Queueing systems with periodic service

Queueing Systems with Periodic Service

M.J.A. van Eenige

Queueing Systems with Periodic Service

Queueing Systems with Periodic Service

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de Rector Magnificus, prof.dr. M. Rem, voor een commissie aangewezen door het College van Dekanen in het openbaar te verdedigen op dinsdag 17 september 1996 om 16.00 uur

door

Michel Johannes Alfonsus van Eenige

geboren te Voorburg

Dit proefschrift is goedgekeurd
door de promotoren:

prof.dr. J. Wessels

en

prof.dr. F.W. Steutel

Copromotor: dr.ir. J. van der Wal

Eenige, Michel Johannes Alfonsus van
Queueing Systems with Periodic Service / Michel Johannes Alfonsus van Eenige. -
Eindhoven: Eindhoven University of Technology
Thesis Technische Universiteit Eindhoven. -
With index, ref. - With summary in Dutch
ISBN 90-386-0348-7

Contents

1

Introduction

1.1 Motivation and objective

In this monograph, we study single-server multi-queue systems with periodic service. For these systems, the time axis consists of intervals of equal length, called cycles. In a cycle, the server attends the different queues to serve customers. The order in which he visits the queues is the same for all cycles. Moreover, the time instants within a cycle at which the server starts switching from one queue to another are fixed and the same for all cycles. Further, switching from one queue to another may take some time, and no customer can be served during these time periods.

As an illustration of such a queueing system, consider the case of two queues and deterministic switch-over times. In Figure 1.1, we give an example of a periodic service policy for this system. For this example, the server attends queue 1 for three time units to serve customers in this queue. After this period, he requires two time units to switch to queue 2. Then, he attends queue 2 for four time units to serve customers, after which he requires one time unit to switch back to queue 1. Finally, the service policy starts over again. So, in this example, the time interval (0, 10) can be considered as a cycle.

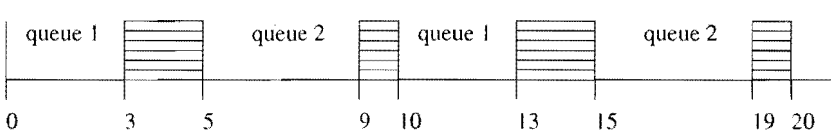

Figure 1.1: *A representation of a periodic service policy for an example of a queueing system with two queues and deterministic switch-over times.*

Various real-life situations are modelled in a natural way by queueing systems with periodic service. We give three examples of such situations. The first example is a fixed-cycle traffic light at an

intersection. The second example is a periodic access scheme to allocate the capacity of communication and computer systems. The third example is the manufacturing of products governed by a periodic production rule.

For the first example, consider a fixed-cycle traffic light to control an intersection. More precisely, for each direction, cars that approach the intersection alternately face red and green time periods of fixed duration. Clearly, this model can be regarded as a queueing system with periodic service with the traffic light as the server, the different directions as the queues, and the cars as the customers. These models were first studied in the 1950's to evaluate the throughput and delay of cars. Nowadays, fixed-cycle traffic lights particularly arise in heavy traffic and as a result of phased traffic lights.

Since the mid 1970's, queueing systems with periodic service are also used in communication and computer systems. The capacity of these systems for transmitting data or executing tasks has to be shared by different types of data or tasks. In real-time computer systems, some tasks have strict time-critical requirements. To meet these requirements, these tasks have priority, and their executions are scheduled periodically. Hence, the execution of ordinary tasks is interrupted periodically. For these systems, the quantities of interest are the probability of a buffer overflow and the probability that ordinary tasks meet their deadlines.

In the last decade, queueing systems with periodic service have been applied to model situations at production centres as well. This application is our main motivation for studying these systems. Consider a machine at a production centre, which manufactures a variety of products. Switching the production from one type of product to another may take a considerable amount of time, that is, it may cause loss of the machine's production capacity. In order to control and restrict the number of switch-overs, the production centre may apply a periodic production rule. More precisely, the centre uses production cycles of equal length. The order in which the different types of products are manufactured is the same for all production cycles. Furthermore, the machine starts switching at fixed time instants in a cycle, and these instants are the same for all cycles. A key performance measure for the centre and its customers is the delivery times of orders. From these times, other interesting performance measures, such as tardiness and the fraction of customers delivered in time, can be obtained. Tardiness is the amount of time an order is delivered too late.

Queue lengths and sojourn times are important performance measures, and many other performance measures, such as the fraction of customers served in time, can be obtained from them. With respect to the three aforementioned examples, think of the delay of cars, the buffer size and response times of tasks, and the delivery times and tardiness of orders, respectively. To evaluate the performance of queueing systems with periodic service, we need techniques to analyse these systems.

Although the server attends more than one queue, the analysis of the joint queue-length process reduces to analysing the queue-length process for each queue separately. The reason for this is that the queue-length processes do not affect each other, because of the fixed switch-over instants. Hence, for each queue, the sojourn times of customers can be determined separately as well. In spite of the fact that the analysis can be reduced to essentially that of a single-queue system, there are not many useful results known, since the analysis is mathematically hard. In Section 1.2, we shall present an overview of the literature, so that we confine ourselves here to the main conclusions.

In the literature, typical analytical approaches for studying queueing systems with periodic service are the generating-function technique, the use of Laplace-Stieltjes transforms, and, more recently, the matrix-geometric approach. Unfortunately, we face both analytical and numerical problems when applying these techniques.

For the generating-function technique, an important and well-known problem is the determination of the solutions of a characteristic equation, which is often a (possibly, high degree) polynomial equation. Further, these solutions have to be substituted into a system of regularity conditions. Since these solutions may be closely clustered, solving this nearly linearly dependent system can lead to numerical difficulties. The use of Laplace-Stieltjes transforms requires the often difficult tasks of solving integral equations and inverting these transforms for obtaining explicit results. If we want to apply the matrix-geometric approach, then we have to solve a polynomial matrix equation. Solving this equation may be rather time consuming when the matrices are large or when the utilisation factor (that is, the quotient of the average amount of work arriving at the system per time unit and the average service capacity per time unit) is close to one. Furthermore, the size of the matrices involved become quite large when this approach is applied to queueing systems with periodic service; this approach faces a 'dimensionality curse'.

Numerical techniques for the determination of the (moments of the) queue-length or sojourn-time distribution can be found in the literature as well. However, these techniques have drawbacks. Firstly, some of these techniques seem to be only applicable when the server visits each queue essentially once every cycle (as for the special example in Figure 1.1). Although this case is important, it can be too restrictive. Secondly, other techniques are rather cumbersome, since they solve the system of equilibrium equations of a periodic Markov chain, describing the queue-length process, by starting with an initial distribution and then iterating these equations to compute the stationary distribution of this chain.

Finally, some attention in the literature is focussed on deriving approximations for the average queue length or sojourn time of customers. However, information about the averages only is often insufficient for evaluating queueing systems. The reason for this is that other performance measures, like the fraction of customers served in time and tardiness, are important too.

So, both the analytical approaches and the numerical approximations, as found in the literature, may not be quite suited or may be too limited for analysing and evaluating queueing systems with periodic service. This raises the question whether there are useful techniques for analysing and evaluating these systems; in particular, whether there are techniques for determining the queue-length and/or sojourn-time *distribution* of customers. The development of such techniques is the main objective of this monograph.

To achieve this objective, we consider the queueing systems in discrete-time. The reason for this is that it reduces the complexity of the analysis considerably. Furthermore, it enables us to use probabilistic arguments to obtain the quantities of interest or approximations for these quantities. We prefer the application of probabilistic arguments to transform techniques, because we can use our intuition with reference to the problem.

In this monograph, we present two techniques for analysing the queue-length processes of customers for discrete-time queueing systems with periodic service. The results of these techniques are used to obtain the sojourn-time distribution of customers. To give a brief description of these two techniques, we recall that the queue-length processes can be studied separately. Therefore, we consider one queue only. Further, the cycles are divided into intervals of equal length, called slots.

The first technique exploits the fact that the stationary distribution of a one-dimensional Markov chain often has an asymptotically geometric tail. For stationary distributions having this tail behaviour, the parameter determining this behaviour can be computed easily and accurately. By imposing this behaviour on the stationary probabilities from a certain state onwards, we solve a finite system of equi-

librium equations of this chain to obtain estimates for the remaining stationary probabilities. To apply this technique for studying the queue-length process, we have to impose some restrictions on the arrival process and service-time distribution of customers. Fortunately, these restrictions are not too severe. For example, customers may arrive according to a periodically time-varying Bernoulli process, and the service-time distribution may be an arbitrary discrete distribution with bounded support or a mixture of a finite number of negative binomial distributions with the same parameter q. Once the stationary queue-length distribution is approximated, the sojourn-time distribution of customers is computed exactly.

This first technique uses detailed information about the service times of customers. In practice, however, one often only has the first two moments of these times (approximately). Therefore, it makes sense to study a second technique, which only uses this limited information, for the determination of the performance measures. With this second technique, we can approximate the performance measures for broader classes of queueing systems than the previous class. More specifically, we derive a periodic system of equations to evaluate the performance. This system describes the queue-length process of customers at two consecutive slots. Each of these equations is related to Lindley's equation for the $D/G/1$ queueing system with discrete service times (see, for instance, Grimmett & Stirzaker [1992]). These equations are solved by an efficient moment-iteration algorithm, which involves a novel procedure for fitting discrete distributions on the first two moments. From these approximations of the queue-length distributions, we compute the sojourn-time distribution.

These two techniques can be used to evaluate a specific queueing system with periodic service. Furthermore, we can analyse some modifications of this system by the same techniques. We consider three modifications of the periodic production rule as described by the third example. These modifications concern the production of a limited number of products to stock (think of fast movers), working overtime, and splitting the arrival process of orders into periodic arrivals with priority and random arrivals. This third modification assumes that it is possible for some customers (think of regular customers) to place orders according to a periodic pattern.

In Section 1.2, we give an overview of the literature related to the analysis of queueing systems with periodic service. The organisation of this monograph will be given in Section 1.3.

1.2 Overview of related literature

In this section, we give an overview of the literature related to the analysis of queueing systems with periodic service. Analytical techniques for such systems are discussed in Section 1.2.1, together with the applicability of these techniques in relation to that of the numerical technique exploiting the geometric tail behaviour (the GT technique, for short). Section 1.2.2 is devoted to approximation techniques. The main conclusions of Sections 1.2.1 and 1.2.2 are listed in Section 1.2.3.

To facilitate the exposition of this overview, we introduce some definitions to obtain a unified terminology instead of the diverse terminology used in specific applications of these queueing systems. As explained earlier, we consider one queue only. The time periods during which the server is attending and not attending this queue are called on-periods and off-periods for this queue. It is assumed that a cycle starts with an off-period and ends with an on-period. Unless stated otherwise, in this section, we assume that the lengths of the on-periods and of the off-periods are both constant, in which case a cycle consists of one on- and off-period. The resulting queueing system constitutes the basic system

considered in this section, as this is the simplest and most studied system with periodic service. In Figure 1.2, we graphically represent these assumptions. Finally, the arrival patterns of customers are assumed to be stochastically identical and independent for each cycle.

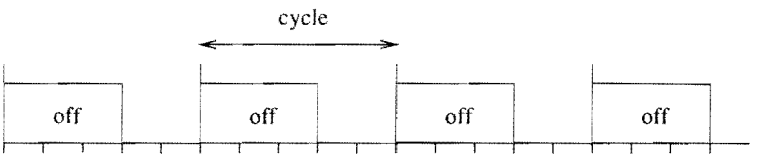

Figure 1.2: *A representation of the basic queueing system considered in Section 1.2.*

1.2.1 Analytical techniques

The major part of the literature treating analytical techniques is devoted to the case that the service of customers can start at certain discrete time instants in an on-period only. Clearly, this holds for discrete-time queueing systems, which naturally arise in modelling of, for instance, communication and computer systems. Moreover, in the study of fixed-cycle traffic-light queues, it appears to be fairly reasonable to assume that cars in a queue pass the traffic light at equally spaced time instants in a green period, and that cars that do not have to queue during a green period experience no delay at all. So, cars in a queue can effectively start with passing the traffic light at certain discrete time instants only. The assumption that customers arriving in an on-period at an empty queue experience no delay is called the traffic-light queue assumption (or, in brief, the TLQ assumption). We remark that arrivals do not have to coincide with the time instants at which a service may begin. A minor part of the literature is concerned with the case that services can begin at arbitrary time instants in an on-period.

For the analytical techniques, we discuss the following issues: main results, problems associated with the implementation of these techniques, approximations found in the literature to avoid these problems, and the relation between the applicability of these techniques and the GT technique. We begin with discussing the analytical techniques for the case that services can begin at discrete time instants only and then treat the case that these instants are arbitrary.

1.2.1.1 Services start at discrete time instants

For the case that services can only begin at discrete time instants, we introduce the following conventions. We discretise the time axis by dividing a cycle into intervals of equal length, and assume that the service times are multiples of these intervals. Such an interval is called a slot. In Figure 1.2, we have subdivided a cycle into five slots.

In this case, three typical techniques for analysing queueing systems with periodic service are the generating-function technique, reducing the analysis to the study of a certain $GI/G/1$ queueing system, and the matrix-geometric approach. These techniques are successively discussed. We conclude this section with mentioning some results for specific queueing systems.

For discrete-time queueing systems with periodic service, the generating-function technique is the technique most frequently used for analysing the queue-length process at the start of cycles. This process is called the imbedded queue-length process. The generating-function technique is also used to

determine the waiting-time or sojourn-time distribution of customers. Applying this technique yields the probability generating function of the stationary imbedded queue-length distribution and that of the waiting-time or sojourn-time distribution, respectively.

For service times equal to one slot, studies of the imbedded queue-length process by this technique have been presented by numerous authors. For instance, Newell [1960] and Meissl [1962] study this process by the generating-function technique for a Bernoulli arrival process; Meissl [1963], Darroch [1964], Anderson, Foschini & Gopinath [1979], Rubin [1979a,1979b], Rubin & Zhang [1988], Steyaert & Bruneel [1991], and Heidemann [1994] for a generally distributed number of arrivals in a slot. In addition, Meissl [1962], Rubin [1979a,1979b], Rubin & Zhang (for several arrival processes), and Steyaert & Bruneel apply this technique for determining the probability generating function of the waiting-time or sojourn-time distribution. We remark that the generating-function technique can also be used to study the transient queue-length process (see, for example, De Smit [1971]). Moreover, we note that Anderson, Foschini & Gopinath even allow for multiple (and different) on- and off-periods in a cycle. Further, we mention that Kaplan [1983] assumes a (time-varying) Poisson arrival process of customers with generally distributed discrete service times. By considering the off-periods as the service times of periodically arriving customers, Kaplan applies the generating-function technique to determine the probability generating function of the virtual waiting time (that is, the amount of work in the system).

In general, the determination of the distributions from probability generating functions leads to numerical problems. First of all, we have to calculate the roots of a characteristic equation. The determination of these roots is often difficult (see, for instance, Kleinrock [1975]). In fact, the accurate computation of these roots seems to be possible only if the parameters of the model (like the number of slots in a cycle) are small or if the characteristic equation possesses a special structure (see, for instance, Newell [1960], Meissl [1962], and Adan & Zhao [1994]). However, even if we can compute all roots accurately, then it is common that these roots are closely clustered. These roots have to be substituted into a system of regularity conditions, so that this system is nearly singular and solving it is quite delicate. We shall show in Chapter 2 that these problems occur already for small cycle lengths in queueing systems with periodic service. Further, the determination of the moments of the distributions from their probability generating function generally requires the roots of the characteristic equation and the solutions to the regularity conditions as well. Only if the length of the on-period is equal to one slot, we do not need these roots and the solutions to the regularity conditions for the calculation of these moments.

Because of these problems, the generating-function technique is often only used to evaluate special queueing systems, or to derive bounds and approximations for the performance measures. Meissl [1962,1963] and Anderson, Foschini & Gopinath [1979] present some numerical results for special queueing systems. Newell [1960] gives an approximation for the average sojourn time, but he does not evaluate the quality of this approximation. Darroch [1964] derives upper and lower bounds for the average sojourn time, but the difference between these bounds is rather large. Rubin & Zhang [1988] and Steyaert & Bruneel [1991] also derive upper and lower bounds, and these bounds can be used to obtain a rough estimate for the average sojourn time. Further, under the TLQ assumption, Ohno [1978] obtains an expression for the average sojourn time for Poisson arrivals. This expression contains the average number of customers at the start of a cycle in terms of the roots of an equation. In order to compute the average sojourn time, he uses the approximations for this average number of customers as derived by Miller [1963], Newell [1965], and McNeil [1968] (we discuss these approximations in Section 1.2.2). Ohno's approximations perform quite well.

So, a fruitful application of the generating-function technique seems to be limited to very special queueing systems with periodic service. As we shall see, the geometric tail (GT) technique is applicable to a broad class of queueing systems with periodic service. For instance, this technique can be applied in the case of multiple on- and off-periods in a cycle, and for many important arrival processes and service-time distributions. Other important situations can be very well approximated by these arrival processes and service-time distributions.

The second analytical technique employed in the literature is based on the observation that the determination of the performance measures reduces to the study of a certain $GI/G/1$ queueing system or of a queueing system with bulk service. This reduction implies that Lindley's integral equation has to be solved (cf. Lindley [1952]). The solution of this equation is the same as the limiting solution for n tending to infinity of the recurrence relation

$$W_{n+1} = \max\{0, W_n + B_n - A_n\}, \tag{1.1}$$

where W_n is the quantity of interest; A_n and B_n are independent random variables with known distributions and independent of W_n. This recurrence relation is also known as Lindley's equation (cf. Grimmett & Stirzaker [1992]). Due to the periodic nature of the queueing system, the random variables A_n and B_n may be periodic.

Beckmann, McGuire & Winsten [1956] study the same model as Newell [1960], and they express the average sojourn time in terms of the average queue length at the start of a cycle. They observe that this latter quantity is equal to the average number of customers in a queueing system with bulk service. This number of customers can be represented by Lindley's equation for a $D/G/1$ queueing system with binomially distributed service times. Chu & Konheim [1972] consider the restrictive case that the length of an on-period is equal to one slot, and they allow for a general distributed number of slots of work arriving in a slot. They observe that the queue-length process can be related to a $D/G/1$ queueing system. To derive the first two moments of the queue-length distribution, Chu & Konheim require the probability of an empty queue at the start of a cycle. They determine this probability by studying the gambler's ruin problem (see also, for instance, Feller [1968]). Kosovych [1978] considers the case of geometrically distributed batches that arrive according to a Poisson process. He relates this model to the analysis of a queueing system with bulk service and expresses the average sojourn time in terms of the roots of an equation. Fredericks [1979] and Fredericks, Farrell & DeMaio [1985] suppose that the slots in an off-period correspond to the service times of higher priority customers. These authors study the case that ordinary customers can only arrive in one fixed slot of the cycle. The waiting-time process for ordinary customers is then described by a $D/G/1$ queueing system with interarrival times equal to the length of an on-period. Ackroyd [1985] assumes that customers arrive at the start of slots. Furthermore, the slots in an off-period are considered as the service times of scheduled customers having preemptive priority. He relates the amount of work (including that of scheduled customers) in two consecutive slots, yielding a periodic system of Lindley's equations (that is, A_n and B_n in (1.1) are periodic).

