation $RH(\alpha) = LH(\alpha)$ (see (3.11) and (3.12)). This concludes the determination of α_+. In the next section we determine C.

3.4. Explicit determination of the normalizing constant

In this section we derive an explicit formula for the normalizing constant C, which however, is *not essential* to the compensation method itself: substitution of the series (3.20) and (3.21), defining $x_{m,n}(\alpha_+)$, into the normalization equation

$$C = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} x_{m,n}(\alpha_+)$$

leads to a series of product forms for C, analogous to the series for $x_{m,n}(\alpha_+)$. The method to obtain the explicit formula, by means of the generating function, is different from the main arguments in this thesis. Therefore we omit details and only sketch the proof. Define the generating function $F(y, z)$ by

$$F(y, z) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} p_{m,n} y^m z^n$$

$$= C^{-1} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} x_{m,n}(\alpha_+) y^m z^n .$$

Substituting of (3.20) and (3.21) in this expression and then changing summations leads to,

$$F(y, z) = C^{-1} \left\{ \sum_{i=0}^{\infty} d_i \left[\frac{c_i}{1-\alpha_i y} + \frac{c_{i+1}}{1-\alpha_{i+1} y} \right] \frac{\beta_i z}{1-\beta_i z} + \frac{c_0 f_0}{1-\alpha_0 y} + \sum_{i=0}^{\infty} \frac{c_{i+1} f_{i+1}}{1-\alpha_{i+1} y} \right\}, \tag{3.33}$$

valid in $|y| < 1/\alpha_0$, $|z| < 1/\beta_0$. The partial fraction decomposition of the generating function is difficult to obtain, at least in this explicit form, from the analysis of Kingman [44] and Flatto and McKean [23]. The equilibrium equations (3.1)-(3.9) reduce to the following functional equation for $F(y, z)$,

$$F(y, z)g(y, z) = F(y, 0)h(y, z) + F(0, z)k(y, z),$$

where

$$g(y, z) = z^2 + y(2\rho y + 1) - 2(\rho + 1)yz,$$

$$h(y, z) = y(2\rho y + 1) - (\rho + 1)yz - \rho y z^2,$$

$$k(y, z) = z(z - y).$$

It follows that, if y and z satisfy $|y| < 1/\alpha_0$, $|z| < 1/\beta_0$ and $g(y, z) = 0$, then $F(y, 0)$ and $F(0, z)$ are related by

$$F(y, 0)h(y, z) + F(0, z)k(y, z) = 0. \tag{3.34}$$

In the analysis of Kingman [44] and Flatto and McKean [23] this relationship between $F(y, 0)$ and $F(0, z)$ eventually leads to the determination of these functions. We use it to establish that

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

First, note that $F(0, 1)$ is the fraction of time server 1 (or 2) is idle. Since 2ρ is the offered load, we obtain, by symmetry, that

$$F(0, 1) = 1 - \rho.$$

Starting with $F(0, 1)$, we subsequently apply relationship (3.34) to the pairs $(y, z) = (1/2\rho, 1)$ and $(1/2\rho, 1/\rho)$, both satisfying $g(y, z) = 0$. This leads to

$$F(0, 1/\rho) = (1 - \rho)(2 - \rho). \tag{3.35}$$

Next, we apply (3.34) to (y, z) satisfying $g(y, z) = 0$, and let $y \uparrow 1/\rho^2$ $(= 1/\alpha_0)$ and $z \to 1/\rho$. Here, note that, by treating y as a parameter, the equation

$$g(y, z(y)) = 0, \quad z(1/\rho^2) = 1/\rho,$$

is solved by

$$z(y) = (\rho + 1)y - \sqrt{y((\rho^2 + 1)y - 1)}.$$

It is easily verified that

$$h(y, z(y)) = \frac{(2 + \rho)(\rho - 1)}{2\rho}(y - 1/\rho^2) + o(y - 1/\rho^2) \quad (y \uparrow 1/\rho^2), \tag{3.36}$$

and from (3.33) we obtain

$$F(y, 0) = C^{-1}\frac{f_0}{1 - \alpha_0 y} + O(1) \qquad (y \uparrow 1/\rho^2 = 1/\alpha_0),$$

which by insertion of (3.25) and $\alpha_0 = \rho^2$ reduces to

$$F(y, 0) = \frac{-1}{C\rho(\rho+1)(y-1/\rho^2)} + O(1) \qquad (y \uparrow 1/\rho^2) . \tag{3.37}$$

Then, inserting $z = z(y)$ into relationship (3.34) and letting $y \uparrow 1/\rho^2$, we finally obtain the desired expression for C by using (3.35)-(3.37). We end this section by restating theorem 3.1.

Theorem 3.2

For all $m \geq 0, n \geq 0,$

$$p_{m,n} = C^{-1} x_{m,n}(\alpha_+) ,$$

where $\alpha_+ = \rho^2$ *and*

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

3.5. Monotonicity of the terms in the series of products

In this section we prove that the terms in the series (3.20) defining $x_{m,n}(\alpha_+)$ for $m \geq 0$ and $n > 0$, are *alternating and monotonically decreasing in modulus.* From a numerical point of view this is a nice property, since then the error of each partial sum can be bounded by the absolute value of its final term.

In section 2.8 we have seen that for all $i \geq 0$,

$$\frac{d_{i+1}}{d_i} < 0 ,$$

and it directly follows from (3.22) and (3.31) that for all $i \geq 0$,

$$\frac{c_{i+1}}{c_i} > 0 .$$

Hence, the coefficients d_i are alternating and the coefficients c_i are positive. Since all α_i and β_i are positive, we can conclude that for $m \geq 0$ and $n > 0$ the terms $d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^m$ are alternating. To prove that these terms are decreasing in modulus, we first rewrite the recursion relation (3.23) for d_{i+1} in a more convenient form by expressing the quotient in (3.23) in terms of the ratios α_{i+1}/β_{i+1} and α_{i+1}/β_i.

Substituting (3.10) into the relations (2.25) and (2.26) yields

$$\beta_i\beta_{i+1} = \frac{\alpha_{i+1}^2}{\alpha_{i+1} + 2\rho} , \tag{3.38}$$

$$\beta_i + \beta_{i+1} = \frac{\alpha_{i+1} 2(\rho + 1)}{\alpha_{i+1} + 2\rho} . \tag{3.39}$$

Rewriting (3.38) as

$$\alpha_{i+1} = \frac{\alpha_{i+1}}{\beta_i} \frac{\alpha_{i+1}}{\beta_{i+1}} - 2\rho ,$$

and then inserting this relation into the denominator of (3.23) yields

$$\left[\frac{\alpha_{i+1}}{\beta_i} - (\rho + 1) \right] \frac{\alpha_{i+1}}{\beta_i} \frac{\alpha_{i+1}}{\beta_{i+1}} + \rho \left[2(\rho + 1) - \frac{\alpha_{i+1}}{\beta_i} \right] . \tag{3.40}$$

Combining the relations (3.38) and (3.39) leads to

$$2(\rho + 1) = \frac{\alpha_{i+1}}{\beta_i} + \frac{\alpha_{i+1}}{\beta_{i+1}} .$$

By insertion this equality into (3.40), the denominator in (3.23) finally reduces to

$$\frac{\alpha_{i+1}}{\beta_{i+1}} \left[\frac{\alpha_{i+1}}{\beta_i} - \rho \right] \left[\frac{\alpha_{i+1}}{\beta_i} - 1 \right] .$$

Since the numerator in (3.23) can be rewritten similarly, we arrive at the following recursion relation

$$d_{i+1} = - \frac{\dfrac{\alpha_{i+1}}{\beta_i} \left[\dfrac{\alpha_{i+1}}{\beta_{i+1}} - \rho \right] \left[\dfrac{\alpha_{i+1}}{\beta_{i+1}} - 1 \right]}{\dfrac{\alpha_{i+1}}{\beta_{i+1}} \left[\dfrac{\alpha_{i+1}}{\beta_i} - \rho \right] \left[\dfrac{\alpha_{i+1}}{\beta_i} - 1 \right]} d_i . \tag{3.41}$$

This relation helps in proving that the terms in (3.20) are decreasing in modulus, at least with rate $R = 4/(4 + 2\rho + \rho^2)$.

Lemma 3.3.

Let $R = 4/(4 + 2\rho + \rho^2) < 1$. *Then for all* $m \geq 0$, $n > 0$ *and* $i \geq 0$,

$$| d_{i+1}(c_{i+1}\alpha_{i+1}^m + c_{i+2}\alpha_{i+2}^m)\beta_{i+1}^n | < R \, | d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n | .$$

Proof.

We first prove the lemma for $m = 0$ and $n = 1$. By (3.22), we obtain for $i \geq 0$,

$$c_{i+1} = - \frac{1 - \dfrac{\alpha_{i+1}}{\beta_i}}{1 - \dfrac{\alpha_i}{\beta_i}} c_i ,$$

from which it follows that

$$\frac{c_{i+1}+c_{i+2}}{c_i+c_{i+1}} = -\frac{\left[1-\frac{\alpha_{i+1}}{\beta_i}\right]\left[\frac{\alpha_{i+2}}{\beta_{i+1}}-\frac{\alpha_{i+1}}{\beta_{i+1}}\right]}{\left[1-\frac{\alpha_{i+1}}{\beta_{i+1}}\right]\left[\frac{\alpha_{i+1}}{\beta_i}-\frac{\alpha_i}{\beta_i}\right]}. \tag{3.42}$$

By virtue of the lemmas 2.15 and 2.16, the ratios β_i/α_i and α_{i+1}/β_i are decreasing, so for $i \geq 0$,

$$\frac{\alpha_i}{\beta_i} \geq \frac{\alpha_0}{\beta_0} = 2+\rho, \quad 0 < \frac{\alpha_{i+1}}{\beta_i} \leq \frac{\alpha_1}{\beta_0} = \frac{2\rho}{2+\rho}. \tag{3.43}$$

Hence, from (3.41)-(3.43) it follows that

$$\frac{|d_{i+1}(c_{i+1}+c_{i+2})\beta_{i+1}|}{|d_i(c_i+c_{i+1})\beta_i|} = \frac{\left[\frac{\alpha_{i+1}}{\beta_i}\right]^2\left[\frac{\alpha_{i+1}}{\beta_{i+1}}-\rho\right]\left[\frac{\alpha_{i+1}}{\beta_{i+1}}-\frac{\alpha_{i+2}}{\beta_{i+1}}\right]}{\left[\frac{\alpha_{i+1}}{\beta_{i+1}}\right]^2\left[\rho-\frac{\alpha_{i+1}}{\beta_i}\right]\left[\frac{\alpha_i}{\beta_i}-\frac{\alpha_{i+1}}{\beta_i}\right]}$$

$$< \frac{\left[\frac{\alpha_{i+1}}{\beta_i}\right]^2}{\left[\rho-\frac{\alpha_{i+1}}{\beta_i}\right]\left[\frac{\alpha_i}{\beta_i}-\frac{\alpha_{i+1}}{\beta_i}\right]} \leq \frac{\left[\frac{\alpha_1}{\beta_0}\right]^2}{\left[\rho-\frac{\alpha_1}{\beta_0}\right]\left[\frac{\alpha_0}{\beta_0}-\frac{\alpha_1}{\beta_0}\right]} = R < 1. \tag{3.44}$$

This proves the lemma for $m = 0$ and $n = 1$. Now consider an arbitrary $m \geq 0$ and $n > 0$. Since α_i and β_i are positive and decreasing (see (3.31)) and the coefficients c_i are positive, it follows from (3.44) that for all $i \geq 0$,

$$\begin{aligned}
|d_{i+1}(c_{i+1}\alpha_{i+1}^m + c_{i+2}\alpha_{i+2}^m)\beta_{i+1}^n| &= |d_{i+1}|(c_{i+1}\alpha_{i+1}^m + c_{i+2}\alpha_{i+2}^m)\beta_{i+1}^n \\
&< |d_{i+1}|(c_{i+1}+c_{i+2})\beta_{i+1}\ \alpha_{i+1}^m\beta_i^{n-1} \\
&< R\ |d_i|(c_i+c_{i+1})\beta_i\ \alpha_{i+1}^m\beta_i^{n-1} \\
&< R\ |d_i|(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n \\
&= R\ |d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n|.
\end{aligned}$$

□

3.6. Asymptotic expansion

We now return to the asymptotic equivalence (3.29). By lemma 3.3 this result can be extended such as to yield a complete asymptotic expansion for $x_{m,n}(\alpha_+)$ in (3.20). First, since α_j and β_j are decreasing, it follows for all $j \geq 1$ that,

$$d_j(c_j\alpha_j^m + c_{j+1}\alpha_{j+1}^m)\beta_j^n = o(d_{j-1}(c_{j-1}\alpha_{j-1}^m + c_j\,\alpha_j^m)\beta_{j-1}^n)\,, \quad (m+n \to \infty,\ n > 0)\,.$$

Thus successive terms in the series (3.20) are indeed refinements. Since the terms in (3.20) are alternating and decreasing in modulus, the error of each partial sum can be bounded by the absolute value of the first term omitted. Hence, we have for all $j \geq 1$,

$$p_{m,n} = C^{-1}\sum_{i=0}^{j-1} d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n + O(d_j(c_j\alpha_j^m + c_{j+1}\alpha_{j+1}^m)\beta_j^n)\,, \quad (m+n \to \infty,\ n > 0)\,.$$

The O-formula for $j = 1$ improves on the asymptotic equivalence (3.29), since

$$C^{-1}d_0(c_0\alpha_0^m + c_1\alpha_1^m)\beta_0^n + O(d_1(c_1\alpha_1^m + c_2\alpha_2^m)\beta_1^n)$$
$$= C^{-1}d_0(c_0\alpha_0^m + c_1\alpha_1^m)\beta_0^n\,(1 + o(1))\,, \quad (m+n \to \infty,\ n > 0)\,.$$

Similarly, the O-formula for $j = 2$ improves on the one for $j = 1$, and so on. The notation $\simeq$ is used in order to represent the whole set of O-formulas for $j = 1, 2, \cdots$ by a single formula (see e.g. de Bruijn [13], section 1.5),

Lemma 3.4.

$$p_{m,n} \simeq C^{-1}\sum_{i=0}^{\infty} d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1})\beta_i^n\,, \quad (m+n \to \infty,\ n > 0)\,.$$

3.7. Product form expressions for the moments of the waiting time

In this section we show that the product form expressions for the probabilities $p_{m,n}$ lead to similar expressions for the first and second moment of the waiting time. The waiting time of a job is given by

$$W = S_1 + S_2 + \cdots + S_M\,,$$

where M is the length of the shortest queue on arrival and $S_1, S_2, \ldots$ are independent exponentially distributed random variables with unit mean and independent of M. By conditioning on M and using the property that Poisson arrivals see time averages (see e.g. Wolff [62]) we find

$$\mathbb{E}W = 2\sum_{m=1}^{\infty}\sum_{n=1}^{\infty} mp_{m,n} + \sum_{m=1}^{\infty} mp_{m,0}\,,$$

$$\mathbb{E}W^2 = 2\sum_{m=1}^{\infty}\sum_{n=1}^{\infty} m(m+1)p_{m,n} + \sum_{m=1}^{\infty} m(m+1)p_{m,0}\,.$$

Substituting (3.5) to eliminate $p_{m,0}$ and then inserting the series for $p_{m,n}$ with $m \geq 0$, $n > 0$, we obtain by changing summations,

$$\mathbb{E}W = C^{-1}\left\{ 2\sum_{i=0}^{\infty} d_i\left[\frac{c_i\alpha_i}{(1-\alpha_i)^2} + \frac{c_{i+1}\alpha_{i+1}}{(1-\alpha_{i+1})^2}\right]\frac{\beta_i}{1-\beta_i} \right.$$
$$\left. + \frac{1}{\rho+1}\sum_{i=0}^{\infty} d_i\left[\frac{c_i(2\rho+\alpha_i)}{(1-\alpha_i)^2} + \frac{c_{i+1}(2\rho+\alpha_{i+1})}{(1-\alpha_{i+1})^2}\right]\beta_i \right\},$$

and a similar expression for $\mathbb{E}W^2$. The terms in both series are alternating and decreasing (cf. the proof of lemma 3.3), so the error of each partial sum can be bounded by the absolute value of its final term. Similar expressions can be obtained for higher moments of the waiting time or other quantities of interest.

3.8. Numerical results

Representation (3.20) is suitable for numerical evaluation. The terms alternate in sign, decrease exponentially fast in modulus, and can easily be calculated. For the calculation of α_i and β_i we have the option to use the formulas in theorem 2.11 or to use the relations for $\alpha_i\alpha_{i+1}$ and $\beta_i\beta_{i+1}$ (cf. (2.23) and (2.25)). The coefficients c_i and d_i can be calculated from the relations (3.22) and (3.23). We finally note that, instead of using the series (3.21), the probabilities $p_{m,0}$ can easily be calculated from the equilibrium equations (3.5). In table 3.1 we list the probabilities $p_{0,1}$, $p_{0,2}$, $p_{1,1}$ and $p_{1,2}$ computed with an accuracy of 0.1%. The numbers in parentheses denote the number of terms in (3.20) needed.

ρ	$p_{0,1}$	$p_{0,2}$	$p_{1,1}$	$p_{1,2}$
0.1	0.0817 (40)	0.0007 (2)	0.0009 (2)	0.0000 (2)
0.3	0.1591 (14)	0.0100 (3)	0.0156 (3)	0.0007 (2)
0.5	0.1580 (10)	0.0233 (3)	0.0441 (4)	0.0047 (2)
0.7	0.1100 (8)	0.0275 (4)	0.0606 (4)	0.0118 (3)
0.9	0.0380 (6)	0.0140 (4)	0.0350 (4)	0.0104 (3)

Table 3.1.
Values of $p_{0,1}$, $p_{0,2}$, $p_{1,1}$ and $p_{1,2}$ with an accuracy of 0.1% for increasing values of ρ. The numbers in parentheses denote the number of terms in (3.20).

Let us investigate the rate of convergence of the terms in the series (3.20) as a function of ρ. From (3.14) and the limits (2.68) in section (2.10), it follows that for all $m \geq 0$ and $n > 0$,

$$\frac{|d_{i+1}(c_{i+1}\alpha_{i+1}^m + c_{i+2}\alpha_{i+2}^m)\beta_{i+1}^n|}{|d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n|} \to \frac{1-A_1}{A_2-1}\left[\frac{A_1}{A_2}\right]^{m+n-1} \tag{3.45}$$

as $i \to \infty$. For $0 < \rho < 1$, the factor $(1 - A_1)/(A_2 - 1)$ is decreasing and A_1/A_2 is increasing, and

$$\lim_{\rho \downarrow 0} \frac{1 - A_1}{A_2 - 1} = 1, \quad \lim_{\rho \uparrow 1} \frac{A_1}{A_2} = \frac{2 - 2^{1/2}}{2 + 2^{1/2}} = 3 - 2^{3/2}.$$

Hence, if $m > 0$ or $n > 1$, convergence of the terms in the series (3.20) is very fast for all ρ, at least with rate $3 - 2^{3/2} = 0.1715....$ If $m = 0$ and $n = 1$, then the rate of convergence is determined by $(1 - A_1)/(A_2 - 1)$ only, so, as table 3.1 illustrates, convergence is slow for small ρ.

In table 3.2 we list values of $\mathbb{E}W$ and $\mathbb{E}W^2$, together with the coefficient of variation $cv(W)$ of the waiting time, for increasing values of ρ. $\mathbb{E}W$ and $\mathbb{E}W^2$ are computed with an accuracy of 0.1%. The numbers of terms needed are shown in parentheses. For comparison we also computed the mean waiting time $\mathbb{E}W_c$ and the coefficient of variation $cv(W_c)$ of the waiting time for the corresponding common-queue system, that is, the $M|M|2$ queue with arrival rate 2ρ and service rate 1 for both servers. Table 3.2 illustrates that the performance of the shortest queue system is close to that of the common queue system.

ρ	$\mathbb{E}W$	$\mathbb{E}W^2$	$cv(W)$	$\mathbb{E}W_c$	$cv(W_c)$
0.1	0.0177 (39)	0.0358 (39)	10.632	0.0101	10.440
0.3	0.1441 (13)	0.3181 (13)	3.7846	0.0989	3.6667
0.5	0.4262 (8)	1.1472 (8)	2.3053	0.3333	2.2361
0.7	1.1081 (6)	4.3842 (5)	1.6032	0.9608	1.5714
0.9	4.4748 (4)	47.208 (3)	1.1652	4.2632	1.1600

Table 3.2.
Values of the first moment $\mathbb{E}W$, the second moment $\mathbb{E}W^2$ and the coefficient of variation $cv(W)$ of the waiting time with an accuracy of 0.1%, together with the first moment $\mathbb{E}W_c$ and the coefficient of variation $cv(W_c)$ of the corresponding common-queue system for increasing values of ρ. The numbers in parentheses denote the number of terms needed.

3.9. Numerical solution of the equilibrium equations

From limit (3.45) it follows that convergence of the series (3.20) is faster for states further away from the origin. This feature is illustrated in table 3.1. In particular, for $\rho = 0.1$ forty terms of the series for $p_{0,1}$ have to be computed, whereas for $p_{0,2}$ and $p_{1,1}$ two terms suffice to

attain the same accuracy. This feature can be used to compute $p_{0,1}$ from the equilibrium equation (3.8) rather then from the series (3.20). By inserting (3.5) and (3.9) to eliminate $p_{1,0}$ and $p_{0,0}$, equation (3.8) reduces to

$$p_{0,1}2\rho^2 = p_{0,2}(\rho + 1) + p_{1,1} .$$

This idea can be generalized as follows: the series (3.20) are used to calculate $p_{m,n}$ for $m + n > M$, where M is some integer, whereas for $m + n \le M$ the probabilities $p_{m,n}$ are calculated from the equilibrium equations. In this section we show that the equilibrium equations in states with $m + n \le M$ can be solved efficiently and numerically stable from the solution for $m + n > M$. The algorithm is based on the special property that the only flow from level l, defined by

$$\text{level } l = \{(0, l), (1, l-1), (2, l-2), ..., (l-1, 1), (l, 0)\} , \quad l \ge 0 ,$$

to level $l+1$ is via state $(l, 0)$. By this property, the problem of simultaneously solving the equations at the levels $l \le M$, given the solution at level $M+1$, can be reduced to that of recursively solving the equations at level $M \to M-1 \to ... \to 1 \to 0$.

We first formulate the equilibrium equations at level $l > 0$ (see (3.3) and (3.1)).

$$p_{0,l}(2\rho + 1) = p_{0,l+1} + p_{1,l-1} , \tag{3.46}$$

$$p_{k,l-k}2(\rho + 1) = p_{k-1,l-k+1}2\rho + p_{k,l-k+1} + p_{k+1,l-k-1} , \qquad 0 < k < l-1. \tag{3.47}$$

The equilibrium equations in the states $(l-1, 0)$ and $(l, 0)$ are replaced by the following two equations. Applying the general balance principle "rate out of A = rate into A" to

$$A = \{ (m, n) \mid m \ge 0, n \ge 0, m + n \le l \} \setminus \{(l, 0)\} ,$$

leads for all $l > 0$ to

$$p_{l-1,1}2\rho = p_{l,0} + \sum_{k=0}^{l-1} p_{k,l-k+1} , \tag{3.48}$$

and applying this principle to

$$A = \{ (m, n) \mid m \ge 0, n \ge 0, m + n \le l \} ,$$

yields for all $l \ge 0$,

$$p_{l,0}\rho = \sum_{k=0}^{l} p_{k,l-k+1} . \tag{3.49}$$

Now the probabilities at level l can be solved from the equations (3.46)-(3.49), given the probabilities at level $l+1$. This scheme can be repeated to recursively compute the probabilities at level $l-1 \to ... \to 1 \to 0$. The equations (3.46)-(3.49) form a *second* order recursion relation for the probabilities at level l. Below we show that these equations can be reduced to a *first* order

recursion relation.

Definition 3.5.

The sequence $x_0, x_1, x_2, \ldots$ *is the solution of*

$$x_{i+1} = x_i 2(\rho + 1) - x_{i-1} 2\rho \,, \quad i \geq 1 \,,$$

with initial values $x_0 = 1$ *and* $x_1 = 2\rho + 1$.

The numbers x_i are solved by

$$x_i = \frac{1 - A_1}{A_2 - A_1} A_1^i + \frac{A_2 - 1}{A_2 - A_1} A_2^i \,,$$

with A_1, A_2 given by (3.14).

Theorem 3.6.

For all $l > 0$,

$$p_{k,l-k} x_{k+1} = p_{k+1,l-k-1} x_k + \sum_{i=0}^{k} p_{i,l-i+1} x_i (2\rho)^{k-i} \quad \textit{for} \quad k = 0, 1, \ldots, l-2 \,. \tag{3.50}$$

Proof.

We prove the recursion relation (3.50) by induction. For $k = 0$ the equations (3.50) and (3.46) are identical. Assume that (3.50) holds for $k = j$. Multiplying (3.50) for $k = j$ by 2ρ and (3.47) for $k = j+1$ by x_{j+1} and adding the resulting equations, yields (3.50) for $k = j+1$. □

Based on theorem 3.6 the probabilities at level l can be computed efficiently, given the probabilities at level $l+1$. First, $p_{l,0}$ follows from (3.49) and $p_{l-1,1}$ from (3.48). Then $p_{l-2,2} \rightarrow p_{l-3,3} \rightarrow \ldots \rightarrow p_{0,l}$ can be successively calculated from (3.50). This recursion is ***numerically stable***, since all coefficients in the recursion relations are nonnegative, so the calculations involve only the multiplication and addition of nonnegative numbers. However, since x_i increases exponentially fast, it is numerically sensible to scale (3.50) by dividing both sides by x_{k+1}.

This concludes the analytical as well as the numerical treatment of the symmetric shortest queue problem. In the next sections we analyse some simple variants.

3.10. Unequal routing probabilities

We now consider a variant of the model of section 3.1; the equal routing probabilities in case of equal queue lengths are replaced by a and $1-a$, where a is an arbitrary number between 0 and 1. The transition-rate diagram is depicted in figure 3.4.

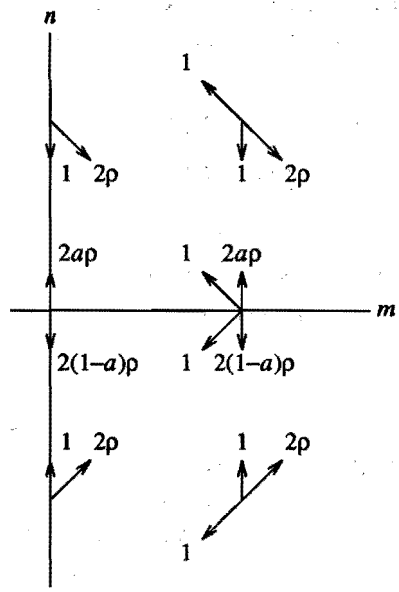


Figure 3.4.
Transition-rate diagram for the shortest queue model with unequal routing probabilities a and 1 – a respectively.

This problem is not symmetric anymore. However, the asymmetry is rather weak in the sense that the regions $n > 0$ and $n < 0$ are still mirror images of each other and the average $(p_{m,n} + p_{m,-n})/2$ satisfies all equilibrium equations (3.1)-(3.9) of the symmetric problem. This suggests that $p_{m,n}$ can be expressed as

$$p_{m,n} = C^{-1} \sum_{i=0}^{\infty} d_i^{+} (c_i \alpha_i^m + c_{i+1} \alpha_{i+1}^m) \beta_i^n \qquad \text{for} \quad m \geq 0\,,\ n > 0\,,$$

$$p_{m,n} = C^{-1} \sum_{i=0}^{\infty} d_i^{-} (c_i \alpha_i^m + c_{i+1} \alpha_{i+1}^m) \beta_i^{-n} \qquad \text{for} \quad m \geq 0\,,\ n < 0\,, \tag{3.51}$$

$$p_{m,n} = C^{-1} \left\{ c_0 f_0 \alpha_0^m + \sum_{i=0}^{\infty} c_{i+1} f_{i+1} \alpha_{i+1}^m \right\} \qquad \text{for} \quad m \geq 0\,,\ n = 0\,,$$

where

$$\frac{1}{2}(d_i^{\dagger} + d_i^{\ddagger}) = d_i \quad (i = 0, 1, ...) . \tag{3.52}$$

The motivation for introducing new coefficients $d_i^{\dagger}$ and $d_i^{\ddagger}$ is the fact that the transition structure at the m-axis is not symmetric anymore. For any choice of $\{d_i^{\dagger}\}$ and $\{d_i^{\ddagger}\}$ satisfying (3.52) the series (3.51) satisfy all equilibrium equations, except for the equations for $|n| = 1$. To also satisfy the latter conditions it is readily verified that $\{d_i^{\dagger}\}$ and $\{d_i^{\ddagger}\}$ have to satisfy the following relations:

$$(d_i + d_{i+1})(\alpha_{i+1} + \rho) = (d_i^{\dagger} + d_{i+1}^{\dagger})(\alpha_{i+1} + 2a\rho) \qquad (i = 0, 1, ...) ,$$

$$(d_i + d_{i+1})(\alpha_{i+1} + \rho) = (d_i^{\ddagger} + d_{i+1}^{\ddagger})(\alpha_{i+1} + 2(1-a)\rho) \qquad (i = 0, 1, ...) ,$$

with initially

$$d_0(\alpha_0 + \rho) = d_0^{\dagger}(\alpha_0 + 2a\rho) = d_0^{\ddagger}(\alpha_0 + 2(1-a)\rho) .$$

The solutions $\{d_i^{\dagger}\}$ and $\{d_i^{\ddagger}\}$ of these relations indeed satisfy (3.52).

3.11. Threshold jockeying

In this section we consider the shortest queue model with a threshold-type jockeying; one job switches from the longest to the shortest queue if the difference between the lengths of both queues exceeds some threshold value T. It appears that the compensation approach also works for this model. In fact, the main term in the series (3.20) already satisfies the boundary conditions, so no compensation arguments are required.

There are several other techniques to analyse this model. The form of the state space suggests to apply the matrix-geometric approach developed by Neuts [51]. Actually, Gertsbakh [29] studies the threshold jockeying model by using this approach. In [8] the relationship between our approach and the matrix-geometric approach has been investigated. It appears that our approach suggests a state space partitioning which is more useful than the one used by Gertsbakh [29]. In [4] it is shown that the matrix-geometric approach can also be used to analyse the threshold jockeying model with c parallel servers. The results in this paper emphasize the importance of a suitable choice of the state space partitioning. Another approach to the jockeying model with c parallel servers can be found in Grassmann and Zhao [63]. They use the concept of modified lumpability for continuous-time Markov processes. It is finally mentioned that the instantaneous jockeying model ($T = 1$) has been addressed by Haight [34] for $c = 2$ and by Disney and Mitchell [17], Elsayed and Bastani [18], Kao and Lin [42] and Zhao and Grassmann [31] for arbitrary c.

The threshold jockeying model has a much simpler state space than the original model, since n only varies between $-T$ and T. The model is symmetric with respect to the m-axis. Therefore the analysis can be restricted to the state space in the first quadrant. The transition rates are depicted in figure 3.5.

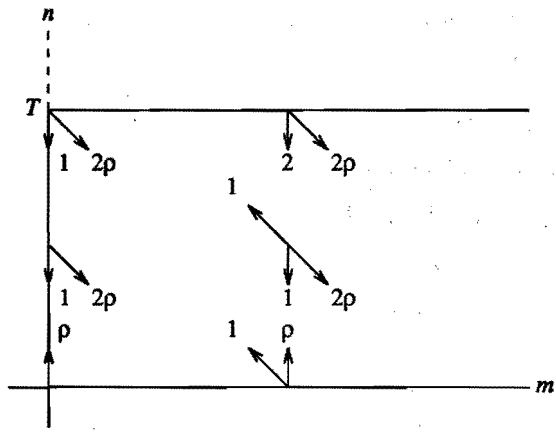


Figure 3.5.
Transition-rate diagram for the shortest queue model with threshold jockeying. The threshold value is T.