However, the limiting solution for n tending to infinity of Lindley's equation cannot be determined analytically except in a few special cases, let alone the periodic solution for a periodic system of these equations. Therefore, several approximations have been proposed. For instance, Fredericks [1979] and Fredericks, Farrell & DeMaio [1985] exploit the fact that, under rather mild conditions, the tail of the waiting-time distribution is asymptotically exponential (see Feller [1971]). This procedure, described in detail in Fredericks [1982], gives reasonable approximations. Ackroyd [1984] suggests to

truncate the state space and to solve the reduced system of the equations in Ackroyd [1985] by Levinson's method (see Levinson [1946] and Robinson [1967]), because this reduced system has a block Töplitz form. In Ackroyd [1985], the author iterates the periodic system of Lindley's equations until the solution converges to a periodic stationary distribution.

The queueing system of Fredericks [1979] and Fredericks, Farrell & DeMaio [1985] is rather restrictive, mainly because customers can arrive in exactly one fixed slot in the cycle. These authors allow, however, more general service times than we allow for the GT technique. At the cost of slightly less general service times, this technique allows us to analyse their queueing system, extended to the case where customers can arrive in each slot of the cycle. Ackroyd [1985] also allows more general service-time distributions. By making the same slight reduction in generality, his system can be analysed by the GT technique too. This technique is more efficient than his iteration method. Furthermore, it is likely that, to obtain the same accuracy, the GT technique may reduce the denumerable system of equilibrium equations more than simple truncation as employed in Ackroyd [1984]. The reason for this is that the GT technique exploits the geometric tail behaviour of the stationary (imbedded) queue-length distribution.

The third analytical approach is the matrix-analytic approach. This approach is extensively discussed in Neuts [1981,1989], and it can be used to determine the stationary distribution of multi-dimensional Markov chains with a discrete state space, which is infinitely large in at most one dimension. For a discrete-time queueing system with a Markovian arrival process, service times equal to one slot, and multiple on- and off-periods in a cycle, Alfa & Neuts [1995] use the matrix-geometric approach of Neuts [1981] to study the queue-length process. We mention that this system is more general than the one we consider for the GT technique, because Alfa & Neuts allow for an arrival process which may not be stochastically identical and independent for each cycle (although this dependency is of a special form).

To apply the matrix-geometric approach, Alfa & Neuts [1995] have to solve a quadratic matrix equation first, and then a system of linear boundary equations. The solution of the matrix equation can in general not be obtained analytically, but it has to be determined by, for instance, successive substitutions. However, if the size of the matrices involved becomes large (due to, for instance, lengthy cycles), if the degree of the polynomial is high, or if the utilisation is close to one, then the determination of this solution is rather time consuming. We finally remark that the matrix-geometric approach can be applied to more general phase-type service-time distributions than we consider for the GT technique. This generalisation, however, increases the computational effort (considerably), because the size of the matrix equation becomes larger and because this equation may not be quadratic, but a higher degree polynomial equation.

To conclude this section, we mention four other results derived by analytical techniques, for the case of a Poisson arrival process. Haight [1959] and Buckley & Wheeler [1964] assume that the service times are equal to one slot, and these authors make the TLQ assumption. Haight obtains the conditional probability of the number of customers at the start of an cycle, given this number at the start of the preceding cycle, in terms of Poisson and Borel-Tanner distributions. Buckley & Wheeler determine the Laplace-Stieltjes transform of the joint distribution of the number of delayed customers and the total delay, and the transforms for the marginal probability distributions. The use of these two models is, due to the TLQ assumption, rather limited, because this assumption is certainly not valid for computer and production systems. Lam [1977] and De Moraes & Rubin [1984] consider the case that on-periods

are equal to one slot. Lam assumes that the service times are also equal to one slot, and he redefines the service time of a customer as the time that this customer is at the head of the queue. Then, his system corresponds to an $M/G/1$ queueing system with an exceptional first service, so that standard results can be used to derive the probability generating function of the number of customers in the system. The moments of the corresponding distribution can easily be obtained. De Moraes & Rubin assume that the service times are a generally distributed number of slots. They transform the service times of customers into the number of cycles customers are at the head of the queue. Moreover, they introduce low priority customers. These low priority customers arrive according to a Poisson process with such a rate that the server is (almost) never idle, and these customers have service times equal to one cycle. By relating this model to an $M/G/1$ queueing system with non-preemptive priority, De Moraes & Rubin derive the Laplace-Stieltjes transform for the waiting-time distribution. The utility of these latter two queueing systems is, of course, rather limited due to the fact that the length of an on-period is equal to one slot. Furthermore, the relation with standard $M/G/1$ queueing systems seems to be neither applicable to the case that on-periods consist of more than one slot, nor to multiple on- and off-periods in a cycle, nor to the modifications mentioned in Section 1.1.

1.2.1.2 Services start at arbitrary time instants

In this section, we discuss four techniques for studying queueing systems with periodic service, when services can start at arbitrary time instants. These techniques involve the determination of the performance measures of interest by using a system of differential or integral equations, reducing the analysis to that of a certain $GI/G/1$ queueing system, the use of Laplace-Stieltjes transforms, and a decomposition of the queueing system.

The first technique to be discussed is the use of differential or integral equations relating the quantities of interest. Newell [1956] considers the case that the service times are deterministic and that the interarrival times of customers, which have a general distribution, are at least equal to a service time. Furthermore, if the queue is empty in an on-period, then it remains empty during the remaining part of the on-period (in other words, Newell makes the TLQ assumption). To determine the sojourn-time distribution, he derives a system of integral equations. Robillard & Naor [1968] study the case that customers arrive according to a Poisson process and have exponentially distributed service times. By using differential equations for the queue-length distribution, they derive an expression for the average queue length in terms of the average queue length at the start of a cycle.

The system of integral equations in Newell [1956] can in general not be solved analytically. Therefore, he suggests to solve this system approximately by successive substitutions, provided that the average queue length at the start of an off-period is small compared to the average queue length at the start of an on-period. Robillard & Naor [1968] determine approximations for the unknown quantity for the case that the lengths of the on- and off-periods are short and for the case that these lengths are long, but they present no numerical results.

The main drawback of the queueing system of Newell [1956] is the simplifying TLQ assumption. A drawback of the analysis in Robillard & Naor [1968] is that this analysis is exclusively focussed on the average queue length as performance measure. By approximating these queueing systems, we can use the GT technique or the moment-iteration technique to obtain approximations of the performance measures for these systems.

As in Section 1.2.1.1, the analysis of some systems reduces to the study of Lindley's (integral) equation. Mine & Ohno [1971] consider a general interarrival-time distribution, and they assume that a service time consists of a lost time and an actual service time, which have general distributions. Under the TLQ assumption, the virtual waiting-time process at the start of cycles can be described by Lindley's integral equation for the following $GI/G/1$ queueing system. The interarrival times of customers are equal to the length of an on-period minus the total lost time in a cycle, and the service times are equal to the total actual service time of customers arriving in this cycle.

The Lindley's integral equation Mine & Ohno [1971] obtain seems to be solvable in a few exceptional cases only. They suggest to solve this equation by successive substitutions for the probability distribution of the virtual waiting time, but they present no numerical results. Despite the generality of the interarrival-time and service-time distributions, the system of Mine & Ohno is only reasonable for studying traffic-light queues. Furthermore, it seems that their technique cannot cope with multiple on- and off-periods in a cycle. In the same way as for the previous two models, we can derive approximations for the performance measures by the two techniques to be discussed in this monograph.

Laplace-Stieltjes transforms are widely and fruitfully applied in the analysis of continuous-time queueing systems (see, for instance, Prabhu [1965] and Cohen [1982]). For a Poisson arrival process of customers with generally distributed service times, Şahin & Bhat [1971] and Schassberger [1974] use these transforms to analyse queueing systems with periodic service. These authors consider the off-periods as service times of periodically arriving customers. Şahin & Bhat study the virtual waiting-time process (including the work of periodic customers). They exploit the circumstance that this process is identical to the transient virtual waiting-time process in an $M/G/1$ queueing system during on-periods. To derive the Laplace-Stieltjes transform of the virtual waiting time, they have to solve an integral equation. This equation is formally solved by a Wiener-Hopf decomposition. Schassberger considers the periodic customers as customers with preemptive priority. Using Laplace-Stieltjes transforms, he derives a system of functional equations, which constitutes a Hilbert problem. He presents a formal solution to this problem, from which performance measures can in principle be obtained.

The use of Laplace-Stieltjes transforms involves the solution of a (system of) integral equation(s). Unfortunately, it does not seem possible to determine this solution explicitly. If the length of the off-period has a rational Laplace-Stieltjes transform, then Şahin & Bhat [1971] can compute the moments of the virtual waiting time in terms of the roots of an equation. As we have already mentioned, it is in general quite hard to determine these roots. Further, it seems that these techniques do not lend themselves to analysing the modifications.

Ott [1987b] studies the same queueing system as Şahin & Bhat [1971]. He shows that the virtual waiting time consists of the virtual waiting time in an $M/G/1$ queueing system (as if there are no off-periods) plus the virtual waiting time in a $D/G/1$ queueing system, with interarrival times equal to the length of a cycle and service times equal to the time to empty an $M/G/1$ queueing system which begins with an initial amount of work equal to the length of an off-period. So, the analysis reduces to that of a $D/G/1$ queueing system and, hence, similar problems arise as in solving Lindley's (integral) equation as mentioned in Section 1.2.1.1. Ott [1987b] presents, however, a numerical approach based on Ott [1987a]. Numerical examples in Ott [1987a,1987b] indicate that this approach works well. Further, Sengupta [1990] uses the decomposition of Ott [1987b] to approximate the average waiting time for independent generally distributed lengths of the on- and off-periods. Note that if the lengths of the on- and off-period are not constant, then service is not offered periodically in his model. For the case

of on- and off-periods of fixed duration, however, he shows no numerical results. This kind of decompositions is well known to exist for $M/G/1$ queueing systems with so-called server vacations (see, for instance, Fuhrmann & Cooper [1985], Shanthikumar [1988], and Zhang, Vickson & Van Eenige [1995]), or for systems with (possibly, non-Markovian) additional inputs (see Ott [1984]).

Unfortunately, this decomposition into $M/G/1$ and other single-server queueing systems does not seem to hold for multiple on- and off-periods in a cycle. Moreover, we see no opportunity to apply this decomposition technique to study the interesting modifications mentioned in Section 1.1.

1.2.2 Approximation techniques

In Section 1.2.1, we discussed analytical techniques; almost all of them leading to mathematical difficulties. Because of these difficulties, some attention in the literature is focussed on the derivation of approximations. In this section, we discuss some of these approximations. Firstly, we discuss approximations that result from analysing fixed-cycle traffic-light queues. Mainly because of the TLQ assumption, the resulting expressions are not useful in the study of the queueing systems arising in other applications. Secondly, we present approximations resulting from analysing communication and computer systems. Thirdly, we give an approximation used to study a periodic production rule. Finally, we mention an approximation used to derive optimal periodic service policies. The last three approximations yield insight into the performance of queueing systems with periodic service, but they do not seem to be extendible to the modifications.

For the results obtained in studies of fixed-cycle traffic-light queues, we mention three results. Firstly, we present results based on fluid approximations, secondly, approximations obtained by fitting an expression to simulated data, and, finally, results which arise from approximating the average queue length at the start of a cycle.

Wardrop [1952] and Newell [1965] analyse queueing systems arising in the study of fixed-cycle traffic lights, and they make the common TLQ assumption. Wardrop considers the case of Poisson arrivals and deterministic service times. Newell allows for general interarrival-time and service-time distributions. By considering customers as a continuous fluid, both Wardrop and Newell derive approximations for the average sojourn time. The approximation of Wardrop tends to underestimate the average sojourn time. Using numerical examples, Newell compares his approximations with those of Webster [1958] (to be discussed next) and shows that his best approximation is within 5% of Webster's approximation.

Webster [1958] studies the same queueing system as Wardrop [1952]. Like Wardrop and Newell [1965], Webster gives an (approximate) expression for the average sojourn time. This expression consists of three terms, namely, the expression of Wardrop, the average sojourn time in an $M/D/1$ queueing system with adjusted service time, and a correction term obtained by fitting simulated data. The expression of Webster compares fairly well with simulations.

Miller [1963], McNeil [1968], and Cowan [1981] also study fixed-cycle traffic-light queues. These authors assume deterministic service times and make the TLQ assumption. To derive an expression for the average sojourn time, these authors need the average queue length at the start of cycles. Miller and McNeil use Lindley's equation for this imbedded queue-length process, while this equation does not really apply. Both authors obtain an expression for the average sojourn time: Miller for the case of a general arrival process and McNeil for a compound Poisson arrival process. In case of Poisson arrivals, numerical results show that both expressions yield approximations close to those of Webster

[1958]. Cowan considers bursty arrivals. More precisely, the interarrival time of customers within a burst is supposed to be equal to one service time and the interarrival time between two successive bursts is assumed to be one service time plus a negative exponentially distributed amount of time. Cowan determines the average number of customers by evaluating a transient Markov chain iteratively, but he does not present any numerical results.

Mainly because of the TLQ assumption, these approximations are not useful in the general setting of this monograph.

In the study of communication and computer systems that allocate their capacity in a periodic fashion, Fischer [1977a,1977b], Ko & Davis [1984], Bruneel [1986], and Keilson & Servi [1990] determine expressions for the average waiting time or sojourn time of customers. For a Poisson arrival process and generally distributed service times, Fischer uses a diffusion approximation to derive a closed-form expression for the (approximate) average waiting time of a customer. This expression gives fairly good results compared to simulation results. Ko & Davis assume a Poisson arrival process and deterministic service times. They use Lindley's equation to describe the queue-length process at the start of on-periods approximately and apply the generating-function technique. To compute the average sojourn time, they obviously face the aforementioned numerical problems arising in this technique. Bruneel generalises the model of Ko & Davis to general arrival processes, and he applies the same technique. Keilson & Servi study the same model as Fischer and assume that the off-periods correspond to periodically arriving customers who have preemptive priority. The work-load process of these periodic customers is approximated by a limit of compound Poisson distributions. For this system, they give an expression for the average and the variance of the sojourn time of ordinary customers. Numerical results are in line with simulated results.

In the context of production centres, Federgruen & Green [1986] and Dellaert [1988] evaluate periodic production rules for the case that customers arrive according to a Poisson process. Federgruen & Green derive bounds and approximations for the average waiting time, the probability of waiting, and the queue-length distribution for general service-time distributions. Simulation results show that the approximations for the average waiting time and the probability of waiting are fairly accurate. Dellaert considers the model of Robillard & Naor [1968], that is, Poisson arrivals and exponentially distributed service times. He decomposes the sojourn time into three parts, namely, the time until the next on-period (if a customer arrives during an off-period), the time to serve the customer and the customers in front of him, and the lengths of the off-periods occurring during this service. From this decomposition, he derives an approximation for the average sojourn time in terms of the average number of customers at the start of cycles. This latter quantity can be studied by means of the transient queue-length process in an $M/M/1$ queueing system. Since Dellaert is interested in minimising costs, he needs a nicer expression for this quantity. He gives two expressions, namely, a shadow approximation and a weighted average of the shadow approximation and the average number of customer waiting in an $M/M/1$ queueing system. The shadow approximation is the average number of customers in an $M/M/1$ queueing system with the same arrival rate as for the original system, but with the mean service times adjusted such that the utilisation of this system is equal to that of the original system. Numerical results show that the shadow approximation performs very poorly and that the weighted average is within 10% of the exact values.

Finally, Borst, Boxma, Harink & Huitema [1994] look for optimal rules for polling systems according

to a fixed-time polling scheme. In addition, they consider gated service policies. More specifically, they assume that only those customers are served in an on-period that were in the system at the start of this on-period. Furthermore, they suppose that all these customers are served within this on-period. They decompose the waiting time of a customer in two parts. The first part is the time period between the arrival instant of the customer and the start of the first on-period after this arrival. The second part is the time period from the start of this on-period until the time instant at which the service of this customer begins. By approximating this second part, they obtain an approximation for the average waiting time.

1.2.3 Conclusions

In the literature, many techniques and approximations have been used and proposed to study queueing systems with periodic service. Most of the analytical techniques are applicable to a limited class of these systems only, or they yield results that are merely interesting from a theoretical point of view. Exceptions are the iteration method in Ackroyd [1985] and the matrix-geometric approach used in Alfa & Neuts [1995]. The drawbacks of these latter two approaches are related to their computational effort. Further, many analytical techniques do not seem to be suitable for studying modifications of these systems or extensions to multiple on- and off-periods in a cycle.

An important part of the approximations is restricted to the study of systems with the TLQ assumption. These systems are too restrictive to act as reasonable models for analysing computer or production systems. Moreover, the approximations are mainly focussed on the average queue length or sojourn time, which is often insufficient.

The GT technique and the moment-iteration technique to be developed in this monograph are applicable to an important and broad class of queueing systems with periodic service and some interesting modifications. These techniques are also efficient as compared to the ones discussed in Sections 1.2.1 and 1.2.2.

To conclude this section, we briefly mention three other systems that are related to queueing systems with periodic service. Firstly, the off-periods as faced by customers of a certain queue can be regarded as server vacation periods. So, queueing systems with periodic service can be viewed as special queueing systems with server vacations. For a survey on these latter systems, we refer to Doshi [1986,1990]. Secondly, the off-periods may correspond to time periods of scheduled maintenance, so that queueing systems with periodic service can be viewed as special preventive maintenance models (see, for instance, Valdez-Flores & Feldman [1989] for a survey on such models). Finally, as pointed out in Sections 1.2.1 and 1.2.2, the off-periods can be viewed as service times of higher priority customers. So, queueing systems with periodic service can be regarded as special priority systems.

1.3 Outline of the monograph

This monograph is concerned with the development of two techniques for the determination of the stationary queue-length distribution in queueing systems with periodic service. Furthermore, an algorithm is developed to compute the sojourn-time distribution of customers, given the stationary queue-length distribution.

To illustrate the techniques for analysing the queue-length process, in Chapter 2, we shall analyse this process in one of the queues for a queueing system with periodic service. The sojourn time of

2

A Queueing System with Periodic Service

2.1 Introduction

In this chapter, we first demonstrate two kinds of numerical problems that can occur when analytical techniques are used for determining the stationary distribution of Markov chains. These problems concern the determination of the roots of a characteristic equation and, if all roots have been computed accurately, the clustering of these roots; substituting them into a system of regularity conditions leads to a nearly linearly dependent system. After that, we present the main ideas of the two numerical techniques, which will be explored further in this monograph. For the examples we consider, it appears that these two techniques are numerically stable.

As an illustration, we consider a discrete-time queueing system with periodic service. As mentioned in Chapter 1, the queue-length processes of customers do not affect each other, so that they can be analysed separately. Therefore, in this chapter, we consider customers of one queue only. Customers are supposed to arrive according to a Bernoulli process, and they are assumed to have deterministic service times of one time unit. The time periods during which the server is attending (on-periods) and not attending (off-periods) this queue are assumed to be of fixed duration. These assumptions imply that, once the amount of work upon arrival is known, the sojourn time of a customer is deterministic. So, it suffices to study the queue-length process only.

We note that this queueing system is the same one as previously introduced by Beckmann, McGuire & Winsten [1956], Newell [1960], and Meissl [1962]. These authors use this model for analysing the delay of cars at an intersection which is governed by a fixed-cycle traffic light.

The number of customers at the start of off-periods can be described by a one-dimensional Markov chain. The equilibrium equations of this chain constitute a homogeneous linear difference equation of finite order with constant coefficients. So, the solution of these equations is a linear combination of powers. Two standard analytical techniques to determine this linear combination are the generating-function technique (as used by the above authors) and an approach (which is called the method of particular solutions in Feller [1968]) that directly seeks a linear combination of powers satisfying this equation. Both techniques face the aforementioned numerical difficulties.

The first numerical technique to be discussed in this chapter exploits the fact that the tail of the stationary distribution of the Markov chain is asymptotically geometric. This tail behaviour will be utilised in the same way as in Tijms & Van de Coevering [1991]. The second numerical technique exploits the circumstance that the Markov chain corresponds to the waiting-time process in a $D/G/1$ queueing system with binomially distributed service times. This waiting-time process (and, hence, the queue-length process) is studied by a discrete version of the moment-iteration method described in De Kok [1989]. This technique uses the first two moments of the service-time distribution only, and, moreover, it uses a novel procedure for fitting discrete distributions on the first two moments.

The outline of this chapter is as follows. In Section 2.2, we present the queueing system with periodic service in detail and introduce the Markov chain describing the number of customers at the start of off-periods. In Section 2.3, the method of particular solutions and the generating-function technique are applied to determine the stationary distribution of this chain, and their numerical instabilities are illustrated. The numerical technique exploiting the geometric tail behaviour (henceforth, abbreviated as GT technique) is discussed in Section 2.4. The moment-iteration technique (or, in brief, the MI technique) is the topic of Section 2.5. Finally, the main conclusions of this chapter are summarised in Section 2.6.

2.2 The model and the imbedded queue-length process

We consider a single-server queueing system in discrete time by dividing the time axis into intervals of equal length. Such an interval is called a slot. Customers are supposed to arrive according to a Bernoulli process with parameter p. This means that in each slot exactly one customer arrives with probability p, and no customer with probability $1 - p$. In other words, the interarrival times have a geometric distribution with parameter p. The service times of customers are deterministic and equal to one slot.

The server renders service periodically: there is service during on-periods and no service during off-periods. The length of the off-periods and of the on-periods are both assumed to be constant. An off-period and the next on-period together are called a cycle. Hence, the length of a cycle is constant. The number of slots in a cycle is denoted by C and the length of an on-period is equal to A slots, with A and C non-negative integers. The slots in a cycle are numbered $1, 2, \ldots, C$. During a slot in an on-period (an on-slot), the server is either idle or rendering service to a customer. In Figure 2.1, we give a representation of the on- and off-periods as faced by the customers.

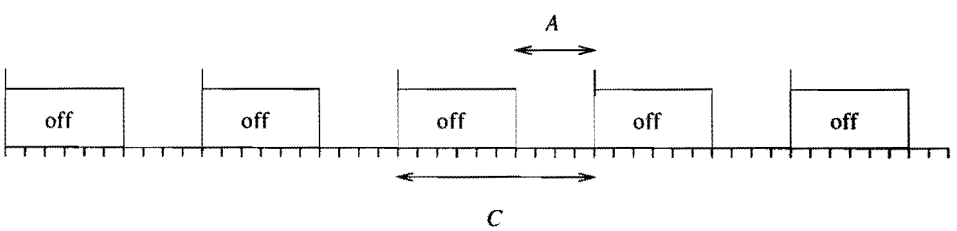

Figure 2.1: *The on- and off-periods for the customers.*

Customers are served in the order of their arrival. Further, we assume that $Cp < A$, so that the av-

erage number of slots of work (that is, the average number of customers) arriving in a cycle is strictly less than the service capacity of the server per cycle. Observe that the queue length cannot increase in an on-period, because in each slot at most one customer can arrive and the service time of a customer is exactly equal to one slot. This observation implies that the traffic-light queue assumption (as introduced in Section 1.2.1) is implicitly made.

Our purposes are the determining of the stationary queue-length distributions at the slot boundaries in the cycle and of the sojourn-time distribution of customers. If the number of customers upon arrival is known, then the sojourn time of an arriving customer is deterministic, because the service times as well as the lengths of the on- and off-periods are deterministic. This observation implies that the determination of the sojourn-time distribution reduces to the determination of the queue-length distribution at arrival instants. Before analysing the queue-length process, we introduce some conventions.

Customer arrivals as well as the start and completion of the service of a customer occur at slot boundaries. For convenience, we assume that arrivals and the start of the service of a customer occur *just after* slot boundaries, and that service completions (that is, customer departures) occur *just before* slot boundaries, see Figure 2.2. In the sequel, a customer departing at the n-th slot boundary in the cycle (that is, the boundary between slot $n-1$ and slot n) is said to depart in slot $n-1$, and a customer arriving at this slot boundary is said to arrive in slot n. Finally, the service of a customer arriving at an empty queue during an on-period (that is, arriving when the server is idle) begins immediately.

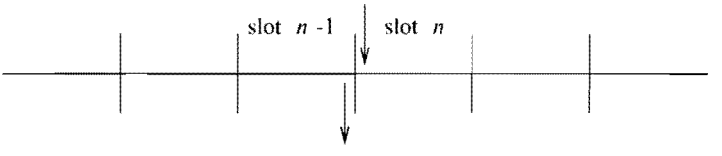

Figure 2.2: *A customer departure in slot $n-1$, and a customer arrival in slot n.*

For the analysis of the queue-length process, we consider the number of customers at the start of cycles. More precisely, we consider the number of customers at the first slot boundary of cycles. So, we look at the system just after a possible departure in the last slot of the preceding cycle, but just before a possible arrival in the first slot of the forthcoming cycle. Let X_k denote the number of customers at the start of the k-th cycle, see Figure 2.3.

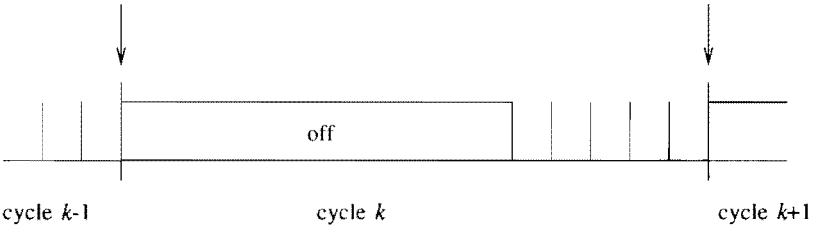

Figure 2.3: *The imbedded time epochs.*

The stochastic process $\{X_k, k = 1, 2, 3, \ldots\}$ is a discrete-time Markov chain with state space the set of non-negative integers $\{0, 1, 2, \ldots\}$ and $X_1 = i$, where i is a possibly random non-negative integer. This process is also called the imbedded queue-length process. The Markov chain is easily seen to be

irreducible and aperiodic, and since the average number of slots of work arriving per cycle is strictly less than the service capacity per cycle (that is, $Cp < A$), the chain is ergodic (cf. Pakes [1969]). Hence, this chain has a unique stationary distribution $\{\pi_j, j = 0, 1, 2, \ldots\}$.

To determine this stationary distribution, we use the recurrence relation

$$X_{k+1} = \max\{0, X_k + N_k - A\}, \tag{2.1}$$

where N_k is the number of arrivals in the k-th cycle. This relation can be derived as follows. As mentioned earlier, the number of customers in the system cannot increase during an on-period. From this observation, X_{k+1} is equal to the maximum of the number of customers at the end of the off-period in the k-th cycle plus the number of customers arriving during the on-period in the k-th cycle minus the service capacity A per cycle, and zero. Since the number of customers at the end of the off-period in the k-th cycle is equal to the sum of X_k and the number of customers arriving during this off-period, relation (2.1) is an immediate consequence. This relation has already been pointed out by Beckmann, McGuire & Winsten [1956] in their study of a fixed-cycle traffic-light queue. Furthermore, this relation naturally appears in the study of queueing systems with bulk service (see, for instance, Bailey [1954b], and Downton [1955,1956]) and in the analysis of a periodic-review (R, S) inventory system (cf. De Kok [1989]).

In Section 2.5, we propose a moment-iteration technique for approximating the stationary distribution of the Markov chain. This technique exploits the fact that relation (2.1) corresponds to Lindley's equation for the waiting time X_{k+1} of the $(k+1)$-st arriving customer in a $D/G/1$ queueing system. In the next section, we present two standard analytical techniques for the determination of this distribution, and we demonstrate that applying these techniques leads to numerical problems, even for relatively small C and A. In Section 2.4, we utilise the fact that the stationary distribution has an asymptotically geometric tail.

2.3 Two analytical techniques

The stationary distribution $\{\pi_j, j = 0, 1, 2, \ldots\}$ is the unique normalised solution of the system of equilibrium equations of the Markov chain. More specifically, when $p_{i,j}$ denotes the stationary transition probability from state i to state j, that is, for all integers $k \geq 1$,

$$p_{i,j} := \Pr\{X_{k+1} = j | X_k = i\}, \qquad i, j = 0, 1, 2, \ldots .$$

this distribution is the unique normalised solution of the equations

$$\pi_j = \pi_0 p_{0,j} + \pi_1 p_{1,j} + \cdots + \pi_{j+A} p_{j+A,j}, \qquad j = 0, 1, 2, \ldots . \tag{2.2}$$

It is clear that, given that there is at least one customer at the start of a cycle, the server cannot have been idle during the on-period of the preceding cycle. The reason for this is that, as mentioned earlier, the queue length cannot increase during an on-period. As a result, the probabilities of transitions to states j, with $j > 0$, depend only on the number of arriving customers in the preceding cycle and not on the arrival pattern inside this cycle.

To show that $p_{i,j}$ depends on $j - i$ only, for $j > 0$, consider the recurrence relation (2.1) and assume that $X_{k+1} = j > 0$. Then, we have

$$p_{i,j} = \Pr\{X_k + N_k - A = j | X_k = i\} = \begin{cases} \Pr\{N_k = j - i + A\}, & 0 \leq j - i + A \leq C, \\ 0, & \text{otherwise}, \end{cases} \tag{2.3}$$

from which it immediately follows that, for j positive, $p_{i,j}$ only depends on $j - i$ and is determined by the number of arrivals in a cycle. Further, the number of arrivals in a cycle has a binomial distribution with parameters C and p, that is, for all $k = 1, 2, 3, \ldots$,

$$\alpha_h := \Pr\{N_k = h\} = \binom{C}{h} p^h (1-p)^{C-h}, \qquad h = 0, 1, 2, \ldots, C. \tag{2.4}$$

The transition probabilities $p_{i,0}$ are given by

$$p_{i,0} = \Pr\{X_k + N_k - A \leq 0 | X_k = i\} = \begin{cases} \sum_{h=0}^{A-i} \alpha_h, & i \leq A, \\ 0, & \text{otherwise.} \end{cases} \tag{2.5}$$

From relations (2.3), (2.4), and (2.5), the equilibrium equations (2.2) can be partitioned as follows

$$\pi_0 = \pi_0 \sum_{h=0}^{A} \alpha_h + \pi_1 \sum_{h=0}^{A-1} \alpha_h + \cdots + \pi_A \alpha_0, \tag{2.6}$$

$$\pi_j = \pi_0 \alpha_{j+A} + \pi_1 \alpha_{j+A-1} + \cdots + \pi_{j+A} \alpha_0, \qquad 1 \leq j \leq C - A - 1, \tag{2.7}$$

$$\pi_j = \pi_{j-(C-A)} \alpha_C + \pi_{j-(C-A-1)} \alpha_{C-1} + \cdots + \pi_{j+A} \alpha_0, \qquad j \geq C - A. \tag{2.8}$$

The equations (2.8) have a constant structure and hold for all states, except for the first $C - A$ at the boundary of the state space. Therefore, we call the equations (2.6) and (2.7) the boundary equations and the equations (2.8) the inner equations.

As already mentioned, the stationary distribution is the unique solution of the equilibrium equations and the normalisation equation

$$\sum_{j=0}^{\infty} \pi_j = 1. \tag{2.9}$$

Taking a closer look at the inner equations, we notice that these equations form a C-th order homogeneous linear difference equation with constant coefficients. Then, we know from the theory of difference equations (see, for instance, Henrici [1968]) that there are C (not necessarily distinct) solutions of the form $\pi_j = z^j$ of this equation. By linearly combining these solutions, we may satisfy the boundary equations. However, only for the solutions z^j with $|z| < 1$ the coefficients in the linear combination can be non-zero, because otherwise this linear combination cannot satisfy equation (2.9). So, the stationary distribution is in essence a linear combination of geometric distributions.

There are two standard techniques for the solving of this difference equation with its boundary equations. The first technique directly seeks solutions of the form $\pi_j = z^j$ satisfying the inner equations and, after that, it uses a linear combination of these solutions z^j with $|z| < 1$ to satisfy the boundary equations and the normalisation equation. Adopting the terminology in Feller [1968], we denote this technique by the method of particular solutions. The other technique is the generating-function technique. We begin with discussing the former technique for solving the equilibrium equations and, thereafter, we discuss the latter one.

2.3.1 The method of particular solutions

For the method of particular solutions, we substitute $\pi_j = z^j$ into the inner equations (2.8) and divide by $z^{j-(C-A)}$ to obtain

$$z^{C-A} = \alpha_0 z^C + \alpha_1 z^{C-1} + \cdots + \alpha_C = (p + (1-p)z)^C, \tag{2.10}$$

where the last equality results from using (2.4) and Newton's binomial formula. This equation is also known as the characteristic polynomial equation of the difference equation.

Equation (2.10) has exactly C solutions in the complex plane. But, only those solutions z^j of the difference equation with $|z| < 1$ can satisfy equation (2.9). As we shall prove for a more general case in Chapter 3, we have the following lemma concerning the number of solutions z of the characteristic polynomial equation with $|z| < 1$.

Lemma 2.3.1 *Equation (2.10) has exactly $C - A$ roots inside the unit circle, if $Cp < A$.*

Proof. See Lemma 3.3.1 in Chapter 3. □

Let $z_1, z_2, \ldots, z_{C-A}$ denote these $C - A$ solutions inside the unit circle of the characteristic polynomial equation (2.10). Clearly, any linear combination of the solutions z_k^j, with $|z_k| < 1$, is a solution of the inner equations (2.8) as well. We now try to find a linear combination of these $C - A$ solutions z_k^j to satisfy the $C - A$ boundary equations (2.6) and (2.7), and the normalisation equation (2.9).

Although we have $C - A$ solutions z_k^j, these solutions may not be all distinct, so that the number of independent solutions may be insufficient for solving the boundary equations and the normalisation equation. The next lemma states how to obtain $C - A$ independent solutions from the $C - A$ roots of equation (2.10), which is a familiar result from the theory of difference equations.

Lemma 2.3.2 *Let equation (2.10) have the K distinct solutions $\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_K$ inside the unit circle, with $K \leq C - A$, and let solution $\hat{z}_k$ have multiplicity m_k, so that $\sum_{k=1}^{K} m_k = C - A$. Then, the K sequences with elements $\hat{z}_k^j, j\hat{z}_k^j, \ldots, j^{m_k-1}\hat{z}_k^j$, with $k = 1, 2, \ldots, K$, constitute a system of $C - A$ independent solutions for the difference equation (2.8).*

Proof. See, for example, Theorem 5.3 in Henrici [1968]. □

To simplify the notation, we assume that the solutions $z_1, z_2, \ldots, z_{C-A}$ are distinct (Feller [1968] remarks that these solutions are distinct in most practical cases). So, we have $C - A$ independent solutions z_k^j which satisfy the inner equations, and any linear combination of these solutions

$$\pi_j = \sum_{k=1}^{C-A} \lambda_k z_k^j, \tag{2.11}$$

is a solution of these equations as well, where the λ_k's denote complex numbers. We now show that this representation of π_j can be used to solve the boundary equations and the normalisation equation.

We substitute the form (2.11) into the boundary equations (2.7), forgetting for the moment the equilibrium equation for state $j = 0$. Recall that the form (2.11) already satisfies the inner equations. This system of boundary equations forms a system of $C - A - 1$ homogeneous linear equations with the $C - A$ unknown coefficients λ_k. Hence, this system has a non-trivial solution, that is, a non-null solution. Notice that, since $|z_k| < 1$ for all k,

$$\sum_{j=0}^{\infty} |\pi_j| = \sum_{j=0}^{\infty} \sum_{k=1}^{C-A} |\lambda_k z_k^j| < \infty. \tag{2.12}$$

Consider any non-trivial solution for the coefficients λ_k of the equilibrium equations (2.7) and (2.8). Then, since adding the equations (2.7) and (2.8) yields the equation (2.6), this solution automatically satisfies the equilibrium equation (2.6) for state $j = 0$. So, we have a solution (2.11) of the equilibrium equations. Since this solution satisfies (2.12), it follows from a Foster's criterion (cf. Foster [1953]) that this solution, after normalising it (using equation (2.9)), is equal to the stationary distribution. Furthermore, this implies that the boundary equations (2.7) and the normalisation equation constitute a system of linearly independent equations, so that the coefficients λ_k are unique and can be obtained directly by solving this system of equations (of course, after substituting the form (2.11)).

So, in principle, we have now determined the stationary imbedded queue-length distribution, since the solution (2.11) satisfies the equilibrium equations and the normalisation equation. However, to use this solution, we have to determine the $C - A$ roots z inside the unit circle of the characteristic polynomial equation (2.10) and to solve a system of $C - A$ linear equations for the determination of the coefficients λ_k of the linear combination of z^j. Unfortunately, applying this analytical technique leads to numerical difficulties for relatively small C already, as we shall see below. Firstly, we show that it is in general hard to determine all roots accurately for polynomial equations, even if these equations are of rather low degree. Secondly, we illustrate that, even if we are able to compute all roots accurately, (some of) these roots may be closely clustered, so that the reduced system of boundary equations (2.7) and the normalisation equation (2.9) are nearly singular. Both numerical problems are well known to be hard to overcome in general (see, for instance, Press, Flannery, Teukolsky & Vetterling [1986]).

To illustrate the first problem, we have to be able to compute the roots of the characteristic polynomial equation (2.10) accurately, at least for some specific problems, and to compare these roots with the roots as computed by applying a standard numerical method. For computing these roots accurately, we exploit the structure of the characteristic polynomial equation.