The transition rates in states with $n < T$ are identical to the rates of the original model (see figure 3.2). Hence, the first terms in the series (3.20) and (3.21), i.e.,

$$\alpha_0^m \beta_0^n \quad \text{for} \quad m \geq 0\,,\ 0 < n < T-1\,, \tag{3.53}$$

$$f_0 \alpha_0^m \quad \text{for} \quad m \geq 0\,,\ n = 0\,, \tag{3.54}$$

satisfy all equilibrium equations for $m > 0$ and $0 \leq n < T-1$. We now investigate whether $\alpha_0^m \beta_0^n$ can be fitted to the equations for $m > 0$ and $T-1 \leq n \leq T$. We define

$$g_0 \alpha_0^m \beta_0^T \quad \text{for} \quad m \geq 0\,,\ n = T\,, \tag{3.55}$$

and try to choose g_0 such that the equilibrium equations for $m > 0$ and $T-1 \leq n \leq T$ are satisfied. These equations state:

$$p_{m,T-1} 2(\rho + 1) = p_{m-1,T} 2\rho + p_{m,T} 2 + p_{m+1,T-2}\,, \quad m > 0\,,$$

$$p_{m,T} 2(\rho + 1) = p_{m+1,T-1}\,, \quad m > 0\,.$$

Insertion of (3.53) and (3.55) in the equations for $m > 0$ and $n = T$ yields

$$g_0 = \frac{\alpha_0}{2(\rho+1)\beta_0} . \tag{3.56}$$

For this choice of g_0, it is easily verified, by using $\alpha_0 = \rho^2$ and $\beta_0 = \rho^2/(2+\rho)$, that the equations for $m > 0$ and $n = T-1$ are also satisfied. The solution (3.53)-(3.55) violates the conditions for $m = 0$, and therefore cannot produce $p_{m,n}$ for all m and n, but we will prove:

Theorem 3.7. (threshold jockeying)

For all $m + n > T$ and for $m = T$ and $n = 0$,

$$\begin{aligned} p_{m,n} &= K^{-1}\alpha_0^m\beta_0^n \quad \textit{for} \quad 0 < n < T-1 , \\ &= K^{-1}f_0\alpha_0^m \quad \textit{for} \quad n = 0 , \\ &= K^{-1}g_0\alpha_0^m\beta_0^T \quad \textit{for} \quad n = T , \end{aligned}$$

where K is the normalizing constant and

$$\alpha_0 = \rho^2 , \quad \beta_0 = \frac{\rho^2}{2+\rho} , \quad f_0 = \frac{\alpha_0}{\alpha_0+\rho} , \quad g_0 = \frac{\alpha_0}{2(\rho+1)\beta_0} .$$

Proof.

Consider the Markov process *restricted to* $\mathcal{V} = \{(m, n) \mid m + n > T,\ 0 \le n \le T\} \cup \{(T, 0)\}$ (cf. section 2.12 where an analogous argument is used to prove theorem 2.33). The transition rates are depicted in figure 3.6. All transitions from states with $m + n = T+1$ and $0 < n \le T$ to state $(T, 0)$ result from excursions to states outside $\mathcal{V}$, which always end at $(T, 0)$. Let $\{v_{m,n}\}$ be the equilibrium distribution of the process restricted to $\mathcal{V}$.

It is readily verified that the products (3.53)-(3.55) with g_0 specified by (3.56), satisfy all equilibrium equations on $\mathcal{V}$. The sum of these products over $\mathcal{V}$ converges, since $\alpha_0 = \rho^2 < 1$. Hence, the products (3.53)-(3.55) are positive and convergent solutions of all equilibrium equations. By a result of Foster (see appendix A), this proves that the restricted process is ergodic, and the products (3.53)-(3.55) can be normalized to produce $\{v_{m,n}\}$. Since the number of states outside $\mathcal{V}$ is finite, the original process is also ergodic, and $p_{m,n}$ and $v_{m,n}$ are related by

$$p_{m,n} = v_{m,n}\, P(\mathcal{V}) \quad \text{for} \quad (m, n) \in \mathcal{V} ,$$

where $P(\mathcal{V})$ is the probability that the original process is in $\mathcal{V}$. The proof of theorem 3.7 is now completed by observing that the products (3.53)-(3.55) produce $v_{m,n}$ up to some multiplicative constant. □

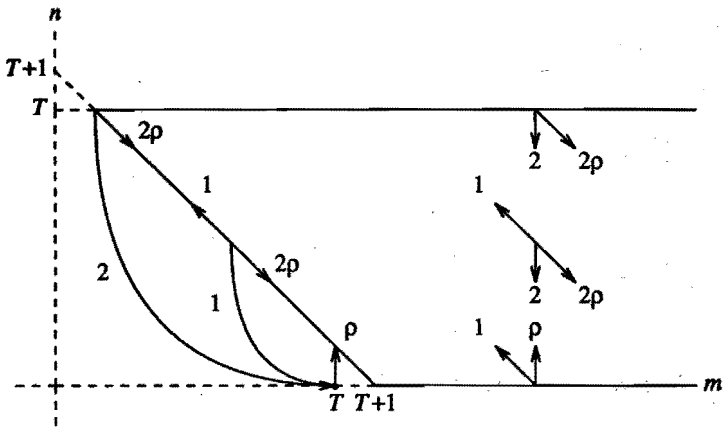


Figure 3.6.
Transition-rate diagram for the shortest queue model with threshold jockeying, restricted to $\mathcal{V} = \{(m, n) | m + n > T, 0 \le n \le T\} \cup \{(T, 0)\}$.

From a computational point of view we note that, given the solution on $\mathcal{V}$, the equations on the complement of $\mathcal{V}$ can be solved efficiently and numerically stable by the recursive algorithm derived in section 3.9. In table 3.3 we list the mean waiting time W for increasing values of ρ and T. For $T = 1$ the mean waiting time W is the same as the mean waiting time W_c of the corresponding common-queue system. The case $T = \infty$ corresponds to the shortest queue problem without jockeying. Table 3.3 illustrates that already for small T the mean waiting time W is very close to the mean waiting for $T = \infty$.

3.12. Conclusion

In this chapter we applied the general theory, developed in chapter 2, to the symmetric shortest queue problem and proved that the equilibrium probabilities can be expressed as a series of products. These products, as well as their coefficients, are found explicitly. From the expressions for the equilibrium probabilities, similar expressions can be derived for the moments of the waiting time, or other quantities of interest. We showed that the product form expressions are useful from a numerical point of view and that the analysis can easily be extended to some variants of the original problem. An important variant of the symmetric shortest queue problem, which has not been discussed yet, is the asymmetric shortest queue

	W				
ρ	$T=1$	$T=2$	$T=4$	$T=6$	$T=\infty$
0.1	0.0101	0.0176	0.0177	0.0177	0.0177
0.3	0.0989	0.1405	0.1440	0.1440	0.1440
0.5	0.3333	0.4091	0.4260	0.4263	0.4263
0.7	0.9608	1.0624	1.1052	1.1081	1.1082
0.9	4.2632	4.3822	4.4614	4.4733	4.4749

Table 3.3.
Values of the mean waiting time W for increasing values of ρ *and T.*

problem, i.e., the shortest queue problem for nonidentical servers. In chapter 5 we show how the analysis can be extended to this problem, but first, in chapter 4 we apply the general theory to a queueing model arising in the field of computer performance analysis.

Chapter 4

Multiprogramming queues

In this chapter we analyse a queueing model for a multiprogramming computer system consisting of an input-output unit (IO) and a central processor (CP). This model was introduced by Hofri [37]. He uses the same generating function approach as Kingman used in [44] for the shortest queue problem. In doing so, Hofri shows that the generating function of the stationary queue length probabilities $p_{m,n}$ is a meromorphic function and he finds explicit expressions for the poles and residues. By decomposing the generating function into partial fractions he is able to prove that the probabilities $p_{m,n}$ can be represented by an infinite linear combination of product form solutions. The decomposition, however, leads to cumbersome formulae for the coefficients. In [1] it is shown that these representations are partly incorrect in the sense that they do not always hold for small m and n. This complication is overlooked by Hofri [37] due to the fact that he uses an incorrect version of Mittag-Leffer's theorem to deduce the partial fraction decomposition.

The multiprogramming system can be modelled as a Markov process satisfying assumption 2.6. Therefore, we may apply the theory in chapter 2 to analyse the multiprogramming system. This approach improves the results obtained by Hofri in the sense that explicit expressions are found for the coefficients in the infinite linear combination of product form solutions. Similar expressions can be derived for global performance measures, such as the mean number of jobs in the system. A new feature not occurring in the analysis of the symmetric shortest queue problem, is that the resulting series of product form solutions may diverge for small m and n. By exploiting the explicit expressions for the product forms and their coefficients, an efficient numerical procedure with tight bounds on the error of each partial sum can be derived. This model can also be treated as a direct application of the compensation approach, i.e., without use of the theory of chapter 2. The details of this direct application are worked out in [5].

This chapter is organised as follows. In section 4.1 we formulate the model and equilibrium equations. In section 4.2 we apply the theory of chapter 2 to this problem. It appears that only the feasible pair $(\alpha_+, X_-(\alpha_+))$ plays a role. We prove that on $\mathcal{A}(N)$ the probabilities $p_{m,n}$ can be expressed as $x_{m,n}(\alpha_+)$ up to some normalizing constant C. Both α_+ and C are found explicitly. In section 4.3 error bounds on each partial sum of $x_{m,n}(\alpha_+)$ are derived. In section 4.4 we develop a recursive algorithm for numerically solving $p_{m,n}$ from the equilibrium equations. Product form expressions for the mean number of jobs at the IO are derived in section 4.5. The next section presents numerical results and the final section is devoted to conclusions.

4.1. Model and equilibrium equations

Consider the queueing model for the multiprogramming system depicted in figure 4.1.

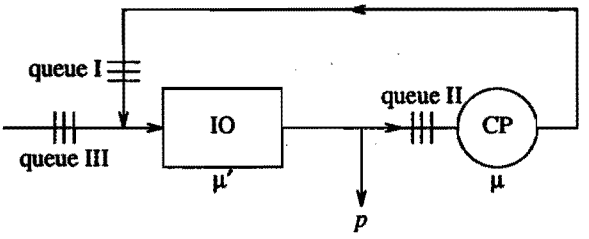


Figure 4.1.
Queueing model for the multiprogramming system.

In the queueing model it is supposed that queue III of incoming jobs provides an infinite source of available jobs. The multiprogramming system consists of an input-output unit and a central processor. Incoming jobs start at the IO unit with an exponentially distributed service time with parameter μ'. Subsequently, the job leaves the system (with probability p) or proceeds to queue II at the CP (with probability $1-p$). At the CP a job has an exponentially distributed service time with parameter μ. Next the job is recycled to the IO unit where it joins queue I. The IO unit treats the jobs in queue I with nonpreemptive priority with respect to the new jobs in queue III. Since the IO unit is always busy, jobs arrive at the CP according to a Poisson process with rate $\lambda = (1-p)\mu'$. Hence, the CP constitutes an $M \mid M \mid 1$ queue with arrival rate λ and service rate μ. Therefore, it is sensible to assume that $\lambda < \mu$.

The system can be represented by a continuous-time Markov process with states (m, n), $m, n = 0, 1, \ldots$ where m is the length of queue II *including* the job being served and n is the length of queue I *excluding* the job being served (the IO unit is constantly busy). Let $\{p_{m,n}\}$ be the equilibrium distribution. The transition rates are depicted in figure 4.2, where $\eta = p\mu'$

The equilibrium equations for $\{p_{m,n}\}$ are formulated below, where $\kappa = \lambda + \mu + \eta$.

$$p_{m,n}\kappa = p_{m-1,n+1}\lambda + p_{m,n+1}\eta + p_{m+1,n-1}\mu, \qquad m > 1, n > 1 \quad (4.1)$$

$$p_{1,n}\kappa = p_{0,n+1}\lambda + p_{1,n+1}\eta + p_{2,n-1}\mu, \qquad n > 1 \quad (4.2)$$

$$p_{0,n}(\lambda + \eta) = p_{0,n+1}\eta + p_{1,n-1}\mu, \qquad n > 1 \quad (4.3)$$

$$p_{m,1}\kappa = p_{m-1,2}\lambda + p_{m,2}\eta + p_{m+1,0}\mu, \qquad m > 1 \quad (4.4)$$

$$p_{m,0}(\lambda + \mu) = p_{m-1,1}\lambda + p_{m,1}\eta + p_{m-1,0}\lambda, \qquad m > 1 \quad (4.5)$$

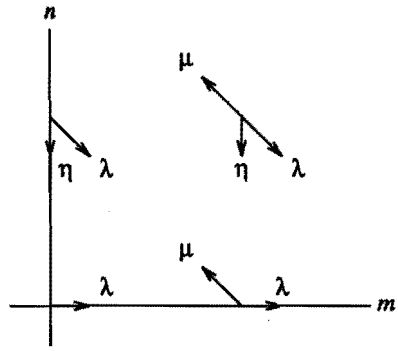


Figure 4.2.
Transition-rate diagram for the multiprogramming model, where $\eta = p\mu'$.

$$p_{1,1}\kappa = p_{0,2}\lambda + p_{1,2}\eta + p_{2,0}\mu\,, \tag{4.6}$$

$$p_{0,1}(\lambda+\eta) = p_{0,2}\eta + p_{1,0}\mu\,, \tag{4.7}$$

$$p_{1,0}(\lambda+\mu) = p_{0,1}\lambda + p_{1,1}\eta + p_{0,0}\lambda\,, \tag{4.8}$$

$$p_{0,0}\lambda = p_{0,1}\eta\,. \tag{4.9}$$

This formulation can be simplified by observing that the equations (4.1) hold for all $m > 0$ and $n > 0$. The reason for not doing so, is that the equations (4.1)-(4.5) correspond with the equations (2.1)-(2.9) and therefore fit the theory of chapter 2. In the next section we investigate how the compensation approach works out here.

4.2. Application of the compensation approach

We first translate the transition rates q_{ij}, ν_{ij}, h_{ij} and r_{ij} in terms of the rates in figure 4.2. It is easily derived that,

$$\begin{aligned}
&q_{-1,1} = \mu\,, \quad q_{0,-1} = \eta\,, \quad q_{1,-1} = \lambda\,, \quad q_{-1,-1} = q_{-1,0} = q_{0,1} = q_{1,1} = q_{1,0} = 0\,, \quad q = \kappa\,,\\
&\nu_{0,-1} = \eta\,, \quad \nu_{1,-1} = \lambda\,, \qquad\qquad \nu_{1,0} = \nu_{1,1} = \nu_{0,1} = 0\,,\\
&h_{-1,1} = \mu\,, \quad h_{1,0} = \lambda\,, \qquad\qquad h_{-1,0} = h_{1,1} = h_{0,1} = 0\,,\\
&r_{1,0} = \lambda\,, \qquad\qquad r_{1,1} = r_{0,1} = 0\,.
\end{aligned} \tag{4.10}$$

For the rates q_{ij} from states (m, n) with $m > 0, n > 0$ we directly obtain

$$q_{0,1} = q_{1,1} = q_{1,0} = 0 ,$$

which is the essential condition for the application of the compensation approach (cf. assumption 2.6). From the boundary behaviour it follows that $(\alpha_+, X_-(\alpha_+))$ is the only feasible pair possible (cf. conclusion 2.14(ii)). This pair exists if and only if condition (2.50) in theorem 2.19 is satisfied. This condition is checked below. Suppose that

$$q_{-1,-1} + q_{0,-1} + q_{1,-1} = \eta + \lambda > \mu = q_{-1,1} . \tag{4.11}$$

Then $RH'(1) > LH'(1)$ must hold. Inserting (4.10) in $RH(\alpha)$ and $LH(\alpha)$ (cf. (2.44)) and using (4.11) yields

$$RH(\alpha) = \frac{\alpha 2\mu(\alpha\eta + \lambda)}{\kappa + \sqrt{\kappa^2 - 4(\alpha\eta + \lambda)\mu}} + \lambda ,$$

$$LH(\alpha) = \alpha(\lambda + \mu) ,$$

from which it follows that

$$RH'(1) = \mu\left[1 + \frac{\eta}{\eta + \lambda - \mu}\right] ,$$

$$LH'(1) = \lambda + \mu .$$

So $RH'(1) > LH'(1)$, since $\lambda < \mu$. Hence, the feasible pair $(\alpha_+, X_-(\alpha_+))$ indeed exists. $\{x_{m,n}(\alpha_+)\}$ formally satisfies the conditions (4.1)-(4.5). Since $v_{1,j} = q_{1,j}$ and $h_{j,1} = q_{j,1}$, we have $e_i = c_i + c_{i+1}$ and $f_{i+1} = d_i + d_{i+1}$ for all i (cf. remark 2.4). So the series (2.56) defining $x_{0,n}(\alpha_+)$ for $n > 0$ is identical to (2.54) with $m = 0$; the series (2.57) defining $x_{m,0}(\alpha_+)$ for $m > 0$ is identical to (2.55) with $n = 0$. Then, by taking the series (2.55) with $m = n = 0$ as definition of $x_{0,0}(\alpha_+)$, it easily follows that $\{x_{m,n}(\alpha_+)\}$ formally satisfies (4.6)-(4.8). The remaining equation (4.9) is satisfied, due to the fact that the equilibrium equations are dependent. Hence, $\{x_{m,n}(\alpha_+)\}$ is a formal solution to all equations. Before investigating properties of $\mathcal{A}(N)$ we restate the definition of $x_{m,n}(\alpha_+)$, which for the present model simplifies considerably.

The series $x_{m,n}(\alpha_+)$ is defined by

$$x_{m,n}(\alpha_+) = \sum_{i=0}^{\infty} d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n , \qquad m \geq 0,\ n > 0 ; \tag{4.12}$$

$$x_{m,0}(\alpha_+) = c_0 d_0 \alpha_0^m + \sum_{i=0}^{\infty} c_{i+1}(d_i + d_{i+1})\alpha_{i+1}^m , \qquad m \geq 0 . \tag{4.13}$$

Substitution of (4.10) into (2.20)-(2.21) leads to the following recursion relations for c_i and d_i. The recursion relation for c_i is simplified by insertion of the relations for $\alpha_i\alpha_{i+1}$ and $\alpha_i + \alpha_{i+1}$ (cf. (2.23)-(2.24)); the one for d_i is simplified by insertion of the analogous relations for $\beta_i\beta_{i+1}$ and $\beta_i + \beta_{i+1}$ (cf. (2.25)-(2.26)).

$$c_{i+1} = -\frac{\beta_i - \alpha_{i+1}}{\beta_i - \alpha_i} c_i \qquad (i = 0, 1, \cdots), \tag{4.14}$$

$$d_{i+1} = -\frac{1-\beta_{i+1}}{1-\beta_i} d_i \qquad (i = 0, 1, \cdots). \tag{4.15}$$

By theorem 2.25 the series $x_{m,n}(\alpha_+)$ converges absolutely on (see section 2.12)

$$\mathcal{A}(N) = \{(m, n) \mid m \geq 0, n \geq 0, m + n > N\} \cup \{(N, 0)\},$$

where N is defined as the smallest nonnegative integer such that (see definition 2.28)

$$|HV|(A_1/A_2)^{N+1} < 1. \tag{4.16}$$

Inserting (4.10) into the definitions of A_1, A_2, H and V (see lemma 2.26) yields

$$A_1 = \frac{\kappa - \sqrt{\kappa^2 - 4\mu\lambda}}{2\mu}, \quad A_2 = \frac{\kappa + \sqrt{\kappa^2 - 4\mu\lambda}}{2\mu},$$

$$H = 1, \qquad V = \frac{A_1^{-1}\lambda - (\lambda + \eta)}{A_2^{-1}\lambda - (\lambda + \eta)} = \frac{1 - A_1}{1 - A_2}. \tag{4.17}$$

The final equality is obtained by substituting the identities $\lambda/\mu = A_1 A_2$ and $\kappa/\mu = A_1 + A_2$. Inserting (4.17) in (4.16) we find that N is the smallest nonnegative integer such that

$$\frac{1 - A_1}{A_2 - 1}\left[\frac{A_1}{A_2}\right]^{N+1} < 1.$$

For $\mu' = 1$, $\mu = 2$ and $p = 3/25$ it follows that $N = 2$. Hence, N is not necessarily zero implying that the series for $x_{m,n}$ may diverge for small m and n. In fact, in remark 2.29 we have seen for $\mu'/\mu = 1/2$ and $p = \delta$ that $N \to \infty$ as $\delta \downarrow 0$.

By restricting the Markov process to $\mathcal{A}(N)$ it is easily shown that $x_{m,n}(\alpha_+)$ produces $p_{m,n}$ for all $(m, n) \in \mathcal{A}(N)$ up to some multiplicative constant C. For the present problem α_+ and C can be found explicitly by observing that the CP constitutes an $M \mid M \mid 1$ queue with arrival rate λ and service rate μ. Hence, the probability p_m of finding m jobs at the CP is given by

$$p_m = (1 - \lambda/\mu)(\lambda/\mu)^m. \tag{4.18}$$

On the other hand, we have for $m > N$

$$p_m = \sum_{n=0}^{\infty} p_{m,n} = C^{-1} \sum_{n=0}^{\infty} x_{m,n}(\alpha_+).$$

The series (4.12) can be rewritten by forming pairs with the same α-factor (which is permitted by theorem 2.25(i)). Inserting this series together with (4.13) in the expression for p_m and using the recursion relation (4.15) leads to (note that $c_0 = d_0 = 1$),

$$p_m = C^{-1}\sum_{n=0}^{\infty}\left\{c_0 d_0 \alpha_0^m \beta_0^n + \sum_{i=0}^{\infty} c_{i+1}(d_i\beta_i^n + d_{i+1}\beta_{i+1}^n)\alpha_{i+1}^m\right\}$$

$$= C^{-1}\left\{\frac{\alpha_0^m}{1-\beta_0} + \sum_{i=0}^{\infty} c_{i+1}\left(\frac{d_i}{1-\beta_i} + \frac{d_{i+1}}{1-\beta_{i+1}}\right)\alpha_{i+1}^m\right\}$$

$$= C^{-1}\frac{\alpha_0^m}{1-\beta_0}, \tag{4.19}$$

where changing of summations is allowed by the absolute convergence stated in theorem 2.25(iv). Combining (4.18) and (4.19) yields $\alpha_0 = \lambda/\mu$, and so $\beta_0 = X_-(\alpha_0) = \lambda/(\eta+\mu)$. It is easily verified that $\alpha = \lambda/\mu$ indeed satisfies $LH(\alpha) = RH(\alpha)$. Further, from (4.18)-(4.19),

$$C^{-1}\frac{1}{1-\beta_0} = 1 - \frac{\lambda}{\mu},$$

so we obtain

$$C = \frac{\mu(\eta+\mu)}{(\mu-\lambda)(\eta+\mu-\lambda)}.$$

Our findings are summarized in the following theorem.

Theorem 4.1.

For all $m \geq 0$, $n \geq 0$ *with* $m + n > N$ *and for* $m = N$ *and* $n = 0$,

$$p_{m,n} = C^{-1}x_{m,n}(\alpha_+),$$

where $\alpha_+ = \lambda/\mu$ *and*

$$C = \frac{\mu(\eta+\mu)}{(\mu-\lambda)(\eta+\mu-\lambda)}.$$

In the next two sections we concentrate on numerical aspects.

4.3. Error bounds on each partial sum of product forms

In this section we derive bounds on the error of each partial sum of the series (4.12) and (4.13) defining $x_{m,n}(\alpha_+)$. First, for $m > N$ the series (4.13) can be rewritten as (4.12) with $n = 0$ (which is allowed by theorem 2.25(i)). Note that for $m = N$ this may lead to a *divergent* series. Hence, we have for all $m \geq 0$, $n \geq 0$ with $m + n > N$,

$$x_{m,n}(\alpha_+) = \sum_{i=0}^{\infty} d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n. \tag{4.20}$$

We now investigate whether the terms in this series are alternating and monotonically decreasing in modulus (just as for the terms in (3.20)). Since (cf. (2.60))

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

it follows from the recursion relations (4.14) and (4.15) that for $i \geq 0$,

$$\frac{c_{i+1}}{c_i} > 0 , \quad \frac{d_{i+1}}{d_i} < 0 .$$

Hence, the coefficients c_i are positive and the coefficients d_i are alternating. Consequently, the terms in the series (4.20) are alternating. However, these terms are not necessarily decreasing in modulus from the beginning. Below we derive bounds on the decrease of the terms in the series (4.20) from which we can decide when these terms are decreasing. To do so, we first need bounds for c_{i+1} / c_i and d_{i+1} / d_i. From lemma 2.27 it follows that as $i \to \infty$,

$$\frac{\alpha_i}{\beta_i} \uparrow A_2 , \quad \frac{\alpha_{i+1}}{\beta_i} \downarrow A_1 . \tag{4.21}$$

Hence, by defining for $i \geq 0$

$$\bar{c}_i = \frac{1 - A_1}{\alpha_i / \beta_i - 1} ,$$

$$\bar{d}_i = \frac{1}{1 - \beta_i} ,$$

we obtain from the recursions relations (4.14) and (4.15) that for $i \geq 0$,

$$\frac{c_{i+1}}{c_i} \leq \bar{c}_i , \quad \frac{|d_{i+1}|}{|d_i|} \leq \bar{d}_i .$$

The bounds $\bar{c}_i$ and $\bar{d}_i$ are decreasing and asymptotically tight, i.e., as $i \to \infty$,

$$\bar{c}_i \downarrow \frac{1 - A_1}{A_2 - 1} = -V , \quad \bar{d}_i \downarrow 1 = H . \tag{4.22}$$

Based on the bounds $\bar{c}_i$ and $\bar{d}_i$ and the monotonicity given in (4.20) and (4.21) we can prove:

Lemma 4.2.

For all $m \geq 0$, $n \geq 0$ and $i \geq 0$,

$$|d_{i+1}| (c_{i+1} \alpha_{i+1}^m + c_{i+2} \alpha_{i+2}^m) \beta_{i+1}^n \leq R(i, m, n) \, |d_i| (c_i \alpha_i^m + c_{i+1} \alpha_{i+1}^m) \beta_i^n ,$$

where

$$R(i, m, n) = \bar{d}_i \bar{c}_i (\alpha_{i+1} / \alpha_i)^m \, (\beta_{i+1} / \beta_i)^n .$$

Proof.

Let $m \geq 0$, $n \geq 0$ and $i \geq 0$. Then

$$|d_{i+1}|(c_{i+1}\alpha_{i+1}^m + c_{i+2}\alpha_{i+2}^m)\beta_{i+1}^n \leq \bar{d}_i |d_i| (\bar{c}_i \frac{\alpha_{i+1}^m}{\alpha_i^m} c_i \alpha_i^m + \bar{c}_{i+1} \frac{\alpha_{i+2}^m}{\alpha_{i+1}^m} c_{i+1} \alpha_{i+1}^m) \frac{\beta_{i+1}^n}{\beta_i^n} \beta_i^n ,$$

and insertion of

$$\bar{c}_{i+1} \leq \bar{c}_i , \quad \frac{\alpha_{i+2}}{\alpha_{i+1}} = \frac{\alpha_{i+2}}{\beta_{i+1}} \frac{\beta_{i+1}}{\alpha_{i+1}} \leq \frac{\alpha_{i+1}}{\beta_i} \frac{\beta_i}{\alpha_i} = \frac{\alpha_{i+1}}{\alpha_i} ,$$

(cf. (4.21), (4.22)) in the right-hand side of this inequality proves the lemma. □

The monotonicity given in (4.21)-(4.22) yields that $R(i, m, n)$ is decreasing in i and asymptotically tight.

Lemma 4.3.

For all $m \geq 0$ and $n \geq 0$, as $i \to \infty$,

$$R(i, m, n) \downarrow |HV| (A_1 / A_2)^{m+n} .$$

We can now conclude that for fixed $m \geq 0$, $n \geq 0$ with $m + n > N$ the bounds $R(i, m, n)$ on the decrease of the i-th term in (4.20) are eventually less than one. From there on the terms in (4.20) are monotonically decreasing in modulus, so, since these terms are also alternating, the error of each partial sum is bounded by the modulus of its final term.

4.4. Numerical solution of the equilibrium equations

If $N > 0$, then the equilibrium equations for $m + n \leq N$ have to be solved numerically from the solution on the complement. These equations can be solved efficiently and numerically stable by an approach similar to the one in section 3.9. Below, this approach is outlined briefly. Define

$$\text{level } l = \{(0, l), (1, l-1), \ldots, (l-1, 1), (l, 0)\} , \quad l \geq 0 .$$

We show that the probabilities at level l can be solved from the solution at level $l+1$. This scheme can then be repeated to subsequently compute the probabilities at level $l \to l-1 \to \ldots \to 1 \to 0$. First, $p_{l,l}$ can be computed from the following equation derived by application of the balance principle to $\{(m, n) \mid m \geq 0, n \geq 0, m + n \leq l\}$.

$$p_{l,l}\lambda = \sum_{k=0}^{\infty} p_{k,l-k+1}\eta \,.$$

Next, the probabilities $p_{l-1,1} \rightarrow p_{l-2,2} \rightarrow \ldots \rightarrow p_{0,l}$ can be computed from the following recursion relation. This relation can be established similarly as the one in theorem 3.6.

Definition 4.4.

The sequence $x_0, x_1, x_2, \ldots$ *is the solution of*

$$x_{i+1} = x_i\kappa - x_{i-1}\lambda\mu \,, \quad i \geq 1 \,,$$

with initial values $x_0 = 1$ *and* $x_1 = \lambda + \eta$.

It is easily verified that the numbers x_i are positive and given by

$$x_i = \frac{(\lambda+\eta)-\tau_1}{\tau_2-\tau_1}\tau_1^i + \frac{\tau_2-(\lambda+\eta)}{\tau_2-\tau_1}\tau_2^i \,,$$

where

$$\tau_1 = \frac{\kappa - \sqrt{\kappa^2 - 4\lambda\mu}}{2} \,, \quad \tau_2 = \frac{\kappa + \sqrt{\kappa^2 - 4\lambda\mu}}{2} \,.$$

Theorem 4.5.

For all $l > 0$,

$$p_{k,l-k}x_{k+1} = p_{k+1,l-k-1}x_k\mu + \sum_{i=0}^{k} p_{i,l-i+1}x_i\lambda^{k-i}\eta \quad \textit{for} \quad k = 0, 1, \ldots, l-1 \,. \tag{4.23}$$

The recursion relation in theorem 4.5 is *numerically stable*, since all coefficients in these recursion relations are nonnegative, so the calculations involve only the multiplication and addition of nonnegative numbers. This concludes the numerical solution of the equilibrium equations for $m + n \leq N$. Finally, from (4.17) and the limits (2.68) in section (2.10), it follows that successive terms in the series (4.20) satisfy

$$\frac{|d_{i+1}(c_{i+1}\alpha_{i+1}^m + c_{i+2}\alpha_{i+2}^m)\beta_{i+1}^n|}{|d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n|} \rightarrow \frac{1-A_1}{A_2-1}\left[\frac{A_1}{A_2}\right]^{m+n} \quad (i \rightarrow \infty) \,,$$

for all $m \geq 0$, $n \geq 0$. Hence, convergence of the series (4.20) is faster for states further away from the origin, so it is numerically sensible to use the series (4.20) only to calculate $p_{m,n}$ for $m + n > M$ where $M > N$ and to use the relations (4.23) to calculate $p_{m,n}$ for $m + n \leq M$.

4.5. Product form expression for the number of jobs in queue I

In this section we show that the product form expressions for the probabilities $p_{m,n}$ lead to similar expressions for the mean number of jobs in queue I. Similar expressions can be derived for other quantities of interest.

Inserting $p_{m,n} = C^{-1}x_{m,n}(\alpha_+)$ for $m + n > N$ in the expression for the mean number of jobs L_I in queue I leads to

$$L_I = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} np_{m,n}$$

$$= \sum_{\substack{m\geq 0, n\geq 1 \\ m+n\leq N}} np_{m,n} + C^{-1}\sum_{n=1}^{N} n\sum_{m=0}^{\infty} x_{m,n}(\alpha_+) + C^{-1}\sum_{n=N+1}^{\infty} n\sum_{m=0}^{\infty} x_{m,n}(\alpha_+)$$

$$= \sum_{\substack{m\geq 0, n\geq 1 \\ m+n\leq N}} np_{m,n} + C^{-1}\sum_{n=1}^{N} n\sum_{i=0}^{\infty} d_i\left(\frac{c_i\alpha_i^{N-n+1}}{1-\alpha_i} + \frac{c_{i+1}\alpha_{i+1}^{N-n+1}}{1-\alpha_{i+1}}\right)\beta_i^n$$

$$+ C^{-1}\sum_{i=0}^{\infty} d_i\left(\frac{c_i}{1-\alpha_i} + \frac{c_{i+1}}{1-\alpha_{i+1}}\right)\frac{N+1-N\beta_i}{(1-\beta_i)^2}\beta_i^{N+1}. \qquad (4.24)$$

The terms in the series above are alternating and the decrease of these terms is also bounded by $R(i, m, n)$. Hence, if $R(i, m, n) < 1$ for some i, then from there on the error of each partial sum is bounded by its final term.

4.6. Numerical examples

In this section we present some numerical results. In table 4.1 we list for $\mu' = 1$, $\mu = 10$ and decreasing values of p the probabilities $p_{0,1}, p_{0,2}, p_{0,3}$ and $p_{0,4}$ computed with an accuracy of 0.1% by using the series (4.20). The results in table 4.1 illustrate that the convergence of the series (4.20) is faster for states further away from the origin and that N increases when p decreases.

In table 4.2 we list values of L_I with an accuracy of 0.1% for $\mu' = 1$, $\mu = 10$ and decreasing values of p. In the second column L_I is calculated by use of (4.24); in the fourth column the same calculations are done by use of (4.24) where N is replaced by a somewhat larger number, M say. Of course, then some extra effort is needed to solve the equilibrium equations for $m + n \leq M$, but this effort is easily compensated by the advantages of efficiently computing the series in the expression (4.24).

p	$p_{0,1}$	$p_{0,2}$	$p_{0,3}$	$p_{0,4}$	N
0.1	0.0989 (4)	0.0010 (2)	0.0000 (2)	0.0000 (2)	0
0.3	0.2876 (8)	0.0109 (3)	0.0003 (2)	0.0000 (2)	0
0.5	0.4540 (40)	0.0396 (4)	0.0020 (3)	0.0001 (2)	0
0.7		0.1130 (5)	0.0086 (3)	0.0006 (2)	1
0.9		0.3462 (19)	0.0508 (4)	0.0048 (3)	1

Table 4.1.
Values of $p_{0,1}$, $p_{0,2}$, $p_{0,3}$ and $p_{0,4}$ with an accuracy of 0.1% for $\mu' = 1$, $\mu = 10$ and decreasing values of p. The numbers in parentheses denote the number of terms in (4.20) needed.

p	L_I	N	L_I	M
0.1	0.1009 (4)	0	0.1010 (2)	1
0.3	0.3113 (8)	0	0.3116 (2)	2
0.5	0.5442 (39)	0	0.5440 (3)	2
0.7	0.8333 (5)	1	0.8332 (3)	2
0.9	1.3701 (18)	1	1.3703 (3)	3

Table 4.2.
Values of the mean number of jobs L_I in queue I with an accuracy of 0.1% for $\mu' = 1$, $\mu = 10$ and decreasing values of p. In the second column L_I is calculated by use of (4.24); in the fourth column the same calculations are done by use of (4.24) where N is replaced by M. The numbers in parentheses denote the number of terms in each of the series in (4.24) needed.

4.7. Conclusion

In the chapters 3 and 4 we treated two queueing problems as an application of the theory in chapter 2. For these two problems we found explicit expressions for the stationary queue length probabilities as well as for some global performance measures. These expressions are useful from a numerical point of view, since they can be calculated efficiently and accurately. In chapter 5 we show how the compensation approach can be extended to the asymmetric

shortest queue problem. This problem differs from the ones studied so far with respect to the form of the state space; this problem can be modelled as a Markov process on two coupled regions $n \geq 0$ and $n \leq 0$.

Chapter 5

The asymmetric shortest queue problem

In chapter 2 as the state space we have chosen the lattice in the first quadrant of $\mathbb{R}^2$. In this chapter we analyse the asymmetric shortest queue problem, which cannot be modelled as a Markov process on such form of state space. This problem is characterized as follows. Jobs arrive in a Poisson stream at a system consisting of two parallel exponential servers with *different* service rates. The jobs require an exponentially distributed workload with unit mean. On arrival a job joins the shortest queue and, in case of equal queue lengths, joins one queue or the other with probability $1-q$ and q respectively, where q is an arbitrary number between 0 and 1. This problem can be modelled as a Markov process on the lattice in the right half-plane of $\mathbb{R}^2$ with different properties in the upper and lower quadrant. It appears that the compensation approach also works for this problem. So extensions of the approach to a more general state space are quite well possible.

Only few analytical results for the asymmetric shortest queue problem are available in the literature. The uniformization approach used by Kingman [44] and Flatto and McKean [23] to analyse the symmetric version does not seem to be generalizable to the asymmetric version. Cohen and Boxma [14] and Fayolle and Iasnogorodski [19,21,40] show that the analysis of the asymmetric shortest queue problem can be reduced to that of a *simultaneous* boundary value problem in two unknowns. This type of boundary value problem stems from the interaction between the upper and lower quadrant of the state space and requires further research. Knessl, Matkowsky, Schuss and Tier [46] derive asymptotic expressions for the stationary queue-length distribution. To our knowledge no further results exist in the literature.

The compensation approach constructs solutions on the upper and lower quadrant. These solutions are infinite sums of products and, due to the interaction at the horizontal axis, these infinite sums have a *binary tree structure*. The expressions for the equilibrium probabilities easily lead to similar ones for the moments of the sojourn time or other quantities of interest. The compensation approach further is easily adapted to small modifications in the model; the approach can be extended to a threshold-type shortest queue problem and to the shortest queue problem with parallel multi-server queues. The analytical results offer efficient numerical algorithms. In fact, all nice numerical properties of the symmetric problem are preserved. The compensation approach yields recursion relations by which the terms in the binary tree can easily be calculated and few terms suffice due to the fact that the convergence is exponentially fast. In addition, bounds for the error on each partial tree are provided. For *highly unbalanced*

systems however, the series of product forms may diverge for small m and n. A system is called *highly unbalanced* if one of the servers is working much faster than the other. Therefore we derive a numerically stable recursive algorithm for solving the equilibrium equations around the origin of the state space from the solution on the complement. This approach can also be used if convergence of the series of product forms in states around the origin is slow compared to the convergence in states further away from the origin.

This chapter is organized as follows. In section 5.1 the model and the equilibrium equations are formulated. In section 5.2 we outline the compensation approach and we define the resulting infinite sum of product form solutions denoted by $x_{m,n}$. Sections 5.3, 5.4 and 5.5 are devoted to prove that $x_{m,n}$ converges absolutely. Section 5.6 presents the main result stating that $p_{m,n}$ equals $x_{m,n}$ up to a normalizing constant C. In sections 5.7 and 5.8 similar series of product form solutions are derived for C and the moments of the sojourn time. Section 5.9 comments on two extensions of the compensation approach and concludes the analytical treatment. The rest of the chapter approaches the results from a computational point of view. In section 5.10 we derive bounds on the contribution of each subtree and the next section presents a basic scheme for the computation of the trees of product forms. In section 5.12 we propose an efficient and numerically stable algorithm for numerically solving the equilibrium equations for states around the origin of the state space from the solution on the complement of the state space. Section 5.13 is devoted to some numerical considerations and presents numerical results. An alternative strategy to the computation of trees is discussed in section 5.14. The final section is devoted to conclusions.

5.1. Model and equilibrium equations

Consider a queueing system consisting of two parallel servers with service rates γ_1 and γ_2 respectively, where $\gamma_1 > 0$, $\gamma_2 > 0$ and $\gamma_1 + \gamma_2 = 2$ (see figure 5.1). Jobs arrive according to a Poisson stream with rate 2ρ where $0 < \rho < 1$. On arrival a job joins the shortest queue and, if queues have equal lengths, joins one queue or the other with probability $1-q$ and q respectively, where q is an arbitrary number between 0 and 1. The jobs require exponentially distributed service times with unit mean, the service times are supposed to be independent and independent of the arrival process. This model is known as the asymmetric shortest queue model. This queueing system can be represented by a continuous-time Markov process, whose state space consists of the pairs (i, j), $i, j = 0, 1, \ldots$ where i and j are the lengths of the two queues. Instead of i and j we use the variables $m = \min(i, j)$ and $n = j - i$. Let $\{p_{m,n}\}$ be the equilibrium distribution. The transition-rate diagram is depicted in figure 5.2. The equilibrium equations for $\{p_{m,n}\}$ are formulated below.

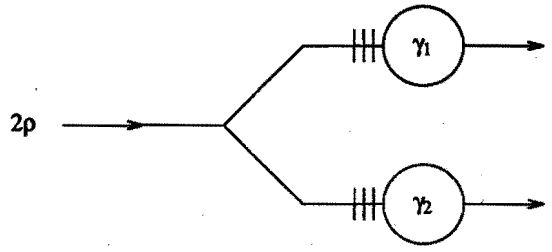


Figure 5.1.
The asymmetric shortest queue model. Arriving jobs join the shortest queue and, if queues have equal lengths, join either queue with probability $1-q$ *and* q *respectively. It is supposed that* $0 < \rho < 1$ *and* $\gamma_1 + \gamma_2 = 2$.

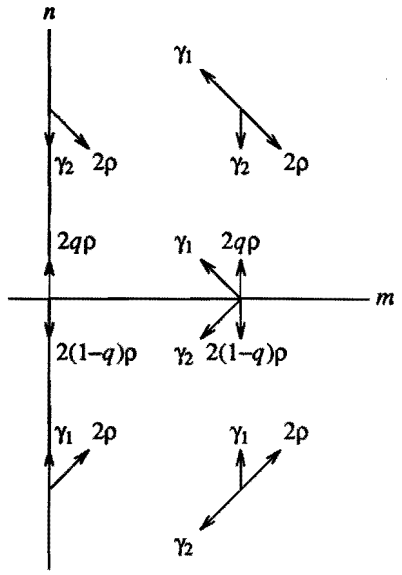


Figure 5.2.
Transition-rate diagram of the asymmetric shortest queue model in figure 5.1.

$$p_{m,n}2(\rho+1) = p_{m-1,n+1}2\rho + p_{m,n+1}\gamma_2 + p_{m+1,n-1}\gamma_1 , \qquad m>0, n>1 \quad (5.1)$$

$$p_{m,1}2(\rho+1) = p_{m-1,2}2\rho + p_{m,2}\gamma_2 + p_{m+1,0}\gamma_1 + p_{m,0}2q\rho , \qquad m>0 \quad (5.2)$$

$$p_{0,n}(2\rho+\gamma_2) = p_{0,n+1}\gamma_2 + p_{1,n-1}\gamma_1 , \qquad n>1 \quad (5.3)$$

$$p_{0,1}(2\rho+\gamma_2) = p_{0,2}\gamma_2 + p_{1,0}\gamma_1 + p_{0,0}2q\rho , \qquad (5.4)$$

$$p_{m,n}2(\rho+1) = p_{m-1,n-1}2\rho + p_{m,n-1}\gamma_1 + p_{m+1,n+1}\gamma_2 , \qquad m>0, n<-1 \quad (5.5)$$

$$p_{m,-1}2(\rho+1) = p_{m-1,-2}2\rho + p_{m,-2}\gamma_1 + p_{m+1,0}\gamma_2 + p_{m,0}2(1-q)\rho , \qquad m>0 \quad (5.6)$$

$$p_{0,n}(2\rho+\gamma_1) = p_{0,n-1}\gamma_1 + p_{1,n+1}\gamma_2 , \qquad n<-1 \quad (5.7)$$

$$p_{0,-1}(2\rho+\gamma_1) = p_{0,-2}\gamma_1 + p_{1,0}\gamma_2 + p_{0,0}2(1-q)\rho , \qquad (5.8)$$

$$p_{m,0}2(\rho+1) = p_{m-1,1}2\rho + p_{m,1}\gamma_2 + p_{m-1,-1}2\rho + p_{m,-1}\gamma_1 , \qquad m>0 \quad (5.9)$$

$$p_{0,0}2\rho = p_{0,1}\gamma_2 + p_{0,-1}\gamma_1 . \qquad (5.10)$$

In figure 5.3 it is illustrated where the different types of conditions hold.

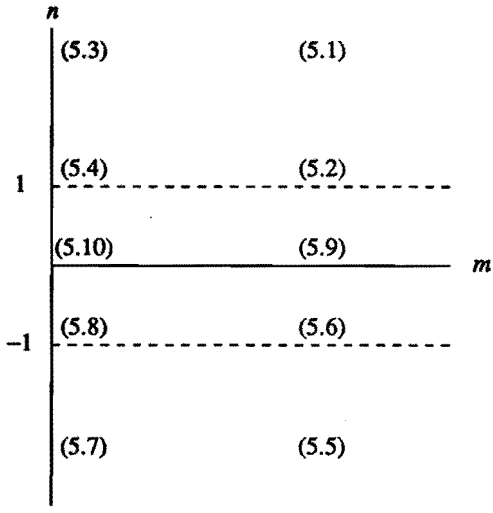


Figure 5.3.
The different types of conditions for the equilibrium distribution $\{p_{m,n}\}$.

For the symmetric problem, i.e., when $\gamma_1 = \gamma_2 = 1$ and $q = 1/2$, the regions $n \geq 0$ and $n \leq 0$ are mirror images of each other, which implies that $p_{m,-n} = p_{m,n}$, so the analysis can then be

restricted to one region. For the asymmetric problem, however, these regions are no longer mirror images of each other, so we have to construct solutions on each of the two regions. In the following sections we try to prove that there are α_i, η_i, ξ_i, and c_i, d_i and e_i such that

$$p_{m,n} = \sum_{i=0}^{\infty} c_i \alpha_i^m \eta_i^n \quad \text{for} \quad m \geq 0\,,\ n > 0\,;$$

$$p_{m,n} = \sum_{i=0}^{\infty} d_i \alpha_i^m \xi_i^{-n} \quad \text{for} \quad m \geq 0\,,\ n < 0\,;$$

$$p_{m,0} = \sum_{i=0}^{\infty} e_i \alpha_i^m \qquad \text{for} \quad m \geq 0\,.$$

After introducing the first terms these series are constructed by adding terms in the upper quadrant satisfying (5.1) and by adding terms in the lower quadrant satisfying (5.5) so as to alternately satisfy the vertical conditions (5.3) and (5.7) and the horizontal conditions (5.2), (5.6) and (5.9). Afterwards it is checked whether the remaining conditions (5.4), (5.8) and (5.10) are satisfied.

5.2. The compensation approach

The compensation approach starts with the solution describing the asymptotic behaviour of $p_{m,n}$ as $m \to \infty$. Numerical experiments suggest that there are α_0, β_1, β_2, d_1 and d_2 such that

$$\begin{aligned} p_{m,n} &\sim K d_1 \alpha_0^m \beta_1^n \quad (m \to \infty\,,\ n > 0)\,; \\ p_{m,n} &\sim K d_2 \alpha_0^m \beta_2^{-n} \quad (m \to \infty\,,\ n < 0)\,; \\ p_{m,0} &\sim K \alpha_0^m \quad (m \to \infty)\,, \end{aligned} \tag{5.11}$$

for some constant K. The question arises: what are α_0, β_1, β_2, d_1 and d_2? First α_0 is found by following the same reasoning as in section 1.1, yielding

$$\alpha_0 = \rho^2\,.$$

The products (5.11) describe the behaviour of $p_{m,n}$ away from the vertical boundary. Therefore they have to satisfy the conditions (5.1), (5.2), (5.5), (5.6) and (5.9). Insertion of $\alpha_0^m \beta_1^n$ in (5.1) and then dividing the resulting equation by $\alpha_0^{m-1} \beta_1^{n-1}$ yields a quadratic equation for β_1. A similar equation is obtained for β_2 by use of (5.5).

Lemma 5.1.

(i) *The product* $\alpha^m\beta^n$ *is a solution of equation (5.1) if and only if*

$$\alpha\beta 2(\rho+1) = \beta^2 2\rho + \alpha\beta^2\gamma_2 + \alpha^2\gamma_1 \; ; \tag{5.12}$$

(ii) *The product* $\alpha^m\beta^{-n}$ *is a solution of equation (5.5) if and only if*

$$\alpha\beta 2(\rho+1) = \beta^2 2\rho + \alpha\beta^2\gamma_1 + \alpha^2\gamma_2 \; . \tag{5.13}$$

For fixed α the quadratic equation (5.12) in β is solved by (cf. (2.41))

$$X_{\pm}(\alpha) = \alpha \frac{\rho+1 \pm \sqrt{(\rho+1)^2 - (2\rho+\alpha\gamma_2)\gamma_1}}{2\rho+\alpha\gamma_2} \; .$$

Let $Y_{\pm}(\beta)$ be the roots of (5.12) for fixed β. Similarly $x_{\pm}(\alpha)$ and $y_{\pm}(\beta)$ are the roots of (5.13) for fixed α and β respectively. Applying lemma 5.1(i) to $\alpha_0^m\beta_1^n$ with $\alpha_0 = \rho^2$ we obtain the roots $\beta_1 = X_+(\rho^2) = \rho$ and $\beta_1 = X_-(\rho^2)$. The first root yields the asymptotic solution $p_{m,n} \sim K\rho^{2m}\rho^n$ for some K, corresponding to the equilibrium distribution of two independent $M \mid M \mid 1$ queues, each with a workload ρ. However, the queues of the shortest queue model are strongly dependent, so the only reasonable choice is

$$\beta_1 = X_-(\rho^2) = \frac{\rho^2\gamma_1}{2+\rho\gamma_2} \; .$$

Similarly we obtain from lemma 5.1(ii)

$$\beta_2 = x_-(\rho^2) = \frac{\rho^2\gamma_2}{2+\rho\gamma_1} \; .$$

The coefficient d_1 is found by substituting $p_{m,n} = d_1\alpha_0^m\beta_1^n$ for $m \geq 0$, $n > 0$ and $p_{m,0} = \alpha_0^m$ for $m \geq 0$ in condition (5.2). This results in the following equation, which is simplified by use of equation (5.12).

$$d_1\alpha_0\gamma_1 = \alpha_0\gamma_1 + 2q\rho \; .$$

A similar equation is obtained for d_2 by insertion of $p_{m,n} = d_2\alpha_0^m\beta_2^{-n}$ for $m \geq 0$, $n < 0$ and $p_{m,0} = \alpha_0^m$ for $m \geq 0$ in condition (5.6). This yields

$$d_2\alpha_0\gamma_2 = \alpha\gamma_2 + 2(1-q)\rho \; .$$

For d_1 and d_2 given by these two equations, it is easily verified that condition (5.9) is also satisfied. Hence, for the values of α_0, β_1, β_2, d_1 and d_2 found, we conclude that the sequence $p_{m,n}$ given by

$$d_1\alpha_0^m\beta_1^n \quad \text{for} \quad m \geq 0,\ n > 0 \; ;$$

$d_2\alpha_0^m\beta_2^{-n}$ for $m \geq 0, n < 0$;

α_0^m for $m \geq 0, n = 0$,

satisfies the conditions (5.1), (5.2), (5.5), (5.6) and (5.9). However, this sequence violates the conditions (5.3) and (5.7) at the n-axis. To compensate for $d_1\alpha_0^m\beta_1^n$ on the positive n-axis we follow the same procedure as in section 1.1:

Try to find c_1, α, β with α, β satisfying (5.12) such that
$d_1\alpha_0^m\beta_1^n + d_1c_1\alpha^m\beta^n$ satisfies (5.3).

To satisfy (5.3) for all $n > 1$ the β-factor of the two terms must be the same, so we have to take

$$\beta = \beta_1 .$$

For $\beta = \beta_1$ equation (5.12) has roots $Y_+(\beta_1)$ and $Y_-(\beta_1)$ satisfying $Y_+(\beta_1) > \beta_1 > Y_-(\beta_1) > 0$ (cf. lemma 2.15). So $\alpha_0 = Y_+(\beta_1)$ and thus we are forced to take the smaller root for α, i.e.,

$$\alpha = \alpha_1 = Y_-(\beta_1) .$$

Insertion of $d_1\alpha_0^m\beta_1^n + d_1c_1\alpha_1^m\beta_1^n$ in (5.3) and dividing by $d_1\beta_1^{n-1}$ leads to an equation for c_1 which is easily solved. The same procedure is applied to compensate for $d_2\alpha_0^m\beta_2^{-n}$ on the negative n-axis: add $d_2c_2\alpha_2^m\beta_2^{-n}$ where $\alpha_2 = y_-(\beta_2)$ and solve c_2 from condition (5.7). This results in the sum

$d_1\alpha_0^m\beta_1^n + d_1c_1\alpha_1^m\beta_1^n$ for $m \geq 0, n > 0$;

$d_2\alpha_0^m\beta_2^{-n} + d_2c_2\alpha_2^m\beta_2^{-n}$ for $m \geq 0, n < 0$;

α_0^m for $m \geq 0, n = 0$,

satisfying the conditions (5.1), (5.3), (5.5) and (5.7). This procedure is generalized in the following lemma (cf. lemma 1.2).

Lemma 5.2.

(i) *Let* $z_{m,n} = Y_+^m(\beta)\beta^n + cY_-^m(\beta)\beta^n$ *for* $m \geq 0, n > 0$.
Then $z_{m,n}$ *satisfies (5.1) and (5.3) if* c *is given by*

$$c = -\frac{Y_-(\beta) - \beta}{Y_+(\beta) - \beta} .$$

(ii) *Let* $w_{m,n} = y_+^m(\beta)\beta^{-n} + cy_-^m(\beta)\beta^{-n}$ *for* $m \geq 0, n < 0$.
Then $w_{m,n}$ *satisfies (5.5) and (5.7) if* c *is given by*

$$c = -\frac{y_-(\beta) - \beta}{y_+(\beta) - \beta} .$$

The new terms $d_1c_1\alpha_1^m\beta_1^n$ and $d_2c_2\alpha_2^m\beta_2^{-n}$ violate the conditions (5.2), (5.6) and (5.9) at the m-axis. Compensation of these errors requires addition of terms with the same α-factor as the two error terms. Since the two error terms have different α-factors, we have to compensate for each of them *separately*. For the compensation of $d_1c_1\alpha_1^m\beta_1^n$ we use the following procedure:

Try to find $d_3, d_4, f_1, \beta_3, \beta_4$ with α_1, β_3 satisfying (5.12)
and α_1, β_4 satisfying (5.13) such that the sequence $p_{m,n}$ given by

$$\begin{aligned} &c_1d_1\alpha_1^m\beta_1^n + c_1d_3\alpha_1^m\beta_3^n && \text{for } m \geq 0,\ n > 0; \\ &c_1d_4\alpha_1^m\beta_4^{-n} && \text{for } m \geq 0,\ n < 0; \\ &c_1f_1\alpha_1^m && \text{for } m \geq 0,\ n = 0, \end{aligned} \tag{5.14}$$

satisfies (5.2), (5.6) and (5.9).

For $\alpha = \alpha_1$ equation (5.12) has roots $X_+(\alpha_1)$ and $X_-(\alpha_1)$ satisfying $X_+(\alpha_1) > \alpha_1 > X_-(\alpha_1) > 0$. So $\beta_1 = X_+(\alpha_1)$ leaving the smaller root for β_3, i.e.,

$$\beta_3 = X_-(\alpha_1).$$

For β_4 we may choose between the roots $x_+(\alpha_1)$ and $x_-(\alpha_1)$ of equation (5.13) with $\alpha = \alpha_1$. It is desirable to keep $d_4\alpha_1^m\beta_4^{-n}$ as small as possible, so we take the smaller root, i.e.,

$$\beta_4 = x_-(\alpha_1).$$

Insertion of the terms (5.14) in (5.2), (5.6) and (5.9) and dividing by $c_1\alpha_1^{m-1}$ leads to three equations for d_3, d_4 and f_1 which are easily solved. The same procedure is applied to compensate for $d_2c_2\alpha_2^m\beta_2^{-n}$: add $d_5c_2\alpha_2^m\beta_5^n$ solution in the upper quadrant, add $d_6c_2\alpha_2^m\beta_6^{-n}$ to the solution in the lower quadrant and add $c_2f_2\alpha_2^m$ to the solution on the m-axis where

$$\beta_5 = X_-(\alpha_2),$$

$$\beta_6 = x_-(\alpha_2),$$

and solve d_5, d_6 and f_2 from the conditions (5.2), (5.6) and (5.9). This results in the sum

$$\begin{aligned} &d_1\alpha_0^m\beta_1^n + d_1c_1\alpha_1^m\beta_1^n + d_3c_1\alpha_1^m\beta_3^n + d_5c_2\alpha_2^m\beta_5^n && \text{for } m \geq 0,\ n > 0; \\ &d_2\alpha_0^m\beta_2^{-n} + d_2c_2\alpha_2^m\beta_2^{-n} + d_4c_1\alpha_1^m\beta_4^{-n} + d_6c_2\alpha_2^m\beta_6^{-n} && \text{for } m \geq 0,\ n < 0; \\ &\alpha_0^m + c_1f_1\alpha_1^m + c_2f_2\alpha_2^m && \text{for } m \geq 0, \end{aligned}$$

satisfying the conditions (5.1), (5.2), (5.5), (5.6) and (5.9). The following lemma generalizes the compensation at the m-axis.

Lemma 5.3.

(i) *Let*

$$z_{m,n}=\begin{cases}\alpha^m X_+^n(\alpha)+d\alpha^m X_-^n(\alpha) & \text{for } m\geq 0,\ n>0,\\ g\alpha^m x_-^{-n}(\alpha) & \text{for } m\geq 0,\ n<0,\\ f\alpha^m & \text{for } m\geq 0,\ n=0.\end{cases}$$

Then $z_{m,n}$ satisfies (5.1), (5.2), (5.5), (5.6) and (5.9) if d, g and f are given by

$$d=-\frac{\dfrac{\alpha\gamma_1+2q\rho}{X_-(\alpha)}+\dfrac{\alpha\gamma_2+2(1-q)\rho}{x_+(\alpha)}-2(\rho+1)}{\dfrac{\alpha\gamma_1+2q\rho}{X_+(\alpha)}+\dfrac{\alpha\gamma_2+2(1-q)\rho}{x_+(\alpha)}-2(\rho+1)},$$

$$g=-\frac{\gamma_1(\alpha\gamma_2+2(1-q)\rho)\left[\dfrac{1}{X_-(\alpha)}-\dfrac{1}{X_+(\alpha)}\right]}{\gamma_2\left[\dfrac{\alpha\gamma_1+2q\rho}{X_+(\alpha)}+\dfrac{\alpha\gamma_2+2(1-q)\rho}{x_+(\alpha)}-2(\rho+1)\right]},$$

$$f=-\frac{\alpha\gamma_1\left[\dfrac{1}{X_-(\alpha)}-\dfrac{1}{X_+(\alpha)}\right]}{\dfrac{\alpha\gamma_1+2q\rho}{X_+(\alpha)}+\dfrac{\alpha\gamma_2+2(1-q)\rho}{x_+(\alpha)}-2(\rho+1)}.$$

(ii) *Let*

$$w_{m,n}=\begin{cases}\alpha^m x_+^{-n}(\alpha)+d\alpha^m x_-^{-n}(\alpha) & \text{for } m\geq 0,\ n<0,\\ g\alpha^m X_-^n(\alpha) & \text{for } m\geq 0,\ n>0,\\ f\alpha^m & \text{for } m\geq 0,\ n=0.\end{cases}$$

Then $w_{m,n}$ satisfies (5.1), (5.2), (5.5), (5.6) and (5.9) if d, g and f are given by

$$d=-\frac{\dfrac{\alpha\gamma_1+2q\rho}{X_+(\alpha)}+\dfrac{\alpha\gamma_2+2(1-q)\rho}{x_-(\alpha)}-2(\rho+1)}{\dfrac{\alpha\gamma_1+2q\rho}{X_+(\alpha)}+\dfrac{\alpha\gamma_2+2(1-q)\rho}{x_+(\alpha)}-2(\rho+1)},$$

$$g=-\frac{\gamma_2(\alpha\gamma_1+2q\rho)\left[\dfrac{1}{x_-(\alpha)}-\dfrac{1}{x_+(\alpha)}\right]}{\gamma_1\left[\dfrac{\alpha\gamma_1+2q\rho}{X_+(\alpha)}+\dfrac{\alpha\gamma_2+2(1-q)\rho}{x_+(\alpha)}-2(\rho+1)\right]},$$

$$f = -\frac{\alpha\gamma_2\left[\dfrac{1}{x_-(\alpha)} - \dfrac{1}{x_+(\alpha)}\right]}{\dfrac{\alpha\gamma_1 + 2q\rho}{X_+(\alpha)} + \dfrac{\alpha\gamma_2 + 2(1-q)\rho}{x_+(\alpha)} - 2(\rho+1)}.$$

We added terms in the upper and lower quadrant to compensate for $d_1 c_1 \alpha_1^m \beta_1^n$ and $d_2 c_2 \alpha_2^m \beta_2^n$ on the horizontal axis and in doing so introduced new errors at the vertical axis, since the terms added violate the conditions (5.3) and (5.7). It is clear how to continue: the compensation procedure consists of adding on terms so as to alternately compensate for the error at the vertical axis, according to lemma 5.2, and for the error at the horizontal axis, according to lemma 5.3. This results in an infinite sum of terms in the upper and lower quadrant. Due to the compensation at the horizontal axis these sums have a *binary tree structure*. Let $x_{m,n}$ be the resulting infinite linear combination of products $\alpha^m\beta^n$. The detailed definition of $x_{m,n}$ is given below. We first formulate the recursion relations defining the α's and β's in this linear combination. These α's and β's can be represented by the binary tree depicted in figure 5.4. This tree is called the *parameter tree*.

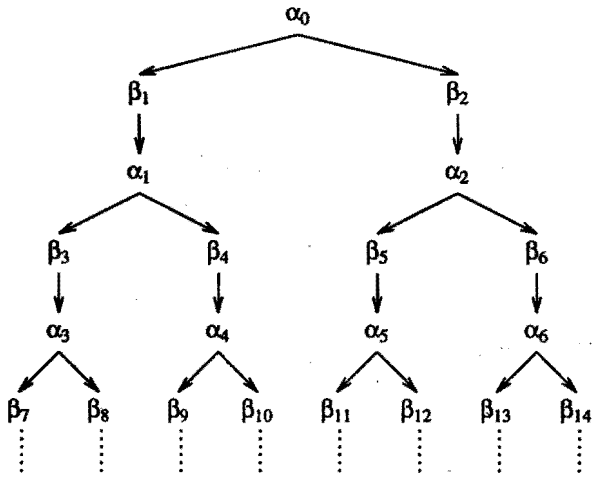


Figure 5.4.
The parameter tree. The sequences $\{\alpha_i\}$ and $\{\beta_i\}$ are generated by use of the quadratic equations in the upper and lower quadrant.

In figure 5.4 the α's and β's are numbered from the root and from left to right. For specifying the recursion relations to generate the parameter tree, we use the following notations:

$\beta_{l(i)}$ = the left descendant of α_i ;

$\beta_{r(i)}$ = the right descendant of α_i ;

$\alpha_{p(i)}$ = the α-parent of β_i .

Further define L as the set of indices i of β_i's that are left descendants and R as the set of indices i of β_i's that are right descendants, i.e.,

$$L = \{l(i) \mid i = 0, 1, 2, \cdots\},$$

$$R = \{r(i) \mid i = 0, 1, 2, \cdots\}.$$

It is easy to check that for the numbering in figure 5.3 we have

$$l(i) = 2i + 1 ; \quad r(i) = 2i + 2 ; \quad p(i) = \left\lfloor \frac{i-1}{2} \right\rfloor ;$$

$$L = \{2i + 1 \mid i = 0, 1, 2, \cdots\} ;$$

$$R = \{2i + 2 \mid i = 0, 1, 2, \cdots\}.$$

As starting value we take

$$\alpha_0 = \rho^2 .$$

Then for all $i \geq 0$ the left descendant $\beta_{l(i)}$ of α_i is defined as the smaller root of equation (5.12) with $\alpha = \alpha_i$ and the right descendant $\beta_{r(i)}$ of α_i is defined as the smaller root of equation (5.13) with $\alpha = \alpha_i$. The descendant $\alpha_{l(i)}$ of $\beta_{l(i)}$ is defined as the smaller root of (5.12) with $\beta = \beta_{l(i)}$ and and the descendant $\alpha_{r(i)}$ of $\beta_{r(i)}$ is defined as the smaller root of (5.13) with $\beta = \beta_{r(i)}$. By lemma 2.15 we have for $0 < \alpha < 1$

$$X_+(\alpha) > \alpha > X_-(\alpha) > 0$$

and similar inequalities hold for $Y_{\pm}(\beta)$, $x_{\pm}(\alpha)$ and $y_{\pm}(\beta)$. Using induction it then follows that for all $i \geq 0$,

$$\beta_{l(i)} = X_-(\alpha_i) , \quad \beta_{r(i)} = x_-(\alpha_i) , \tag{5.15}$$

$$\alpha_{l(i)} = Y_-(\beta_{l(i)}) , \; \alpha_{r(i)} = y_-(\beta_{r(i)}) ,$$

and that

$$\alpha_i > \beta_{l(i)} > \alpha_{l(i)} > 0 ,$$

$$\alpha_i > \beta_{r(i)} > \alpha_{r(i)} > 0 .$$

So $\{\alpha_i\}$ and $\{\beta_i\}$ form a *decreasing positive tree*. For all $i \in L$ the products $\alpha_{p(i)}^m \beta_i^n$ and $\alpha_i^m \beta_i^n$

satisfy condition (5.1) by lemma 5.1(i); for all $i \in R$ the products $\alpha_{p(i)}^m \beta_i^{-n}$ and $\alpha_i^m \beta_i^{-n}$ satisfy condition (5.5) by lemma 5.1(ii). Let us define the infinite sum $x_{m,n}$ by

$$x_{m,n} = \sum_{i \in L} d_i (c_{p(i)} \alpha_{p(i)}^m + c_i \alpha_i^m) \beta_i^n \quad \text{for} \quad m \geq 0, n > 0, \tag{5.16}$$

$$x_{m,n} = \sum_{i \in R} d_i (c_{p(i)} \alpha_{p(i)}^m + c_i \alpha_i^m) \beta_i^{-n} \quad \text{for} \quad m \geq 0, n < 0. \tag{5.17}$$

By linearity the sum $x_{m,n}$ satisfies the conditions (5.1) and (5.5). Below we define the coefficients c_i and d_i such that $x_{m,n}$ also satisfies the boundary conditions (5.2), (5.3), (5.6), (5.7) and (5.9). First we note that for each $m \geq 0$, $n > 0$ the sums $x_{m,n}$ and $x_{m,-n}$ can be represented in one binary tree derived from the parameter tree. This tree is called the *compensation tree* and depicted in figure 5.5. The indices i of left descendants $d_i(c_{p(i)} \alpha_{p(i)}^m + c_i \alpha_i^m) \beta_i^n$ in the compensation tree run through L. The indices i of right descendants run through R. So $x_{m,n}$ is the sum of all left descendants and $x_{m,-n}$ is the sum of all right descendants in the compensation tree.