We recall that the characteristic polynomial equation (2.10) reads

$$z^{(1-A/C)C} = (p + (1-p)z)^C. \tag{2.13}$$

Raising both sides of this equation to the power $1/C$, the C roots of this equation can be obtained by solving the C equations

$$z^{1-A/C} = \varphi_k(p + (1-p)z), \quad k = 0, 1, \ldots, C-1, \tag{2.14}$$

with φ_k satisfying $\varphi_k^C = 1$, so that $\varphi_k = e^{2\pi ik/C}$ with $i = \sqrt{-1}$. If we divide A and C by their greatest common divisor, so that A/C becomes q/r, say, and raise both sides of equation (2.14) to the power r, then this equation becomes

$$z^{r-q} = \varphi_k^r(p + (1-p)z)^r. \tag{2.15}$$

If r is rather small, then the r roots of this equation can easily be computed accurately by, for instance, the modified Laguerre method. Hence, the determination of the C roots of the characteristic polynomial equation reduces to solving C/r polynomial equations of the degree r.

To illustrate that the finding of all roots of a polynomial equation generally leads to numerical instabilities, we use the following example. Suppose that 20% of the service capacity is reserved for rendering service to customers, that is, $A = 0.2C$. Customers arrive at a rate $p = 0.17$, so that the effective utilisation Cp/A is 0.85. We consider four cases, namely, the cycle length C is equal to 10, 20, 40, and 80 slots (so, the length A of the on-period is equal to 2, 4, 8, and 16 slots, respectively).

Since the solutions of equation (2.13) are also solutions of the same equation after raising both sides to the same power, the solutions of the equation for the case $C = 10$ are also solutions for the case $C = 20$, the solutions for the case $C = 20$ are solutions for the case $C = 40$ as well, and the solutions for the case $C = 40$ are solutions for the case $C = 80$ too. In Figure 2.4, we depict the computed values of solutions inside the unit circle of the characteristic polynomial equation (2.10) for the four examples. These roots are computed by applying the modified Laguerre method to this equation using double machine precision. In this figure, z_1 denotes the largest of these roots in absolute value, which, as we shall prove later on in Lemma 2.4.1, is the unique positive root inside the unit circle. The ovally shaped contours in this figure denote the solutions of the characteristic polynomial equation when C tends to infinity. These roots are given by the relation

$$|z|^{r-q} = |p + (1-p)z)|^r.$$

For all examples, the computed values should lie on this ovally shaped contour. The figure shows that the computed values lie on the contour when C is small only. However, when C gets larger, the number of the computed values which are not lying on this contour increases rapidly, which indicates that the precision is insufficient.

From Figure 2.4, we conclude that the determination of all roots (inside the unit circle) of polynomial equations of rather low degree is hard already. However, for the queueing system of Section 2.2, we can still determine all roots of the characteristic polynomial equation accurately for certain parameter settings (recall equation (2.15)). We now show that, although the roots can be computed accurately, (some of) these can be closely clustered, so that the reduced system of boundary equations (2.7) and the normalisation equation (2.9) are nearly linearly dependent.

For the examples $C = 20$ and $C = 40$, we depict in Figure 2.5 the roots inside the unit circle of the corresponding characteristic polynomial equation, which are computed by applying the modified Laguerre method to equation (2.15). This figure shows that, already for a relatively small number of roots, many of these roots can be very closely clustered.

We solved the reduced system of boundary equations (2.7) and the normalisation equation (2.9) (after substituting the form (2.11) and the computed roots) for the coefficients λ_k of the linear combination. In Table 2.1, we list some of the 'stationary probabilities' resulting from the method of particular solutions (MPS) for the two examples by using double machine precision, and we compare them with the exact values. The exact stationary probabilities are obtained analytically by utilising the structure of this reduced system of boundary equations, which yields (see Lemma 3.3.3 in Chapter 3)

$$\lambda_k = (1 - z_k) \frac{\prod_{\substack{j=1 \\ j \neq k}}^{C-A} (z_j^{-1} - 1)}{\prod_{\substack{j=1 \\ j \neq k}}^{C-A} (z_j^{-1} - z_k^{-1})}.$$

For the case $C = 20$, the numerical results are satisfactory. But already for $C = 40$, some of the roots are so closely clustered that the reduced system of boundary equations and the normalisation equation are nearly singular, which leads to numerical problems.

From these examples, we conclude that solving the equilibrium equations by the method of particular solutions may be suitable from a numerical point of view for small problems only. We remark

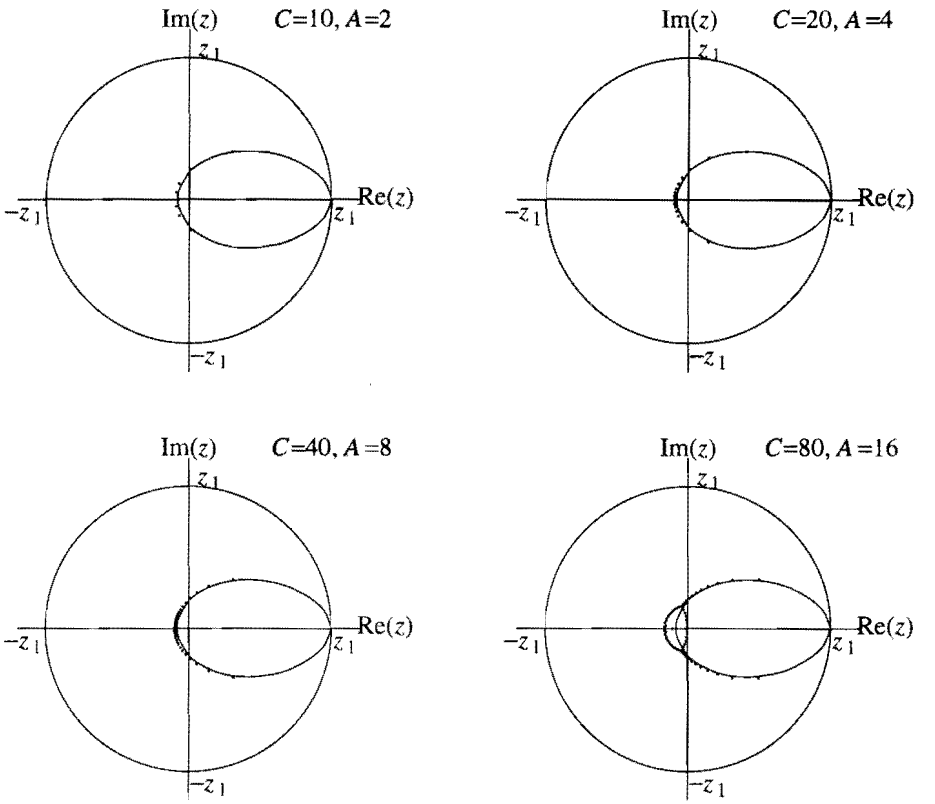

Figure 2.4: *The computed values of the roots of equation (2.10) inside the unit circle for the four examples with* $p = 0.17$ *and with* $z_1 = 0.676$ *the unique positive root inside the unit circle.*

that increasing the machine precision enables one to evaluate larger numerical examples accurately by this method than the examples above, but the numerical problems will occur eventually. For instance, lengthening the cycle leads to similar numerical problems.

2.3.2 The generating-function technique

Another standard technique for solving the equilibrium equations which form a homogeneous linear difference equation of finite order with constant coefficients is the generating-function technique. As mentioned before, Beckmann, McGuire & Winsten [1956], Newell [1960], and Meissl [1962] apply this technique to this model which, for their case, describes the queue-length process of cars at an intersection governed by a fixed-cycle traffic light.

For this technique, we define the probability generating function

$$\Pi(z) := \sum_{j=0}^{\infty} \pi_j z^j,$$

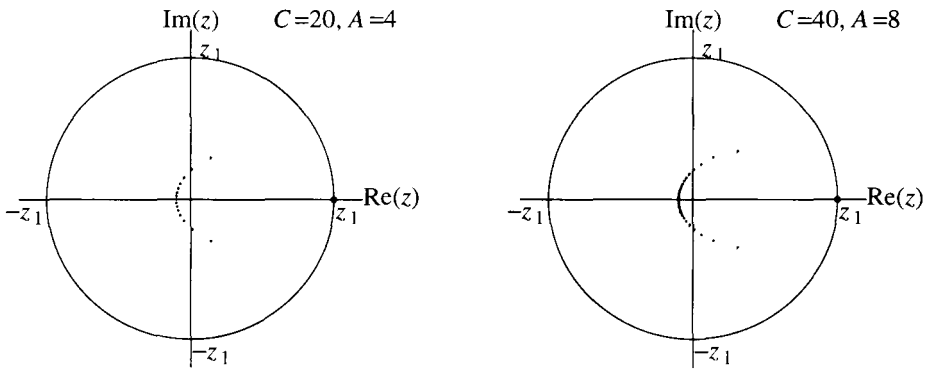

Figure 2.5: *The roots of equation (2.10) inside the unit circle for two examples with* $p = 0.17$ *and with* $z_1 = 0.676$ *the unique positive root inside the unit circle.*

		π_j	
C	j	Exact	MPS
20	0	0.50948	0.50948
	1	0.14456	0.14456
	2	0.10896	0.10896
	3	0.07643	0.07643
40	0	0.60476	0.60156−0.01563i
	1	0.11049	0.11133+0.00391i
	2	0.08594	0.08594+0.00044i
	3	0.06211	0.06211+0.00098i

Table 2.1: *Some 'stationary probabilities' computed by the method of particular solutions (MPS) and the exact values.*

which is well defined for $|z| \leq 1$, since the Markov chain has a unique stationary distribution. Multiplying the equations (2.6), (2.7), and (2.8) by z^j, and adding, we obtain after some algebra

$$\Pi(z) = \frac{z^A \sum_{j=0}^{A-1} \left(\sum_{h=0}^{A-j} (1 - z^{j+h-A}) \alpha_h \right) \pi_j}{z^A - \sum_{h=0}^{C} \alpha_h z^h} = \frac{z^A \sum_{j=0}^{A-1} \left(\sum_{h=0}^{A-j} (1 - z^{j+h-A}) \alpha_h \right) \pi_j}{z^A - (pz + (1-p))^C}. \tag{2.16}$$

Since a probability generating function converges for $|z| \leq 1$, it may not have singularities inside and on the unit circle. Hence, the numerator of (2.16) must have the same roots inside and on the unit circle as the denominator of (2.16), and with the same multiplicity.

Lemma 2.3.3 *The denominator of (2.16) has exactly A roots inside and on the unit circle, if* $Cp < A$.

Proof. See Lemma 3.3.4 in Chapter 3. □

This result also follows from Lemma 2.3.3, because the roots of the characteristic polynomial equation (2.10) are the reciprocals of the zeroes of the denominator of (2.16). This is easily verified by dividing the characteristic polynomial equation by z^C and comparing the resulting equation with the equation that the denominator is equal to zero.

We denote the A roots in Lemma 2.16 by $\xi_{C-A+1}, \xi_{C-A+2}, \ldots, \xi_C$ and show, by similar arguments as in Bailey [1954b] and Meissl [1962], that they are distinct.

Lemma 2.3.4 *The roots $\xi_{C-A+1}, \xi_{C-A+2}, \ldots, \xi_C$ are distinct.*

Proof. Suppose that ξ is a multiple root of the denominator of (2.16) with $|\xi| \leq 1$. Then, ξ is a solution of

$$\xi^A = (p\xi + (1-p))^C,$$

and its derivative

$$A\xi^{A-1} = Cp(p\xi + (1-p))^{C-1}.$$

Dividing these two equations and rewriting yields

$$\xi = \frac{A(1-p)}{(C-A)p}.$$

Clearly, the right-hand side of this latter equality is positive. Furthermore, ξ is a root inside or on the unit circle, so that we must have

$$\frac{A(1-p)}{(C-A)p} \leq 1.$$

However, this inequality implies $A \leq Cp$, which is a contradiction with the ergodicity assumption for the Markov chain. Thus, the assumption that the root ξ has multiplicity larger than one is incorrect, so that we conclude that the roots $\xi_{C-A+1}, \xi_{C-A+2}, \ldots, \xi_C$ are distinct. □

The stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{A-1}$ must satisfy the conditions that the numerator and denominator of (2.16) have the same zeroes inside and on the unit circle. These conditions are also known as regularity conditions. Clearly, $z = 1$ is a root of the denominator, and this is also a root of the numerator, irrespective of the values of π_j, with $j = 0, 1, \ldots, A-1$. Hence, this root, ξ_{C-A+1}, say, does not restrict the values of these stationary probabilities. Substituting the other roots into the numerator leads to a system of $A-1$ homogeneous linear equations with the A unknowns $\pi_0, \pi_1, \ldots, \pi_{A-1}$. By adding the normalisation equation, that is, $\lim_{z \uparrow 1} \Pi(z) = 1$, to this system, the solution of these A equations yields the stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{A-1}$, since these A equations are linearly independent.

Lemma 2.3.5 *The system of $A-1$ equations*

$$\xi_k^A \sum_{j=0}^{A-1} \left(\sum_{h=0}^{A-j} (1 - \xi_k^{h+j-A}) \alpha_h \right) \pi_j = 0, \quad k = C-A+2, C-A+3, \ldots, C,$$

with the normalisation equation $\lim_{z \uparrow 1} \Pi(z) = 1$ is a system of A linearly independent equations with the A unknowns $\pi_0, \pi_1, \ldots, \pi_{A-1}$.

Proof. For convenience, we order this system of A equations such that the first condition is the normalisation equation. This ordered system of A equations can be written in matrix notation as $\boldsymbol{A}\boldsymbol{\pi}_A = \boldsymbol{b}$, where $\boldsymbol{\pi}_A$ denotes the column vector $(\pi_0, \pi_1, \ldots, \pi_{A-1})$ and $\boldsymbol{b}$ the column vector with first entry $A - Cp$ (after elaborating the normalisation equation) and the other entries equal to zero. The matrix $\boldsymbol{A}$ has as first row the row

$$\left(\sum_{h=0}^{A-1}(A-h)\alpha_h, \quad \cdots, \quad \sum_{h=0}^{1}(2-h)\alpha_h, \quad \alpha_0 \right),$$

and, for $m = 2, 3, \ldots, A$, as m-th row the row

$$\left(\sum_{h=0}^{A-1}(\xi_{C-A+m}^{A} - \xi_{C-A+m}^{h})\alpha_h, \quad \cdots, \quad \sum_{h=0}^{1}(\xi_{C-A+m}^{A} - \xi_{C-A+m}^{h+A-2})\alpha_h, \quad (\xi_{C-A+m}^{A} - \xi_{C-A+m}^{A-1})\alpha_0 \right).$$

If this system of equations is linearly independent, then the matrix $\boldsymbol{A}$ should be nonsingular. We show that linearly combining the columns of this matrix yields a nonsingular matrix, so that the matrix $\boldsymbol{A}$ is nonsingular.

We initialise $k = A$ and execute the following steps. Firstly, we multiply the k-th column with $1/\alpha_0$. Secondly, we subtract, for $m = 1, 2, \ldots, k-1$, from the m-th column α_{k-m} times the new k-th column (that is, the k-th column after multiplying). Finally, we set $k := k - 1$, and repeat these two steps, if $k > 0$, and otherwise, we stop.

After executing the above steps, we have the matrix

$$\begin{pmatrix} A & A-1 & \cdots & 1 \\ \xi_{C-A+2}^{A} - 1 & \xi_{C-A+2}^{A} - \xi_{C-A+2} & \cdots & \xi_{C-A+2}^{A} - \xi_{C-A+2}^{A-1} \\ \xi_{C-A+3}^{A} - 1 & \xi_{C-A+3}^{A} - \xi_{C-A+3} & \cdots & \xi_{C-A+3}^{A} - \xi_{C-A+3}^{A-1} \\ \vdots & \vdots & & \vdots \\ \xi_{C}^{A} - 1 & \xi_{C}^{A} - \xi_{C} & \cdots & \xi_{C}^{A} - \xi_{C}^{A-1} \end{pmatrix}.$$

Subtracting column $k+1$, for each $k = 1, 2, \ldots, A-1$, from the column k, and then dividing, for $k = 2, 3, \ldots, A$, row k by $\xi_{C-A+k} - 1$, yields the matrix

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & \xi_{C-A+2} & \cdots & \xi_{C-A+2}^{A-1} \\ 1 & \xi_{C-A+3} & \cdots & \xi_{C-A+3}^{A-1} \\ \vdots & \vdots & & \vdots \\ 1 & \xi_{C} & \cdots & \xi_{C}^{A-1} \end{pmatrix}.$$

This matrix is of a Vandermonde-type (see, for instance, Bellman [1970]), because by Lemma 2.3.4 the roots ξ_k, $k = C-A+2, C-A+3, \ldots, C$, are distinct. Hence, this matrix is nonsingular, so that the matrix $\boldsymbol{A}$ is nonsingular and the proof is complete. □

Since we can now determine the stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{A-1}$, the final step of the generating-function technique is to invert the probability generating function $\Pi(z)$. To invert the probability generating function (2.16), we first divide both the numerator and denominator of (2.16) by $-\alpha_C$, and we denote the resulting numerator and denominator by $N(z)$ and $D(z)$, respectively. Further, the $C - A$ roots of $D(z)$ outside the unit circle are denoted by $\xi_1, \xi_2, \ldots, \xi_{C-A}$.

Since the terms $(z - \xi_k)$, for $k = C - A + 1, C - A + 2, \ldots, C$, appear in the numerator $N(z)$ as well as in the denominator $D(z)$, they cancel out. So, under the assumption that all roots ξ_k are distinct, with $k = 1, 2, \ldots, C - A$, $\Pi(z)$ can be expanded into partial fractions

$$\Pi(z) = \frac{N(z)}{D(z)} = \sum_{k=1}^{C-A} \frac{\eta_k}{\xi_k - z}, \tag{2.17}$$

where the η_k's denote the coefficients of the partial-fraction expansion. The coefficients η_k can be determined explicitly.

If the root ξ_k has multiplicity m, with $m > 1$, then we should expand the probability generating function into the partial fractions $1/(\xi_k - z), 1/(\xi_k - z)^2, \ldots, 1/(\xi_k - z)^{m-1}$.

For the determination of η_k (see, for instance, Jury [1964]), we multiply relation (2.17) by $z - \xi_k$, let z tend to ξ_k, and employ l'Hôpital's rule, to obtain

$$\eta_k = \lim_{z \to \xi_k} \frac{-N(z)(z - \xi_k)}{D(z)} = \frac{-N(\xi_k)}{D'(\xi_k)}.$$

The derivative of $D(z)$ is obviously

$$D'(z) = \sum_{i=1}^{C} \prod_{\substack{j=1 \\ j \neq i}}^{C} (z - \xi_j),$$

so that

$$D'(\xi_k) = \prod_{\substack{j=1 \\ j \neq k}}^{C} (\xi_k - \xi_j). \tag{2.18}$$

The probability generating function $\Pi(z)$ can now easily be inverted, so that the stationary distribution is (again) expressed as a linear combination of geometric distributions

$$\pi_j = \sum_{k=1}^{C-A} \frac{\eta_k}{\xi_k} \left(\frac{1}{\xi_k}\right)^j. \tag{2.19}$$

To apply the generating-function technique, we again have to determine the roots of a polynomial equation. But, now we have to compute all these roots: inside, on, and outside the unit circle. Further, we have to solve a system of A equations for the determination of the stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{A-1}$. The coefficients η_k of the partial-fraction expansion can be given explicitly. Unfortunately, applying this technique leads to similar numerical problems as the previous technique. Since we determine the roots of an equation that is basically the same as the characteristic polynomial equation (2.10), we omit illustrating the problems resulting from numerically computing these roots. We only show that the clustering of the roots leads to a nearly singular system of equations for the determination of the stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{A-1}$. To show this, we again have to be able to compute all roots accurately.