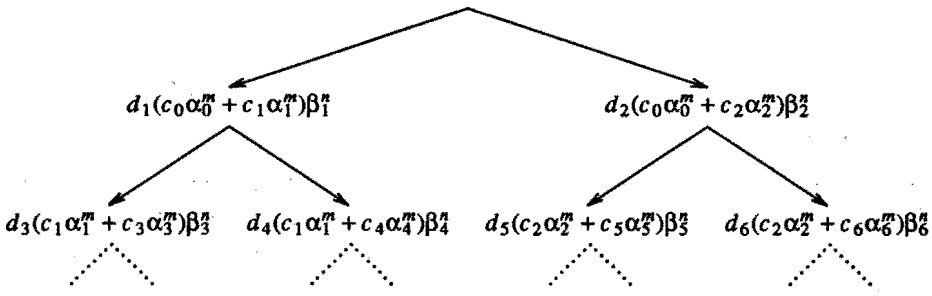


Figure 5.5.
The compensation tree. For each $m \geq 0$, $n > 0$ the left descendants add up to $x_{m,n}$ and the right descendants add up to $x_{m,-n}$.

The coefficients c_i are such that $(c_{p(i)} \alpha_{p(i)}^m + c_i \alpha_i^m) \beta_i^n$ satisfies (5.3) for all $i \in L$ and such that $(c_{p(i)} \alpha_{p(i)}^m + c_i \alpha_i^m) \beta_i^{-n}$ satisfies (5.7) for all $i \in R$. Application of lemma 5.2 and using the relations (5.15) yields that c_i can be obtained from $c_{p(i)}$ by

$$c_i = -\frac{Y_-(\beta_i) - \beta_i}{Y_+(\beta_i) - \beta_i} c_{p(i)} \qquad (i \in L),$$

$$c_i = -\frac{y_-(\beta_i) - \beta_i}{y_+(\beta_i) - \beta_i} c_{p(i)} \qquad (i \in R),$$

with initial condition

$$c_0 = 1 .$$

By linearity of (5.3) and (5.7) it then follows that $x_{m,n}$ satisfies (5.3) and (5.7). Compensation at the horizontal axis requires pairs with the same α-factor. Therefore we rewrite $x_{m,n}$ as

$$x_{m,n} = c_0 d_1 \beta_1^n \alpha_0^m + \sum_{i \in L} c_i (d_i \beta_i^n + d_{l(i)} \beta_{l(i)}^n) \alpha_i^m + \sum_{i \in R} c_i d_{l(i)} \beta_{l(i)}^n \alpha_i^m$$

$$\text{for} \quad m \geq 0 , \, n > 0 , \qquad (5.18)$$

$$x_{m,n} = c_0 d_2 \beta_2^{-n} \alpha_0^m + \sum_{i \in L} c_i d_{r(i)} \beta_{r(i)}^{-n} \alpha_i^m + \sum_{i \in R} c_i (d_i \beta_i^{-n} + d_{r(i)} \beta_{r(i)}^{-n}) \alpha_i^m$$

$$\text{for} \quad m \geq 0 , \, n < 0 , \qquad (5.19)$$

and define $x_{m,n}$ on the m-axis by

$$x_{m,0} = \sum_{i=0}^{\infty} c_i f_i \alpha_i^m = c_0 f_0 \alpha_0^m + \sum_{i \in L} c_i f_i \alpha_i^m + \sum_{i \in R} c_i f_i \alpha_i^m \qquad \text{for} \quad m \geq 0 . \qquad (5.20)$$

The coefficients d_i and f_i are such that for $i \in L$ the terms

$$(d_i \beta_i^n + d_{l(i)} \beta_{l(i)}^n) \alpha_i^m \qquad \text{for} \quad m \geq 0 , \, n > 0 ;$$

$$d_{r(i)} \beta_{r(i)}^{-n} \alpha_i^m \qquad \text{for} \quad m \geq 0 , \, n < 0 ;$$

$$f_i \alpha_i^m \qquad \text{for} \quad m \geq 0 , \, n = 0 ,$$

satisfy (5.2), (5.6) and (5.9) and such that for $i \in R$ the same conditions are satisfied by

$$d_{l(i)} \beta_{l(i)}^n \alpha_i^m \qquad \text{for} \quad m \geq 0 , \, n > 0 ;$$

$$(d_i \beta_i^{-n} + d_{r(i)} \beta_{r(i)}^{-n}) \alpha_i^m \qquad \text{for} \quad m \geq 0 , \, n < 0 ;$$

$$f_i \alpha_i^m \qquad \text{for} \quad m \geq 0 , \, n = 0 .$$

Application of lemma 5.3 and using the relations (5.15) yields that $d_{l(i)}$, $d_{r(i)}$, f_i can be obtained from d_i by

$$d_{l(i)} = - \frac{\dfrac{\alpha_i \gamma_1 + 2q\rho}{X_-(\alpha_i)} + \dfrac{\alpha_i \gamma_2 + 2(1-q)\rho}{x_+(\alpha_i)} - 2(\rho + 1)}{\dfrac{\alpha_i \gamma_1 + 2q\rho}{X_+(\alpha_i)} + \dfrac{\alpha_i \gamma_2 + 2(1-q)\rho}{x_+(\alpha_i)} - 2(\rho + 1)} d_i \qquad (i \in L) ,$$

$$d_{r(i)} = - \frac{\gamma_1 (\alpha_i \gamma_2 + 2(1-q)\rho) \left[\dfrac{1}{X_-(\alpha_i)} - \dfrac{1}{X_+(\alpha_i)} \right]}{\gamma_2 \left[\dfrac{\alpha_i \gamma_1 + 2q\rho}{X_+(\alpha_i)} + \dfrac{\alpha_i \gamma_2 + 2(1-q)\rho}{x_+(\alpha_i)} - 2(\rho + 1) \right]} d_i \qquad (i \in L) ,$$

$$f_i = -\frac{\alpha_i\gamma_1\left[\dfrac{1}{X_-(\alpha_i)} - \dfrac{1}{X_+(\alpha_i)}\right]}{\dfrac{\alpha_i\gamma_1 + 2q\rho}{X_+(\alpha_i)} + \dfrac{\alpha_i\gamma_2 + 2(1-q)\rho}{x_+(\alpha_i)} - 2(\rho+1)}\, d_i \qquad (i\in L),$$

$$d_{l(i)} = -\frac{\gamma_2(\alpha_i\gamma_1 + 2q\rho)\left[\dfrac{1}{x_-(\alpha_i)} - \dfrac{1}{x_+(\alpha_i)}\right]}{\gamma_1\left[\dfrac{\alpha_i\gamma_1 + 2q\rho}{X_+(\alpha_i)} + \dfrac{\alpha_i\gamma_2 + 2(1-q)\rho}{x_+(\alpha_i)} - 2(\rho+1)\right]}\, d_i \qquad (i\in R),$$

$$d_{r(i)} = -\frac{\dfrac{\alpha_i\gamma_1 + 2q\rho}{X_+(\alpha_i)} + \dfrac{\alpha_i\gamma_2 + 2(1-q)\rho}{x_-(\alpha_i)} - 2(\rho+1)}{\dfrac{\alpha_i\gamma_1 + 2q\rho}{X_+(\alpha_i)} + \dfrac{\alpha_i\gamma_2 + 2(1-q)\rho}{x_+(\alpha_i)} - 2(\rho+1)}\, d_i \qquad (i\in R),$$

$$f_i = -\frac{\alpha_i\gamma_2\left[\dfrac{1}{x_-(\alpha_i)} - \dfrac{1}{x_+(\alpha_i)}\right]}{\dfrac{\alpha_i\gamma_1 + 2q\rho}{X_+(\alpha_i)} + \dfrac{\alpha_i\gamma_2 + 2(1-q)\rho}{x_+(\alpha_i)} - 2(\rho+1)}\, d_i \qquad (i\in R),$$

with initially

$$d_1 = \frac{\alpha_0\gamma_1 + 2q\rho}{\alpha_0\gamma_1},$$

$$d_2 = \frac{\alpha_0\gamma_2 + 2(1-q)\rho}{\alpha_0\gamma_2},$$

$$f_0 = 1.$$

By linearity of (5.2), (5.6) and (5.9) it then follows that $x_{m,n}$ satisfies these conditions. This completes the definition of $x_{m,n}$. We may conclude that $\{x_{m,n}\}$ is a formal solution to all equilibrium equations if we show that the conditions (5.4), (5.8) and (5.10) are also satisfied. To show that (5.4) is satisfied, we first rewrite this condition as

$$p_{0,1}(2\rho + \gamma_2) - p_{0,2}\gamma_2 = p_{1,0}\gamma_1 + p_{0,0}2q\rho\,. \tag{5.21}$$

Insertion of the sum (5.16) in the left-hand side of (5.21) yields

$$x_{0,1}(2\rho + \gamma_2) - x_{0,2}\gamma_2 = \sum_{i\in L} d_i\left[(c_{p(i)} + c_i)\beta_i(2\rho + \gamma_2) - (c_{p(i)} + c_i)\beta_i^2\gamma_2\right]. \tag{5.22}$$

For $i\in L$ the coefficient c_i is such that $(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$ satisfies (5.3) or equivalently (5.21). Substituting this term in (5.21) leads to

$$(c_{p(i)} + c_i)\beta_i^n(2\rho + \gamma_2) - (c_{p(i)} + c_i)\beta_i^{n+1}\gamma_2 = (c_{p(i)}\alpha_{p(i)} + c_i\alpha_i)\beta_i^{n-1}\gamma_1 .$$

Dividing both sides of this equality by β_i^{n-1} and then inserting in (5.22) yields

$$x_{0,1}(2\rho + \gamma_2) - x_{0,2}\gamma_2 = \sum_{i\in L} d_i(c_{p(i)}\alpha_{p(i)} + c_i\alpha_i)\gamma_1$$

$$= c_0 d_1 \alpha_0 \gamma_1 + \sum_{i\in L} c_i(d_i + d_{l(i)})\alpha_i\gamma_1 + \sum_{i\in R} c_i d_{l(i)}\alpha_i\gamma_1 . \qquad (5.23)$$

On the other hand, inserting the sum (5.20) in the right-hand side of (5.21) yields

$$x_{1,0}\gamma_1 + x_{0,0}2q\rho = \sum_{i=0}^{\infty} c_i f_i(\alpha_i\gamma_1 + 2q\rho) . \qquad (5.24)$$

From the definition of f_i it is readily verified that

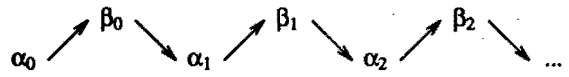
$$f_i = (d_i + d_{l(i)})\frac{\alpha_i\gamma_1}{\alpha_i\gamma_1 + 2q\rho} \qquad (i\in L) ,$$

$$f_i = d_{l(i)}\frac{\alpha_i\gamma_1}{\alpha_i\gamma_1 + 2q\rho} \qquad (i\in R) .$$

These relations reduce the right-hand side of (5.24) to (5.23). So $\{x_{m,n}\}$ indeed satisfies (5.4). Similarly it follows that $\{x_{m,n}\}$ satisfies (5.8). The final condition (5.10) is also satisfied due to the dependence of the equilibrium equations. Hence we can now conclude that $\{x_{m,n}\}$ is a formal solution to all equilibrium equations. The next problem is to prove the convergence of the infinite sum $x_{m,n}$. This is the object in the following three sections.

Remark 5.4.

The analysis simplifies if $\gamma_1 = \gamma_2$. In this case the quadratic equations in the upper and lower quadrant are the same, so in the parameter tree all α's at the same depth and all β's at the same depth are identical and the tree thus simplifies to a sequence with the structure



Consequently $x_{m,n}$ has a linear structure. In fact, up to some multiplicative constant $x_{m,n}$ is equal to $x_{m,n}(\alpha_+)$ defined by (3.20) and (3.21) in chapter 3.

Remark 5.5.

In this section α_0 is found by a heuristic argument. However α_0 can also be found by requiring that the initial products

$$d_1\alpha_0^m\beta_1^n \quad \text{for} \quad m \geq 0 , n > 0 ,$$

$d_2\alpha_0^m\beta_2^{-n}$ for $m \geq 0, n < 0$,

α_0^m for $m \geq 0, n = 0$,

satisfy the conditions (5.1), (5.2), (5.5), (5.6), (5.9) and

$$0 < \beta_1 < \alpha_0, \quad 0 < \beta_2 < \alpha_0.$$

The final inequalities are imposed to ensure that compensation terms are generated in the decreasing direction only (cf. definition 2.9 and conclusion 2.10). It is easily shown that to meet these requirements α_0 has to be the solution in (0, 1) of the equation (cf. equation (2.44))

$$\alpha 2(\rho + 1) = \frac{\alpha\gamma_1 + 2q\rho}{X_+(\alpha)} + \frac{\alpha\gamma_2 + 2(1-q)\rho}{x_+(\alpha)}.$$

The solution in (0, 1) of this equation is indeed given by $\alpha = \rho^2$.

5.3. Absolute convergence of the formal solution

We now prove that the series (5.16), (5.17) and (5.20) defining $x_{m,n}$ *converge absolutely*. Absolute convergence is needed to guarantee equality of (5.16) and (5.18) and of (5.17) and (5.19). These series, however, may diverge for small m and n, but we will prove:

Theorem 5.6 (Absolute convergence).

There is an integer N (to be specified later on) such that:

(i) *The series (5.16) defining $x_{m,n}$ for $m \geq 0$ and $n > 0$, converges absolutely for all $m \geq 0$, $n \geq 0$ with $m + n > N$;*

(ii) *The series (5.17) defining $x_{m,n}$ for $m \geq 0$ and $n < 0$, converges absolutely for all $m \geq 0$, $n \leq 0$ with $m - n > N$;*

(iii) *The series (5.20) defining $x_{m,0}$ for $m \geq 0$, converges absolutely for all $m \geq N$;*

(iv) $$\sum_{\substack{m \geq 0 \\ m + |n| > N}} |x_{m,n}| < \infty.$$

5.4. Preliminary results for the proof of theorem 5.6

To prove the absolute convergence of (5.16), (5.17) and (5.20) we need information about the asymptotic behaviour of α_i, β_i, c_i, d_i, f_i as $i \to \infty$. We first prove that α_i and β_i decrease *exponentially* fast. The next lemma follows directly from lemma 2.15.

Lemma 5.7.

For all $0 < \alpha < 1$

(i) *the ratio* $X_+(\alpha)/\alpha$ *is decreasing and* $X_-(\alpha)/\alpha$ *is increasing;*

(ii) $X_+(\alpha) > \alpha > X_-(\alpha) > 0$.

The same properties hold for $Y_\pm(\beta)$, $x_\pm(\alpha)$ *and* $y_\pm(\beta)$.

Corollary 5.8.

For all $i = 0, 1, 2, \cdots$

$$\alpha_0 \geq \alpha_i > \beta_{l(i)} > \alpha_{l(i)} > 0\,,$$

$$\alpha_0 \geq \alpha_i > \beta_{r(i)} > \alpha_{r(i)} > 0\,,$$

where

$$\beta_{l(i)} \leq \frac{\gamma_1}{2+\rho\gamma_2}\,\alpha_i\,, \quad \alpha_{l(i)} \leq \frac{2\rho}{2+\rho\gamma_2}\,\beta_{l(i)}$$

$$\beta_{r(i)} \leq \frac{\gamma_2}{2+\rho\gamma_1}\,\alpha_i\,, \quad \alpha_{r(i)} \leq \frac{2\rho}{2+\rho\gamma_1}\,\beta_{r(i)}\,.$$

Proof.

The corollary is proved by induction: we descend the parameter tree by starting at the root. Assume that $0 < \alpha_i \leq \alpha_0$, which trivially holds for $i = 0$. Then by lemma 5.7,

$$0 < \beta_{l(i)} = X_-(\alpha_i) \leq \frac{X_-(\alpha_0)}{\alpha_0}\,\alpha_i = \frac{\gamma_1}{2+\rho\gamma_2}\,\alpha_i$$

and $\beta_{l(i)} = X_-(\alpha_i) \leq X_-(\alpha_0) = \beta_1$, so again by lemma 5.7,

$$\alpha_{l(i)} = Y_-(\beta_{l(i)}) \leq \frac{Y_-(\beta_1)}{\beta_1}\,\beta_{l(i)} = \frac{2\rho}{2+\rho\gamma_2}\,\beta_{l(i)}\,.$$

The inequalities for $\beta_{r(i)}$ and $\alpha_{r(i)}$ are proved similarly. □

Corollary 5.8 states that α_i and β_i decrease exponentially fast as $i \to \infty$. The asymptotic behaviour of α_i and β_i is stated in

Lemma 5.9.

As $i \to \infty$, then

$$\frac{\beta_{l(i)}}{\alpha_i} \to \frac{1}{A_2}, \quad \frac{\beta_{r(i)}}{\alpha_i} \to \frac{1}{a_2}$$

$$\frac{\alpha_{l(i)}}{\beta_{l(i)}} \to A_1, \quad \frac{\alpha_{r(i)}}{\beta_{r(i)}} \to a_1,$$

where

$$A_1 = \frac{\rho + 1 - \sqrt{(\rho+1)^2 - 2\rho\gamma_1}}{\gamma_1}, \quad A_2 = \frac{\rho + 1 + \sqrt{(\rho+1)^2 - 2\rho\gamma_1}}{\gamma_1}$$

$$a_1 = \frac{\rho + 1 - \sqrt{(\rho+1)^2 - 2\rho\gamma_2}}{\gamma_2}, \quad a_2 = \frac{\rho + 1 + \sqrt{(\rho+1)^2 - 2\rho\gamma_2}}{\gamma_2}.$$

Proof.

We prove the first limit. The other limits are proved similarly. As $i \to \infty$, then $\alpha_i \to 0$ by corollary 5.8, so

$$\frac{\beta_{l(i)}}{\alpha_i} = \frac{X_-(\alpha_i)}{\alpha_i} \to \frac{A_1 \gamma_1}{2\rho} = \frac{1}{A_2}. \qquad \square$$

The asymptotic behaviour of the coefficients c_i is stated in

Lemma 5.10.

The coefficients c_i are positive for all $i = 0, 1, 2, \cdots$; and as $i \to \infty$, then

$$\frac{c_i}{c_{p(i)}} \to C_l \quad \textit{if } i \textit{ runs through } L;$$

$$\frac{c_i}{c_{p(i)}} \to C_r \quad \textit{if } i \textit{ runs through } R,$$

where

$$C_l = \frac{1 - A_1}{A_2 - 1}, \quad C_r = \frac{1 - a_1}{a_2 - 1}.$$

Proof.

Lemma 5.7 implies that $c_i / c_{p(i)} > 0$ for all $i > 0$. Since $c_0 > 0$ it follows by using induction that $c_i > 0$ for all $i \geq 0$. We now prove the first limit. The other limits are proved similarly. As $i \to \infty$, then $\beta_i \to 0$ by corollary 5.8, so if i runs through L

$$\frac{c_i}{c_{p(i)}} = \frac{1 - Y_-(\beta_i)/\beta_i}{Y_+(\beta_i)/\beta_i - 1} \to \frac{1 - A_1}{A_2 - 1} = C_l \,. \qquad \square$$

Before stating the asymptotic behaviour of d_i and f_i we investigate whether the common denominator in the definitions of $d_{l(i)}$, $d_{r(i)}$ and f_i never vanishes. The denominator

$$\frac{(\alpha\gamma_1 + 2q\rho)\,\alpha}{X_+(\alpha)} + \frac{(\alpha\gamma_2 + 2(1-q)\rho)\,\alpha}{x_+(\alpha)} - \alpha 2(\rho + 1)$$

vanishes at $\alpha = \alpha_0$ and $\alpha = 1$ and is strictly convex for $0 < \alpha < 1$ (cf. the proof of lemma 2.17). Hence the denominator is positive for $0 < \alpha < \alpha_0$. Since $0 < \alpha_i < \alpha_0$ for $i \geq 1$, it then follows that the common denominator in the definitions of $d_{l(i)}$, $d_{r(i)}$ and f_i is positive for $i \geq 1$. Using lemma 5.7 it is easily shown that the numerators in these definitions are also positive. So d_i and f_i are alternating of sign with respect to the depth in the compensation tree, i.e, for all $i \geq 1$,

$$\frac{d_{l(i)}}{d_i} < 0\,, \quad \frac{d_{r(i)}}{d_i} < 0\,,$$

$$\frac{f_{l(i)}}{f_i} < 0\,, \quad \frac{f_{r(i)}}{f_i} < 0\,.$$

The next lemma describes the asymptotic behaviour of the coefficients d_i.

Lemma 5.11.

As $i \to \infty$, then

$$\frac{d_{l(i)}}{d_i} \to -D_{ll}\,, \quad \frac{d_{r(i)}}{d_i} \to -D_{lr} \qquad \textit{if } i \textit{ runs through } L;$$

$$\frac{d_{l(i)}}{d_i} \to -D_{rl}\,, \quad \frac{d_{r(i)}}{d_i} \to -D_{rr} \qquad \textit{if } i \textit{ runs through } R,$$

where

$$D_{ll} = \frac{qA_2 + (1-q)a_1}{qA_1 + (1-q)a_1}\,, \qquad D_{lr} = \frac{(1-q)(A_2 - A_1)\gamma_1}{(qA_1 + (1-q)a_1)\gamma_2}\,,$$

$$D_{rl} = \frac{q(a_2 - a_1)\gamma_2}{(qA_1 + (1-q)a_1)\gamma_1}\,, \qquad D_{rr} = \frac{qA_1 + (1-q)a_2}{qA_1 + (1-q)a_1}\,.$$

Proof.

We prove the first limit. The other limits are proved similarly. Multiplying the numerator and denominator in the definition of $d_{l(i)}$ for $i \in L$ by α_i and letting $i \to \infty$, so $\alpha_i \to 0$ by corollary 5.8, we obtain

$$\frac{d_{l(i)}}{d_i} \to -\frac{qA_2 + (1-q)a_1}{qA_1 + (1-q)a_1} = -D_{ll} . \qquad \square$$

The asymptotics of f_i is stated in the lemma below. Its proof is similar to that of lemma 5.11.

Lemma 5.12.

As $i \to \infty$, *then*

$$\frac{f_i}{d_i\alpha_i} \to -F_l \quad \text{if } i \text{ runs through } L;$$

$$\frac{f_i}{d_i\alpha_i} \to -F_r \quad \text{if } i \text{ runs through } R,$$

where

$$F_l = \frac{\gamma_1(A_2 - A_1)}{2\rho(qA_1 + (1-q)a_1)} , \quad F_r = \frac{\gamma_2(a_2 - a_1)}{2\rho(qA_1 + (1-q)a_1)} .$$

The lemmas 5.9-5.12 are the ingredients to prove theorem 5.6.

5.5. Proof of theorem 5.6

We can now prove that the series (5.16), (5.17) and (5.20) converge absolutely. First consider a fixed $m \geq 0$ and $n \geq 0$. Since α_i, β_i, c_i are positive for all $i \geq 0$, it follows that the series (5.17) with $-n$ replaced by n and the series (5.16) converge absolutely if and only if

$$\sum_{i=1}^{\infty} |d_i| (c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n < \infty .$$

Below it is shown that the terms in this infinite sum converge exponentially fast to zero. First,

$$\frac{|d_{l(i)}|(c_i\alpha_i^m + c_{l(i)}\alpha_{l(i)}^m)\beta_{l(i)}^n}{|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n} \text{ is abbreviated by } \begin{cases} R_{ll}(i, m, n) & \text{for } i \in L ; \\ R_{rl}(i, m, n) & \text{for } i \in R ; \end{cases}$$

$$\frac{|d_{r(i)}|(c_i\alpha_i^m + c_{r(i)}\alpha_{r(i)}^m)\beta_{r(i)}^n}{|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n} \text{ is abbreviated by } \begin{cases} R_{lr}(i, m, n) & \text{for } i \in L ; \\ R_{rr}(i, m, n) & \text{for } i \in R . \end{cases}$$

By the lemmas 5.9-5.12 we obtain that as $i \to \infty$ and i runs through L,

$$R_{ll}(i,\ m,\ n) \to R_{ll}(m,\ n) := D_{ll}C_l\left[\frac{A_1}{A_2}\right]^{m+n},$$

$$R_{lr}(i,\ m,\ n) \to R_{lr}(m,\ n) := D_{lr}C_l\,\frac{1+C_r\,(a_1/a_2)^m}{1+C_l\,(A_1/A_2)^m}\left[\frac{A_1}{A_2}\right]^m\left[\frac{A_1}{a_2}\right]^n,$$

and as $i \to \infty$ and i runs through R,

$$R_{rl}(i,\ m,\ n) \to R_{rl}(m,\ n) := D_{rl}C_r\,\frac{1+C_l\,(A_1/A_2)^m}{1+C_r\,(a_1/a_2)^m}\left[\frac{a_1}{a_2}\right]^m\left[\frac{a_1}{A_2}\right]^n,$$

$$R_{rr}(i,\ m,\ n) \to R_{rr}(m,\ n) := D_{rr}C_r\left[\frac{a_1}{a_2}\right]^{m+n}.$$

Hence, in the limit the terms behave geometrically. To formulate necessary and sufficient conditions for the convergence of the infinite sum we need the notion of a *positive geometrical binary tree.*

Definition 5.13.

The numbers $n_1, n_2, n_3, \ldots$ form a positive geometrical binary tree if:

(i) *The numbers n_i have a binary tree structure as depicted in figure 5.6;*

(ii) *The initial values n_1 and n_2 are positive;*

(iii) *The geometrical behaviour is determined by the nonnegative matrix* $\begin{bmatrix} R_{ll} & R_{rl} \\ R_{lr} & R_{rr} \end{bmatrix}$ *such that:*

$$n_{l(i)} = R_{ll}\,n_i\,,\quad n_{r(i)} = R_{lr}\,n_i \quad \text{if } n_i \text{ is a left descendant};$$

$$n_{l(i)} = R_{rl}\,n_i\,,\quad n_{r(i)} = R_{rr}\,n_i \quad \text{if } n_i \text{ is a right descendant.}$$

Notice that the tree of numbers $n_1, n_2, n_3, \ldots$ is of the same structure as the compensation tree depicted in figure 5.5. Let $\sigma(A)$ denote the spectral radius of the matrix A, then in particular,

$$\sigma\begin{bmatrix} R_{ll} & R_{rl} \\ R_{lr} & R_{rr} \end{bmatrix} = \frac{R_{ll}+R_{rr}+\sqrt{(R_{ll}-R_{rr})^2+4R_{lr}R_{rl}}}{2}. \tag{5.25}$$

The next lemma provides a necessary and sufficient condition for the convergence of $\sum_{i=1}^{\infty} n_i$.

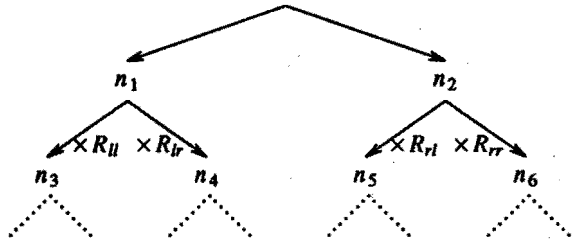


Figure 5.6.
The binary tree structure of the numbers n_i.

Lemma 5.14.

$$\sum_{i=1}^{\infty} n_i < \infty \iff \sigma\begin{bmatrix} R_{ll} & R_{rl} \\ R_{lr} & R_{rr} \end{bmatrix} < 1.$$

Proof.

Define for all $m \geq 0$,

$W_l(m)$ = the sum of all numbers n_i at depth m, which are a left descendant,

$W_r(m)$ = the sum of all numbers n_i at depth m, which are a right descendant.

Then for all $m \geq 0$,

$$\begin{bmatrix} W_l(m+1) \\ W_r(m+1) \end{bmatrix} = \begin{bmatrix} R_{ll} & R_{rl} \\ R_{lr} & R_{rr} \end{bmatrix} \begin{bmatrix} W_l(m) \\ W_r(m) \end{bmatrix} = \ldots = \begin{bmatrix} R_{ll} & R_{rl} \\ R_{lr} & R_{rr} \end{bmatrix}^m \begin{bmatrix} W_l(1) \\ W_r(1) \end{bmatrix}, \tag{5.26}$$

where $W_l(1) = n_1$ and $W_r(1) = n_2$. Hence

$$\sum_{i=1}^{\infty} n_i = \sum_{m=0}^{\infty} \left[W_l(m+1) + W_r(m+1) \right] = (1,\ 1) \sum_{m=0}^{\infty} \begin{bmatrix} R_{ll} & R_{rl} \\ R_{lr} & R_{rr} \end{bmatrix}^m \begin{bmatrix} W_l(1) \\ W_r(1) \end{bmatrix}.$$

If $\sigma\begin{bmatrix} R_{ll} & R_{rl} \\ R_{lr} & R_{rr} \end{bmatrix} < 1$, then $\begin{bmatrix} R_{ll} & R_{rl} \\ R_{lr} & R_{rr} \end{bmatrix}^m$ converges exponentially fast to zero, so $\sum_{i=1}^{\infty} n_i < \infty$.

If on the other hand $\sum_{i=1}^{\infty} n_i < \infty$, then, since $W_l(1)$ and $W_r(1)$ are positive and $\begin{bmatrix} R_{ll} & R_{rl} \\ R_{lr} & R_{rr} \end{bmatrix} \geq 0$,

$$\begin{bmatrix} R_{ll} & R_{rl} \\ R_{lr} & R_{rr} \end{bmatrix}^m \to 0 \quad \text{as} \quad m \to \infty,$$

which holds if and only if

$$\sigma\begin{bmatrix} R_{ll} & R_{rl} \\ R_{lr} & R_{rr} \end{bmatrix} < 1 .$$ □

Hence, the convergence of a positive geometrical binary tree is determined by the spectral radius of the matrix of rates. Since the compensation tree of terms $|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$ behaves *asymptotically* as a positive geometrical binary tree with rates

$$\begin{bmatrix} R_{ll} & R_{rl} \\ R_{lr} & R_{rr} \end{bmatrix} = \begin{bmatrix} R_{ll}(m, n) & R_{rl}(m, n) \\ R_{lr}(m, n) & R_{rr}(m, n) \end{bmatrix},$$

we expect that the convergence is also determined by the spectral radius of this matrix. First, let us define:

Definition 5.15.

For all $n \geq 0$, $\sigma(n)$ is defined by the following equation:

$$\sigma(n) = \frac{1}{2}\left[D_{ll}C_l(A_1/A_2)^n + D_{rr}C_r(a_1/a_2)^n + \sqrt{(D_{ll}C_l(A_1/A_2)^n - D_{rr}C_r(a_1/a_2)^n)^2 + 4D_{lr}C_lD_{rl}C_r(A_1/A_2)^n(a_1/a_2)^n}\right].$$

From this definition we conclude (cf. (5.25))

$$\sigma\begin{bmatrix} R_{ll}(m, n) & R_{rl}(m, n) \\ R_{lr}(m, n) & R_{rr}(m, n) \end{bmatrix} = \sigma(m + n) .$$

Since $0 < a_1 < 1 < a_2$ and $0 < A_1 < 1 < A_2$, it follows that $R_{ll}(0, n)$, $R_{lr}(0, n)$, $R_{rl}(0, n)$ and $R_{rr}(0, n) \downarrow 0$ as $n \to \infty$. Hence, since $0 \leq A \leq B$ implies $\sigma(A) \leq \sigma(B)$ (see e.g. [28]), we obtain that $\sigma(n) \downarrow 0$ as $n \to \infty$ So it is sensible to define (cf. definition 2.28):

Definition 5.16.

Let N be the smallest nonnegative integer such that $\sigma(N+1) < 1$.

Below we prove that if $\sigma(m + n) < 1$, or equivalently $m + n > N$, then

$$\sum_{i=1}^{\infty} |d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n < \infty$$

and otherwise, if $\sigma(m + n) > 1$, then this series diverges.

First assume that $\sigma(m+n) < 1$. In this case there exist rates R_{lr}, R_{ll}, R_{rl} and R_{rr} such that

$$R_{ll} > R_{ll}(s, t), \quad R_{rl} > R_{rl}(s, t),$$
$$R_{lr} > R_{lr}(s, t), \quad R_{rr} > R_{rr}(s, t)$$

and

$$\sigma\begin{bmatrix} R_{ll} & R_{rl} \\ R_{lr} & R_{rr} \end{bmatrix} < 1.$$

The indices i of the terms $|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$ at depth d in the compensation tree run from $i = 2^d-1$ to $i = 2^{d+1}-2$ (cf. figure 5.5). By taking d sufficiently large we have for $i \geq 2^d-1$,

$$R_{ll}(i, m, n) < R_{ll}, \quad R_{lr}(i, m, n) < R_{lr} \quad \text{if} \quad i \in L, \tag{5.27}$$
$$R_{rl}(i, m, n) < R_{rl}, \quad R_{rr}(i, m, n) < R_{rr} \quad \text{if} \quad i \in R.$$

Consider the positive geometrical binary tree of numbers n_i with rates $\begin{bmatrix} R_{ll} & R_{rl} \\ R_{lr} & R_{rr} \end{bmatrix}$ and initial values $n_1 = n_2 = K$, where K is taken sufficiently large such that for $i \leq 2^{d+1}-2$,

$$|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n \leq n_i.$$

In particular, this inequality holds for $i = 2^d-1, \ldots, 2^{d+1}-2$, so by (5.27) it follows that this inequality holds for *all* i. By lemma 5.14, the sum of n_i converges, so

$$\sum_{i=1}^{\infty} |d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n \leq \sum_{i=1}^{\infty} n_i < \infty.$$

Now assume that $\sigma(m+n) > 1$. Then the compensation tree of terms $|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$ can be bounded *below* by a *divergent* geometrical tree, so

$$\sum_{i=1}^{\infty} |d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n = \infty.$$

Finally, if $\sigma(m+n) = 1$, nothing can be said in general. This completes the proof of parts (i) and (ii) of theorem 5.6. We now prove part (iii), that is, for all $m \geq N$ the series

$$\sum_{i=0}^{\infty} c_i|f_i|\alpha_i^m$$

converges. This series can also be represented in a binary tree of terms $c_i|f_i|\alpha_i^m$. From the lemmas 5.9-5.12 it follows that as $i \to \infty$ and i runs through L,

$$\frac{c_{l(i)}|f_{l(i)}|\alpha_{l(i)}^m}{c_i|f_i|\alpha_i^m} \to D_{ll}C_l\left[\frac{A_1}{A_2}\right]^{m+1},$$

$$\frac{c_{r(i)}|f_{r(i)}|\alpha_{r(i)}^m}{c_i|f_i|\alpha_i^m} \to D_{lr}C_r\frac{F_r}{F_l}\left[\frac{a_1}{a_2}\right]^{m+1},$$

and as $i \to \infty$ and i runs through R,

$$\frac{c_{l(i)}|f_{l(i)}|\alpha_{l(i)}^m}{c_i|f_i|\alpha_i^m} \to D_{rl}C_l\frac{F_l}{F_r}\left[\frac{A_1}{A_2}\right]^{m+1},$$

$$\frac{c_{r(i)}|f_{r(i)}|\alpha_{r(i)}^m}{c_i|f_i|\alpha_i^m} \to D_{rr}C_r\left[\frac{a_1}{a_2}\right]^{m+1}.$$

Hence, the tree of terms $c_i|f_i|\alpha_i^m$ behaves asymptotically as a positive geometrical tree for which the spectral radius of the matrix of rates is given by $\sigma(m+1)$. If $\sigma(m+1) < 1$, or equivalently $m \geq N$, it then follows similar to the proof of parts (i)-(ii) of theorem 5.6 that

$$\sum_{i=0}^{\infty} c_i|f_i|\alpha_i^m < \infty.$$

This completes the proof of part (iii) of theorem 5.6. We finally prove part (iv) stating that

$$\sum_{\substack{m\geq 0\\ m+|n|>N}} |x_{m,n}|$$

converges. Inserting the series (5.16), (5.17) and (5.20) into this sum yields

$$\begin{aligned}\sum_{\substack{m\geq 0\\ m+|n|>N}} |x_{m,n}| &\leq \sum_{m=0}^{N-1}\sum_{n=N+1-m}^{\infty}\sum_{i=1}^{\infty}|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n \\ &+ \sum_{m=N}^{\infty}\sum_{n=1}^{\infty}\sum_{i=1}^{\infty}|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n + \sum_{m=N}^{\infty}\sum_{i=0}^{\infty} c_i|f_i|\alpha_i^m \\ &= \sum_{m=0}^{N-1}\sum_{i=1}^{\infty}|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\frac{\beta_i^{N+1-m}}{1-\beta_i} \\ &+ \sum_{i=1}^{\infty}|d_i|\left[\frac{c_{p(i)}\alpha_{p(i)}^N}{1-\alpha_{p(i)}} + \frac{c_i\alpha_i^N}{1-\alpha_i}\right]\frac{\beta_i}{1-\beta_i} + \sum_{i=0}^{\infty}\frac{c_i|f_i|\alpha_i^N}{1-\alpha_i} < \infty,\end{aligned}$$

since the spectral radius of the matrix of limiting rates for each of the infinite sums is equal to $\sigma(N+1) < 1$. □

Remark 5.17.

In general, the integer N is small. In case $\gamma_1 = \gamma_2$, it follows that $A_1 = a_1$ and $A_2 = a_2$, so $\sigma(n)$ simplifies to (cf. (3.13))

$$\sigma(n) = \frac{1-A_1}{A_2-1}\left[\frac{A_1}{A_2}\right]^{n-1}.$$

Hence, if $\gamma_1 = \gamma_2$, then $\sigma(1) < 1$ and thus $N = 0$. Only for highly unbalanced systems, that is, as

$\gamma_1 \to 0$ or $\gamma_1 \to 2$, the integer N is somewhat larger. In table 5.1 we list N for fixed $q = ½$ and increasing values of ρ and γ_1.

N		γ_1		
		0.2	0.5	0.8
	0.1	1	1	0
ρ	0.5	1	0	0
	0.9	0	0	0

Table 5.1.
Values of N for fixed q = 1/2 and increasing values of ρ *and* γ_1.

5.6. Main result

We now have all ingredients to prove our main theorem stated below. The proof of this theorem, however, is similar to that of theorem 2.33 and therefore it is omitted.

Theorem 5.18. (main result)

For all m, n with $m \geq 0$ *and* $m + |n| > N$ *and for* $m = N$ *and* $n = 0$,

$$p_{m,n} = C^{-1} x_{m,n} ,$$

where C is the normalizing constant.

In the next two sections we show that the expressions for $p_{m,n}$ lead to similar ones for the normalizing constant and the moments of the sojourn time.

5.7. Product form expression for the normalizing constant

To derive an expression for C in the form of product forms we use the equation stating that the number of arrivals per unit time balances the number of departures per unit time, i.e.,

$$2\rho = \rho_1\gamma_1 + \rho_2\gamma_2 ,$$

where $\rho_{1(2)}$ is the fraction of time server 1(2) is busy. Insertion of the identities

$$\rho_1 = 1 - \sum_{n=0}^{\infty} p_{0,n}, \quad \rho_2 = 1 - \sum_{n=0}^{\infty} p_{0,-n}.$$

in this balance equation yields

$$2(1-\rho) = \sum_{n=0}^{\infty} p_{0,n}\gamma_1 + \sum_{n=0}^{\infty} p_{0,-n}\gamma_2. \tag{5.28}$$

We now define for $m \geq 0$, $n \geq 0$ the *unnormalized quantities* $\bar{p}_{m,n}$ by

$$\bar{p}_{m,n} = Cp_{m,n}. \tag{5.29}$$

By substituting (5.29) in equation (5.32) and inserting for $n > N$ the series (5.16) for $\bar{p}_{0,n} = x_{0,n}$ and the series (5.17) for $\bar{p}_{0,-n} = x_{0,-n}$ we obtain

$$2(1-\rho)C = \sum_{n=0}^{N} \bar{p}_{0,n}\gamma_1 + \sum_{n=0}^{N} \bar{p}_{0,-n}\gamma_2$$

$$+ \sum_{n=N+1}^{\infty} \sum_{i \in L} d_i(c_{p(i)} + c_i)\beta_i^n\gamma_1 + \sum_{n=N+1}^{\infty} \sum_{i \in R} d_i(c_{p(i)} + c_i)\beta_i^n\gamma_2.$$

Interchanging of summations finally yields the following equation for C:

$$2(1-\rho)C = \sum_{n=0}^{N} \bar{p}_{0,n}\gamma_1 + \sum_{n=0}^{N} \bar{p}_{0,-n}\gamma_2$$

$$+ \sum_{i \in L} d_i(c_{p(i)} + c_i)\frac{\beta_i^{N+1}}{1-\beta_i}\gamma_1 + \sum_{i \in R} d_i(c_{p(i)} + c_i)\frac{\beta_i^{N+1}}{1-\beta_i}\gamma_2. \tag{5.30}$$

5.8. Product from expressions for the moments of the sojourn time

In this section we indicate how expressions in the form of series of products are found for the first and second moment of the sojourn time S. We do not work out all details. By conditioning on the number of jobs in the two queues on arrival of a job and using the property that Poisson arrivals see time averages (see e.g. Wolff [62]) we find

$$\mathbb{E}S = \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} (m+1)\left[p_{m,n}/\gamma_1 + p_{m,-n}/\gamma_2 \right] + \sum_{m=0}^{\infty} (m+1)p_{m,0}(q/\gamma_2 + (1-q)/\gamma_1),$$

$$\mathbb{E}S^2 = \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} (2(m+1) + (m+1)m)\left[p_{m,n}/\gamma_1^2 + p_{m,-n}/\gamma_2^2 \right]$$

$$+ \sum_{m=0}^{\infty} (2(m+1) + (m+1)m)p_{m,0}(q/\gamma_2^2 + (1-q)/\gamma_1^2).$$

So we need to evaluate series as, for instance,

$$\sum_{m=0}^{\infty}\sum_{n=1}^{\infty}(m+1)p_{m,n} = C^{-1}\sum_{m=0}^{\infty}\sum_{n=1}^{\infty}(m+1)\bar{p}_{m,n}\ .$$

By inserting for all $m+n>N$ the series (5.16) for $\bar{p}_{m,n}=x_{m,n}$ in the right-hand side of this equation we obtain

$$\begin{aligned}\sum_{m=0}^{\infty}\sum_{n=1}^{\infty}(m+1)\bar{p}_{m,n} &= \sum_{m=0}^{N-1}\sum_{n=1}^{N-m}(m+1)\bar{p}_{m,n}\\ &+\sum_{m=0}^{N-1}\sum_{n=N-m+1}^{\infty}(m+1)\sum_{i\in L}d_i(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n\\ &+\sum_{m=N}^{\infty}\sum_{n=1}^{\infty}(m+1)\sum_{i\in L}d_i(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n\ .\end{aligned}$$

Interchanging of summations and inserting the equality

$$\sum_{m=N}^{\infty}(m+1)\alpha^m = \frac{\alpha^N(1+N(1-\alpha))}{(1-\alpha)^2}\ ,$$

valid for $0<\alpha<1$, finally yields

$$\begin{aligned}\sum_{m=0}^{\infty}\sum_{n=1}^{\infty}(m+1)\bar{p}_{m,n} &= \sum_{m=0}^{N-1}\sum_{n=1}^{N-m}(m+1)\bar{p}_{m,n}\\ &+\sum_{m=0}^{N-1}(m+1)\sum_{i\in L}d_i(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\frac{\beta_i^{N-m+1}}{1-\beta_i}\\ &+\sum_{i\in L}d_i\left[\frac{c_{p(i)}\alpha_{p(i)}^N(1+N(1-\alpha_{p(i)}))}{(1-\alpha_{p(i)})^2}+\frac{c_i\alpha_i^N(1+N(1-\alpha_i))}{(1-\alpha_i)^2}\right]\frac{\beta_i}{1-\beta_i}\ .\end{aligned} \tag{5.31}$$

Similar expressions can be derived for the other series in the equations for $\mathbb{E}S$ and $\mathbb{E}S^2$.

In the next section we conclude the theoretical treatment by considering two extensions.

5.9. Two extensions

The compensation approach bears flexibility towards small modifications in the model. In this section we comment on the extension to a threshold-type shortest queue problem and on the extension to the shortest queue problem for two parallel multi-server queues.

The threshold-type shortest queue problem is characterized as follows. Consider a queueing system consisting of two parallel servers with service rates γ_1 and γ_2 respectively, where $\gamma_1>\gamma_2>0$ and $\gamma_1+\gamma_2=2$. Jobs arrive according to a Poisson stream with rate 2ρ where $0<\rho<1$. If an arriving job finds i jobs in queue 1 and j jobs in queue 2, then the job joins queue 1 if $i\le j+T$ and otherwise queue 2. The jobs require exponentially distributed service

times with unit mean, the service times are supposed to be independent.

In this model arriving jobs are always sent to the faster queue, unless this queue is much longer than the slower one. The slower queue functions as a dynamic overflow queue. The value T may be used as a parameter to balance the utilization of both servers. This queueing system can be represented by a continuous-time Markov process, whose state space consists of the pairs (i, j), $i, j = 0, 1, \ldots$ where i and j are the lengths of the two queues. However, the variables $m = \min(i, j+T)$ and $n = j+T-i$ are more suited to application of the compensation approach. The transition-rate diagram is depicted in figure 5.7 (cf. figure 5.2).

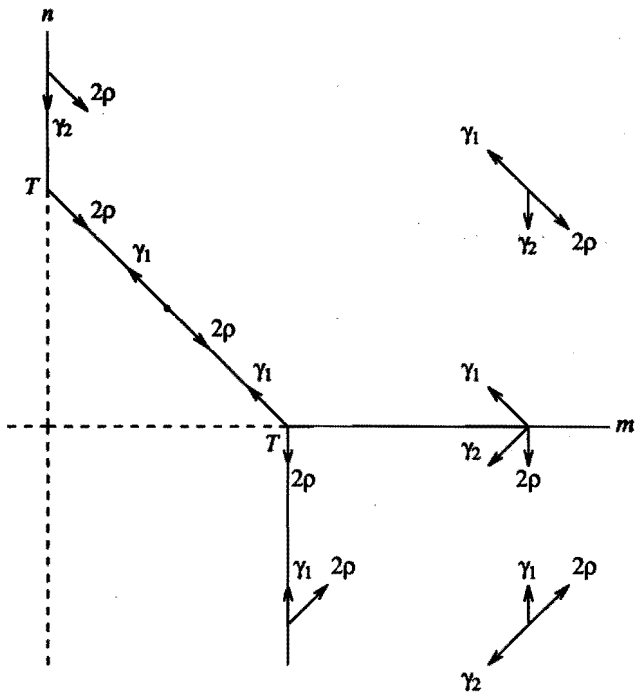


Figure 5.7.
Transition-rate diagram for the threshold-type shortest queue problem with threshold value T. It is supposed that $\gamma_1 > \gamma_2$.

The analysis of this problem is similar to that of the original shortest queue problem, except that compensation on the negative n-axis is now replaced by compensation on the negative part of the line $m = T$. The conditions for $m = T$ and $n < -1$ are (cf. (5.7)):

$$p_{T,n}(2\rho + \gamma_1) = p_{T,n-1}\gamma_1 + p_{T+1,n+1}\gamma_2 \cdot \qquad n < -1 \quad (5.32)$$

The following lemma summarizes the compensation on the negative part of the line $m = T$. The proof of this lemma is similar to that of lemma 5.2.

Lemma 5.19.

Let $w_{m,n} = y_+^m(\beta)\beta^{-n} + cy_-^m(\beta)\beta^{-n}$ *for* $m \geq 0, n < 0$.
Then $w_{m,n}$ *satisfies (5.5) and (5.32) if c is given by*

$$c = -\frac{y_+^T(\beta)}{y_-^T(\beta)} \frac{y_-(\beta) - \beta}{y_+(\beta) - \beta} \cdot$$

We now comment on the shortest queue problem with two parallel multi-server queues. This problem is characterized as follows. Consider a queueing system consisting of two parallel multi-server queues with M_1 and M_2 servers respectively. In the first queue the servers work with rate γ_1 / M_1 and in the second queue the servers with rate γ_2 / M_2, where $\gamma_1 > 0$, $\gamma_2 > 0$ and $\gamma_1 + \gamma_2 = 2$. Jobs arrive according to a Poisson stream with rate 2ρ where $0 < \rho < 1$. On arrival a job joins the shortest queue and, if queues have equal lengths, joins either queue with probability $1 - q$ and q respectively, where q is an arbitrary number between 0 and 1. The jobs require exponentially distributed service times with unit mean, the service times are supposed to be independent.

This system can be represented by a continuous-time Markov process, whose state space consists of the pairs (i, j), $i, j = 0, 1, \ldots$ where i and j are the lengths of the two queues. The transformation $m = \min(i, j)$ and $n = j - i$ leads to a state space more suited to application of the compensation approach. The transition structure at the vertical axis is more complicated than that of the original model with single servers. However, once the equivalent of lemma 5.2 for multi-servers is established, the analysis is further similar to that of the original model with single servers. For more details the reader is referred to [6].

5.10. The bounding geometrical trees

The compensation approach is constructive in nature. Therefore, a natural question is how these results are used for numerical purposes. This section and the subsequent ones are devoted to numerical aspects of the approach. It will be shown that the compensation tree can be computed efficiently with bounds on the error of each partial tree. In section 5.5 we showed that for each $m \geq 0$ and $n > 0$ the compensation tree of terms $|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$ behaves *asymptotically* as a positive geometrical binary tree with nonnegative rates

$$R(m, n) := \begin{bmatrix} R_{ll}(m, n) & R_{rl}(m, n) \\ R_{lr}(m, n) & R_{rr}(m, n) \end{bmatrix}.$$

Due to the exponential convergence, a few terms suffice to obtain an accurate approximation for $x_{m,n}$ and $x_{m,-n}$ where $x_{m,n}$ is the sum of the left descendants in the compensation tree and $x_{m,-n}$ is the sum of the right descendants. The question that arises is: how accurate is a partial compensation tree? In this section we derive an upper bound on the contribution of the subtrees below the leaves of each partial compensation tree. This upper bound is obtained by bounding these subtrees by geometrical trees.

In appendix B we define for $m \geq 0$, $n > 0$ and all nodes $i \geq 1$ in the compensation tree of terms $|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$, the nonnegative matrix

$$B(i, m, n) := \begin{bmatrix} B_{ll}(i, m, n) & B_{rl}(i, m, n) \\ B_{lr}(i, m, n) & B_{rr}(i, m, n) \end{bmatrix},$$

and we prove that this matrix $B(i, m, n)$ yields a *uniform bound on the rate of convergence* of the terms in the subtree below $|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$ (see figure 5.8), i.e., for all terms $|d_j|(c_{p(j)}\alpha_{p(j)}^m + c_j\alpha_j^m)\beta_j^n$ in the subtree below $|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$ it holds that

$$R_{ll}(j, m, n) \leq B_{ll}(i, m, n), \quad R_{lr}(j, m, n) \leq B_{lr}(i, m, n) \quad \text{if} \quad i \in L,$$

$$R_{rl}(j, m, n) \leq B_{rl}(i, m, n), \quad R_{rr}(j, m, n) \leq B_{rr}(i, m, n) \quad \text{if} \quad i \in R.$$

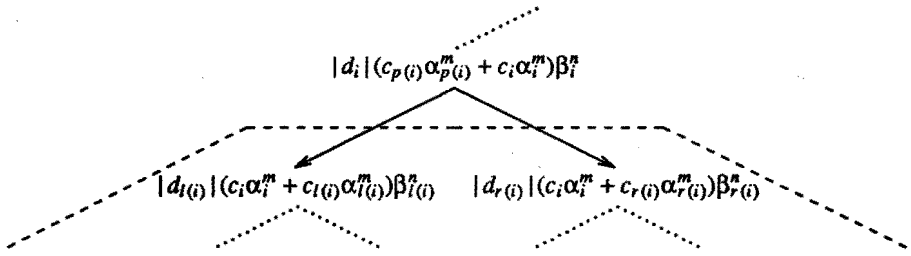


Figure 5.8.
The subtree below $|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$ is the part of the compensation tree below the dashed line.

So the subtree below $|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$ is bounded by the positive geometrical binary tree with the same initial values as this subtree and with rates $B(i, m, n)$ (see figure 5.9). The following theorem summarizes the bounding properties of $B(i, m, n)$. This theorem is proved in appendix B.

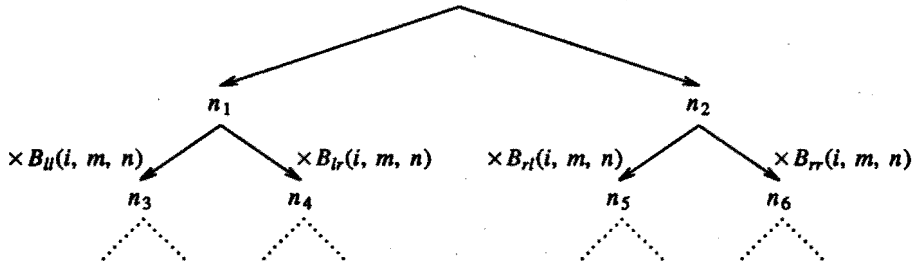


Figure 5.9.
The bounding geometrical tree for the subtree below $|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$, *where* $n_1 = |d_{l(i)}|(c_i\alpha_i^m + c_{l(i)}\alpha_{l(i)}^m)\beta_{l(i)}^n$ *and* $n_2 = |d_{r(i)}|(c_i\alpha_i^m + c_{r(i)}\alpha_{r(i)}^m)\beta_{r(i)}^n$.

Theorem 5.20.

For all $m \geq 0$, $n > 0$ *and all nodes* $i \geq 1$ *in the tree of terms* $|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$, *the subtree below node i is bounded by the positive geometrical binary tree with the same initial values as this subtree and with rates* $B(i, m, n)$.

Let for all $m \geq 0$, $n > 0$ and all $i \geq 1$,

$W_l(i, m, n)$ = weight of all left descendants in the subtree below $|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$;

$W_r(i, m, n)$ = weight of all right descendants in the subtree below $|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$.

By theorem 5.20, the weights $W_l(i, m, n)$ and $W_r(i, m, n)$ are bounded by the weight of all left, respectively all right descendants in the positive geometrical tree with the initial values $n_1 = |d_{l(i)}|(c_i\alpha_i^m + c_{l(i)}\alpha_{l(i)}^m)\beta_{l(i)}^n$ and $n_2 = |d_{r(i)}|(c_i\alpha_i^m + c_{r(i)}\alpha_{r(i)}^m)\beta_{r(i)}^n$ and with rates $B(i, m, n)$. By (5.26) these bounds are easily calculated, yielding

$$\begin{bmatrix} W_l(i, m, n) \\ W_r(i, m, n) \end{bmatrix} \leq \sum_{k=0}^{\infty} B(i, m, n)^k \begin{bmatrix} |d_{l(i)}|(c_i\alpha_i^m + c_{l(i)}\alpha_{l(i)}^m)\beta_{l(i)}^n \\ |d_{r(i)}|(c_i\alpha_i^m + c_{r(i)}\alpha_{r(i)}^m)\beta_{r(i)}^n \end{bmatrix}.$$

This bound is finite if and only if $\sigma(B(i, m, n)) < 1$ in which case it simplifies to:

Theorem 5.21.

For all $m \geq 0$, $n > 0$ *and all* $i \geq 1$,
if $\sigma(B(i, m, n)) < 1$, *then*

$$\begin{bmatrix} W_l(i, m, n) \\ W_r(i, m, n) \end{bmatrix} \leq \left[I - B(i, m, n) \right]^{-1} \begin{bmatrix} |d_{l(i)}| (c_i \alpha_i^m + c_{l(i)} \alpha_{l(i)}^m) \beta_{l(i)}^n \\ |d_{r(i)}| (c_i \alpha_i^m + c_{r(i)} \alpha_{r(i)}^m) \beta_{r(i)}^n \end{bmatrix},$$

where I *denotes the identity matrix.*

If $\sigma(B(i, m, n)) < 1$ holds for all nodes i in the compensation tree, then an upper bound on the error of each partial compensation tree is provided by the sum of the upper bounds on the weights of the subtrees below the leaves. To find out whether this condition for the spectral radii holds for all nodes, we need the following properties of $B(i, m, n)$ proved in appendix B.

Lemma 5.22.

(i) $B(i, m, n)$ *decreases monotonically along each path in the compensation tree for fixed* m *and* n;

(ii) $B(i, m, n) \to R(m, n)$ *as* $i \to \infty$;

(iii) $B(i, m, n)$ *decreases monotonically and exponentially fast as* $m \to \infty$ *for fixed* i *and* n;

(iv) $B(i, m, n)$ *decreases monotonically and exponentially fast as* $n \to \infty$ *for fixed* i *and* m.

Since $0 \leq A \leq D$ implies $\sigma(A) \leq \sigma(D)$ (see e.g. [28]), it follows by lemma 5.22(i) that if $\sigma(B(i, m, n)) < 1$ for $i = 1$ and $i = 2$, then this inequality holds for all nodes i. By lemma 5.22(iii) and 5.22(iv) the spectral radii $\sigma(B(1, m, n))$ and $\sigma(B(2, m, n))$ decrease exponentially fast as $m + n \to \infty$, so $\sigma(B(1, m, n))$ and $\sigma(B(2, m, n))$ are less than one for $m + n$ sufficiently large. Now we have all ingredients for the computation of the compensation tree.

5.11. Basic scheme for the computation of the compensation tree

Below we formulate a basic scheme for the computation of $x_{m,n}$ and $x_{m,-n}$ with a relative error of ε for each $m \geq 0$ and $n > 0$. We assume that bounds can be computed right at the beginning of the compensation tree, i.e., both $\sigma(B(1, m, n))$ and $\sigma(B(2, m, n))$ are less than one. This assumption will be relaxed later on.

Step 0. (Initialization)

Compute

$$\tilde{x}_{m,n} = d_1(c_0\alpha_0^m + c_1\alpha_1^m)\beta_1^n ,$$

$$\tilde{x}_{m,-n} = d_2(c_0\alpha_0^m + c_2\alpha_2^m)\beta_2^n ,$$

and set $k = 1$.

Step 1. (Compute all terms at depth k+1 in the compensation tree and compute the bounds)

Compute for each term $d_i(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$ at depth k in the compensation tree its immediate successors $d_{l(i)}(c_i\alpha_i^m + c_{l(i)}\alpha_{l(i)}^m)\beta_{l(i)}^n$ and $d_{r(i)}(c_i\alpha_i^m + c_{r(i)}\alpha_{r(i)}^m)\beta_{r(i)}^n$ and add these terms to $\tilde{x}_{m,n}$ and $\tilde{x}_{m,-n}$ respectively.

Compute the upper bounds $U_l(i, m, n)$ and $U_r(i, m, n)$ on the weight of all left, respectively right descendants in the subtree below $d_i(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$, *excluding* its initial values $d_{l(i)}(c_i\alpha_i^m + c_{l(i)}\alpha_{l(i)}^m)\beta_{l(i)}^n$ and $d_{r(i)}(c_i\alpha_i^m + c_{r(i)}\alpha_{r(i)}^m)\beta_{r(i)}^n$ (since we already added these terms to $\tilde{x}_{m,n}$ and $\tilde{x}_{m,-n}$ respectively). That is, by theorem 5.21,

$$\begin{bmatrix} U_l(i, m, n) \\ U_r(i, m, n) \end{bmatrix} = B(i, m, n)\left[I - B(i, m, n)\right]^{-1} \begin{bmatrix} |d_{l(i)}|(c_i\alpha_i^m + c_{l(i)}\alpha_{l(i)}^m)\beta_{l(i)}^n \\ |d_{r(i)}|(c_i\alpha_i^m + c_{r(i)}\alpha_{r(i)}^m)\beta_{r(i)}^n \end{bmatrix}.$$

Step 2. (Convergence)

Let I_k be the set of indices i of the terms $d_i(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n$ at depth k in the compensation tree, i.e., $I_k = \{2^k-1, 2^k, ..., 2^{k+1}-2\}$. If the following two inequalities are satisfied:

$$\sum_{i\in I_k} U_l(i, m, n) \le \varepsilon\left\{\tilde{x}_{m,n} - \sum_{i\in I_k} U_l(i, m, n)\right\},$$

$$\sum_{i\in I_k} U_r(i, m, n) \le \varepsilon\left\{\tilde{x}_{m,-n} - \sum_{i\in I_k} U_r(i, m, n)\right\},$$

then the relative accuracy of ε is attained, so stop and approximate $x_{m,n}$ by $\tilde{x}_{m,n}$ and $x_{m,-n}$ by $\tilde{x}_{m,-n}$; otherwise repeat step 1 with $k = k + 1$.

This scheme computes the compensation tree with error bounds. These bounds are based on theorem 5.20. The analogue of theorem 5.20 also holds for the tree of product forms in expression (5.30) for the normalizing constant C (which is due to the fact that the extra factor $1/(1-\beta)$ is increasing for $0 < \beta < 1$).

Lemma 5.23.

For all $m \geq 0$, $n > 0$ and all nodes $i \geq 1$ in the tree of terms $|d_i|(c_{p(i)}\alpha_{p(i)}^m + c_i\alpha_i^m)\beta_i^n/(1-\beta_i)$, the subtree below node i is bounded by the positive geometrical binary tree with the same initial values as this subtree and with rates $B(i, m, n)$.

Proof.

By corollary 5.8, $\beta_{l(j)} < \beta_j < 1$ for all $j \geq 1$. Hence, by theorem 5.20, we obtain for all left descendants j in the subtree below node i,

$$|d_{l(j)}|(c_j\alpha_j^m + c_{l(j)}\alpha_{l(j)}^m)\frac{\beta_{l(j)}^n}{1-\beta_{l(j)}} \leq |d_{l(j)}|(c_j\alpha_j^m + c_{l(j)}\alpha_{l(j)}^m)\beta_{l(j)}^n\frac{1}{(1-\beta_j)}$$

$$\leq B_{ll}(i, m, n)\,|d_j|(c_{p(j)}\alpha_{p(j)}^m + c_j\alpha_j^m)\frac{\beta_j^n}{(1-\beta_j)}.$$

The other inequalities are obtained similarly. □

Similar results hold for the trees of product forms in the expressions for the moments of the sojourn time (cf. (5.31)). Hence, to compute the trees in the expression (5.30) for C and in the expressions for the moments of the sojourn time, we can use the same scheme as for the computation of the compensation tree. We conclude this section with some remarks.

Remark 5.24.

In the convergence test we implicitly use that $\{x_{m,n}\}$ is a positive solution. This follows by observing that $\{x_{m,n}\}$ has constant sign by virtue of theorem 5.18 and that for fixed n the first term in $x_{m,n}$ is dominating as $m \to \infty$.

Remark 5.25.

The quality of the computation scheme depends on the rate at which the upper bound

$$\sum_{i \in I_k}\begin{bmatrix} U_l(i, m, n) \\ U_r(i, m, n)\end{bmatrix}$$

converges to zero as $k \to \infty$. It follows from lemma 5.22(ii) that the rate of convergence of this upper bound is determined by the rate at which the weight

$$\sum_{i \in I_k}\begin{bmatrix} |d_{l(i)}|(c_i\alpha_i^m + c_{l(i)}\alpha_{l(i)}^m)\beta_{l(i)}^n \\ |d_{r(i)}|(c_i\alpha_i^m + c_{r(i)}\alpha_{r(i)}^m)\beta_{r(i)}^n\end{bmatrix}$$

converges to zero as $k \to \infty$. Since the compensation tree behaves asymptotically as a

geometrical tree with rates $R(m, n)$, it is easily shown that the rate of convergence of this weight is determined by $\sigma(R(m, n)) = \sigma(m+n)$. Hence, since $\sigma(m+n)$ decreases exponentially fast as $m+n \to \infty$, the convergence of the upper bound is faster for states further away from the origin. This aspect is exploited in section 5.12.

Remark 5.26.

In each cycle the immediate successors of *all* leaves of the current partial tree are computed. So the number of computed terms doubles in each cycle. Luckily, few cycles usually suffice to obtain an accurate approximation. In section 5.13 we propose an alternative computation strategy in which a better use is made of the relative importance of the branches of the tree.

Remark 5.27.

The basic scheme assumes that $\sigma(B(i, m, n)) < 1$ for $i = 1, 2$ to be able to compute bounds at the beginning of the compensation tree. This can be relaxed to computing bounds as soon as $\sigma(B(i, m, n)) < 1$. If $m+n > N$, then $\sigma(B(i, m, n)) < 1$ for i sufficiently large. This follows from lemma 5.22(ii) and the fact that $\sigma(R(m, n)) = \sigma(m+n) < 1$ for $m+n > N$.

Remark 5.28.

The equilibrium equations for $n = 0$ may be used to calculate $x_{m,0}$ for $m \geq 0$. These equations state that (see (5.9) and (5.10)):

$$x_{m,0}2(\rho+1) = x_{m-1,1}2\rho + x_{m,1}\gamma_2 + x_{m-1,-1}2\rho + x_{m,-1}\gamma_1 , \qquad m > 0$$

$$x_{0,0}2\rho = x_{0,1}\gamma_2 + x_{0,-1}\gamma_1 .$$

All quantities at the right hand side can be computed by the basic scheme.

5.12. Numerical solution of the equilibrium equations

If $N > 0$, then the equilibrium equations for $m + |n| \leq N$ have to be solved numerically from the solution on the complement. These equations can be solved efficiently and numerically stable by an approach similar to the ones in the sections 3.9 and 4.4. This approach is based on the special property that the only flow from level l, defined by

$$\text{level } l = \{(m, n) \mid m \geq 0,\ m + |n| = l\} , \quad l \geq 0 ,$$

to level $l+1$ is via state $(l, 0)$. By this property, the problem of simultaneously solving the equilibrium equations at the levels $l \leq N$, given the (unnormalized) solution at level $N+1$, can be reduced to that of recursively solving the equations at the levels $N \to N-1 \to \ldots \to 1 \to 0$.