Using the same arguments as for the derivation of equation (2.14), one easily verifies that the determination of the C roots of the denominator of (2.16) reduces to solving the C equations

$$z^{A/C} = \varphi_k(pz + (1 - p)), \quad k = 0, 1, \ldots, C - 1,$$

with $\varphi_k = e^{2\pi ik/C}$ and $i = \sqrt{-1}$. So, if A and C are divided by their greatest common divisor, then we have that A/C becomes q/r, say. Consequently, solving a polynomial equation of the degree C reduces to solving C/r polynomial equations of the degree r

$$z^q = \varphi_k^r(pz + (1-p))^r. \tag{2.20}$$

Suppose that 62.5% of the service capacity is reserved for servicing customers, that is, $A = 0.625C$. The arrival rate of customers is 0.5, so that the effective utilisation Cp/A is equal to 0.8. We consider three cases, namely, $C = 16$, $C = 40$, and $C = 80$ (and so, A is equal to 10, 25, and 50 slots, respectively).

In Figure 2.6, we depict the roots inside the unit circle of the denominator of relation (2.16) for the case $C = 40$ and $C = 80$. These roots are computed by applying the modified Laguerre method to (2.20) with $r = 5$ and $q = 1$.

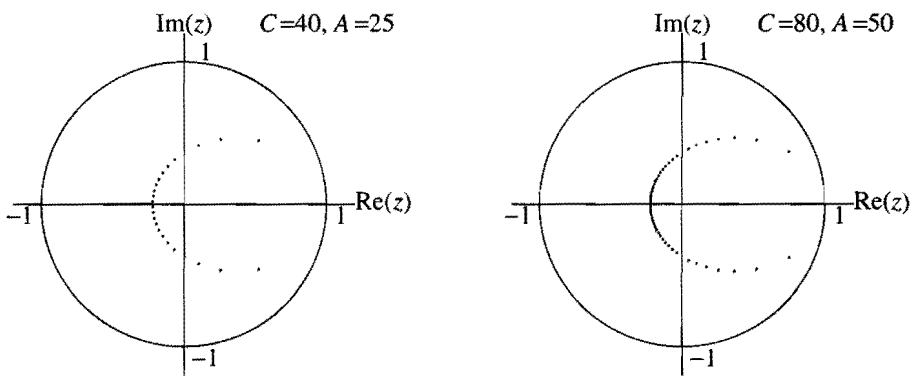

Figure 2.6: *The roots of the denominator of (2.16) inside the unit circle for two examples with* $p = 0.50$.

The roots inside the unit circle and the normalisation equation $\lim_{z\to 1^-} \Pi(z) = 1$ are used to compute the stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{A-1}$. In Table 2.2, we list the resulting stationary probabilities (GFT-a), using double machine precision, and the exact values. For small C, the results are satisfactory. When enlarging C, only the computed stationary probabilities for small values of j are good, but for j close to $A-1$, the results are of no practical value. Increasing C further, the system of equations for the determination of these probabilities is almost singular, so that it is not possible to compute these probabilities in a numerically stable way.

From the computed roots and the stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{A-1}$, we compute the coefficients η_k. Using expression (2.19) and double machine precision, the resulting stationary probabilities (GFT-b) are listed in Table 2.2. For small C, the results are excellent, even if not all stationary probabilities π_j, $j = 0, 1, \ldots, A-1$, are computed accurately (that is, the probabilities from GFT-a). The reason is that the erroneous probabilities π_j, with j close to $A-1$, are multiplied by a very small number, as can be verified from (2.16), so that the total effect of the error is negligible. Evidently, for larger C, none of the coefficients is correct, so that the resulting 'probabilities' do not make any sense at all. Finally, as we mentioned for the method of particular solutions, increasing the machine precision only allows us to apply the generating-function technique successfully to larger queueing systems, but the numerical difficulties will occur eventually.

		π_j		
C	j	Exact	GFT-a	GFT-b
16	0	0.86057	0.86057	0.86057
	1	0.07774	0.07774	0.07774
	2	0.03791	0.03791	0.03791
	3	0.01540	0.01540	0.01540
40	0	0.95519	0.95519	0.95519
	1	0.02262	0.02262	0.02262
	2	0.01210	0.01210	0.01210
	3	0.00587	0.00587	0.00587
	10	0.00001	0.00269+0.00015i	0.00001
	15	0.00000	−15.89236−0.33263i	0.00000
	20	0.00000	4310.84341−590.25444i	0.00000
80	0	0.99059	1.03608−0.02210i	1.01722+0.02135i
	1	0.00440	−0.04101−0.00695i	0.00454+0.00013i
	2	0.00247	0.03590+0.04445i	0.00254+0.00007i
	3	0.00131	0.13196−0.12954i	0.00136+0.00002i

Table 2.2: *The 'stationary probabilities' obtained by implementing the generating-function technique and the exact values.*

2.3.3 Conclusions

For the two presented analytical techniques, we have pointed out two numerical problems, which arise when implementing them, and both problems can hardly be overcome (see, for instance, Press, Flannery, Teukolsky & Vetterling [1986]). The first problem is the determination of all solutions of a polynomial equation. For polynomial equations of relatively small degree, significant numerical problems occur already. Furthermore, even if we are able to compute all these solutions accurately, these solutions may be closely clustered, so that the system of equations (after substituting these solutions), for determining the coefficients λ_k of the linear combination or the first A stationary probabilities, is nearly singular. Only if C is small, that is, the cycle length is small, these techniques give accurate results. Therefore, we present two numerical approaches for approximating the stationary distribution. The first of these approaches exploits the tail behaviour of this distribution as previously suggested by Tijms & Van de Coevering [1991].

2.4 A numerical approach exploiting the tail behaviour

Because of the structure of the equilibrium equations, the stationary distribution can be expressed by the form (2.11) or (2.19), with possibly some of the z_k^j or $1/\xi_k^j$ multiplied by powers of j. It is clear from this form that the largest z_k or $1/\xi_k$ in absolute value determine(s) the tail behaviour of the stationary distribution. Since both techniques lead to the same results, and because $z_k = 1/\xi_k$, we use the notation and formulae of the method of particular solutions of Section 2.3.1 for deriving a numerical approach.

The next lemmas state that the largest root in absolute value of the characteristic polynomial equation (2.10) is unique, and that it is real and positive. These results are well known from the theory of

branching processes (see, for instance, Athreya & Ney [1972]).

Lemma 2.4.1 *Equation (2.10) has exactly one root in the interval* $(0, 1)$, *if* $Cp < A$.

Proof. See Lemma 3.4.1 in Chapter 3. □

Let this positive root be denoted by z_1.

Lemma 2.4.2 *For all* $k = 2, 3, \ldots, C - A$, *we have* $|z_k| < z_1$.

Proof. See Lemma 3.4.2 in Chapter 3. □

Thus, from these two lemmas and the form (2.11), we have

$$\lim_{j\to\infty} \frac{\pi_{j+1}}{\pi_j} = z_1, \tag{2.21}$$

so that the tail of the stationary distribution is asymptotically geometric. Since the positive root z_1 completely determines the tail behaviour of this distribution, we try to exploit it for numerical purposes. That is, we try to use it for approximating the solution to the equilibrium equations (2.7) and (2.8), and the normalisation equation (2.9). Equation (2.6) is omitted, because it is redundant, as we noted in Section 2.3.1. To use this tail behaviour, we should be able to compute this root accurately.

By equation (2.15), the positive root z_1 can be computed accurately for $\varphi_0 = 1$, if r is small, by, for instance, the modified Laguerre method. Moreover, for $r = 1, 2, 3, 4$, this root can even be computed analytically. If r is not rather small, then this root z_1 can easily be determined by, for example, bisection. For the examples considered in Section 2.3, applying bisection to the equation (2.10) directly does not give numerical problems. As we shall show in Chapter 3, calculating z_1 by bisection from a general characteristic equation can give numerical problems. These problems are caused by the fact that the differences between the left-hand and the right-hand side of this equation may not be distinguished from zero within the machine precision in the neighbourhood of z_1. Therefore, we shall suggest to divide both the left-hand and right-hand side of this equation by z^{C-A} first, and then take logarithms to avoid these problems.

Since the root z_1, which completely determines the tail behaviour of the stationary distribution, can be computed accurately, we use it for the numerical determination of this distribution. By (2.21), we have for large j approximately

$$\pi_{j+1} \approx z_1 \pi_j.$$

Tijms & Van de Coevering [1991] propose an algorithm, which uses this approximation to compute the stationary distribution numerically.

For this algorithm, we choose an integer J and set, for $j \geq J$,

$$\pi_{j+1} = z_1 \pi_j = z_1^{j-J+1} \pi_J, \tag{2.22}$$

so that we approximate the quotient π_{j+1}/π_j for these j by z_1. In other words, we reduce the denumerable system of equilibrium equations to a finite system of equations by assuming that (2.22) holds for $j \geq J$. As noted earlier, we omit the equilibrium equation (2.6) for state $j = 0$. Then, the form (2.22), for $j \geq J$, is substituted into the equilibrium equations for the states $1, 2, \ldots, J$ and the normalisation equation, which now reads

$$\sum_{j=0}^{J-1} \pi_j + \frac{1}{1 - z_1} \pi_J = 1.$$

Solving the reduced system of J equilibrium equations and the normalisation equation yields an approximation for the stationary probabilities $\pi_0, \pi_1, \ldots, \pi_J$, and hence for the stationary distribution.

The quality of this numerical approach depends of course on the rate at which the quotient π_{j+1}/π_j can be approximated well by z_1. Hence, the quotient of the second-largest root in absolute value of the characteristic polynomial equation, denoted by z_2, and z_1 plays an important role. For the numerical examples of Section 2.3, we list the quotient of these roots in absolute value in Table 2.3. This table suggests that this quotient $|z_2|/|z_1|$ increases with C (keeping A/C fixed). Actually, if C tends to infinity, then this quotient tends towards one. Hence, to attain the same accuracy for the resulting stationary probabilities, we have to increase J when we increase C. In Table 2.3, we list the quotient to the power $C - A$ as well. This gives an indication as to whether for this method we have to solve a smaller or larger system of equations than the number of boundary equations to be solved by the method of particular solutions. As this table shows, the effect of the second-largest root is practically nil for $j \geq C - A$. Indeed, the results of Table 2.4 confirm that for our examples z_1 is a fairly good approximation for the quotient π_{j+1}/π_j for relatively small values of j. Recall that for these examples z_1 is the same.

C	$\frac{\lvert z_2\rvert}{\lvert z_1\rvert}$	$\left(\frac{\lvert z_2\rvert}{\lvert z_1\rvert}\right)^{C-A}$
10	0.212	$4.03 \cdot 10^{-6}$
20	0.327	$1.68 \cdot 10^{-8}$
40	0.463	$2.03 \cdot 10^{-11}$
80	0.609	$8.66 \cdot 10^{-15}$

Table 2.3: *The quotient of the second-largest and the largest root in absolute value.*

j	$C = 10$	$C = 20$	$C = 40$	$C = 80$
1	0.7247	0.7537	0.7773	0.7915
2	0.6826	0.7014	0.7292	0.7560
3	0.6756	0.6807	0.6994	0.7279
4	0.6761	0.6757	0.6835	0.7067
5	0.6764	0.6757	0.6768	0.6919
10	0.6764	0.6764	0.6764	0.6752
15	0.6764	0.6764	0.6764	0.6765
20	0.6764	0.6764	0.6764	0.6764
25	0.6764	0.6764	0.6764	0.6764
∞	0.6764	0.6764	0.6764	0.6764

Table 2.4: *The quotient π_{j+1}/π_j for the examples.*

According to Tijms & Van de Coevering [1991], their numerical approach works well in practice, since z_1 is a fairly good approximation for the quotient π_{j+1}/π_j even for small values of j. The theoretical foundation of this approach, however, is still incomplete. Therefore, it is unclear why this method works so well, and it is not possible to give rules for the choice of J. So, J has to be determined experimentally, by, for instance, comparing the difference between the approximation for different values

of J or evaluating whether the quotient π_{j+1}/π_j is (nearly) equal to z_1 for $j \geq J$. Of course, the value of J depends on the required accuracy.

So, we have now considered the stationary imbedded queue-length distribution. We conclude this section with returning to one of its main purposes, namely, the determination of the stationary queue-length distribution at arrival instants. These distributions can easily be computed recursively.

Let Y_n denote the number of customers in the system at the n-th slot boundary in the cycle in statistical equilibrium, for $n = 1, 2, \ldots, C$. Then, the probability distribution of Y_1 is $\{\pi_j, j = 0, 1, 2, \ldots\}$. Suppose that the probability distribution of Y_n is known for some n, with $n = 1, 2, \ldots, C-1$. It is easily seen that

$$Y_{n+1} = \max\{0, Y_n + B_n - \delta_n\},$$

where B_n denotes the number of arrivals in slot n and δ_n indicates whether slot n is an off-slot ($\delta_n := 0$) or an on-slot ($\delta_n := 1$). From this relation, it is clear that, if slot n is an off-slot,

$$\begin{aligned} \Pr\{Y_{n+1} = 0\} &= \Pr\{Y_n = 0\}(1-p), \\ \Pr\{Y_{n+1} = j\} &= \Pr\{Y_n = j-1\}p + \Pr\{Y_n = j\}(1-p), \qquad j \geq 1, \end{aligned}$$

and that, if slot n is an on-slot,

$$\begin{aligned} \Pr\{Y_{n+1} = 0\} &= \Pr\{Y_n = 0\}(1-p) + \Pr\{Y_n = 1\}(1-p), \\ \Pr\{Y_{n+1} = j\} &= \Pr\{Y_n = j\}p + \Pr\{Y_n = j+1\}(1-p), \qquad j \geq 1. \end{aligned}$$

Hence, in this way, we can recursively compute the stationary queue-length distribution at slot boundaries.

By the Bernoulli-arrivals-see-time-average property (cf. Halfin [1983]), the stationary distribution of the queue length as seen by a arbitrary customer arriving in slot n is equal to the stationary queue-length distribution at the n-th slot boundary.

2.5 A moment-iteration technique

In practice, one often only has information about the first two moments of the service-time distribution of customers. Think, for instance, of the demand of orders. Therefore, we develop a moment-iteration method for approximating the performance measures of interest, which uses this limited information only. For the main idea of this approximation, we return to the stochastic process $\{X_k, k = 1, 2, 3, \ldots\}$ of the number of customers at the first slot boundary of cycles, as introduced in Section 2.2.

For this process, we derived the recurrence relation (cf. equation (2.1))

$$X_{k+1} = \max\{0, X_k + N_k - A\}, \qquad k = 1, 2, 3, \ldots \tag{2.23}$$

where N_k denotes the number of customers arriving in the k-th cycle consisting of C slots. We recognise this relation as Lindley's equation for the waiting time X_{k+1} of the $(k+1)$-st customer arriving in a $D/G/1$ queueing system with interarrival time A and discrete service-time distribution $\{\alpha(n), n = 0, 1, \ldots, C\}$ (see, for instance, Grimmett & Stirzaker [1992]). Hence, the analysis of the imbedded queue-length process reduces to the analysis of the waiting-time process of customers in a $D/G/1$ queueing system.

Relation (2.23) and, more generally, Lindley's equation for the waiting-time process of customers in a $GI/G/1$ queueing system can be solved, for k tending to infinity, analytically in a few exceptional cases only. Therefore, we develop a numerical method for approximating the limiting solution of this equation for k tending to infinity.

For the continuous-time $GI/G/1$ queueing system, De Kok [1989] presents a simple and quite accurate algorithm for the determination of the waiting-time characteristics by using the corresponding version of Lindley's equation. We briefly describe this algorithm for approximating the waiting-time characteristics for the case of a continuous-time $D/G/1$ queueing system. More precisely, we use the equation

$$W_{k+1} = \max\{0, W_k + B_k - A\}, \tag{2.24}$$

where the random variables B_k and W_k denote the service time and the waiting time of the k-th arriving customer, to approximate

$$\mathrm{E}\{W\} := \lim_{k\to\infty} \mathrm{E}\{W_k\}, \quad \text{and} \quad \Pi_W := \lim_{k\to\infty} \Pr\{W_k > 0\}.$$

These limits are well defined, if the average service time is smaller than the average interarrival time and if the first two moments of the service-times are finite (see, for example, Asmussen [1987]).

Let the generic random variable B have the same distribution as B_k, with $k = 1, 2, 3, \ldots$, and define σ_k as the standard deviation of W_k. We initialise $\mathrm{E}\{W_1\} = 0$ and $\mathrm{E}\{W_1^2\} = 0$, so that $\sigma_1 = 0$, and set $k = 1$. The iteration step consists of two parts. Firstly, we compute $\mathrm{E}\{W_k + B\}$ and $\mathrm{E}\{(W_k + B)^2\}$ and fit a tractable distribution $F_k(\cdot)$ on the first two moments of $W_k + B$. Below, we return to the issue of tractable distributions. Secondly, using relation (2.24) and the distribution $F_k(\cdot)$, we compute the approximations for the first two moments of W_{k+1}, that is,

$$\mathrm{E}\{W_{k+1}\} = \int_{x=A}^{\infty} (x - A)\mathrm{d}F_k(x), \quad \text{and} \quad \mathrm{E}\{W_{k+1}^2\} = \int_{x=A}^{\infty} (x - A)^2\mathrm{d}F_k(x), \tag{2.25}$$

and compute σ_{k+1}. If both $|\mathrm{E}\{W_{k+1}\} - \mathrm{E}\{W_k\}|$ and $|\sigma_{k+1} - \sigma_k|$ are small, then we stop and have the approximations $\mathrm{E}\{W_{k+1}\}$ and $1 - F_k(A)$ for $\mathrm{E}\{W\}$ and Π_W, respectively. Otherwise, we set $k := k + 1$ and repeat the iteration step.

To apply this moment-iteration algorithm, we have to find tractable distributions for fitting probability distributions on the first two moments of a non-negative random variable (note that the random variable $X_k + B$ is always non-negative). With tractable distributions, we mean distributions such that the formulae in (2.25) can easily be evaluated numerically. A commonly used way to fit a tractable distribution on the first moment $\mathrm{E}\{Y\}$ and on the coefficient of variation c_Y (that is, the quotient of the standard deviation and mean) of a continuous non-negative random variable Y is the following (see, for instance, Tijms [1986]). If $c_Y < 1$, then one fits a mixture of two Erlang distributions with the same scale parameter μ. If $c_Y \geq 1$, then one fits a mixture of two exponential distributions (also known as a hyperexponential distribution) with balanced means. This way of fitting is used for $F_k(\cdot)$ in De Kok [1989].