We first formulate the equilibrium equations at level $l > 0$.

$$p_{0,l}(2\rho + \gamma_2) = p_{0,l+1}\gamma_2 + p_{1,l-1}\gamma_1 , \tag{5.33}$$

$$p_{k,l-k}2(\rho + 1) = p_{k-1,l-k+1}2\rho + p_{k,l-k+1}\gamma_2 p_{k+1,l-k-1}\gamma_1 , \qquad 0 < k < l \tag{5.34}$$

$$p_{l-1,1}2(\rho + 1) = p_{l-2,2}2\rho + p_{l-1,2}\gamma_2 + p_{l,0}\gamma_1 + p_{l-1,0}2q\rho , \tag{5.35}$$

$$p_{0,-l}(2\rho + \gamma_1) = p_{0,-l-1}\gamma_1 + p_{1,-l+1}\gamma_2 , \tag{5.36}$$

$$p_{k,-l+k}2(\rho + 1) = p_{k-1,-l+k-1}2\rho + p_{k,-l+k-1}\gamma_1 p_{k+1,-l+k+1}\gamma_2 , \qquad 0 < k < l \tag{5.37}$$

$$p_{l-1,-1}2(\rho + 1) = p_{l-2,-2}2\rho + p_{l-1,-2}\gamma_1 + p_{l,0}\gamma_2 + p_{l-1,0}2(1-q)\rho . \tag{5.38}$$

The equilibrium equation in state $(l, 0)$ is replaced by the following two equations. Applying the balance principle "rate out of A = rate into A" to

$$A = \{(m, n) | m \geq 0, m + |n| \leq l\} \cup \setminus \{(l, 0)\} ,$$

leads for all $l > 0$ to

$$p_{l-1,1}2\rho + p_{l-1,-1}2\rho = p_{l,0}2 + \sum_{k=0}^{l-1} p_{k,l-k+1}\gamma_2 + \sum_{k=0}^{l-1} p_{k,-l+k-1}\gamma_1 , \tag{5.39}$$

and applying the balance principle to

$$A = \{(m, n) | m \geq 0, m + |n| \leq l\} ,$$

yields for all $l \geq 0$,

$$p_{l,0}2\rho = \sum_{k=0}^{l} p_{k,l-k+1}\gamma_2 + \sum_{k=0}^{l} p_{k,-l+k-1}\gamma_1 . \tag{5.40}$$

By eliminating $p_{l-1,0}$ in the equations (5.35) and (5.38) we obtain a set of linear equations for the probabilities at level l, given the probabilities at level $l+1$. These equations form a second order recursion relation for the probabilities at level l. Below we show that these equations can be reduced to a first order recursion relation.

Definition 5.29.

The sequence $x_0, x_1, x_2, \ldots$ is the solution of

$$x_{i+1} = x_i 2(\rho + 1) - x_{i-1}2\rho\gamma_1 , \quad i \geq 1 ,$$

with initial values $x_0 = 1$ and $x_1 = 2\rho + \gamma_2$.
The sequence $y_0, y_1, y_2, \ldots$ is defined similarly with γ_1 and γ_2 interchanged.

Theorem 5.30.

For all $l > 0$,

$$p_{k,l-k}x_{k+1} = p_{k+1,l-k-1}x_k\gamma_1 + \sum_{i=0}^{k} p_{i,l-i+1}x_i(2\rho)^{k-i}\gamma_2 , \tag{5.41}$$

$$p_{k,-l+k}y_{k+1} = p_{k+1,-l+k+1}y_k\gamma_2 + \sum_{i=0}^{k} p_{i,-l+i-1}y_i(2\rho)^{k-i}\gamma_1 \quad \textit{for} \quad k = 0, 1, ..., l-2 , \tag{5.42}$$

where the initial values $p_{l-1,1}$ and $p_{l-1,-1}$ follow from the equations

$$\begin{aligned} p_{l-1,1}2\rho(x_{l-1}y_l + x_l y_{l-1}(1-q)) &= p_{l,0}2x_{l-1}(y_{l-1}\gamma_1(1-q) + (y_l - y_{l-1}\gamma_2\rho)q) \\ &\quad + \sum_{i=0}^{l-1} p_{i,l-k+1}(x_i y_{l-1}(2\rho)^{l-i}(1-q)\gamma_2 + x_{l-1}y_l q\gamma_2) \\ &\quad + \sum_{i=0}^{l-1} p_{i,-l+k-1}x_{l-1}(y_l - y_i(2\rho)^{l-i})q\gamma_1 , \end{aligned} \tag{5.43}$$

$$\begin{aligned} p_{l-1,-1}2\rho(y_{l-1}x_l + y_l x_{l-1}q) &= p_{l,0}2y_{l-1}(x_{l-1}\gamma_2 q + (x_l - x_{l-1}\gamma_1\rho)(1-q)) \\ &\quad + \sum_{i=0}^{l-1} p_{i,-l+k-1}(y_i x_{l-1}(2\rho)^{l-i}q\gamma_1 + y_{l-1}x_l(1-q)\gamma_1) \\ &\quad + \sum_{i=0}^{l-1} p_{i,l-k+1}y_{l-1}(x_l - x_i(2\rho)^{l-i})(1-q)\gamma_2 . \end{aligned} \tag{5.44}$$

Proof.

We prove the recursion relation (5.41) by induction. For $k = 0$ the equations (5.41) and (5.33) are identical. Assume that (5.41) holds for $k = j$. Multiplying (5.41) for $k = j$ by 2ρ and (5.34) for $k = j+1$ by x_{j+1} and adding the two equations yields (5.41) for $k = j+1$. This proves (5.41) for $k = 0, 1, ..., l-2$ by induction. The recursion relation (5.42) is proved similarly. It remains to prove (5.43) and (4.44). Multiplying (5.41) for $k = l-2$ by 2ρ and (5.35) by x_{l-1} and adding the two equations yields

$$p_{l-1,1}x_l = p_{l,0}x_{l-1}\gamma_1 + \sum_{i=0}^{l-1} p_{i,l-i+1}x_i(2\rho)^{l-1-i}\gamma_2 + p_{l-1,0}x_{l-1}2q\rho ,$$

and similarly we obtain

$$p_{l-1,-1}y_l = p_{l,0}y_{l-1}\gamma_2 + \sum_{i=0}^{l-1} p_{i,-l+i-1}y_i(2\rho)^{l-1-i}\gamma_1 + p_{l-1,0}y_{l-1}2(1-q)\rho .$$

Eliminating $p_{l-1,0}$ in these two equations leads to

$$p_{l-1,1}x_l y_{l-1}(1-q) - p_{l-1,-1}x_{l-1}y_l q = p_{l,0}x_{l-1}y_{l-1}(\gamma_1(1-q) - \gamma_2 q)$$

$$+\sum_{i=0}^{l-1} p_{i,l-i+1}x_i y_{l-1}(2\rho)^{l-1-i}\gamma_2(1-q)$$

$$+\sum_{i=0}^{l-1} p_{i,-l+i-1}x_{l-1} y_i(2\rho)^{l-1-i}\gamma_1 q\ .$$

Together with equation (5.39) we now have two equations for $p_{l-1,1}$ and $p_{l-1,-1}$. The solution is given by (5.43) and (5.44). □

By first calculating the series for $\bar{p}_{0,N+1}, \bar{p}_{1,N}, \ldots, \bar{p}_{N,1}$ and for $\bar{p}_{0,-N-1}, \bar{p}_{1,-N}, \ldots, \bar{p}_{-N,1}$, the equations at level N are solved efficiently by use of the recursion relations in theorem 5.30. First $\bar{p}_{N,0}$ follows from (5.40) and $\bar{p}_{N-1,1}$ and $\bar{p}_{N-1,-1}$ follow from (5.43) and (5.44). Then $\bar{p}_{N-2,2} \rightarrow \bar{p}_{N-3,3} \rightarrow \ldots \rightarrow \bar{p}_{0,N}$ are subsequently computed from (5.41) and $\bar{p}_{N-2,-2} \rightarrow \bar{p}_{N-3,-3} \rightarrow \ldots \rightarrow \bar{p}_{0,-N}$ from (5.42). Once the solution at level N is computed, we repeat this scheme to subsequently compute the solution at level $N-1 \rightarrow N-2 \rightarrow \ldots \rightarrow 1 \rightarrow 0$. The following result is required to establish that the recursion relations in theorem 5.30 are *numerically stable*.

Lemma 5.31.

For all $i \geq 0$ *and* $j \geq 0$

$$x_{i+j} \geq x_i(2\rho)^j \geq 0\ , \qquad y_{i+j} \geq y_i(2\rho)^j \geq 0\ . \tag{5.45}$$

Proof.

We first prove the lemma for fixed $j = 1$ and $i \geq 0$ by induction. For $i = 0$ and $j = 1$, inequality (5.45) trivially holds. Assume that (5.45) holds for $i = k-1$ and $j = 1$, then

$$x_{k+1} = x_k 2(\rho+1) - x_{k-1}2\rho\gamma_1 = x_k 2\rho + x_k 2 - x_{k-1}2\rho\gamma_1 \geq x_k 2\rho + x_{k-1}2\rho(2-\gamma_1) \geq x_k 2\rho\ ,$$

which proves (5.45) for $i = k$ and $j = 1$. By induction we can now conclude that (5.45) holds for all $i \geq 0$ and $j = 1$. Now consider an arbitrary $i \geq 0$ and $j \geq 0$. Then,

$$x_{i+j} \geq x_{i+j-1}2\rho \geq \ldots \geq x_i(2\rho)^j\ ,$$

which completes the proof of this lemma for x. This lemma is proved similarly for y. □

From lemma 5.31 it follows that all coefficients in the recursion relations in theorem 5.30 are nonnegative. So the calculations involve only the addition and multiplication of nonnegative numbers and thus can cause no loss of significant digits. Hence, if the series for all $\bar{p}_{m,n}$ at level $N+1$ are computed with a relative accuracy of ε say, then repeated application of the recursion relations in theorem 5.30 yields all $\bar{p}_{m,n}$ at the lower levels with the same accuracy.

Remark 5.32.

By lemma 5.31, the numbers x_i and y_i increase exponentially fast. Therefore, to avoid possible overflow problems, it is numerically sensible to scale the recursion relations in theorem 5.30. For example, relation (5.41) is scaled by dividing both sides by x_{k+1}. The resulting recursion relation requires the calculation of the ratios $x_i(2\rho)^{k+1-i}/x_{k+1}$ for $i = 0, 1, ..., k$. For these ratios an explicit formula is easily derived. Moreover, from lemma 5.31 it follows that these ratios are all bounded by one.

5.13. Numerical results

This section is devoted to some numerical aspects and results. The number of cycles of the algorithm in section 5.11, and thereby the size of the partial compensation tree required to approximate $x_{m,n}$ and $x_{m,-n}$ sufficiently close, depends on the convergence of the upper bounds. From remark 5.25 it follows that the rate of convergence of the upper bounds is determined by $\sigma(m+n)$ which decreases exponentially fast as $m+n \to \infty$. So convergence is faster for states further away from the origin. This is illustrated in table 5.2. We list values of $x_{0,1}$, $x_{1,1}$ and $x_{2,1}$ with an accuracy of 0.1% together with $\sigma(1)$, $\sigma(2)$ and $\sigma(3)$ and the depth D of the partial trees needed for $\gamma_1 = 0.8$, $q = 0.7$ and increasing values of ρ. To approximate $x_{0,1}$ sufficiently close for $\rho = 0.1$ and $\rho = 0.3$, partial compensation trees are needed with depth ≥ 10. Therefore the computation has been aborted in these two cases.

ρ	$x_{0,1}$	D	$\sigma(1)$	$x_{1,1}$	D	$\sigma(2)$	$x_{2,1}$	D	$\sigma(3)$
0.1		≥10	0.763	0.000686	3	0.035	0.000007	2	0.002
0.3		≥10	0.529	0.017916	3	0.061	0.001676	2	0.008
0.5	0.262860	9	0.369	0.081151	3	0.058	0.021328	2	0.010
0.7	0.363568	7	0.262	0.219552	3	0.047	0.113707	2	0.009
0.9	0.465454	6	0.190	0.462543	3	0.035	0.396861	2	0.007

Table 5.2.
Values of $x_{0,1}$, $x_{1,1}$ and $x_{2,1}$ with an accuracy of 0.1% together with $\sigma(1)$, $\sigma(2)$, $\sigma(3)$ and the depth D of the partial trees needed for $\gamma_1 = 0.8$, $q = 0.7$ and increasing values of ρ.

Table 5.2 shows that $\sigma(m+n)$, and thereby the size of the partial tree needed to approximate $x_{m,n}$ sufficiently close, decreases fast for states further away from the origin. Hence, it is

numerically sensible to use the compensation tree to calculate $\bar{p}_{m,n} = x_{m,n}$ for $m + |n| > M$ where $M > N$ and to use the recursion relations in theorem 5.30 to calculate $\bar{p}_{m,n}$ for $m + |n| \leq M$. In fact, M must be such that $\sigma(M+1)$ is sufficiently small. Of course, then some extra effort is needed to solve the equations for $m + |n| \leq M$, but this effort is easily compensated by the advantages of efficiently computing the compensation tree in states further away from the origin.

In table 5.3 we list values of $p_{0,0}$, $p_{0,-1}$ and $p_{0,1}$ with a relative accuracy of 0.1% for fixed $q = 0.7$ and increasing values of ρ and γ_1. To obtain these probabilities we first used the compensation tree to calculate $\bar{p}_{m,n}$ with an accuracy of 0.1% for all $m + |n| = M+1$, then we calculated $\bar{p}_{m,n}$ for $m + |n| \leq M$ by use of the recursion relations in theorem 5.30 and finally we used (5.30) with N replaced by M to calculate C. The number D denotes the maximal depth of the partial trees needed to approximate $\bar{p}_{m,n}$ with an accuracy of 0.1% for $m + |n| = M+1$. All partial trees are computed up to the same depth, so some partial trees might be more accurate than strictly necessary. The examples in table 5.3 show that the equilibrium probabilities are computed efficiently.

ρ	γ_1	$p_{0,0}$	$p_{0,-1}$	$p_{0,1}$	M	$\sigma(M+1)$	D	N
0.2	0.2	0.458891	0.326071	0.065746	2	0.077	3	1
	0.6	0.650480	0.153709	0.119976	1	0.082	3	0
	0.9	0.667731	0.104980	0.156919	1	0.051	3	0
0.6	0.2	0.053825	0.131667	0.021253	3	0.101	5	1
	0.6	0.201925	0.162969	0.103235	1	0.090	4	0
	0.9	0.232253	0.123756	0.152112	1	0.048	3	0
0.9	0.2	0.001841	0.007135	0.001048	3	0.069	5	0
	0.6	0.030726	0.039101	0.022748	1	0.057	4	0
	0.9	0.042493	0.035494	0.040494	1	0.032	3	0
0.95	0.2	0.000539	0.002222	0.000322	3	0.060	5	0
	0.6	0.013810	0.018673	0.010739	1	0.052	4	0
	0.9	0.020045	0.017774	0.020081	1	0.030	3	0

Table 5.3.
Values of $p_{0,0}$, $p_{0,-1}$, $p_{0,1}$ with an accuracy of 0.1% together with $\sigma(M+1)$ and the depth D of the partial trees needed for $q = 0.7$ and increasing values of ρ and γ_1.

We conclude this section by illustrating the effect of the unbalance in the service rates on $\mathbb{E}S$ and the coefficient of variation $cv(S)$. Table 5.4 lists values of $\mathbb{E}S$ and $cv(S)$ with an accuracy of 0.1%. For comparison we also computed values of $\mathbb{E}S_c$ and $cv(S_c)$ for a *common-queue*, but further identical, system.

ρ	γ_1	$\mathbb{E}S$	$cv(S)$	$\mathbb{E}S_c$	$cv(S_c)$
0.1	0.6	1.178	1.158	1.162	1.149
	0.8	1.054	1.040	1.044	1.034
	1.0	1.018	1.000	1.010	0.995
0.3	0.6	1.294	1.168	1.203	1.102
	0.8	1.178	1.037	1.123	1.000
	1.0	1.144	0.996	1.099	0.968
0.6	0.6	1.876	1.156	1.628	0.984
	0.8	1.726	1.018	1.578	0.938
	1.0	1.682	0.975	1.563	0.925
0.9	0.6	5.817	1.132	5.308	0.958
	0.8	5.552	1.007	5.274	0.958
	1.0	5.475	0.970	5.263	0.959

Table 5.4.
Values of $\mathbb{E}S$ and $cv(S)$ with an accuracy of 0.1% for the parallel-queue system, together with values of $\mathbb{E}S_c$ and $cv(S_c)$ for the "corresponding" common-queue system, for $q = 0.5$ and increasing values of ρ and γ_1.

Table 5.4 shows that the performance of the parallel-queue system is close to that of the common-queue system and that the performance of both systems is fairly insensitive to the unbalance in service rates, except at light traffic, since then the service time forms the main part of the sojourn time.

5.14. Alternative strategy to compute the compensation tree

The basic scheme in section 5.11 computes in each cycle the immediate successors of *all leaves of the current partial tree*. For highly unbalanced trees, however, this strategy is inefficient. Trees are highly unbalanced for systems where one server is working much faster than the other one. For example, to approximate $x_{3,1}$ and $x_{3,-1}$ with an accuracy of 0.1% for $\rho = 0.9$, $\gamma_1 = 0.2$ and $q = 0.7$ the basic scheme computes the compensation tree up to depth 5. In

figure 5.10 we depict the *relevant* part of this partial tree, i.e., the sum of the absolute values of the other terms is roughly 10^{-8}. The pair in each node i stands for $(i,\ d_i(c_{p(i)}\alpha_{p(i)}^3 + c_i\alpha_i^3)\beta_i)$.

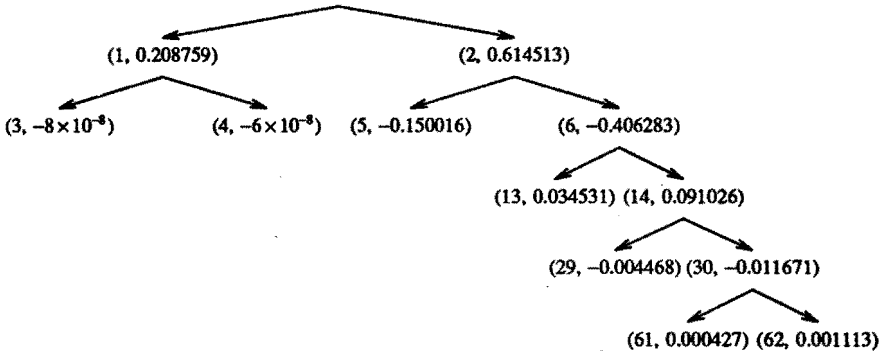


Figure 5.10.
A part of the compensation tree for $x_{3,1}$ *and* $x_{3,-1}$ *for* $\rho = 0.9,\ \gamma_1 = 0.2,\ q = 0.7$

Figure 5.10 illustrates that the compensation tree is highly unbalanced: the weight is concentrated at the very right side of the tree. It is not sensible to compute all 62 terms if only 12 terms are relevant. Therefore we propose an alternative strategy, which makes a better use of the relative importance of the branches of the compensation tree by computing in each cycle the immediate successors of the leaf the subtree of which has *maximum* weight, or more precisely, *maximum upper bound* for its weight. The quantity $U_l(i,\ m,\ n) + U_r(i,\ m,\ n)$ provides an upper bound on the weight of the subtrees below node $l(i)$ and $r(i)$ together, but the new strategy requires upper bounds on the weight of the subtrees below $l(i)$ and $r(i)$ *separately*. Therefore, we decompose the upper bound $U_l(i,\ m,\ n)$ in the contribution $U_{ll}(i,\ m,\ n)$ of the subtree below the left descendant $l(i)$, and the contribution $U_{rl}(i,\ m,\ n)$ of the subtree below the right descendant $r(i)$. Similarly, the upper bound $U_r(i,\ m,\ n)$ is decomposed in the contribution $U_{lr}(i,\ m,\ n)$ and $U_{rr}(i,\ m,\ n)$, yielding

$$\begin{bmatrix} U_{ll}(i,\ m,\ n) \\ U_{lr}(i,\ m,\ n) \end{bmatrix} = B(i,\ m,\ n)\left[I - B(i,\ m,\ n)\right]^{-1} \begin{bmatrix} |d_{l(i)}|(c_i\alpha_i^m + c_{l(i)}\alpha_{l(i)}^m)\beta_{l(i)}^n \\ 0 \end{bmatrix},$$

$$\begin{bmatrix} U_{rl}(i,\ m,\ n) \\ U_{rr}(i,\ m,\ n) \end{bmatrix} = B(i,\ m,\ n)\left[I - B(i,\ m,\ n)\right]^{-1} \begin{bmatrix} 0 \\ |d_{r(i)}|(c_i\alpha_i^m + c_{r(i)}\alpha_{r(i)}^m)\beta_{r(i)}^n \end{bmatrix}.$$

Then the weight of the subtree below node $l(i)$ is bounded by $U_{ll}(i,\ m,\ n) + U_{lr}(i,\ m,\ n)$ and the weight of the subtree below node $r(i)$ is bounded by $U_{rl}(i,\ m,\ n) + U_{rr}(i,\ m,\ n)$. Based on

these decomposed bounds, we are able to determine in each cycle the leaf the subtree of which has maximum upper bound for its weight. To compute approximations for $x_{3,1}$ and $x_{3,-1}$ with an accuracy of 0.1% for $\rho = 0.9$, $\gamma_1 = 0.2$ and $q = 0.7$ this new strategy exactly calculates the partial tree depicted in figure 5.10. Hence, to obtain the desired approximations, this new strategy computes 12 terms, whereas the strategy in section 5.11 computes 62 terms.

5.15. Conclusion

In this chapter we analysed the shortest queue problem with nonidentical servers. This problem can be modelled as a Markov process on the lattice in the right-half plane of $\mathbb{R}^2$ with different properties in the upper and lower quadrant. We showed that the compensation approach also works for this model. It leads to solutions in the upper and lower quadrant in the form of series of product forms. We further derived efficient algorithms for the computation of these solutions as well as global performance measures such as the moments of the sojourn time with the advantage of error bounds. Hence, we can conclude that extensions of the compensation approach with regard to the form of the state space are quite well possible. In fact, the asymmetric shortest queue problem represents a typical example of a Markov process on two or more adjacent quadrants with different properties in each quadrant. The extension to such a class of problems seems straightforward and therefore we do not present the details.

Chapter 6

Conclusions and comments

In section 1.1 we showed how the equilibrium probabilities $p_{m,n}$ of the symmetric shortest queue problem can be found by using a compensation method, which, after introducing the first term, consists of adding on terms of the form $c\alpha^m\beta^n$ so as to alternately satisfy the vertical and horizontal boundary conditions. This method exploits the asymptotic behaviour of the probabilities $p_{m,n}$ in the sense that the product $\alpha_0^m\beta_0^n$ which is the dominant term in the asymptotic behaviour of $p_{m,n}$ as $m \to \infty$ and $n > 0$, is taken as the first term in the series generated by the compensation method. It is well known however, that for the coupled processor problem (see [20,47,58]) and for the problem of two $M \mid M \mid 1$ queues with coupled arrivals (see [24,45]) the equilibrium probabilities $p_{m,n}$ have more complicated asymptotic behaviour involving extra factors $m^{-1/2}$ or $n^{-1/2}$. Therefore, it seems unlikely that the compensation approach also works for these problems. So the question arises for what problems exactly the compensation approach works.

As a first attempt to answer this question we extended in chapter 2 the compensation approach to a class of Markov processes on the pairs (m, n) of nonnegative integers. We considered Markov processes for which the transition rates are constant in the interior points of the state space and on each of the axes. To simplify the analysis, we assumed that transitions are possible to neighbouring states only. This class of Markov processes contains sense that the queueing problems mentioned above (by choosing an appropriate model). The compensation approach first characterizes the set of product form solutions $\alpha^m\beta^n$ satisfying the equilibrium equations in the interior points and then, by confronting these solutions with the boundary conditions, builds up an infinite linear combination of product form solutions that also satisfies these boundary conditions. The essence of this construction is a compensation idea: after introducing the main term, the approach consists of adding terms so as to alternately compensate for the vertical and horizontal boundary conditions. This construction leads to a formal, and thus possibly not useful (*divergent*) solution of the equilibrium equations. Therefore we derived conditions guaranteeing that this approach leads to useful results, that is, to a *convergent* infinite linear combination of products. The crucial condition appeared to be that no transitions are possible from the interior states to the north, north-east and east. Other conditions were either not relevant (but imposed for convenience only) or imposed to guarantee the ergodicity of the Markov process. The general theory developed in chapter 2 was applied to the symmetric shortest queue problem in chapter 3 and to a queueing model for a multiprogramming system in chapter

4. Both problems can be modelled by a Markov process of the type studied in chapter 2 and satisfying the condition on the transitions in the interior points. The compensation approach does indeed fail for the coupled processor problem and the problem of two $M|M|1$ queues with coupled arrivals, since these problems violate the latter condition.

It is important to remark that the compensation approach is constructive in nature and that therefore this approach is well suited for numerical purposes. This was demonstrated for the two specific problems in the chapters 3 and 4. It appeared that the compensation approach leads to efficient numerical procedures for the calculation of the equilibrium probabilities $p_{m,n}$ as well as other quantities of interest, such as the moments of the waiting time, with the advantage of tight error bounds.

As mentioned before, the analysis in chapter 2 should be regarded as a first attempt to characterize the Markov processes for which the compensation approach works. Further extensions are possible in several directions. Some of these extensions will be discussed in the subsequent sections.

6.1. Form of the state space

The analysis in chapter 2 is restricted to Markov processes in the first quadrant. Extensions to a more general form of state space are definitely possible, as shown in chapter 5. The main subject of chapter 5 is the analysis of the asymmetric shortest queue problem. This problem can be modelled as a Markov process on the pairs of integers (m, n) with m nonnegative, which behaves differently on each of the regions $n > 0$ and $n < 0$. Obviously this problem does not fit in the class of problems treated in chapter 2, but it is shown that the compensation approach also works for this problem. It leads to a series of product forms for the probabilities $p_{m,n}$ in the region $n > 0$ and a similar series for the probabilities in the region $n < 0$. These analytic results are exploited to construct efficient numerical procedures. Fayolle and Iasnogorodski [19,40] and Cohen and Boxma [14] show that the analysis of the generating function can be reduced to that of a *simultaneous* Riemann-Hilbert boundary value problem. This type of boundary value problem, however, requires further research. Knessl, Matkowsky, Schuss and Tier [46] derive asymptotic expressions for the stationary queue length distribution. To our knowledge no further results are available in the literature.

The asymmetric shortest queue problem represents a typical example of a Markov process on two adjacent quadrants (or on two coupled regions of possibly different form; cf. the threshold-type shortest queue problem in section 5.9) with different properties in each quadrant. The extension to such a class of problems seems straightforward and therefore we do not present the details.

Currently, we try to extend the compensation approach to Markov processes on higher dimensional state spaces. Although no definitive results are available yet, it seems likely that extensions in this direction are possible. Recent results indicate that the conditions required for n dimensional Markov processes, can be expressed in terms of conditions for $n-1$ dimensional Markov processes.

6.2. Complex boundary behaviour

The transitions at the horizontal and vertical boundary of the Markov processes treated in chapter 2 have a fairly simple structure. Extensions to more complex transition structures at the boundaries are feasible. In fact, an important case for which the transition structure at the vertical boundary is more complicated than the one treated in chapter 2, is the shortest queue problem with two parallel multi-server queues. The compensation approach is also applicable to this model (cf. section 5.9). In the next two sections we present two models with more complex behaviour at the boundaries as well as in the interior points.

6.3. The symmetric shortest delay problem for Erlang servers

One severe limitation of the models studied in the chapters 3, 4 and 5 is the assumption of exponential service times. In this section we study the problem of chapter 3 with Erlang servers and shortest delay routing. We sketch the essential features of the extension of the compensation approach to this problem. A paper on the detailed analysis of this problem is forthcoming. The models involved are not skipfree to the south, which is a basic assumption for the problems studied in the book by Cohen and Boxma [14]. Up to now, no analytical results seem to be available in the literature for these types of problems.

Consider a system with two identical parallel servers. The service times are Erlang-l distributed with mean l. Jobs arrive in a Poisson stream with intensity 2λ, where we assume that $\lambda l < 1$. Intuitively, this condition guarantees that the system can handle the offered load. An arriving job can be thought of as consisting of l identical subjobs, where each subjob requires an exponentially distributed service time with unit mean. Arriving jobs join the queue with the smallest number of *subjobs*, and in case the number of subjobs in the two queues is equal, join either queue with probability 1/2. This routing policy is called *shortest delay routing*, since arriving jobs join the queue promising the shortest delay, and it is optimal for parallel Erlang servers, see e.g. Hordijk and Koole [39] and Weber [60].

This queueing system can be represented by a continuous-time Markov process, whose natural state space consists of the pairs (i, j) where i and j are the numbers of subjobs in each queue. Instead of i and j we use the variables m and n where $m = \min(i, j)$ and $n = j - i$. Let

$\{p_{m,n}\}$ be the equilibrium distribution. For simplicity of presentation we restrict the analysis to the problem with Erlang-2 servers. The extension to Erlangian-l servers is briefly discussed at the end of this section. In fact, the Erlang-2 problem contains in its treatment already all ingredients, needed for the general problem. The transition-rate diagram for the shortest delay problem with Erlang-2 servers is depicted in figure 6.1. The main difference with respect to the simple model in chapter 3 is that transitions are not to neighbouring states only. Consequently, the behaviour at the boundaries is more complicated. By symmetry we have $p_{m,n} = p_{m,-n}$. Therefore, the analysis is further restricted to the probabilities in the first quadrant.

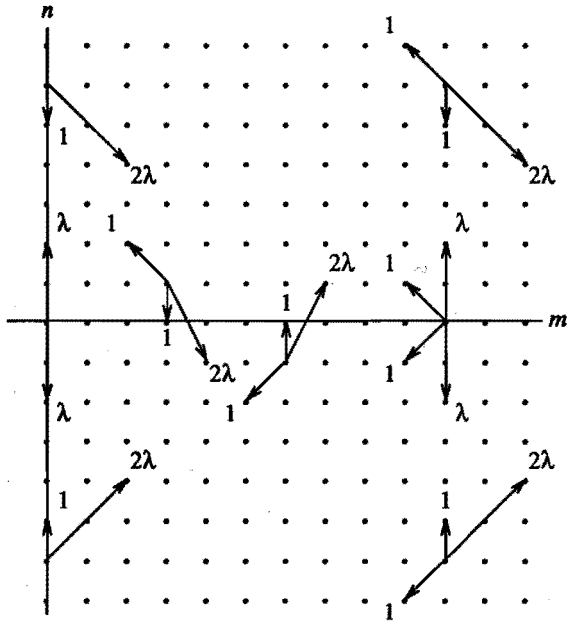


Figure 6.1.
Transition-rate diagram for the shortest delay problem with Erlang-2 servers.

Inspired by chapter 2, we start to look for feasible initial pairs α_0, β_0. That is, we try to find products $\alpha_0^m \beta_0^n$ satisfying the equilibrium equations in the interior points and satisfying the equations at the horizontal boundary (if $1 > |\alpha_0| > |\beta_0|$) or the vertical boundary (if $1 > |\beta_0| > |\alpha_0|$). Inserting the product $\alpha_0^m \beta_0^n$ into the equations for the interior points, that is, the points with $m > 1$ and $n > 2$, we find that α_0 and β_0 have to be roots of the cubic

equation

$$\alpha^2\beta 2(\lambda+1) = \beta^3 2\lambda + \alpha^2\beta^2 + \alpha^3 \,. \tag{6.1}$$

This cubic equation is the analogue of the quadratic equation (2.9). By Rouché's theorem it is easily shown that for each fixed α with $0 < |\alpha| < 1$, equation (6.1) has exactly *one* root β with $0 < |\beta| < |\alpha|$; and for each fixed β with $0 < |\beta| < 1$, equation (6.1) has exactly *two* (simple) roots α with $0 < |\alpha| < |\beta|$ (cf. lemma 2.7).

To find the solutions $\alpha_0^m\beta_0^n$ that satisfy the equations at the vertical boundary, that is, the points with $m \le 1$, note that the behaviour at this boundary is just the truncation of the behaviour at the interior points (cf. remark 2.4) and so we need not to introduce extra coefficients for the solution on the vertical boundary. Furthermore, since for fixed $\beta = \beta_0$ equation (6.1) has two roots α_0 and α_1 with $|\alpha| < |\beta_0|$, it seems more sensible to look for linear combinations $\alpha_0^m\beta_0^n + c_1\alpha_1^m\beta_0^n$ satisfying the vertical boundary conditions. It can easily be shown however that no such linear combinations exist. This result is of course suggested by the transition structure at the vertical boundary (cf. conclusion 2.14(iii)). Hence, feasible initial solutions stem from the horizontal boundary only.

To find the solutions $\alpha_0^m\beta_0^n$ that satisfy the equations at the horizontal boundary, that is, the points with $n \le 2$, we need to introduce extra coefficients for the solution on this boundary. Let $f_0\alpha_0^m$ be the solution for $n = 0$ and $g_0\alpha_0^m$ be the one for $n = 1$. Insertion of these solutions into the equations for $n \le 2$ leads to three nonlinear equations for α_0, f_0 and g_0. The analysis of this set of equations is difficult, since these equations also involve the parameter β_0 which can be regarded as a (complicated) function of α_0 (note that β_0 is the root with $|\beta| < |\alpha_0|$ of the cubic equation (6.1) for fixed $\alpha = \alpha_0$). Therefore we propose the following approach, leading to the desired feasible values of α_0 without use of β_0.

In section 3.11 we showed that the first term of the solution of the shortest queue problem gives the solution of the threshold jockeying problem. This suggests that the first terms for the shortest delay problem may be found by analysing the shortest delay problem with threshold jockeying. First consider the shortest delay problem with Erlang-1 servers and assume that one subjob jumps to the shortest queue as soon as the difference between the number of subjobs in the two queues exceeds 1. In fact, this is the problem in section 3.11 with $T = 1$. It is called the *instantaneous* jockeying problem. The solution of this problem is of the form

$$p_{m,n} = u_n\gamma^m \,, \quad m > 0 \,, \; 0 \le n \le 1 \,, \tag{6.2}$$

and is easily proved as follows. Inserting (6.2) in the equilibrium equations for states with $m > 1$ and $0 \le n \le 1$ leads to

$$u_0\gamma(\lambda+1) = u_1 2(\gamma+\lambda) \,, \tag{6.3}$$

$$u_1 2(\lambda+1) = u_0(\gamma+\lambda) \,.$$

These equations have a nonnull solution if and only if the determinant is zero, i.e., if

$$2(\gamma-\lambda^2)(\gamma-1)=0\,.$$

Hence, since the absolute value of γ must be less than one (necessary for normalization), we have to set $\gamma=\lambda^2$ and then u_0 and u_1 are solved from (6.3) (up to some multiplicative constant). Note that $\gamma=\lambda^2$ yields the feasible initial α_0 for the problem *without jockeying*, that is, the classical shortest queue problem.

This suggests that also for the Erlangian-2 servers the feasible initial values of α_0 may be found by analysing the related jockeying problem for which it is assumed that one subjob jumps to the shortest queue as soon as the difference between the number of subjobs in the two queues exceeds 2. This jockeying problem is solved by a linear combination of *three* geometric terms of the form (6.2), i.e.,

$$p_{m,n}=u_n\gamma_1^m+v_n\gamma_2^m+w_n\gamma_3^m\,,\quad m>1\,,\ 0\le n\le 2\,. \tag{6.4}$$

The proof of this result is omitted. We only state that $\gamma_1=\eta_1^2$, $\gamma_2=\eta_2^2$ and $\gamma_3=-\lambda/(1+\lambda)$, where η_1 and η_2 are the roots with $|\eta|<1$ of

$$\eta^2 2(\lambda+1)=2\lambda+\eta^3 2\,.$$

The parameters γ_1 and γ_2 can also be found as follows. Consider the jockeying process on the aggregate states k where k is the total number of subjobs in the system. Let P_k be the probability of being in state k. Equating the rate out and the rate into state $k>2$ yields

$$P_k 2(\lambda+1)=P_{k-2}2\lambda+P_{k+1}2\,,$$

from which we can conclude that for some constants c_1 and c_2,

$$P_k=c_1\eta_1^k+c_2\eta_2^k\,. \tag{6.5}$$

By using (6.4) and (6.5) and the fact that $P_{2k+1}=p_{k,1}$, we find two parameters, γ_1 and γ_2 say. This approach, however, does not lead to the determination of γ_3.

It can be shown that γ_1, γ_2 and γ_3 are indeed feasible values for α_0. That is, if α_0 is given by one of these values and β_0 is the root with $|\beta|<|\alpha_0|$ of equation (6.1) for $\alpha=\alpha_0$, then there exist coefficients f_0 and g_0 such that the horizontal boundary conditions are satisfied by

$$\begin{aligned}
&\alpha_0^m\beta_0^n && \text{for}\quad m\ge 0\,,\ n>1\,,\\
&g_0\alpha_0^m && \text{for}\quad m\ge 0\,,\ n=1\,,\\
&f_0\alpha_0^m && \text{for}\quad m\ge 0\,,\ n=0\,.
\end{aligned}$$

For each feasible pair α_0, β_0 the initial product $\alpha_0^m\beta_0^n$ violates the vertical boundary conditions. To compensate for this error we add the two products $c_1\alpha_1^m\beta_0^n$ and $c_2\alpha_2^m\beta_0^n$ where α_1 and α_2 are the roots with $|\alpha|<|\beta_0|$ of equation (6.1) for fixed $\beta=\beta_0$ and then try to choose

c_1 and c_2 such that the linear combination $\alpha_0^m\beta_0^n + c_1\alpha_1^m\beta_0^n + c_2\alpha_2^m\beta_0^n$ satisfies the vertical boundary conditions. The new term $c_1\alpha_1^m\beta_0^n$ however, violates the horizontal boundary conditions. To compensate for this error we add for $n > 1$ the term $d_1c_1\alpha_1^m\beta_1^n$, where β_1 is the root with $|\beta| < |\alpha_1|$ of equation (6.1) for fixed $\alpha = \alpha_1$; for $n = 0$ the term $f_1\alpha_1^m$; and for $n = 1$ the term $g_1\alpha_1^m$. Then we try to choose the coefficients d_1, f_1 and g_1 such that the horizontal boundary conditions are satisfied. The error of $c_2\alpha_2^m\beta_0^n$ on the horizontal boundary is compensated similarly. By repeating this procedure we generate an infinite sequence of compensation terms, which grows, due to the compensation on the vertical boundary, as a *binary tree*. The final solution $x_{m,n}(\alpha_0, \beta_0)$ is given by ($c_0 = d_0 = 1$ by definition)

$$x_{m,n}(\alpha_0, \beta_0) = \sum_{i=0}^{\infty} d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m + c_{i+2}\alpha_{i+2}^m)\beta_i^n \quad \text{for} \quad m \geq 0\,,\ n > 1\,,$$

and on the horizontal boundary by

$$x_{m,1}(\alpha_0, \beta_0) = \sum_{i=0}^{\infty} g_i\alpha_i^m \qquad \text{for} \quad m \geq 0\,,$$

$$x_{m,0}(\alpha_0, \beta_0) = \sum_{i=0}^{\infty} f_i\alpha_i^m \qquad \text{for} \quad m \geq 0\,.$$

The tree structure of $x_{m,n}(\alpha_0, \beta_0)$ and the corresponding structure of the parameters α_i and β_i are depicted in the figures 6.3 and 6.2 respectively. The series $x_{m,n}(\alpha_0, \beta_0)$ is a formal solution of the equilibrium equations.

It can be shown that for each feasible pair α_0, β_0 the construction of $x_{m,n}(\alpha_0, \beta_0)$ indeed succeeds, and furthermore, that there exists an integer $N > 1$ such that for each feasible pair α_0, β_0 the series $x_{m,n}(\alpha_0, \beta_0)$ converges absolutely for all states with $m \geq 0$, $n \geq 0$ and $m + n > N$, and for the three boundary states $(N-2, 0)$, $(N-2, 1)$ and $(N-1, 0)$. Finally, by restricting the Markov process on this convergence region (note that the Markov process always reenters the states with $m + n > N$ via one of the three boundary states), we can prove that there exist coefficients $k(\alpha_0, \beta_0)$ such that for all states in the convergence region,

$$p_{m,n} = \sum_{(\alpha_0, \beta_0)} k(\alpha_0, \beta_0)x_{m,n}(\alpha_0, \beta_0)\,,$$

where (α_0, β_0) runs through the three feasible initial pairs.

This concludes the analysis for Erlang-2 servers. The analysis can be extended to Erlang-l servers, in which case the equilibrium probabilities $p_{m,n}$ can be expressed as a linear combination of $l(l+1)/2$ series of product forms. Due to the compensation on the vertical boundary, each of these series has the structure of a *l-fold tree*. The extension to Erlang-l servers however, is still a little bit incomplete in the sense that the analogue of the properties stated in condition 2.24 has not been established yet (apart from the case $l = 2$).

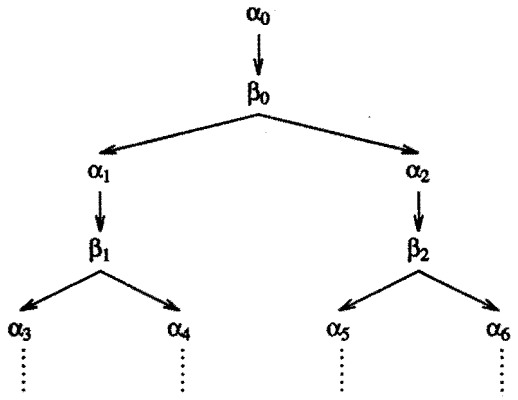


Figure 6.2.
The binary tree structure of the sequences $\{\alpha_i\}$ *and* $\{\beta_i\}$ *in* $x_{m,n}(\alpha_0, \beta_0)$. *These sequences are generated by the cubic equation (6.1).*

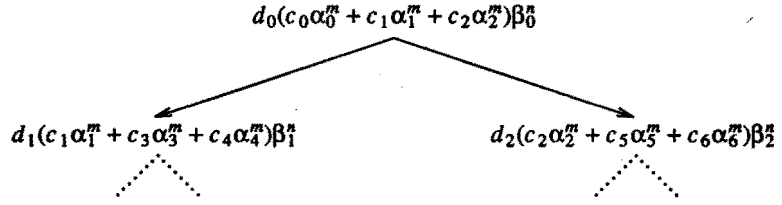


Figure 6.3.
The binary tree structure of the terms in the formal solution $x_{m,n}(\alpha_0, \beta_0)$.

In the following section we sketch the analysis of the $M\,|E_r|\,c$ queue, which is also an example of a Markov process, that can make larger jumps. Furthermore, this problem is a special example of the extension mentioned in section 6.1, since the state space is infinite in one direction only.

6.4. Analysis of the M|E$_r$|c queue

In this section we sketch the analysis of the $M \mid E_r \mid c$ queue. Consider a system with c parallel identical servers and a common queue. Jobs arrive according to a Poisson stream with intensity λ. On arrival each job requires an Erlang-r distributed service time with mean r/μ. The service discipline is first-come first-served. We assume that

$$\frac{r\lambda}{c\mu} < 1 .$$

Intuitively, this condition guarantees that the system can handle the offered load. This system can be modelled as a continuous-time Markov process, whose state space consists of the vectors $(n_0, n_1, ..., n_c)$, where n_0 is the number of waiting jobs and n_i is the number of remaining service phases for server i, $i = 1, ..., c$. This problem is a special example of the extension in 6.1, since the state space is infinite in the n_0-direction only.

The $M \mid E_r \mid c$ queue has been extensively studied in the literature. We mention that Mayhugh and McCormick [49] and Heffer [36] treated this queueing problem by using generating functions. Their analysis, however, does not lead to the explicit determination of the equilibrium probabilities. Shapiro [57] studied the $M \mid E_2 \mid c$ queue. His approach has some affinity with the one that is described below.

We first try to characterize the set of products $\alpha^{n_0}\beta_1^{n_1} \cdots \beta_c^{n_c}$ that satisfy the equilibrium equations in the interior points, that is, the points with $n_0 > 0$. It turns out that this set is *finite*. However, it contains a subset, a linear combination of which also satisfies the boundary conditions. Contrary to most of the problems treated before, this construction needs no compensation.

By inserting the products $\alpha^{n_0}\beta_1^{n_1} \cdots \beta_c^{n_c}$ into the equilibrium equations in states with $n_0 > 0$, we obtain a set of 2^c equations for the parameters $\alpha, \beta_1, ..., \beta_c$. Luckily, it can be shown that these equations are equivalent to a set of $c + 1$ equations, which are given below. That is, each equation in the original large set can be written as a linear combination of the equations in the following set:

$$\alpha(\lambda + c\mu) = \lambda + \sum_{i=1}^{c} \alpha\beta_i\mu , \tag{6.6}$$

$$\alpha(\lambda + c\mu) = \lambda + \sum_{i \neq j}^{c} \alpha\beta_i\mu + \frac{\alpha^2\mu}{\beta_j^{r-1}} , \quad j = 1, ..., c . \tag{6.7}$$

By subtracting (6.6) from (6.7) it readily follows that $\alpha = \beta_j^r$ for $j = 1, ..., c$. This implies that $\beta_j^r = \beta_i^r$ for all i and j. Hence, by introducing the parameters x_i satisfying

$$x_0 = 1 , \quad x_i^r = 1 , \quad i = 2, ..., c ,$$

we may write $\beta_i = x_i\beta_1$ for all i. Inserting this relation and $\alpha = \beta_1^r$ into (6.6) leads to

$$\beta_1^r(\lambda + c\mu) = \lambda + \beta_1^{r+1} \sum_{i=1}^{c} x_i \mu .$$

By using Rouché's theorem, it can be shown that for each feasible choice of the parameters x_i this equation has exactly r simple roots β_1 with $|\beta_1| < 1$. The parameters $\alpha, \beta_2, ..., \beta_c$ then follow from

$$\alpha = \beta_1^r , \quad \beta_i = x_i \beta_1 , \quad i = 2, ..., c .$$

Hence, we can conclude that there exist r^c products $\alpha^{n_0} \beta_1^{n_1} \cdots \beta_c^{n_c}$ satisfying the equilibrium equations in the states with $n_0 > 0$. Furthermore, it can be shown that these products are linearly independent on the set of states with $n_0 > 0$. Then, by restricting the Markov process to this set of states, we can prove that there exist coefficients k_i such that

$$p_{n_0, ..., n_c} = \sum_{i=1}^{r^c} k_i \alpha_i^{n_0} \beta_{1i}^{n_1} \cdots \beta_{ci}^{n_c} \quad \text{for all } (n_0, ..., n_c) \text{ with } n_0 > 0 .$$

For a detailed description of the results the reader is referred to [59]. We finally remark that the results can be extended to the $E_k | E_r | c$ queue. A paper on the analysis of the $E_k | E_r | c$ queue is forthcoming.

In the introduction we mentioned that our interest in shortest queue problems arose from problems in the design of flexible assembly systems (see[2, 7]). In the following section we describe these problems in more detail.

6.5. A class of queueing models for flexible assembly systems

In this section we introduce a class of queueing models, which may be used for the modelling of flexible assembly systems with a job-type dependent parallel structure. These models have not been studied in the literature, except for some special examples, see Schwarz [56], Roque [53] and Green [33]. The job-type dependent parallelism however, gives rise to analytical complications, for which no satisfactory mathematical solution techniques yet exist. In this section some of these complications will be illustrated by an example.

A typical production structure, which is encountered in several situations, is the one depicted in figure 6.4. This structure consists of a set of parallel machines and an incoming stream of several job-types. Each machine can treat a restricted set of job types. Incoming jobs are routed to one of the feasible machines according to some policy. If the operations for the different job types are of the same order of magnitude, then 'joining the shortest feasible queue' seems to be a sensible routing.

As an example of this kind of structure one may think of a number of parallel insertion machines, which have to mount vertical components on several types of printed circuit boards

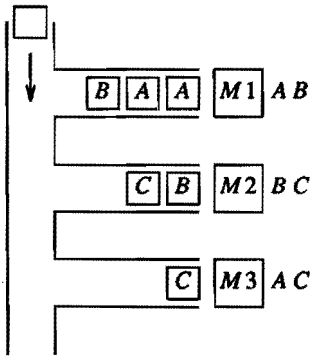


Figure 6.4.
A job-type dependent parallel production structure. M1, M2 and M3 denote machines, that treat the job-types A, B and C. Each machine can treat a restricted set of job types; M1 can treat A and B jobs, M2 can treat B and C jobs, and M3 can treat A and C jobs.

(PCB). For a technical description of these machines the reader is referred to Zijm [65]. Due to the *limited storage capacity* for components, each machine contains components for a restricted set of PCB types. Hence, the limited storage capacity gives rise to the job-type dependent parallelism. One of the important decision aspects in these systems is how to divide the necessary components among the machines. To find an optimal assignment of the components, we should be able to efficiently evaluating models containing the essential features of this system. Since the PCB production is characterized by large production batches and small processing times, queueing models seem to be appropriate. From a modelling point of view, the following class of queueing models contains some of the essential features.

> N types of jobs arrive at a system consisting of M identical parallel exponential servers. Each server can treat a subset of job-types. On arrival jobs join the shortest feasible queue, and in case of equal shortest queue lengths, ties are broken with equal probabilities.

This class of queueing models is of course a strong simplification of reality. But if one wants to be able to efficiently evaluate more realistic models, then one must first be able to do this for simple models. Even these simple models however, cannot be analyzed exactly by existing mathematical techniques. So we tried to develop heuristic evaluation methods for these models, see [2]. However, we did not succeed in developing methods that were both accurate and efficient. Therefore we decided to obtain a better understanding of the process by first considering a further simplified process. By restricting the analysis to one job type and two servers we

arrived at the shortest queue problem. For this problem we developed the compensation approach yielding efficient and accurate evaluation methods and we characterized a class of two-dimensional problems for which this approach works. Unfortunately, the following example illustrates that even models with a fairly simple job-type dependent parallel structure are not contained in this class.

A and *B* jobs arrive at a system consisting of two parallel servers (see figure 6.5). The service times are exponentially distributed with mean μ^{-1}. One server can treat both job-types, whereas the other one serves *B* jobs only. Arriving *A* jobs always join the *AB* queue and *B* jobs join the shortest queue.

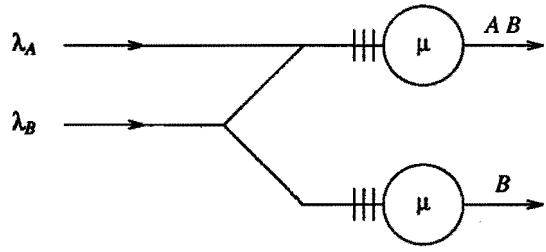


Figure 6.5.
Queueing model with a job-type dependent parallel structure. Arriving A jobs always join the general AB queue. Arriving B jobs join the shortest queue, and in case of equal queues, join either queue with probability 1/2.

This system can be represented by a continuous-time Markov process on the pairs (m, n) where m is the length of the shortest queue and n is the difference in length of the queues. The transition-rate diagram is shown in figure 6.6. This problem is of the same type as the asymmetric shortest queue problem in the sense that the Markov process has different properties in the regions $n > 0$ and $n < 0$ (cf. figure 5.2). However, due to the possibility of transition to the south in states with $n < 0$, the compensation approach does not work for this problem. On the other hand numerical experiments show some nice features. For the model in figure 6.5 numerical experiments suggests that there are $\alpha_0, \beta_1, \beta_2, d_1$ and d_2 such that (cf. (5.11))

$$p_{m,n} \sim K d_1 \alpha_0^m \beta_1^n \quad (m \to \infty, n > 0);$$

$$p_{m,n} \sim K d_2 \alpha_0^m \beta_2^{-n} \quad (m \to \infty, n < 0);$$

$$p_{m,0} \sim K \alpha_0^m \quad (m \to \infty),$$

for some constant K. In some way this empirical finding should be exploited to develop satisfactory evaluation methods.

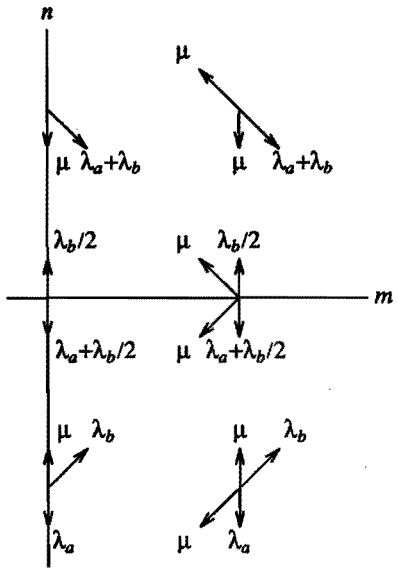


Figure 6.6.
Transition-rate diagram of the queueing model in figure 6.5.

We finally remark that Schwartz [56] considered queueing models with N parallel queues and N job-types, with a hierarchical job-type dependent parallelism. That is, an arriving type-i job may choose among the queues from 1 to i only. His analysis, however, turned out to contain errors, see Roque [53]. Green [33] studied a model, related to the one in figure 6.5, but with a common queue for the servers. The service discipline is first-come first-served, except that B jobs may pass A jobs if the B server becomes available. By truncating an appropriate state variable, she calculated approximations for the stationary probabilities by using the matrix-geometric theory developed by Neuts [51].

References

1. ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., Analysing multiprogramming queues by generating functions. (submitted for publication).
2. 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.
3. 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.
4. ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "Matrix-geometric analysis of the shortest queue problem with threshold jockeying," Memorandum COSOR 91-24, Eindhoven University of Technology, Dep. of Math. and Comp. Sci., 1991.
5. ADAN, I.J.B.F., HOUTUM, G.J. VAN, WESSELS, J., AND ZIJM, W.H.M., "A compensation procedure for multiprogramming queues," Memorandum COSOR 91-13, Eindhoven University of Technology, Dep. of Math. and Comp. Sci., 1991. (submitted for publication).
6. 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.
7. ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "Flexible assembly and shortest queue problems," in *Modern production concepts; theory and applications*, ed. G. Fandel, G. Zaepfel, pp. 644-659, Springer-Verlag, Berlin, 1991.
8. 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, 1991 (to appear).
9. BASKETT, F., CHANDY, K.M., MUNTZ, R.R., AND PALACIOS, F.G., "Open, closed and mixed networks of queues with different classes of customers," *J.ACM*, vol. 22, pp. 248-260, 1975.
10. BLANC, J.P.C., "On a numerical method for calculating state probabilities for queueing systems with more than one waiting line," *J. Comput. Appl. Math.*, vol. 20, pp. 119-125, 1987.
11. BLANC, J.P.C., "The power-series algoritm applied to the shortest-queue model," Memorandum 379, Tilburg University, Department of Economics, 1989.
12. BLANC, J.P.C., "The power-series algoritm applied to cyclic polling systems," *Stochastic Models*, vol. 7, 1991.
13. BRUIJN, N.G. DE, *Asymptotic methods in analysis*, Dover Publications, Inc., New York, 1981.

14. COHEN, J.W. AND BOXMA, O.J., *Boundary value problems in queueing system analysis*, North-Holland, Amsterdam, 1983.

15. COHEN, J.W., "A two-queue, one-server model with priority for the longer queue," *Queueing systems*, vol. 2, pp. 261-283, 1987.

16. CONOLLY, B.W., "The autostrada queueing problem," *J. Appl. Prob.*, vol. 21, pp. 394-403, 1984.

17. 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.

18. ELSAYED, E.A. AND BASTANI, A., "General solutions of the jockeying problem," *Eur. J. Oper. Res.*, vol. 22, pp. 387-396, 1985.

19. FAYOLLE, G., *Méthodes analytiques pour les files d'attente couplées*, Thesis, Univ. de Paris VI, Paris, 1979.

20. 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.

21. FAYOLLE, G., "On functional equations of one and two variables arising in the analysis of stochastic models," in *Math. Comp. Perform/Reliability*, ed. G. Iazeolla, P.J. Courtois, A. Hordijk, pp. 55-75, North-Holland, Amsterdam, 1984.

22. FLATTO, L. AND MCKEAN, H.P., "Two parallel queues with equal servicing rates," Science Report, RC5916, IBM, 1977.

23. FLATTO, L. AND MCKEAN, H.P., "Two queues in parallel," *Comm. Pure Appl. Math.*, vol. 30, pp. 255-263, 1977.

24. 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.

25. FLATTO, L., "The longer queue model," *Prob. Engineer. Inform. Sci.*, vol. 3, pp. 537-559, 1989.

26. FOSCHINI, G. J. AND SALZ, J., "A basic dynamic routing problem and diffusion," *IEEE Trans. Commun.*, vol. COM-26, pp. 320-327, 1978.

27. FOSTER, F.G., "On the stochastic matrices associated with certain queuing processes," *Ann. Math. Stat.*, vol. 24, pp. 355-360, 1953.

28. GANTMACHER, F.R., *The theory of matrices*, vol. II (translated by K.A. Hirsch), Chelsea, New York, 1959.

29. GERTSBAKH, I., "The shorter queue problem: A numerical study using the matrix-geometric solution," *Eur. J. Oper. Res.*, vol. 15, pp. 374-381, 1984.

30. GRASSMANN, W.K., "Transient and steady state results for two parallel queues," *OMEGA Int. J. of Mgmt Sci.*, vol. 8, pp. 105-112, 1980.

31. GRASSMANN, W.K. AND ZHAO, Y., "The shortest queue model with jockeying," *Naval Res. Log.*, vol. 37, pp. 773-787, 1990.

32. GRASSMANN, W.K. AND ZHAO, Y., "A numerically stable algorithm for two server queue models," *Queueing Systems*, vol. 8, pp. 59-80, 1991.

33. GREEN, L., "A queueing system with general-use and limited-use servers," *Opns. Res.*, vol. 33, pp. 168-182, 1985.

34. HAIGHT, F.A., "Two queues in parallel," *Biometrica*, vol. 45, pp. 401-410, 1958.

35. HALFIN, S., "The shortest queue problem," *J. Appl. Prob.*, vol. 22, pp. 865-878, 1985.

36. HEFFER, J.C, "Steady-state solution of the $M \mid E_k \mid c$ (∞, FIFO) queueing system," *INFOR* , vol. 7, pp. 16-30, 1969.

37. HOFRI, M., "A generating-function analysis of multiprogramming queues," *International Journal of Computer and Information Sciences*, vol. 7, pp. 121-155, 1978.

38. 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.

39. HORDIJK, A. AND KOOLE, G., "On the optimality of the generalized shortest queue policy," *Prob. Engineer. Inform. Sci.*, vol. 4, pp. 477-487, 1990.

40. IASNOGORODSKI, R., *Problèmes-frontières dans les files d'attente*, Thesis, Univ. de Paris VI, Paris, 1979.

41. JACKSON, J., *Classical electrodynamics*, 2nd edition, J. Wiley & Sons, New York, 1975.

42. 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.

43. KEILSON, J., *Markov chain models- rarity and exponentiality*, Springer-Verlag, Berlin, 1979.

44. KINGMAN, J.F.C., "Two similar queues in parallel," *Ann. Math. Statist.*, vol. 32, pp. 1314-1323, 1961.

45. KLEIN, J. DE, *Fredholm integral equations in queueing analysis*, Thesis, Rijksuniversiteit Utrecht, Utrecht, 1988.

46. 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.

47. KONHEIM, A.G., MEILIJSON, I., AND MELKMAN, A., "Processor-sharing of two parallel lines," *J. Appl. Prob.*, vol. 18, pp. 952-956, 1981.

48. MAXWELL, J.C., *A treatise on electricity and magnetism*, vol. I, 3rd edition, Oxford, 1904.

49. MAYHUGH, J.O. AND MCCORMICK, R.E., "Steady state solution of the queue $M|E_k|r$," *mgmt. Sci.*, vol. 14, pp. 692-712, 1968.

50. NELSON, R.D. AND PHILIPS, T.K., "An approximation to the response time for shortest queue routing," *Performance Evaluation Review*, vol. 17, pp. 181-189, 1989.

51. NEUTS, M.F., *Matrix-geometric solutions in stochastic models*, Johns Hopkins University Press, Baltimore, 1981.

52. 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.

53. ROQUE, D.R., "A note on "Queueing models with lane selection"," *Opns. Res.*, vol. 28, pp. 419-420, 1980.

54. SCHASSBERGER, R., "Ein wartesystem mit zwei parallelen warteschlangen," *Computing*, vol. 3, pp. 110-124, 1968.

55. SCHASSBERGER, R., "A service system with two parallel queues," *Computing*, vol. 4, pp. 24-29, 1969.

56. SCHWARTZ, B.L., "Queuing models with lane selection: a new class of problems," *Opns. Res.*, vol. 22, pp. 331-339, 1974.

57. SHAPIRO, S., "The M-server queue with Poisson input and Gamma-distributed service of order two," *Opns. Res.*, vol. 14, pp. 685-694, 1966.

58. WAARD, E.N. DE, "A study of the equilibrium probabilities of the shorter queue model and the symmetric coupled processor model," Master's thesis, Delft University of Technology, Fac. of Math. and Inf., 1991.

59. WAARSENBURG, W.A. VAN DE, *The stationary state distribution of multi-server queueing systems with Erlangian distributed srevice times*, Master's Thesis, Eindhoven University of Technology, Eindhoven, 1991.

60. WEBER, R.R., "On the optimal assignment of customers to parallel queues," *J. Appl. Prob.*, vol. 15, pp. 406-413 , 1978.

61. WINSTON, W., "Optimality of the shortest line discipline," *J. Appl. Prob.*, vol. 14, pp. 181-189, 1977.

62. WOLFF, R.W., "Poisson arrivals see time averages," *Opns. Res.*, vol. 30, pp. 223-231, 1982.

63. ZHAO, Y. AND GRASSMANN, W.K., "Solving a parallel queueing model by using modified lumpability," Research paper, Queen's University, Dep. of Math. and Stat., 1991.

64. ZHENG, Y.S. AND ZIPKIN, P., "A queuing model to analyze value of centralized inventory information," Research working paper no. 86-6, Columbia University Business School, New York, 1986.

65. ZIJM, W.H.M., "Operational control of automated PCB assembly lines," in *Modern production concepts; theory and applications*, ed. G. Fandel, G. Zaepfel, Springer-Verlag, Berlin, 1991.

Appendix A

Below we formulate a result of Foster ([27], theorem 1). Let $P = [p_{ij}]$ be the transition matrix of an irreducible, aperiodic Markov chain on the states $\{0, 1, 2, ...\}$ and denote by $P^{(n)} = [p_{ij}^{(n)}]$ its nth power.

Theorem A.1.

If there exists a nonnull solution $\{x_i\}$ *of the equilibrium equations*

$$\sum_{i=0}^{\infty} x_i p_{ij} = x_j \qquad (j = 0, 1, 2, ...) \tag{A.1}$$

such that $\sum |x_i| < \infty$, *then* P *is ergodic and normalization of* $\{x_i\}$ *produces* $\{\pi_i\}$.

Proof.

It is known that $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$ always exists and is independent of i; and further that *either* $\pi_j > 0$ for all j *or* $\pi_j \equiv 0$. For any nonnull solution $\{x_i\}$ of (A.1)

$$\sum_{i=0}^{\infty} x_i p_{ij}^{(n)} = x_j \qquad (j = 0, 1, 2, ...)$$

for all n, and so by the absolute convergence of $\sum x_i$ it follows by letting $n \to \infty$ that

$$\sum_{i=0}^{\infty} x_i \pi_j = x_j \qquad (j = 0, 1, 2, ...). \tag{A.2}$$

Therefore $\pi_j > 0$ (for otherwise $\{x_i\}$ would be null), and so P is ergodic. Moreover, (A.2) states that the ratio of x_j and π_j is independent of j, so normalization of $\{x_i\}$ produces $\{\pi_i\}$. □

This result can be extended to continuous time Markov processes. Let $Q = [q_{ij}]$ be the generator of an irreducible Markov process on the states $\{0, 1, 2, ...\}$ with $\sup_{i\geq 0} -q_{ii} < \infty$; so there is a $\Delta > 0$ such that $\Delta^{-1} > \sup_{i\geq 0} -q_{ii}$. The continuous time equivalent of theorem A.1 is obtained by applying theorem A.1 to the irreducible aperiodic Markov chain $P = I + \Delta Q$ and using that

$$\sum_{i=0}^{\infty} x_i p_{ij} = x_j \iff \sum_{i=0}^{\infty} x_i q_{ij} = 0 \qquad (j = 0, 1, 2, ...).$$

Corollary A.2.

If there exists a nonnull solution $\{x_i\}$ *of the equilibrium equations*

$$\sum_{i=0}^{\infty} x_i q_{ij} = 0 \qquad (j = 0, 1, 2, ...)$$

such that $\sum |x_i| < \infty$, *then* Q *is ergodic and normalization of* $\{x_i\}$ *produces* $\{\pi_i\}$.

Appendix B

In this appendix we prove theorem 5.20 and lemma 5.22.

Proof of theorem 5.20.

For $m \geq 0$, $n > 0$ and all nodes i in the compensation tree we first derive upper bounds $\bar{R}_{ll}(i, m, n)$ and $\bar{R}_{lr}(i, m, n)$ on $R_{ll}(i, m, n)$ and $R_{lr}(i, m, n)$ for $i \in L$ and upper bounds $\bar{R}_{rl}(i, m, n)$ and $\bar{R}_{rr}(i, m, n)$ on $R_{rl}(i, m, n)$ and $R_{rr}(i, m, n)$ for $i \in R$. We prove that these upper bounds are monotone in the sense that for all nodes j in the subtree below node i,

$$\bar{R}_{ll}(j, m, n) \leq \bar{R}_{ll}(l(i), m, n), \quad \bar{R}_{lr}(j, m, n) \leq \bar{R}_{lr}(l(i), m, n) \quad \text{if} \quad j \in L, \tag{B.1}$$

$$\bar{R}_{rl}(j, m, n) \leq \bar{R}_{rl}(r(i), m, n), \quad \bar{R}_{rr}(j, m, n) \leq \bar{R}_{rr}(r(i), m, n) \quad \text{if} \quad j \in R.$$

The proof of theorem 5.20 is then completed by defining the matrix $B(i, m, n)$ as follows.

Definition B.1.