As already mentioned, this moment-iteration algorithm performs well for continuous-time queueing systems. However, since we have discrete distributions and because for computing tail probabilities we have to discretise, it seems more natural to apply a discrete version of this algorithm. For fitting discrete distributions on the first moment and on the coefficient of variation, several procedures are known. However, some of these procedures do not capture all possible combinations of the first

two moments of non-negative discrete random variables, whereas others are not that useful for the moment-iteration algorithm. Therefore, we develop a novel procedure that is a discrete analogue to the fitting method described above. In order to fit all possible values of the first moment $E\{Z\}$ and the coefficient of variation c_Z of a discrete random variable Z on the non-negative integers, we use four classes of distributions instead of two: a mixture of two binomial distributions, a Poisson distribution, a mixture of two negative binomial distributions, and a mixture of two geometric distributions. The reason for four instead of two probability distributions is that the discrete analogues to mixtures of Erlang and exponential distributions are not sufficient to cover all possible values of $E\{Z\}$ and c_Z. By adding the Poisson and binomial distribution, this gap is filled. Further, in contrast to a continuous random variable, not all combinations of $(E\{Z\}, c_Z)$ are possible for a discrete random variable on the non-negative integers. For example, a discrete random variable on the non-negative integers with mean 1.5 and coefficient of variation 0 is obviously not possible. A formal description of all possible pairs $(E\{Z\}, c_Z)$ is given in the next theorem.

Theorem 2.5.1 *For a pair of non-negative real numbers (m, c), there exists a random variable on the non-negative integers with mean m and coefficient of variation c, if, and only if,*

$$c^2 \geq \frac{(n+1-m)(m-n)}{m^2},$$

where n is the unique integer satisfying $n \leq m < n+1$.

Proof. See Theorem 5.5.1 in Chapter 5. □

In Figure 2.7, the 'impossible' regions in the (m, c)-plane are shaded.

Before presenting the method for fitting discrete distributions on a discrete non-negative random variable, we introduce the following notation. $\text{Geo}(q)$ denotes a random variable having probability distribution $\{(1-q)q^i, i = 0, 1, 2, \ldots\}$, $\text{NB}(n, q)$ a random variable that is the sum of n independent $\text{Geo}(q)$ variables, and $\text{Bin}(n, q)$ a random variable that is binomially distributed, with n the number of trials and q the success probability. Then, the method for fitting discrete distributions is briefly given in the next theorem.

Theorem 2.5.2 *Let Z be a random variable on the non-negative integers with mean m and coefficient of variation $c > 0$, and let $\theta = c^2 - 1/m$. Then, the random variable Y matches the first two moments of Z if Y is chosen as follows:*

1. *If $-1/n \leq \theta < -1/(n+1)$ for certain $n = 1, 2, 3, \ldots$, then Y is a mixture of a $Bin(n, q)$ and a $Bin(n+1, q)$ random variable.*
2. *If $\theta = 0$, then Y is a Poisson distribution.*
3. *If $1/(n+1) \leq \theta < 1/n$ for certain $n = 1, 2, 3, \ldots$, then Y is a mixture of an $NB(n, q)$ and an $NB(n+1, q)$ random variable.*
4. *If $\theta \geq 1$, then Y is a mixture of a $Geo(q_1)$ and a $Geo(q_2)$ random variable with balanced means.*

Proof. See Theorem 5.5.2 in Chapter 5. □

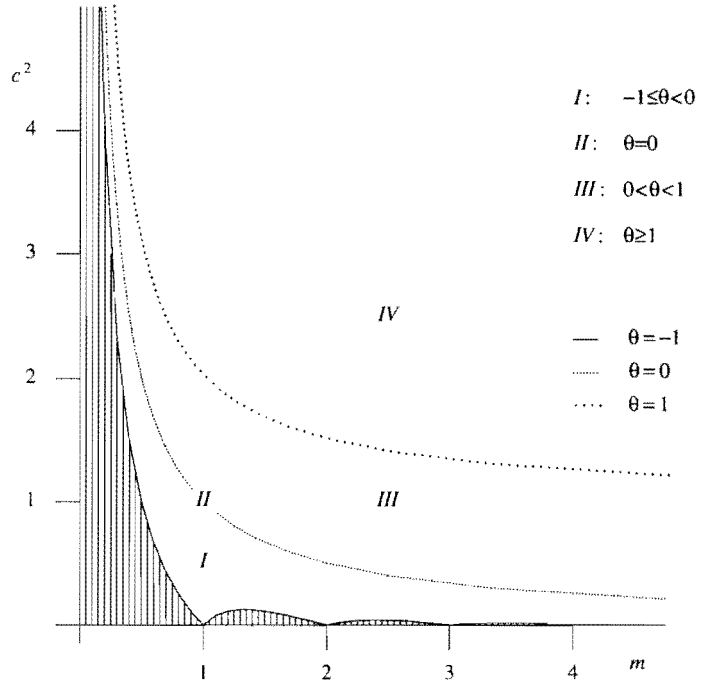

Figure 2.7: *The shaded regions denote the impossible regions for a discrete random variable on the non-negative integers, and the other regions are the four regions for θ indicating which distribution is used to match the first two moments of this random variable.*

For more details, we refer to Chapter 5. The last two cases of Theorem 2.5.2 are the discrete analogues to the mixture of two Erlang distributions with the same parameter μ and to the hyperexponential distribution with balanced means. The results of Theorem 2.5.1 and Theorem 2.5.2 are illustrated in Figure 2.7.

The discrete fits can now be used for the moment-iteration method described above for approximating the stationary queue-length distribution at the start of cycles. We use only the first two moments of the number of arrivals in a cycle. Let the generic random variable N have the same distribution as N_k, with $k = 1, 2, 3, \ldots$, and σ_k the standard deviation of X_k, for $k = 1, 2, 3, \ldots$. By the method described in Theorem 2.5.2, fitting a discrete distribution on the first two moments of N yields the $\text{Bin}(C, p)$ distribution. So, for this special case, we do not really use an approximation for the random variable N. Further, we use a refined version of the moment-iteration algorithm in the sense that we fit a distribution on the first two moments of $(X_k | X_k > 0) - 1$ instead of $X_k + N$. In this way, we reduce the impact of the probability that the queue is empty at the start of a cycles.

Step 1. Initialisation. Set $\text{E}\{X_1\} = \text{E}\{X_1^2\} = 0$, so that $\sigma_1 = 0$ and $X_2 = \max\{0, N - A\}$. Compute $\Pr\{X_2 > 0\}$ and the first two moments of X_2 from relation (2.23). Set $k := 2$.

Step 2. Iteration.

(*i*) Set $Y_k \stackrel{d}{=} (X_k | X_k > 0) - 1$, and compute the first two moments of Y_k.

(*ii*) Fit a tractable distribution to the first two moments of Y_k according to the procedure in Theorem 2.5.2.

(*iii*) Compute $\Pr\{X_{k+1} > 0\}$ and the mean and standard deviation of X_{k+1} from the relation

$$X_{k+1} = \max\{0, N - A\}\Pr\{X_k = 0\} + \max\{0, 1 + Y_k + N - A\}\Pr\{X_k > 0\}.$$

Step 3. If both $|\mathrm{E}\{X_{k+1}\} - \mathrm{E}\{X_k\}|$ and $|\sigma_{k+1} - \sigma_k|$ are sufficiently small, then execute Step 4. Otherwise, set $k := k + 1$ and repeat Step 2.

Step 4. Approximate the probability of a non-empty queue by $\Pr\{X_{k+1} > 0\}$ and the mean and standard deviation of the imbedded queue-length distribution by $\mathrm{E}\{X_{k+1}\}$ and σ_{k+1}.

In Adan, Van Eenige & Resing [1995], the discrete fits have been developed and used for this refined version. This moment-iteration technique shows excellent performance for the examples of Section 2.3, as can be seen in Table 2.5. Furthermore, the computational effort is rather small.

	$\Pr\{X > 0\}$		$\mathrm{E}\{X\}$	
C	Exact	MI	Exact	MI
10	0.5653	0.5655	1.7769	1.7768
20	0.4905	0.4909	1.5693	1.5692
40	0.3952	0.3955	1.3004	1.3002
80	0.2851	0.2852	0.9775	0.9773

Table 2.5: *Exact results and results of the moment-iteration technique (MI) for the examples with* $A = 0.2C$ *and* $p = 0.17$.

Further, we remark that once we start with a random variable X_1 which satisfies the condition of Theorem 2.5.1, then during the iteration step it is not possible to obtain pairs $(\mathrm{E}\{X_k\}, c^2_{X_k})$, which belong to the shaded region in Figure 2.7.

From the approximated stationary distribution for the number of customers at the start of the cycle, we can compute the queue-length distribution at the other slot boundaries in the cycle similarly as described at the end of Section 2.4.

2.6 Conclusions

In this chapter, we used two analytical techniques to study a queueing system with periodic service, namely, the method of particular solutions and the generating-function technique. As we saw, these techniques unfortunately face two numerical problems. The first problem is the determination of the roots of a (polynomial) equation. The second problem results from the (possible) clustering of roots. Except for some specific cases, it is not clear under what conditions and properties of the model these numerical problems do and do not occur. Since these analytical techniques are not generally applicable, we presented two numerical approaches, which appear to be numerically stable and which give excellent results for this queueing system.

The first numerical technique (GT technique, for short) exploits the geometric tail behaviour of the stationary imbedded queue-length distribution. The second numerical technique (the MI technique) utilises the link between the Markov chain and the $D/G/1$ queueing system by approximating the limiting solution of Lindley's equation for this latter system by a moment-iteration approach.

The GT technique can be applied when the equilibrium equations of a one-dimensional Markov chain constitute a system of homogeneous linear difference equations with constant coefficients. As we shall see, the equilibrium equations of a Markov chain describing the imbedded queue-length process of a broad class of queueing systems with periodic service form such a difference equation. Therefore, we shall explore the GT technique further in Chapter 3. We shall apply it to queueing systems with periodic service in Chapter 4 and to modifications of these systems in Chapter 6 and Chapter 7.

The MI technique is applicable to an even broader class of queueing systems with periodic service than the GT technique, because the MI technique only requires the first two moments of the service time instead of the complete distribution. This technique will be explored further and applied to these systems in Chapter 5. In Chapter 6 and Chapter 7, we shall apply this technique to approximate the performance measures of modifications of these systems.

3

A Numerical Technique for a Class of One-Dimensional Markov Chains

3.1 Introduction

In Chapter 2, we proposed two numerical techniques for analysing the queue-length process for a queueing system with periodic service. This analysis, after imbedding the queue-length process at the start of cycles, reduces to the determination of the stationary distribution of a Markov chain with a one-dimensional state space. The first technique (that is, the geometric tail (GT) technique) utilises the geometric tail behaviour of this distribution. The other technique (that is, the moment-iteration (MI) technique) exploits the observation that this Markov chain has the same structure as the waiting-time process of customers in a special discrete $D/G/1$ queueing system. More precisely, the MI technique is a moment-iteration method for determining the limiting solution of Lindley's equation for this process and, hence, for approximating the stationary distribution. This technique will be explored further in Chapter 5.

The present chapter is devoted to the GT technique. In Chapter 2, the asymptotically geometric tail of the stationary distribution resulted from the fact that the equilibrium equations of the Markov chain, describing the imbedded queue-length process, constitute a homogeneous linear difference equation with constant coefficients. As we shall see in Chapter 4, the equilibrium equations of the Markov chain, describing this process, form such a difference equation for a broad class of queueing systems with periodic service. So, it seems obvious to use techniques from the theory of difference equations to solve these equations.

The use of analytical techniques from the theory of difference equations for solving the equilibrium equations of Markov chains is very natural as these equations often constitute a homogeneous linear difference equation with constant coefficients. Therefore, these techniques are applied in many papers and books to a variety of problems that require the determination of the stationary distribution of Markov chains. In fact, A.A. Markov, the initiator of the study of these chains, used such techniques

already (cf. Romanovsky [1970]). By these techniques, we can in principle determine the stationary distribution and its tail behaviour. However, since the main objective of this monograph is to develop techniques for analysing queueing systems with periodic service, we are not only interested in how to determine the stationary (imbedded) queue-length distribution in principle, but also in the usefulness of these techniques for computational purposes.

As is well known from the theory of difference equations, the solution of a homogeneous linear difference equation of finite order with constant coefficients is a linear combination of powers. There are two standard analytical techniques for the determination of this linear combination. One is (what Feller [1968] calls) the method of particular solutions, and the other is the generating-function technique. These two techniques are closely related, and, of course, they give the same results. Both techniques basically consist of the following two steps. Firstly, the roots of an equation, related to the constant equations, have to be determined. These roots are then used to solve a system of linear equations, related to the boundary equations, to obtain the coefficients of the linear combination. Unfortunately, (application of) these analytical techniques can lead to numerical difficulties, as we saw in Chapter 2. Further, even if the parameters can be computed accurately, the form of the solution (that is, a linear combination of geometric distributions) may easily cause loss of accuracy when it is implemented, because terms with opposite signs and of very different magnitude may have to be added up.

Except for some special cases, it is unclear under what conditions and properties of the model these techniques are useful from a computational point of view. For example, if the degree of a polynomial is high, then it is more likely that the accurate determination of all zeroes is harder and that these zeroes are clustered. But whether they are indeed clustered can only be observed and cannot be predicted. Since we are interested in techniques that are generally applicable to solve equilibrium equations of Markov chains, which constitute a homogeneous linear difference equation with constant coefficients, we use a numerical approach.

Nevertheless, the analytical techniques provide structural results, and these results can be exploited for computational purposes. For Markov chains with equilibrium equations constituting a homogeneous linear difference equation with constant coefficients, the stationary distribution is a linear combination of a finite number of geometric distributions (some of them possibly with a parameter which is complex). Hence, the tail behaviour of this distribution is determined by the geometric distribution(s) with the largest parameter in absolute value. The parameter of one of these geometric distributions with the largest parameter in absolute value is positive and can be computed accurately. We exploit this observation in a similar way as in Tijms & Van de Coevering [1991].

The outline of this chapter is as follows. In Section 3.2, we give a formal description of the Markov chains for which the equilibrium equations form a homogeneous linear difference equation with constant coefficients. In Section 3.3, we use the two analytical techniques mentioned to show that the stationary distribution of these chains can be expressed as a linear combination of a finite number of geometric distributions. If the number of boundary equations of the difference equation is larger than the number of geometric distributions of the linear combination, then a finite number of stationary probabilities do not have this form. These stationary probabilities belong to states at the boundary of the state space.

Since the two analytical techniques may lead to computational difficulties, we present in Section 3.4 the GT technique to approximate the stationary distribution. This technique exploits the tail behaviour of the stationary distribution. This tail behaviour is asymptotically geometric, and it is completely determined by the unique positive root inside the unit circle of a non-linear equation. This

equation is related to the constant equations, and it is known as the characteristic equation of the difference equation. Moreover, this root can be determined accurately in a stable way without numerical difficulties. This circumstance is then exploited by an adapted version of the numerical approach in Tijms & Van de Coevering [1991].

The performance of the GT technique depends on the rate at which this largest root dominates the linear combination of geometric distributions. In other words, the performance depends on the quotient of the largest and second-largest root in absolute value of the characteristic equation. To get an idea of this quotient and of how fast the largest root dominates this linear combination, we study the queue-length process of the $GI/E_r/1$ queueing system in Section 3.5. For this system, Adan & Zhao [1994] derive sufficient conditions for the method of particular solutions to be successfully applicable. Numerical examples show that the largest root dominates for states fairly close to the boundary of the state space already. As a result, the computational effort of the GT technique, to obtain accurate results, appears to be fairly low.

Finally, Section 3.6 gives a summary of this chapter and looks ahead at the subsequent chapters in which the GT technique will be used to study queueing systems with periodic service and modifications of these systems.

3.2 The Markov chain and the equilibrium equations

Let the stochastic process $\{X_k, k = 1, 2, 3, \ldots\}$ be a discrete-time Markov chain. We assume that the state space of this chain consists of the non-negative integers $\{0, 1, 2, \ldots\}$ and that $X_1 = n$, with n a (possibly random) non-negative integer. For $i, j = 0, 1, 2, \ldots$, the stationary transition probability from state i to state j of the Markov chain is denoted by $p_{i,j}$, that is, for all $k = 1, 2, 3, \ldots$,

$$p_{i,j} := \Pr\{X_{k+1} = j | X_k = i\}, \qquad i, j, = 0, 1, 2, \ldots .$$

This chain is assumed to be irreducible and aperiodic. Moreover, we assume that it is positive recurrent by supposing that (cf. Pakes [1969])

$$\sum_{j=0}^{\infty} j p_{i,j} < \infty, \qquad i = 0, 1, 2, \ldots , \tag{3.1}$$

and that

$$\limsup_{i \to \infty} \mathrm{E}\{X_{k+1} - X_k | X_k = i\} = \limsup_{i \to \infty} \sum_{j=0}^{\infty} (j - i) p_{i,j} < 0. \tag{3.2}$$

These assumptions imply that the Markov chain is ergodic, so that it has a unique stationary distribution $\{\pi_j, j = 0, 1, 2, \ldots\}$.

For the Markov chain in Chapter 2, inequality (3.1) holds, because in a cycle at most C slots of work can arrive. For this chain, the inequality (3.2) is fulfilled as well, since, for $i \geq C$,

$$\mathrm{E}\{X_{k+1} - X_k | X_k = i\} = Cp - A < 0,$$

because we assumed that the average number of slots of work arriving in a cycle is strictly less than the average service capacity of the server per cycle.

The stationary distribution of the Markov chain is the unique solution of the equilibrium equations, which are in general

$$\pi_j = \pi_0 p_{0,j} + \pi_1 p_{1,j} + \pi_2 p_{2,j} + \cdots, \qquad j = 0, 1, 2, \ldots, \tag{3.3}$$

and the normalisation equation

$$\sum_{j=0}^{\infty} \pi_j = 1. \tag{3.4}$$

In this chapter, we suppose that the equilibrium equations constitute a homogeneous linear difference equation with constant coefficients. For this purpose, we assume the following structure for the Markov chain.

Firstly, the transition probabilities $p_{i,j}$ depend on i and j only through their difference $j - i$, for $i \geq D$ and $j \geq 1$, with D a fixed non-negative integer. Secondly, the jump sizes from state i, with $i \geq D$, to higher states are uniformly bounded by some constant T_H, with T_H a positive integer (that is, the jumps to the right are limited). In other words, the highest state that can be reached by a single transition from a state i, with $i \geq D$, is state $i + T_H$, so that $p_{i,j} = 0$, for $j > i + T_H$. The subscript H is used to indicate transitions to higher states, since, in Section 3.3.2, we shall consider the case that the jumps to lower states are also uniformly bounded by some constant, denoted by T_L. Finally, for states $i < D$, the highest state that can be reached by a single transition from this state i is state $D + T_H - 1$. To sum up, we make the following assumptions

$$\begin{aligned}
&(i) && p_{i,j} =: q_{j-i}, && i \geq D \text{ and } j \geq 1,\\
&(ii) && q_h = 0, && h > T_H,\\
&(iii) && q_{T_H} > 0,\\
&(iv) && \sum_{h=-\infty}^{T_H} q_h = 1,\\
&(v) && p_{i,0} = 1 - \sum_{h=-\infty}^{-i} q_h, && i \geq D,\\
&(vi) && p_{i,j} = 0, && i < D \text{ and } j \geq D + T_H.
\end{aligned}$$

Notice that the inequality (3.1) is now automatically fulfilled, for $i = 0, 1, \ldots, D-1$, and that it reads, for $i \geq D$,

$$\sum_{j=1}^{i+T_H} j q_{j-i} < \infty.$$

Further, the inequality (3.2) reduces to

$$\sum_{h=-\infty}^{T_H} h q_h < 0. \tag{3.5}$$

To illuminate the structure of the Markov chain, we display the transition matrix $\boldsymbol{P}$ of this chain

$$\boldsymbol{P} = \begin{pmatrix}
p_{0,0} & p_{0,1} & p_{0,2} & \cdots & p_{0,D+T_H-1} & 0 & 0 & 0 & 0 & \cdots \\
p_{1,0} & p_{1,1} & p_{1,2} & \cdots & p_{1,D+T_H-1} & 0 & 0 & 0 & 0 & \cdots \\
p_{2,0} & p_{2,1} & p_{2,2} & \cdots & p_{2,D+T_H-1} & 0 & 0 & 0 & 0 & \cdots \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\
p_{D-1,0} & p_{D-1,1} & p_{D-1,2} & \cdots & p_{D-1,D+T_H-1} & 0 & 0 & 0 & 0 & \cdots \\
p_{D,0} & q_{1-D} & q_{2-D} & \cdots & q_{T_H-1} & q_{T_H} & 0 & 0 & 0 & \cdots \\
p_{D+1,0} & q_{-D} & q_{1-D} & \cdots & q_{T_H-2} & q_{T_H-1} & q_{T_H} & 0 & 0 & \cdots \\
p_{D+2,0} & q_{-1-D} & q_{-D} & \cdots & q_{T_H-3} & q_{T_H-2} & q_{T_H-1} & q_{T_H} & 0 & \cdots \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \ddots &
\end{pmatrix}.$$

From this structure of the Markov chain, the equilibrium equations (3.3) can be partitioned as follows

$$\pi_0 = \pi_0 p_{0,0} + \pi_1 p_{1,0} + \pi_2 p_{2,0} + \cdots, \tag{3.6}$$

$$\pi_j = \pi_0 p_{0,j} + \cdots + \pi_{D-1} p_{D-1,j} + \pi_D q_{j-D} + \pi_{D+1} q_{j-(D+1)} + \cdots, \quad 1 \le j < D + T_H, \tag{3.7}$$

$$\pi_j = \pi_{j-T_H} q_{T_H} + \pi_{j-(T_H-1)} q_{T_H-1} + \pi_{j-(T_H-2)} q_{T_H-2} + \cdots, \quad j \ge D + T_H. \tag{3.8}$$

As in Chapter 2, we call the equations (3.6) and (3.7) the boundary equations and the equations (3.8) the inner equations. The constant structure of the inner equations is exploited in the next sections.

As we have already shown, the equilibrium equations (2.2) of the Markov chain in Chapter 2 can be partitioned in this way, namely, the boundary equations (2.6) and (2.7) and the inner equations (2.8). For this chain, we clearly have $D = 0$, $T_H = C - A$, and $q_h = \alpha_{h+A}$, for $h = -A, -A+1, \ldots, C-A$, and $q_h = 0$ otherwise. In the next chapter, we shall show that the equilibrium equations of the Markov chain, describing the imbedded queue-length process, can be partitioned in a similar way for a broad class of queueing systems with periodic service. Further, for many other queueing systems, the equilibrium equations of the Markov chain, describing the queue-length process, possess this structure. A nice example is the number of uncompleted service phases just prior to an arrival in a $GI/Ph/1$ queueing system with service times having a distribution that is mixture of Erlang distributions with the same scale parameter μ.

3.3 Solving the equilibrium equations analytically

In this section, we solve the equilibrium equations (3.6), (3.7), and (3.8), and the normalisation equation (3.4) analytically by exploiting that the equations (3.8) constitute a homogeneous linear difference equation with constant coefficients. To solve such a difference equation, there are two standard and widely used techniques.

Firstly, we can seek solutions of the form $\pi_j = z^j$ and try to linearly combine these solutions to satisfy the boundary equations and the normalisation equation. In other words, we try to express the stationary distribution as a linear combination of geometric distributions. Adopting the terminology of Feller [1968], we call this approach the method of particular solutions. This technique is discussed in Section 3.3.1. Notice that the order of the difference equation is not necessarily finite. The reason that the method of particular solutions can cope with an infinitely large order is that we only use solutions

$\pi_j = z^j$ with $|z| < 1$, because otherwise equation (3.4) cannot be satisfied, and the number of these solutions is finite.

In the literature, the method of particular solutions has been applied to Markov chains and queueing systems by, for example, Conolly [1958a,1958b], Morse [1958], Feller [1968], Romanovsky [1970], and Gross & Harris [1974]. This method is also exposed in many books on the theory of difference equations.

The second technique we can use to solve the equilibrium equations and the normalisation equation is the generating-function technique. Section 3.3.2 is devoted to this technique.

The generating-function technique is applied more widely to analyse Markov chains and queueing systems than the method of particular solutions. Applications to Markov chains and queueing systems of this technique are given by, for example, A.A. Markov (cf. Romanovsky [1970]), Bailey [1954a], Luchak [1958], Prabhu [1965], Giffin [1975], and Cohen [1982]. This technique is also widely used in books and papers on the theory of difference equations.

Before presenting these two techniques, we note that Winsten [1959] studies the Markov chain with transition matrix $\boldsymbol{P}$ in Section 3.2, for which $T_H = 1$. By sample-path arguments, he shows that the stationary distribution of this chain is geometric, except for the stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{D-1}$.

3.3.1 The method of particular solutions

In this section, we try to express the stationary distribution of the Markov chain as a linear combination of geometric distributions directly. We begin with treating the inner equations (3.8) to seek solutions of the form $\pi_j = z^j$. After that, we linearly combine these solutions to satisfy the boundary equations (3.6) and (3.7), and the normalisation equation (3.4). Further, we give for a special case the coefficients of this linear combination explicitly. Finally, we mention a recently obtained result that extends the idea of the method of particular solutions to higher dimensional Markov chains.

3.3.1.1 Solving the inner equations

To solve the inner equations (3.8) by the method of particular solutions, we first substitute the form $\pi_j = z^j$ into these equations, and then we divide these equations by z^{j-T_H} to obtain

$$z^{T_H} = q_{T_H} + q_{T_H-1}z + q_{T_H-2}z^2 + \cdots . \tag{3.9}$$

This equation is called the characteristic equation of the difference equation.

We only use solutions z in the complex plane of the characteristic equation (3.9) with $|z| < 1$, since otherwise the normalisation equation (3.4) cannot be satisfied. The next lemma states that, if the Markov chain is ergodic, this characteristic equation has exactly T_H solutions z inside the unit circle. This lemma can be found in Takács [1962], and it is proved by using Rouché's Theorem.

Theorem 3.3.1 (Rouché's Theorem)
If the functions $f(z)$ and $g(z)$ are analytic inside and on a closed contour C, and $|g(z)| < |f(z)|$ on C, then $f(z)$ and $f(z) + g(z)$ have the same number of zeroes inside C.

Proof. See, for instance, pages 116-117 in Titchmarsh [1960]. □

Lemma 3.3.1 *Equation (3.9) has exactly T_H roots inside the unit circle, if (3.5) holds.*

Proof. Let $f(z) := z^{T_H}$ and $g(z) := q_{T_H} + q_{T_H-1}z + q_{T_H-2}z^2 + \cdots$. Since the Markov chain is positive recurrent, we have the inequality (3.5), so that

$$g'(1) = \sum_{h=-\infty}^{T_H} (T_H - h)q_h = T_H - \sum_{h=-\infty}^{T_H} hq_h > T_H = f'(1).$$

Hence, since $f(1) = g(1)$, it holds for all sufficiently small $\varepsilon > 0$ that

$$f(1-\varepsilon) > g(1-\varepsilon). \tag{3.10}$$

Fix some $\varepsilon > 0$, satisfying inequality (3.10). Then, for all z with $|z| = 1 - \varepsilon$, we have by the triangle inequality that

$$\begin{aligned} |g(z)| &\leq q_{T_H} + q_{T_H-1}|z| + q_{T_H-2}|z^2| + \cdots \\ &= q_{T_H} + q_{T_H-1}(1-\varepsilon) + q_{T_H-2}(1-\varepsilon)^2 + \cdots = g(1-\varepsilon), \\ |f(z)| &= |z^{T_H}| = (1-\varepsilon)^{T_H} = f(1-\varepsilon). \end{aligned}$$

Thus, together with inequality (3.10), we have

$$|-g(z)| = |g(z)| < |f(z)|, \quad |z| = 1 - \varepsilon.$$

Applying Rouché's Theorem (cf. Theorem 3.3.1) to the circle $|z| = 1 - \varepsilon$ yields that equation (3.9) has exactly T_H roots inside this circle. Finally, letting ε tend to zero completes the proof. □

The characteristic equation (3.9) for the Markov chain in Chapter 2 reduces to

$$z^{C-A} = \alpha_C + \alpha_{C-1}z + \cdots + \alpha_0 z^C,$$

with $\{\alpha_h, h = 0, 1, \ldots, C\}$ the binomial distribution with C the number of trials and p the success probability (see equation (2.10)). Lemma 3.3.1 states that this equation has exactly $C - A$ roots inside the unit circle, so that this lemma proves Lemma 2.3.1.

Let the T_H solutions inside the unit circle of the characteristic equation (3.9) be denoted by $z_1, z_2, \ldots, z_{T_H}$. By the linearity of the inner equations (3.8), any linear combination of the solutions z_k^j is a solution of these equations as well. So, we try to satisfy the boundary equations (3.6) and (3.7), and the normalisation equation (3.4) by linearly combining these solutions. Since the solutions z_k may not be all distinct, the number of independent solutions for this combination may be less than T_H. The next lemma shows how to construct T_H independent solutions from the T_H roots inside the unit circle of the characteristic equation. This lemma is a well-known result from the theory of difference equations.

Lemma 3.3.2 *Let equation (3.9) have the K distinct roots inside the unit circle $\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_K$, with $K \leq T_H$, and let root $\hat{z}_k$ have multiplicity m_k, so that $\sum_{k=1}^{K} m_k = T_H$. Then, the K sequences with elements $\hat{z}_k^j, j\hat{z}_k^j, \ldots, j^{m_k-1}\hat{z}_k^j$, with $k = 1, 2, \ldots, K$, form a system of T_H independent solutions for the difference equation (3.8).*

Proof. See, for instance, Theorem 5.3 in Henrici [1968]. □

For notational convenience, we assume that the solutions $z_1, z_2, \ldots, z_{T_H}$ are distinct. According to Feller [1968], these solutions are distinct indeed for most practical cases. So, we have T_H independent solutions z_k^j, which fulfil the inner equations (3.8), and any linear combination of these solutions

$$\pi_j = \sum_{k=1}^{T_H} \lambda_k z_k^j, \tag{3.11}$$

is a solution of the inner equations as well, where the λ_k's denote complex numbers. Hence, we represent the stationary distribution as a linear combination of geometric distributions with parameters z_k. In the next section, we use this representation to satisfy the boundary equations (3.6) and (3.7), and the normalisation equation (3.4).

3.3.1.2 Solving the boundary equations and the normalisation equation

Before using the representation (3.11) for solving the boundary equations (3.6) and (3.7), and the normalisation equation (3.4), we make the following observation. By seeking solutions of the form z^j satisfying the inner equations, we impose this form not only on the stationary probabilities π_j, with $j \geq D + T_H$, but also on the T_H stationary probabilities $\pi_D, \pi_{D+1}, \ldots, \pi_{D+T_H-1}$, since these latter probabilities appear in the inner equations (3.8). However, the equilibrium equations for the corresponding states j, with $j = D, D+1, \ldots, D+T_H-1$, belong to the boundary equations (3.6) and (3.7). In general, the solutions z_k^j do not satisfy the equilibrium equations for these states. Hence, the T_H coefficients λ_k have to be used to satisfy these equations. Finally, we notice that the stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{D-1}$ do not appear in the inner equations, so that we did not impose the form z^j on these probabilities. Therefore, these stationary probabilities can be regarded as unknowns, which have to be the solution of the equilibrium equations and the normalisation equation.

The representation (3.11) is now used to satisfy the boundary equations (3.6) and (3.7), and the normalisation equation (3.4). We substitute this form, for $j \geq D$, into the boundary equations (3.7) and treat, if $D > 0$, the stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{D-1}$ as unknowns. Thus, for the moment, we omit the equilibrium equation for state $j = 0$. Recall that the form (3.11) already satisfies the inner equations (3.8). This system of $D + T_H - 1$ equations (3.7) (after substituting the form (3.11)) is a system of $D + T_H - 1$ homogeneous linear equations with the D unknown stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{D-1}$ and the T_H unknown coefficients λ_k. So, this system has a non-trivial (that is, a non-null) solution.

Consider a non-trivial solution of the boundary equations (3.7) for the coefficients λ_k and the stationary probabilities π_j, for $j = 0, 1, \ldots, D-1$. Then, since adding the equations (3.7) and (3.8) gives the equilibrium equation for state $j = 0$, this solution automatically satisfies the equation (3.6). Thus, we have a solution of the equilibrium equations with

$$\sum_{j=0}^{\infty} |\pi_j| = \sum_{j=0}^{\infty} \sum_{k=1}^{T_H} |\lambda_k z_k^j| < \infty,$$

because $|z_k| < 1$, for all k. So, by a Foster's criterion (cf. Foster [1953]), this solution, after normalising it by using (3.4), is unique and, hence, it is the stationary distribution. We note that this implies that the boundary equations (3.7) and the normalisation equation (3.4) form a system of D linearly independent equations. Consequently, the coefficients λ_k and the stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{D-1}$ can be determined directly from solving the system of equations (3.7) (after substituting the form (3.11)) and the normalisation equation (3.4).

3.3.1.3 An explicit solution for the coefficients if $D = 0$

For the special case that the roots z_k are distinct and $D = 0$, the coefficients λ_k can even be given explicitly. It appears that these coefficients are the solution of a system of Vandermonde equations.

Define $\pi_i := 0$, for $i = -T_H + 1, -T_H + 2, \ldots, -1$. Then, it is easily verified that the equilibrium equations for the states j, with $j = 1, 2, 3, \ldots$, all have the constant structure

$$\pi_j = \pi_{j-\tau_H} q_{\tau_H} + \pi_{j-(\tau_H-1)} q_{\tau_H-1} + \pi_{j-(\tau_H-2)} q_{\tau_H-2} + \cdots .$$

From Section 3.3.1.1, we know that any linear combination of z_k^j, with $k = 1, 2, \ldots, T_H$, satisfies these equations. However, we also have the conditions that $\pi_j = 0$, for $j = -T_H + 1, -T_H + 2, \ldots, -1$, that is,

$$\pi_j = \sum_{k=1}^{T_H} \lambda_k z_k^j = 0, \qquad j = -T_H + 1, -T_H + 2, \ldots, 1. \tag{3.12}$$

To these equations, we add the normalisation equation

$$\sum_{j=0}^{\infty} \pi_j = \sum_{j=0}^{\infty} \sum_{k=1}^{T_H} \lambda_k z_k^j = \sum_{k=1}^{T_H} \frac{\lambda_k}{1 - z_k} = 1. \tag{3.13}$$

In the next lemma, we solve the system of equations (3.12) and (3.13) by Cramer's rule, from which it appears that this system can be transformed into a Vandermonde system, so that we can give its solution explicitly.

Lemma 3.3.3 *If the roots z_k, with $k = 1, 2, \ldots, T_H$, are distinct and $D = 0$, then the coefficient λ_k is given by*

$$\lambda_k = (1 - z_k) \frac{\prod_{\substack{j=1 \\ j \neq k}}^{T_H} (z_j^{-1} - 1)}{\prod_{\substack{j=1 \\ j \neq k}}^{T_H} (z_j^{-1} - z_k^{-1})}, \qquad k = 1, 2, \ldots, T_H.$$

Proof. For convenience, we define $\hat{\lambda}_k := \lambda_k/(1 - z_k)$ and $\tau_k := 1/z_k$, for $k = 1, 2, \ldots, T_H$, and write the system of T_H equations (3.12) and (3.13) in matrix notation as $A\hat{\Lambda} = \boldsymbol{b}$, with

$$A = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \tau_1 - 1 & \tau_2 - 1 & \cdots & \tau_{T_H} - 1 \\ \tau_1(\tau_1 - 1) & \tau_2(\tau_2 - 1) & \cdots & \tau_{T_H}(\tau_{T_H} - 1) \\ \vdots & \vdots & & \vdots \\ \tau_1^{T_H-2}(\tau_1 - 1) & \tau_2^{T_H-2}(\tau_2 - 1) & \cdots & \tau_{T_H}^{T_H-2}(\tau_{T_H} - 1) \end{pmatrix},$$

the column vector $\hat{\Lambda}$ consisting of the coefficients $\hat{\lambda}_k$, and $\boldsymbol{b}$ the column vector with first entry one and the other entries equal to zero. We now solve this system of equations by Cramer's rule.

By Cramer's rule, we have

$$\hat{\lambda}_k = \frac{\det(A_k)}{\det(A)}, \qquad k = 1, 2, \ldots, T_H.$$

where the matrix A_k denotes the matrix A for which the k-th column is replaced by the column vector b. We first determine the determinant of the matrix A, and then the determinant of the matrix A_k. Both determinants appear to be Vandermonde determinants.

For the matrix A, we add to each row k, for $k = 2, 3, \ldots, T_H$, the rows l with $l < k$, to obtain the Vandermonde matrix

$$V := \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \tau_1 & \tau_2 & \cdots & \tau_{T_H} \\ \tau_1^2 & \tau_2^2 & \cdots & \tau_{T_H}^2 \\ \vdots & \vdots & & \vdots \\ \tau_1^{T_H-1} & \tau_2^{T_H-1} & \cdots & \tau_{T_H}^{T_H-1} \end{pmatrix}.$$

By addition of rows, the determinant of the matrix A is transformed into a Vandermonde determinant, so that (see, for instance, Bellman [1970])

$$\det(A) = \det(V) = \prod_{1 \le i < j \le T_H} (\tau_j - \tau_i). \tag{3.14}$$

To compute the determinant of the matrix A_k, we expand this matrix on the k-th column. So,

$$\det(A_k) = (-1)^{k+1} \prod_{\substack{j=1 \\ j \neq k}}^{T_H} (\tau_j - 1) \begin{vmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 \\ \tau_1 & \cdots & \tau_{k-1} & \tau_{k+1} & \cdots & \tau_{T_H} \\ \vdots & & \vdots & \vdots & & \vdots \\ \tau_1^{T_H-2} & \cdots & \tau_{k-1}^{T_H-2} & \tau_{k+1}^{T_H-2} & \cdots & \tau_{T_H}^{T_H-2} \end{vmatrix}. \tag{3.15}$$

Since the determinant at the right-hand side of (3.15) is a Vandermonde determinant, combining this observation with (3.14) yields, for $k = 1, 2, \ldots, T_H$,

$$\hat{\lambda}_k = (-1)^{k+1} \frac{\prod_{\substack{j=1 \\ j \neq k}}^{T_H} (\tau_j - 1) \prod_{\substack{1 \le i < j \le T_H \\ i,\, j \neq k}} (\tau_j - \tau_i)}{\prod_{1 \le i < j \le T_H} (\tau_j - \tau_i)} = \frac{(-1)^{k+1} \prod_{\substack{j=1 \\ j \neq k}}^{T_H} (\tau_j - 1)}{\prod_{1 \le i < k \le T_H} (\tau_k - \tau_i) \prod_{1 \le k < j \le T_H} (\tau_j - \tau_k)}$$

$$= \frac{\prod_{\substack{j=1 \\ j \neq k}}^{T_H} (\tau_j - 1)}{\prod_{\substack{j=1 \\ j \neq k}}^{T_H} (\tau_j - \tau_k)},$$

so that the proof is complete. □

In Section 3.3.2, we use the generating-function technique for solving the equilibrium equations of the Markov chain. Before that, we mention a recent extension of the idea of the method of particular solutions to solve the equilibrium equations of a class of higher dimensional Markov chains.