For all $m \geq 0, n > 0$ *and* $i \geq 1$,

$$B(i, m, n) = \begin{bmatrix} B_{ll}(i, m, n) & B_{rl}(i, m, n) \\ B_{lr}(i, m, n) & B_{rr}(i, m, n) \end{bmatrix} = \begin{bmatrix} \bar{R}_{ll}(l(i), m, n) & \bar{R}_{rl}(r(i), m, n) \\ \bar{R}_{lr}(l(i), m, n) & \bar{R}_{rr}(r(i), m, n) \end{bmatrix}.$$

Let $m \geq 0$ and $n > 0$. Then for all $i \in L$ we derive an upper bound $\bar{R}_{lr}(i, m, n)$ on $R_{lr}(i, m, n)$ satisfying the monotonicty given in (B.1). The other bounds are derived similarly. For $i \in L$ the ratio $R_{lr}(i, m, n)$ may be written as (see section 5.5)

$$R_{lr}(i, m, n) = \frac{c_i}{c_{p(i)}} \frac{|d_{r(i)}|}{|d_i|} \frac{1 + (c_{r(i)} / c_i)(\alpha_{r(i)} / \alpha_i)^m}{1 + (c_i / c_{p(i)})(\alpha_i / \alpha_{p(i)})^m} \left[\frac{\alpha_i}{\alpha_{p(i)}}\right]^m \left[\frac{\beta_{r(i)}}{\beta_i}\right]^n.$$

By substituting (cf. (5.15))

$$\frac{\alpha_{r(i)}}{\alpha_i} = \frac{y_-(\beta_{r(i)})}{y_+(\beta_{r(i)})}, \quad \frac{\alpha_i}{\alpha_{p(i)}} = \frac{Y_-(\beta_i)}{Y_+(\beta_i)}, \quad \frac{\beta_{r(i)}}{\beta_i} = \frac{x_-(\alpha_i)}{X_+(\alpha_i)},$$

in the right-hand side of this equality we obtain

$$R_{lr}(i, m, n) = \frac{c_i}{c_{p(i)}} \frac{|d_{r(i)}|}{|d_i|} \frac{1 + (c_{r(i)} / c_i)(y_-(\beta_{r(i)}) / y_+(\beta_{r(i)}))^m}{1 + (c_i / c_{p(i)})(Y_-(\beta_i) / Y_+(\beta_i))^m} \left[\frac{Y_-(\beta_i)}{Y_+(\beta_i)}\right]^m \left[\frac{x_-(\alpha_i)}{X_+(\alpha_i)}\right]^n.$$

Now we first need bounds on $c_i / c_{p(i)}$, $c_{r(i)} / c_i$ and $|d_{r(j)}| / |d_j|$.

Definition B.2.

For all $0 < \alpha < \alpha_0$, *define*

$$\overline{D}_{ll}(\alpha) = \frac{\dfrac{(\alpha\gamma_1 + 2q\rho)\alpha}{X_-(\alpha)} + \dfrac{(\alpha\gamma_2 + 2(1-q)\rho)\,\alpha}{x_+(\alpha)}}{\dfrac{(\alpha\gamma_1 + 2q\rho)\alpha}{X_+(\alpha)} + \dfrac{(\alpha\gamma_2 + 2(1-q)\rho)\alpha}{x_+(\alpha)} - 2(\rho+1)\alpha},$$

$$\overline{D}_{lr}(\alpha) = \frac{\gamma_1(\alpha\gamma_2 + 2(1-q))(A_2 - A_1)}{\gamma_2\left[\dfrac{(\alpha\gamma_1 + 2q\rho)\alpha}{X_+(\alpha)} + \dfrac{(\alpha\gamma_2 + 2(1-q)\rho)\alpha}{x_+(\alpha)} - 2(\rho+1)\alpha\right]},$$

$$\overline{D}_{rl}(\alpha) = \frac{\gamma_2(\alpha\gamma_1 + 2q)(a_2 - a_1)}{\gamma_1\left[\dfrac{(\alpha\gamma_1 + 2q\rho)\alpha}{X_+(\alpha)} + \dfrac{(\alpha\gamma_2 + 2(1-q)\rho)\alpha}{x_+(\alpha)} - 2(\rho+1)\alpha\right]},$$

$$\overline{D}_{rr}(\alpha) = \frac{\dfrac{(\alpha\gamma_1 + 2q\rho)\alpha}{X_+(\alpha)} + \dfrac{(\alpha\gamma_2 + 2(1-q)\rho)\alpha}{x_-(\alpha)}}{\dfrac{(\alpha\gamma_1 + 2q\rho)\alpha}{X_+(\alpha)} + \dfrac{(\alpha\gamma_2 + 2(1-q)\rho)\alpha}{x_+(\alpha)} - 2(\rho+1)\alpha}$$

and for all $0 < \beta < 1$, *define*

$$\underline{C}_l(\beta) = \frac{1 - Y_-(\beta)/\beta}{A_2 - 1}, \quad \overline{C}_l(\beta) = \frac{1 - A_1}{Y_+(\beta)/\beta - 1},$$

$$\underline{C}_r(\beta) = \frac{1 - y_-(\beta)/\beta}{a_2 - 1}, \quad \overline{C}_r(\beta) = \frac{1 - a_1}{y_+(\beta)/\beta - 1},$$

where

$$A_1 = \frac{\rho + 1 - \sqrt{(\rho+1)^2 - 2\rho\,\gamma_1}}{\gamma_1}, \quad A_2 = \frac{\rho + 1 + \sqrt{(\rho+1)^2 - 2\rho\,\gamma_1}}{\gamma_1}$$

$$a_1 = \frac{\rho + 1 - \sqrt{(\rho+1)^2 - 2\rho\,\gamma_2}}{\gamma_2}, \quad a_2 = \frac{\rho + 1 + \sqrt{(\rho+1)^2 - 2\rho\,\gamma_2}}{\gamma_2}.$$

Lemma B.3.

$$0 < \underline{C}_l(\beta_i) < \frac{c_i}{c_{p(i)}} < \overline{C}_l(\beta_i), \quad \frac{|d_{l(i)}|}{|d_i|} < \overline{D}_{ll}(\alpha_i), \quad \frac{|d_{r(i)}|}{|d_i|} < \overline{D}_{lr}(\alpha_i) \quad \textit{for} \quad i \in L;$$

$$0 < \underline{C}_r(\beta_i) < \frac{c_i}{c_{p(i)}} < \overline{C}_r(\beta_i), \quad \frac{|d_{l(i)}|}{|d_i|} < \overline{D}_{rl}(\alpha_i), \quad \frac{|d_{r(i)}|}{|d_i|} < \overline{D}_{rr}(\alpha_i) \quad \textit{for} \quad i \in R.$$

Proof.

We prove the bounds for $i \in L$. The proof is similar for $i \in R$. The bounds on $c_i / c_{p(i)}$ follow by dividing the denominator and numerator of the quotient defining $c_i / c_{p(i)}$ by β_i and inserting the inequalities (cf. lemma 5.7)

$$Y_+(\beta_i)/\beta_i < \lim_{\beta \downarrow 0} Y_+(\beta)/\beta = A_2 \,, \quad Y_-(\beta_i)/\beta_i > \lim_{\beta \downarrow 0} Y_-(\beta)/\beta = A_1$$

in this quotient. Multiplying the denominator and numerator of the quotient defining $d_{l(i)} / d_i$ by α_i and adding $2(\rho + 1)\alpha_i$ to the numerator of this quotient yields the bound on $|d_{l(i)}| / |d_i|$. Finally, multiplying the denominator and numerator of the quotient defining $d_{r(i)} / d_i$ by α_i and inserting the inequalities

$$X_+(\alpha_i)/\alpha_i < \lim_{\alpha \downarrow 0} X_+(\alpha)/\alpha = A_1^{-1} \,, \quad X_-(\alpha_i)/\alpha_i > \lim_{\alpha \downarrow 0} X_-(\alpha)/\alpha = A_2^{-1}$$

in the numerator of this quotient yields the bound on $|d_{r(i)}| / |d_i|$. □

Application of lemma B.3 yields for $i \in L$

$$R_{lr}(i,\, m,\, n) < \overline{C}_l(\beta_i)\, \overline{D}_{lr}(\alpha_i)\, \frac{1 + \overline{C}_r(\beta_{r(i)})\,(y_-(\beta_{r(i)})/y_+(\beta_{r(i)}))^m}{1 + \underline{C}_l(\beta_i)\,(Y_-(\beta_i)/Y_+(\beta_i))^m} \left[\frac{Y_-(\beta_i)}{Y_+(\beta_i)}\right]^m \left[\frac{x_-(\alpha_i)}{X_+(\alpha_i)}\right]^n .$$

The right-hand side of this inequality is the desired bound $\overline{R}_{lr}(i,\, m,\, n)$.

Definition B.4.

For all $m \geq 0$ and $n > 0$, define for $i \in L$

$$\overline{R}_{ll}(i,\, m,\, n) = \overline{C}_l(\beta_i)\, \overline{D}_{ll}(\alpha_i)\, \frac{1 + \overline{C}_l(\beta_{l(i)})\,(Y_-(\beta_{l(i)})/Y_+(\beta_{l(i)}))^m}{1 + \underline{C}_l(\beta_i)\,(Y_-(\beta_i)/Y_+(\beta_i))^m} \left[\frac{Y_-(\beta_i)}{Y_+(\beta_i)}\right]^m \left[\frac{X_-(\alpha_i)}{X_+(\alpha_i)}\right]^n ,$$

$$\overline{R}_{lr}(i,\, m,\, n) = \overline{C}_l(\beta_i)\, \overline{D}_{lr}(\alpha_i)\, \frac{1 + \overline{C}_r(\beta_{r(i)})\,(y_-(\beta_{r(i)})/y_+(\beta_{r(i)}))^m}{1 + \underline{C}_l(\beta_i)\,(Y_-(\beta_i)/Y_+(\beta_i))^m} \left[\frac{Y_-(\beta_i)}{Y_+(\beta_i)}\right]^m \left[\frac{x_-(\alpha_i)}{X_+(\alpha_i)}\right]^n ,$$

and define for $i \in R$

$$\overline{R}_{rl}(i,\, m,\, n) = \overline{C}_r(\beta_i)\, \overline{D}_{rl}(\alpha_i)\, \frac{1 + \overline{C}_l(\beta_{l(i)})\,(Y_-(\beta_{l(i)})/Y_+(\beta_{l(i)}))^m}{1 + \underline{C}_r(\beta_i)\,(y_-(\beta_i)/y_+(\beta_i))^m} \left[\frac{y_-(\beta_i)}{y_+(\beta_i)}\right]^m \left[\frac{X_-(\alpha_i)}{x_+(\alpha_i)}\right]^n ,$$

$$\overline{R}_{rr}(i,\, m,\, n) = \overline{C}_r(\beta_i)\, \overline{D}_{rr}(\alpha_i)\, \frac{1 + \overline{C}_r(\beta_{r(i)})\,(y_-(\beta_{r(i)})/y_+(\beta_{r(i)}))^m}{1 + \underline{C}_r(\beta_i)\,(y_-(\beta_i)/y_+(\beta_i))^m} \left[\frac{y_-(\beta_i)}{y_+(\beta_i)}\right]^m \left[\frac{x_-(\alpha_i)}{x_+(\alpha_i)}\right]^n .$$

The following properties are required to establish that these upper bounds are monotone.

Lemma B.5.

(i) *For* $0 < \alpha < \alpha_0$, *the functions* $\overline{D}_{ll}(\alpha)$, $\overline{D}_{lr}(\alpha)$, $\overline{D}_{rl}(\alpha)$ *and* $\overline{D}_{rr}(\alpha)$ *are increasing in* α;

(ii) *For* $0 < \beta < 1$, *the functions* $\overline{C}_l(\beta)$ *and* $\overline{C}_r(\beta)$ *are increasing in* β *and the functions* $\underline{C}_j(\beta)$ *and* $\underline{C}_r(\beta)$ *are decreasing in* β;

(iii) *For* $0 < \alpha < \alpha_0$, *the ratios* $X_-(\alpha)/X_+(\alpha)$, $x_-(\alpha)/X_+(\alpha)$, $X_-(\alpha)/x_+(\alpha)$ *and* $x_-(\alpha)/x_+(\alpha)$ *are increasing in* α;

(iv) *For* $0 < \beta < 1$, *the ratios* $Y_-(\beta)/Y_+(\beta)$ *and* $y_-(\beta)/y_+(\beta)$ *are increasing in* β.

Proof.

(i): We prove the monotonicity for $\overline{D}_{ll}(\alpha)$. The proof is similar for the other functions. The denominator of $\overline{D}_{ll}(\alpha)$ vanishes at $\alpha = \alpha_0$ and $\alpha = 1$ and is strictly convex for $0 < \alpha < 1$ (cf. the proof of lemma 2.17). Hence the denominator of $\overline{D}_{ll}(\alpha)$ is positive and decreasing in α for $0 < \alpha < \alpha_0$. By lemma 5.7 the two terms in the numerator are positive and the second one is increasing in α. Now it remains to prove that the first one is also increasing. Since

$$\frac{d}{d\alpha}\frac{(\alpha\gamma_1 + 2q\rho)\,\alpha}{X_-(\alpha)} = \frac{\gamma_1\,\alpha}{X_-(\alpha)} - \frac{(\alpha\gamma_1 + 2q\rho)\,\gamma_2}{2\sqrt{(\rho+1)^2 - (2\rho + \alpha\gamma_2)\,\gamma_1}}$$

is decreasing in α, we obtain for $0 < \alpha < \alpha_0$

$$\frac{d}{d\alpha}\frac{(\alpha\gamma_1 + 2q\rho)\,\alpha}{X_-(\alpha)} > \frac{\gamma_1\,\alpha_0}{X_-(\alpha_0)} - \frac{(\alpha_0\gamma_1 + 2q\rho)\,\gamma_2}{2\sqrt{(\rho+1)^2 - (2\rho + \alpha_0\gamma_2)\,\gamma_1}}$$

$$= 2 + \rho\gamma_2 - \frac{(\rho\gamma_1 + 2q)\,\rho\gamma_2}{2(1 + \rho(1-\gamma_1))} > 2 + \rho\gamma_2 - \frac{(\rho\gamma_1 + 2q)}{2} > 0,$$

proving that the first term in the numerator is increasing in α for $0 < \alpha < \alpha_0$.

(ii), (iii) and (iv): Immediately from lemma 5.7. □

We can now prove the monotonicity property (B.1). Let node j be a *left* descendant in the subtree below node i in the compensation tree. In the parameter tree α_j is a member of the subtree below α_i, so $\alpha_j < \alpha_{p(j)} \le \alpha_i$ by corollary 5.8. From lemma 5.7 it then follows that

$$\beta_j = X_-(\alpha_{p(j)}) \le X_-(\alpha_i) = \beta_{l(i)},$$

$$\alpha_j = Y_-(\beta_j) \le Y_-(\beta_{l(i)}) = \alpha_{l(i)},$$

$$\beta_{r(j)} = x_-(\alpha_j) \le x_-(\alpha_{l(i)}) = \beta_{r(l(i))}.$$

By using these inequalities and lemma B.5 and the fact that $x^m/(1+cx^m)$ is increasing in x for $x \geq 0$ and $c \geq 0$ we obtain the desired inequality

$$\bar{R}_{lr}(j, m, n) \leq \bar{R}_{lr}(l(i), m, n) .$$

Since the other inequalities are proved similarly, this completes the proof of theorem 5.20. □

Proof of lemma 5.22

(i): Immediately from the definition of $B(i, m, n)$ and (B.1).

(ii): We show that as $i \to \infty$

$$B_{ll}(i, m, n) = \bar{R}_{ll}(l(i), m, n) \to R_{ll}(m, n) . \tag{B.2}$$

The limits of the other elements in $B(i, m, n)$ are derived similarly. To establish (B.2) it suffices to show that the bounds on $c_{l(i)}/c_i$, $c_{l(l(i))}/c_{l(i)}$ and $|d_{l(l(i))}|/|d_{l(i)}|$ are asymptotically tight. As $i \to \infty$, then $\beta_{l(i)} \to 0$ by the corollary 5.8. Since

$$Y_-(\beta)/\beta \to A_1 , \quad Y_+(\beta)/\beta \to A_2$$

as $\beta \to 0$, we obtain (cf. lemma 5.10)

$$\underline{C}_l(\beta_{l(i)}) \to C_l , \quad \bar{C}_l(\beta_{l(i)}) \to C_l$$

as $i \to \infty$. So the bounds on $c_{l(i)}/c_i$ are asymptotically tight. Similarly it follows that the bounds on the quotients $c_{l(l(i))}/c_{l(i)}$ and $|d_{l(l(i))}|/|d_{l(i)}|$ are asymptotically tight.

(iii) and (iv): The ratios $Y_-(\beta_{l(i)})/Y_+(\beta_{l(i)})$, $x_-(\alpha_{l(i)})/X_+(\alpha_{l(i)})$, ... appearing in the definition of $B(i, m, n)$ are positive and less than one. For example, by lemma 5.7,

$$0 < \frac{x_-(\alpha_{l(i)})}{X_+(\alpha_{l(i)})} < \frac{x_-(\alpha_0)}{X_+(\alpha_0)} = \frac{\rho\gamma_2}{2+\rho\gamma_1} < 1 .$$

Hence, $B(i, m, n)$ decreases monotonically and exponentially fast as $m \to \infty$ for fixed i and n and as $n \to \infty$ for fixed i and m. □

Samenvatting

Bij de bestudering van wachtrijsystemen speelt de analyse van het evenwichtsgedrag van Markov processen een belangrijke rol. Voor de analyse van het evenwichtsgedrag van Markov processen op een een-dimensionale toestandsruimte is reeds veel wiskundig gereedschap ontwikkeld. De situatie is geheel anders voor Markov processen op een twee-dimensionale toestandsruimte. Pas de laatste decennia is de ontwikkeling van analyse methoden voor twee-dimensionale Markov processen op gang gekomen. De meeste methoden zijn gebaseerd op de analyse van de relevante genererende functie. In het bijzonder noemen we de methode om de functionaal vergelijking voor de genererende functie te reduceren tot een standaard-type randwaarde probleem. Deze methode is toepasbaar op een algemene klasse van twee-dimensionale Markov processen. Echter, een nadeel van de genererende functie aanpak is dat deze aanpak meestal niet leidt tot expliciete resultaten voor de evenwichtskansen en dat voor numerieke berekeningen niet-triviale algoritmen nodig zijn.

Het doel van dit boek is om een bijdrage te leveren aan de ontwikkeling van methoden voor de analyse van het evenwichtsgedrag van Markov processen op een twee-dimensionale toestandsruimte. Het onderzoek is gestart met de analyse van een klassiek probleem uit de wachtrijtheorie, nl. het symmetrische kortste rij probleem. Voor dit probleem ontwikkelen we een methode die leidt tot een expliciete karakterisering van de evenwichtskansen in de vorm van reeksen van produkt-vorm oplossingen. De gedachtengang bij de ontwikkeling van deze methode wordt vrij uitvoerig besproken in hoofdstuk 1. De constructie van de reeksoplossingen is gebaseerd op een compensatie idee. Vandaar dat we deze methode 'de compensatie methode' noemen. De resultaten kunnen direct worden gebruikt voor numerieke berekeningen.

In hoofdstuk 2 wordt de compensatie methode gegeneraliseerd naar een vrij algemene klasse van Markov processen op de roosterpunten in het positieve quadrant in $\mathbb{R}^2$. We beschouwen Markov processen waarvoor de overgangsintensiteiten constant zijn in de inwendige punten van de toestandsruimte en ook constant zijn op de twee assen. Om de analyse te vereenvoudigen hebben we verder aangenomen dat de transities vanuit ieder punt beperkt zijn tot de buurpunten. We onderzoeken onder welke condities de compensatie methode werkt. Het blijkt dat de essentiele conditie is dat er vanuit inwendige punten geen transities mogen zijn naar het noorden, noord-oosten en het oosten.

In hoofdstuk 3 en 4 wordt de algemene theorie die ontwikkeld is in hoofdstuk 2 toegepast op een tweetal wachtrijproblemen. Het symmetrische kortste rij probleem wordt behandeld in hoofdstuk 3 en een model voor een multi-programmerings computer systeem wordt behandeld in hoofdstuk 4. Voor deze twee problemen worden ook extra eigenschappen afgeleid die benut

worden bij het ontwikkelen van efficiente numerieke algoritmen.

In hoofdstuk 2 is als toestandsruimte de verzameling van roosterpunten in het positieve quadrant in $\mathbb{R}^2$ gekozen. In hoofdstuk 5 wordt het asymmetrische kortste rij probleem bestudeerd. Dit probleem kan worden gemodelleerd als een Markov proces op de roosterpunten in het rechter halfvlak in $\mathbb{R}^2$. We laten zien dat de compensatie methode kan worden uitgebreid tot dit probleem en dat de gevonden oplossingen, hoewel ze ingewikkelder zijn dan die voor het symmetrische probleem, leiden tot efficiente numerieke algoritmen. Hieruit kan worden geconcludeerd dat uitbreidingen van de compensatie methode naar algemenere toestandsruimten en algemenere processtructuren zeker mogelijk zijn.

In hoofdstuk 6 worden de verkregen resulaten samengevat. Verder wordt in hoofdstuk 6 in het kort de uitbreiding van het kortste rij probleem naar Erlang-verdeelde bedieningstijden en de uitbreiding naar de $M \mid E_r \mid c$ wachtrij besproken. Voor beide problemen geldt dat de transities vanuit de inwendige punten van de toestandsruimte niet beperkt zijn tot de buurpunten.

Geconcludeerd mag worden dat de compensatie methode een krachtige techniek is waarvan alle toepassingsmogelijkheden nog zeker niet onderzocht zijn. Op dit moment wordt onderzoek gedaan naar de mogelijkheden om deze techniek verder uit te breiden naar Markov processen op meer dan twee-dimensionale toestandsruimten.

Curriculum vitae

De schrijver van dit proefschrift werd op 9 maart 1962 geboren te Roosendaal, Noord-Brabant. Van 1974 tot 1980 bezocht hij de Rijksscholengemeenschap Scheldemond te Vlissingen. Na het behalen van het Atheneum-diploma, begon hij zijn studie aan de Technische Universiteit Eindhoven in de richting wiskunde. In september 1987 werd het ingenieursexamen wiskunde behaald (met lof), met als afstudeerrichting besliskunde. Het afstudeerwerk betrof een studie naar monotonie eigenschappen in netwerken van wachtrijen. Bij dit werk werd hij begeleid door dr.ir. J. van der Wal.

Sinds oktober 1987 is de schrijver als assistent in opleiding verbonden aan de faculteit wiskunde en informatica van de Technische Universiteit Eindhoven. Dit proefschrift is een weerspiegeling van het onderzoek dat de schrijver de afgelopen vier jaar heeft verricht onder begeleiding van prof.dr. J. Wessels en prof.dr. W.H.M. Zijm.

Stelling 1.

Steutel [2] analyseert een stochastisch model voor de beweging van elektronen door een gas tussen twee elektroden. In dit model vertrekt op $t = 0$ een elektron van de kathode en begint met snelheid 1 te bewegen in de richting van de anode, die zich op afstand 1 van de kathode bevindt. Gedurende negatief exponentieel verdeelde perioden is het elektron afwisselend in beweging en wordt het vastgehouden door een stilstaand gasdeeltje. De gemiddelde bewegingsduur is λ^{-1}, de gemiddelde duur dat het elektron wordt vastgehouden is μ^{-1}. De functie $C(t)$ stelt de kans voor dat het elektron op tijdstip t in beweging is en de anode nog niet heeft bereikt. Door partiele integratie van formule (14) in [2] wordt de volgende uitdrukking voor C verkregen,

$$C(t+1) = \frac{\mu}{\lambda+\mu}\Big[1 - J(\lambda, \mu t)\Big] + \frac{\lambda}{\lambda+\mu} e^{-(\lambda+\mu)(t+1)}\Big[1 - G(\mu, \lambda t)\Big] \quad (t > 0),$$

waarin

$$J(x, y) = 1 - e^{-y}\int_0^x I_0(2\sqrt{ys})e^{-s}ds\,; \quad G(x, y) = e^{y}\int_0^x I_0(2\sqrt{ys})e^{s}ds - 1\,.$$

De methode die Goldstein [1] gebruikt om een asymptotische ontwikkeling af te leiden voor de Bessel-functie integraal $J(x, y)$ (zie [1]) voor $xy \to \infty$, kan ook gebruikt worden voor het verkrijgen van een asymptotische ontwikkeling voor $G(x, y)$ voor $xy \to \infty$. De eerste twee termen uit de ontwikkelingen voor $J(x, y)$ en $G(x, y)$ voor $xy \to \infty$ geven een benadering voor C die aanzienlijk nauwkeuriger is, met name voor zeer grote t, dan de heuristische benadering, o.g.v. de centrale limietstelling, die door Steutel [2] wordt gebruikt.

1. GOLDSTEIN, S., "On the mathematics of exchange processes in fixed columns, I. Mathematical solutions and asymptotic expansions," *Proc. Roy. Soc. A.*, vol. 219, pp. 151-171, 1953.
2. STEUTEL, F.W., "Avalanches of electrons in a gas," *J. Appl. Prob.*, vol. 23, pp. 867-879, 1986.

Stelling 2.

De evenwichtsverdeling van een $M \mid C_2 \mid 1 \mid K$ wachtrij wordt expliciet gegeven door een som van twee produkt-vorm oplossingen (vergelijk [2] en hoofdstuk 6 in dit proefschrift).

In [1] ontwikkelt Klaren een iteratieve benaderingsmethode voor de bepaling van de doorzet en de gemiddelde doorlooptijden voor een produktielijn bestaande uit drie machines. In iedere iteratie wordt de evenwichtsverdeling van een $M \mid C_2 \mid 1 \mid K$ wachtrij en die van een $C_2 \mid M \mid 1 \mid K$ wachtrij numeriek berekend. De efficientie van de benaderingsmethode kan aanzienlijk worden verbeterd door de expliciete voorstellingen voor deze evenwichtsverdelingen te gebruiken.

1. KLAREN, M., *Een analyse-methode voor produktielijnen met eindige buffers en storingen*, Afstudeerverslag, Universiteit Twente, Twente, 1987.
2. WAARSENBURG, W.A. VAN DE, *The stationary state distribution of multi-server queueing systems with Erlangian distributed service times*, Master's Thesis, Eindhoven University of Technology, Eindhoven, 1991.

Stelling 3.

Met behulp van genererende functies leidt Hofri [3] expliciete voorstellingen af voor de evenwichtskansen van de rijlengten van een eenvoudig model voor het gedrag van een computer met multiprogrammering, in de vorm van reeksen van produkt-vorm oplossingen. Deze reeksvoorstellingen zijn echter niet geheel correct in die zin dat ze niet noodzakelijk overal convergeren. Deze complicatie is door Hofri over het hoofd gezien, omdat hij een foutieve versie gebruikt van Mittag-Leffler's stelling (zie [1]). De genererende-functie aanpak is echter zeer geschikt om af te leiden wanneer de reeksvoorstellingen gelden en wanneer niet (zie [1, 2]).

1. ADAN, I.J.B.F., WESSELS, J. AND ZIJM, W.H.M., "An error note on "A generating-function analysis of multiprogramming queues"," Memorandum COSOR 90-47, Eindhoven University of Technology, Dep. of Math. and Comp. Sci., 1990.
2. ADAN, I.J.B.F., WESSELS, J. AND ZIJM, W.H.M., "Analysing multiprogramming queues by generating functions," Memorandum COSOR 91-25, Eindhoven University of Technology, Dep. of Math. and Comp. Sci., 1991, (submitted for publication).
3. HOFRI, M, "A generating-function analysis of multiprogramming queues," *International Journal of Computer and Information Sciences*, vol.7, pp. 121-155, 1978.

Stelling 4.

Bij het gebruik van de matrix-geometrische methode voor de berekening van evenwichtsverdelingen van Markovprocessen is de keuze van de partitie van de toestandsruimte een stap die van essentieel belang is voor het al of niet verkrijgen van een bruikbaar resultaat.

1. ADAN, I.J.B.F., WESSELS, J. AND ZIJM, W.H.M., "Matrix-geometric analysis of the shortest queue problem with threshold jockeying," Memorandum COSOR 91-24, Eindhoven University of Technology, Dep. of Math. and Comp. Sci., 1991.
2. NEUTS, M.F., *Matrix-geometric solutions in stochastic models*, Johns Hopkins University Press, Baltimore, 1981.
3. ZHAO, Y. AND GRASSMANN, W.K., "Solving a parallel queueing model by using modified lumpability," Research paper, Queen's University, Dep. of Math. and Stat., 1991.

Stelling 5.

Gertsbakh [2] gebruikt de matrix-geometrische methode voor de analyse van het het kortste-rij probleem met 2 parallele rijen, waarbij het is toegestaan dat een klant in de langste rij overspringt naar de kortste rij als het verschil in lengte tussen deze twee rijen een zekere drempelwaarde overschrijdt. Hij kiest de partitie van de toestandsruimte zodanig dat een zo eenvoudig mogelijke structuur aan de rand wordt verkregen. Echter, de matrix-geometrische aanpak leidt tot meer bruikbare resultaten als men zich bij de keuze van de partitie laat leiden door de speciale overgangsstructuur in het inwendige van de toestandsruimte (zie [1] en [3]).

1. 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, 1991 (to appear).
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. RAMASWAMI, V. AND LATOUCHE, W.K., "A general class of Markov processes with explicit matrix-geometric solutions," *OR Spectrum*, vol. 8, pp. 209-218, 1986.

Stelling 6.

Een $M \mid E_r \mid c$ wachtrij kan worden gemodelleerd als een kontinue-tijd Markov proces op de toestanden $(n_0, n_1, ..., n_c)$, waarin n_0 het aantal wachtende klanten is en n_i het aantal resterende bedieningsfasen voor bediende i, $i = 1, ..., c$. Voor $M \mid E_r \mid 1$ is bekend dat de evenwichtskansen van deze Markovketen kunnen worden gerepresenteerd door een lineaire combinatie van r verschillende produkten $\alpha^{n_0+n_1}$, waarin de α's worden voortgebracht door de nulpunten van een polynoom van de graad $r+1$ (zie b.v. Kleinrock [1]). De analyse van $M \mid E_r \mid c$ is niet wezenlijk moeilijker in die zin dat de evenwichtskansen nu kunnen worden gerepresenteerd door een lineaire combinatie van r^c verschillende produkten $\alpha_0^{n_0}\alpha_1^{n_1} \cdots \alpha_c^{n_c}$, waarin de α's worden voortgebracht door de nulpunten van polynomen sterk verwant aan het polynoom voor $c = 1$ (zoals onlangs is aangetoond in [2], zie ook hoofdstuk 6 in dit proefschrift). Een soortgelijk resultaat geldt ook voor de $E_k \mid E_r \mid c$ wachtrij.

1. KLEINROCK, L., *Queueing systems*, vol. 1, Wiley, New York, 1975.
2. WAARSENBURG, W.A. VAN DE, *The stationary state distribution of multi-server queueing systems with Erlangian distributed service times*, Master's Thesis, Eindhoven University of Technology, Eindhoven, 1991.

Stelling 7.

In [1] wordt bewezen dat de doorzet van een gesloten netwerk, bestaande uit multi-server stations, monotoon niet-dalend is in het aantal klanten in het netwerk. Dit monotonie resultaat kan worden uitgebreid naar een gesloten netwerk met single-server stations, waarvoor de bedieningssnelheid een monotoon niet-dalende functie is van het aantal klanten in de rij.

1. ADAN, I.J.B.F., AND WAL, J. VAN DER, "Monotonicity of the throughput of a closed queueing network in the number of jobs," *Opns. Res.*, vol. 37, pp. 953-957, 1989.

Stelling 8.

Om op een verantwoorde wijze te tillen wordt meestal het advies gegeven om door de knieen te gaan. Bij veel mensen wordt de balans bij het door de knieen gaan verstoord door een verminderde enkelbewegelijkheid. Om de balans te bewaren wordt het dan noodzakelijk om op de voorvoeten te steunen, waarbij de rug een bolle i.p.v. een holle vorm aanneemt. Tillen met een bolle rug betekent echter gevaar voor rugklachten. Dit probleem wordt opgelost door tijdens het tillen te scharnieren in de heupgewrichten i.p.v. in de onderrug. Hierbij wordt men niet gehinderd door een verminderde enkelbewegelijkheid en blijft de balans bewaard.

Het advies voor het vermijden van rugklachten bij tillen zou moeten luiden "ga door de liezen" in plaats van "ga door de knieen".

1. GEERTS, P. AND BERENDSEN, F., "Limited hip and thoracic mobility as a source of vertebral strain," *Orthopaedic Division Newsletter*, January-February, pp. 22-24, 1991.

Stelling 9.

Het normalisatieprincipe van Nirje voor de verzorging van geestelijk gehandicapten kan als volgt worden omschreven:

Streef ernaar om geestelijk gehandicapten leefpatronen en dagelijkse levensomstandigheden aan te bieden, die zo dicht mogelijk de levenswijzen en gewone omstandigheden van de maatschappij benaderen.

Het is een concept, dat niet gebaseerd is op een overdreven idealisering en praktisch zeer goed is toe te passen.

1. NIRJE, B. AND PERRIN, B., "Setting the record straight: a critique of some frequent misconceptions of the normalization principle," *Australia and New Zealand Journal of Developmental Disabilities*, vol. 11, pp. 69-74, 1985.

Stelling 10.

Vaak wordt de indruk gewekt dat bureaucraten niet streven naar zo kort mogelijke rijen, maar er juist op toezien dat de rijen zo lang mogelijk zijn. Een reden hiervoor kan zijn dat ze lange rijen niet opvatten als een teken van inefficientie, maar als een teken dat ze blijkbaar belangrijk werk verrichten.

Stelling 11.

In een ruimte waar de deur naar buiten toe opengaat, dient men ook te kloppen bij het verlaten van deze ruimte.