3.3.1.4 The compensation approach

The idea of the method of particular solutions for solving the equilibrium equations is to try to express the stationary distribution as a linear combination of powers. More specifically, we look for solutions $\pi_j = z^j$ of the inner equations with $|z| < 1$, and then we try to linearly combine these solutions to satisfy the boundary equations and the normalisation equation. Adan [1991] extends this idea to the determination of the stationary distribution of a class of Markov chains with a two-dimensional state space, which is denumerable in both dimensions. This technique is appropriately called the compensation approach, and the basic idea for this approach stems from numerical studies for the stationary distribution. Van Houtum [1995] extends this approach to a class of Markov chains with a higher dimensional state space. For completeness, we remark that the results of Adan [1991] have recently been interpreted in properties of generating functions by Cohen [1994].

For the compensation approach, the starting point is again the partitioning of the equilibrium equations into inner and boundary equations. The inner equations lead to a set of powers, which are linearly combined to satisfy the boundary equations and the normalisation equation. The difference with the method of particular solutions is that this set contains infinitely many powers. Furthermore, the method of particular solutions uses all powers of this set, whereas the compensation approach uses only a subset of the set of powers. Finally, the compensation approach selects these powers in a special way. For a comprehensive exposition of the compensation approach we refer to the aforementioned authors and to Adan, Wessels & Zijm [1993].

3.3.2 The generating-function technique

From Section 3.3.1, we know that the stationary distribution of the Markov chain is a linear combination of powers. Hence, the probability generating function of this distribution is a rational function. However, when transitions to lower states are not uniformly bounded by some constant, the use of the generating-function technique seems to be rather difficult for solving the equilibrium equations of this chain. To simplify the analysis, we make the following additional assumption. For $i \geq D$, the jump sizes from state i to lower states are uniformly bounded by a constant, just like the jump sizes to higher states. More precisely, we assume that the lowest state that can be reached by a single transition from state i is state $i - T_L$, with T_L a fixed positive integer, so that $p_{i,j} = q_{j-i} = 0$, for $j - i < -T_L$. Thus, the equilibrium equations (3.3) now read

$$\pi_0 = \pi_0 p_{0,0} + \pi_1 p_{1,0} + \cdots + \pi_{\max\{D-1,T_L\}} p_{\max\{D-1,T_L\},0}, \tag{3.16}$$

$$\pi_j = \pi_0 p_{0,j} + \cdots + \pi_{D-1} p_{D-1,j} + \pi_D q_{j-D} + \cdots + \pi_{j+T_L} q_{-T_L}, \quad 1 \leq j < D + T_H, \tag{3.17}$$

$$\pi_j = \pi_{j-T_H} q_{T_H} + \pi_{j-(T_H-1)} q_{T_H-1} + \cdots + \pi_{j+T_L} q_{-T_L}, \quad j \geq D + T_H. \tag{3.18}$$

We notice that, to determine the T_H solutions inside the unit circle of the characteristic equation (3.9) numerically, we sometimes make this additional assumption implicitly, since the right-hand side of this equation may have to be truncated. Further, note that we have $T_L = A$ for the Markov chain in Chapter 2.

Since the Markov chain has a unique stationary distribution, the probability generating function

$$\Pi(z) := \sum_{j=0}^{\infty} \pi_j z^j$$

is well defined, for $|z| \leq 1$. Moreover, this function is uniquely determined by the stationary distribution $\{\pi_j, j = 0, 1, 2, \ldots\}$ and vice versa. Let $Q(z)$ denote the shifted probability generating function of the probability distribution $\{q_h, h = -T_L, -T_L + 1, \ldots, T_H\}$. More specifically,

$$Q(z) := z^{T_L} \sum_{h=-T_L}^{T_H} q_h z^h.$$

To apply the generating-function technique, we multiply the equilibrium equation (3.18) for state j by z^j, with $j \geq D + T_H$, and we add these equations. Then, we obtain after some algebra

$$\Pi(z) = \frac{z^{T_L}\left(\sum_{j=0}^{D+T_H-1} \pi_j z^j - \sum_{h=-T_L}^{T_H} \sum_{j=0}^{D+T_H-h-1} q_h \pi_j z^{h+j}\right)}{z^{T_L} - Q(z)}. \tag{3.19}$$

Since $\Pi(z)$ is a probability generating function, it is convergent inside and on the unit circle, so that the roots inside and on the unit circle of the denominator of (3.19) must also be roots of the numerator of (3.19), and with the same multiplicity. This denominator has $T_H + T_L$ roots z in the complex plane. The next lemma (cf. Takács [1962]) states that exactly T_L of these roots are inside and on the unit circle. These roots will be denoted by $\xi_{T_H+1}, \xi_{T_H+2}, \ldots, \xi_{T_H+T_L}$.

Lemma 3.3.4 *The function $z^{T_L} - Q(z)$ has exactly T_L roots inside and on the unit circle, if (3.5) holds.*

Proof. Like Lemma 3.3.1, we prove this lemma by Rouché's Theorem (cf. Theorem 3.3.1). Define $f(z) := z^{T_L}$ and $g(z) := Q(z)$. From the inequality (3.5), it follows that

$$g'(1) = \sum_{h=-T_L}^{T_H} (T_L + h) q_h = T_L + \sum_{h=-T_L}^{T_H} h q_h < T_L = f'(1).$$

Because $f(1) = g(1)$, it holds for all sufficiently small $\varepsilon > 0$ that

$$f(1 + \varepsilon) > g(1 + \varepsilon). \tag{3.20}$$

Fix some $\varepsilon > 0$, satisfying the inequality (3.20). Then, for all z with $|z| = 1 + \varepsilon$, we have by the triangle inequality that

$$\begin{aligned} |g(z)| &\leq q_{-T_L} + q_{-T_L+1}|z| + \cdots + q_{T_H}|z|^{T_H+T_L} = g(1 + \varepsilon), \\ |f(z)| &= |z|^{T_L} = (1 + \varepsilon)^{T_L} = f(1 + \varepsilon). \end{aligned}$$

Thus, together with the inequality (3.20),

$$|-g(z)| = |g(z)| < |f(z)|, \qquad |z| = 1 + \varepsilon.$$

Applying Rouché's Theorem to the circle $|z| = 1 + \varepsilon$ yields that $z^{T_L} - Q(z)$ has exactly T_L roots inside or on this circle. Finally, letting ε tend to zero completes the proof. □

So, the denominator of (3.19) has T_H zeroes outside the unit circle. As can be verified, these zeroes are the reciprocals of the zeroes inside the unit circle of the characteristic equation (3.9), when the additional assumption is made that jump sizes to lower states are bounded by T_L.

For the probability generating function $\Pi(z)$ to be convergent inside and on the unit circle, the numerator of (3.19) must have the same zeroes as the denominator of (3.19) and with the same multiplicity, for $|z| \leq 1$. These conditions are also known as regularity conditions for this probability generating function. Clearly, $z = 1$ is a root of the denominator, and it can be verified that this root has multiplicity one. Thus, this root, ξ_{T+1}, say, is also a root of the numerator of (3.19), irrespective of the values of π_j, for $j = 0, 1, \ldots, D + T_H - 1$. So, the root ξ_{T_H+1} does not restrict the values of these stationary probabilities at all. Substituting the other roots into this numerator and using that ξ_k is a root of the denominator give a system of $T_L - 1$ homogeneous linear equations with $D + T_H$ unknowns.

Further, the stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{D+T_H-1}$ have to satisfy the boundary equations (3.16) and (3.17) and the normalisation equation as well. The normalisation equation is given by

$$\lim_{z \uparrow 1^-} \Pi(z) = 1.$$

Thus, we have $T_L - 1$ regularity conditions for the probability generating function $\Pi(z)$, the $D + T_H$ boundary equations (3.16) and (3.17), and the normalisation equation, and we have the $D + T_H + T_L$ unknown stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{D+T_H+T_L-1}$ appearing in these equations. Since the Markov chain has a unique stationary distribution, we have that this system of equations has a unique solution for $\pi_0, \pi_1, \ldots, \pi_{D+T_H+T_L-1}$.

So, we have the unique stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{D+T_H+T_L-1}$ and the probability generating function $\Pi(z)$, which is convergent inside and on the unit circle. The final step for the determination of the stationary distribution is to invert this probability generating function.

We divide the numerator and denominator of (3.19) by $-q(T_H)$, and cancel out the common terms $z - \xi_k$, for $k = T_H + 1, T_H + 2, \ldots, T_H + T_L$. The resulting numerator $N(z)$ is a polynomial of degree $D + T_H - 1$, and the resulting denominator $D(z)$ is a polynomial of degree T_H. Since the degree of the polynomial of the denominator is not strictly larger than the degree of the polynomial of the numerator, for $D > 0$, we cannot directly expand the probability generating function $\Pi(z)$ into partial fractions.

To expand $\Pi(z)$ into partial fractions, we write the numerator $N(z)$ as $P(z)D(z) + R(z)$, where $P(z)$ is a polynomial of degree D and $R(z)$ a polynomial of degree less than T_H. So, we have

$$\Pi(z) = P(z) + \frac{R(z)}{D(z)}.$$

The quotient $R(z)/D(z)$ can now be expanded into partial fractions, so that

$$\frac{R(z)}{D(z)} = \sum_{k=1}^{T_H} \frac{\eta_k}{\xi_k - z}, \tag{3.21}$$

because the roots are assumed to be distinct (if root ξ has multiplicity $m > 1$, then we should have expanded this quotient into the partial fractions $1/(\xi - z), 1/(\xi - z)^2, \ldots, 1/(\xi - z)^{m-1}$).

The coefficients η_k of the partial-fraction expansion (3.21) can be determined explicitly (see, for instance, Jury [1964]). To determine the coefficient η_k, we multiply both sides of equation (3.21) by $z - \xi_k$ and let z tend to ξ_k, so that

$$\eta_k = \lim_{z \to \xi_k} (z - \xi_k) \frac{-R(z)}{D(z)} = \frac{-R(\xi_k)}{D'(\xi_k)} = \frac{-R(\xi_k)}{\prod_{\substack{j=1 \\ j \neq k}}^{T_H} (\xi_k - \xi_j)}.$$

Now, the probability generating function can be inverted, and from this inversion it is easily verified that the stationary distribution is a linear combination of geometric distributions, except for the stationary probabilities $\pi_0, \pi_1, \ldots, \pi_{D-1}$.

3.3.3 Numerical difficulties of the analytical techniques

To apply the method of particular solutions or the generating-function technique, we basically have to determine the roots of an equation first, and then we have to solve a system of linear equations into which these roots are substituted. Both the computation of these roots and the determination of the solution of this system can give computational difficulties, as we have observed in Chapter 2.

The determination of roots of an equation plays a fundamental role in the study of many queueing systems and Markov chains (see, for example, page 129 in Kleinrock [1975]). This determination is well known for its numerical instabilities (see, for instance, Press, Flannery, Teukolsky & Vetterling [1986]). In general, it is hard to determine the required roots accurately. It seems that these roots can be computed accurately only if the number of roots to be computed is small, if this equation can be reduced to solving a system of polynomial equations of low degree (as, for example, in Chapter 2), or if this equation can be reduced to solving a system contraction equations (as, for instance, in Adan & Zhao [1994]).

Further, even if we are able to compute all the required roots accurately, these roots may be closely clustered. However, we cannot link the clustering of roots to properties of the model, but we can only observe that roots are clustered. If the roots are clustered, then the system of equations into which these roots are substituted may become nearly singular (see, for instance, Table 2.1 and Table 2.2 in Chapter 2, Section 1.6 in Neuts [1981], and Press, Flannery, Teukolsky & Vetterling [1986]). Only for some specific cases, the solution of this system of equations may be expressed explicitly into the computed roots (see, for instance, Lemma 3.3.3).

In general, we cannot give (sufficient) conditions for the properties of the probability distribution $\{q_h, h = \ldots, T_H - 1, T_H\}$ such that the roots can be computed accurately and that these roots are not closely clustered. Therefore, in the next section, we present a numerical approach for determining the stationary distribution. This approach exploits the tail behaviour of the stationary distribution and is an adapted version of the algorithm of Tijms & Van de Coevering [1991].

3.4 Solving the equilibrium equations numerically

In this section, we present a numerical approach for the determination of the stationary distribution of the Markov chains as introduced in Section 3.2. This approach approximates the solution of the equilibrium equations (3.7) and (3.8), and the normalisation equation (3.4). Equation (3.6) is omitted, because this equation is redundant, as noted in Section 3.3. This approximation is based on exploiting the tail behaviour of the stationary distribution. For the determination of this tail behaviour, we use the notation and results from the method of particular solutions in Section 3.3.1. The reason for this is that this method is applicable to the class of Markov chains introduced in Section 3.2, whereas for the generating-function technique we restricted this class slightly. Furthermore, the method of particular solutions directly yields the desired form for the stationary distribution, that is, a linear combination of geometric distributions.

From the form (3.11) for the stationary probabilities π_j, it directly follows that the largest root

in absolute value of the characteristic equation (3.9) determines the tail behaviour of the stationary distribution. Note that this result also holds if the roots inside the unit circle of this equation are not all distinct. We first prove that there is exactly one positive root inside the unit circle of the characteristic equation, and then that this is the largest root in absolute value amongst the roots inside the unit circle. These results are well known in queueing theory, but they are usually stated without reference or proof. Further, this kind of results also appears in the theory of branching processes (see, for instance, Athreya & Ney [1972]).

Lemma 3.4.1 *Exactly one of the roots of equation (3.9) lies in the interval* $(0, 1)$.

Proof. To prove this lemma, we divide equation (3.9) by z^{T_H}, to get

$$1 = q_{T_H} z^{-T_H} + q_{T_H-1} z^{-T_H+1} + q_{T_H-2} z^{-T_H+2} + \cdots, \tag{3.22}$$

and we denote the right-hand side of this equation by $f(z)$. Since $f''(x) > 0$, for all $x > 0$, the function $f(x)$ is strictly convex on $(0, \infty)$. Hence, equation (3.22) has at most two positive solutions. Further, since

$$\lim_{x \downarrow 0} f(x) = \infty, \quad f(1) = \sum_{h=-\infty}^{T_H} q_h = 1, \quad \text{and} \quad f'(1) = \sum_{h=-\infty}^{T_H} -hq_h > 0,$$

where the limit results from $q_{T_H} > 0$ and the inequality from the inequality (3.5), the desired result follows immediately. □

This unique positive root of the characteristic equation (3.9) is denoted by z_1. The next lemma states that all the other roots inside the unit circle have absolute value at most z_1. But, before giving this lemma, we introduce d as the greatest common divisor of T_H and the powers of z having positive coefficients at the right-hand side of equation (3.9).

Lemma 3.4.2 *If* $d = 1$, *then* $|z_k| < z_1$, *for* $k = 2, 3, \ldots, T_H$. *Otherwise,* $|z_k| = z_1$, *for* $k = 2, 3, \ldots, d$, *and* $|z_k| < z_1$ *for* $k = d + 1, d + 2, \ldots, T_H$.

Proof. We first show that $|z_k| \leq z_1$, for $k = 2, 3, \ldots, T_H$. Let $f(z)$ denote the right-hand side of equation (3.22). By the triangle inequality and the strict convexity of $f(x)$ on $(0, \infty)$,

$$|f(z)| \leq f(|z|) < 1, \qquad z_1 < |z| < 1.$$

Hence, the equation $f(z) = 1$ has no roots z with $z_1 < |z| < 1$, so that $|z_k| \leq z_1$, for $k = 2, 3, \ldots, T_H$.

Next, we complete the proof for the case $d = 1$. By the triangle inequality,

$$f(|z_k|) = \sum_{h=-\infty}^{T_H} q_h |z_k|^{-h} \geq \left| \sum_{h=-\infty}^{T_H} q_h z_k^{-h} \right| = |f(z_k)| = 1, \qquad k = 2, 3, \ldots, T_H. \tag{3.23}$$

If equality holds in (3.23), then $|z_k| = z_1$, because the equation $f(|z|) = 1$ has only one solution on $|z| \in (0, 1)$, namely, $|z| = z_1$. Since $q_{T_H} > 0$ and because 1 is the greatest common divisor of the powers of $f(z)$, equality in (3.23) holds only if z_k^{-h} is real and positive for all h with $q_h > 0$. Suppose that $z_k^{-h} > 0$, for all h with $q_h > 0$, and so, assume that $|z_k| = z_1$. Then, we have that $z_k = \varphi z_1$ for some φ which is not equal to 1 and which satisfies $\varphi^{-h} = 1$ for all h with $q_h > 0$ (so, φ lies on the unit circle). However, since 1 is the greatest common divisor of the powers of $f(z)$, $\varphi^{-h} \neq 1$ for all h with

$q_h > 0$, except for $\varphi = 1$. Consequently, φz_1 is not a root of the equation $f(z) = 1$, and so the equality of (3.23) does not hold. Hence, we have proved for $d = 1$ that

$$|z_k| < z_1, \quad k = 2, 3, \ldots, T_H.$$

For the case $d > 1$, we can rewrite equation (3.9) by substituting $y = z^d$ as

$$y^{T_H/d} = q_{T_H} + q_{T_H-d}y + q_{T_H-2d}y^2 + \cdots. \tag{3.24}$$

By Lemma 3.3.1, this equation has T_H/d roots inside the unit circle, and these roots are denoted by $y_1, y_2, \ldots, y_{T_H/d}$. From Lemma 3.4.1 and the first part of the present lemma, we have

$$|y_k| < y_1, \quad k = 2, 3, \ldots, T_H/d.$$

The proof is completed by noticing that (since $y = z^d$) with each root y_k of equation (3.24), d roots z of equation (3.9) correspond, and these d roots have the same absolute value. □

From Lemma 3.4.1, Lemma 3.4.2, and the form (3.11), it is easily seen that the tail of the stationary distribution is asymptotically

$$\lim_{j\to\infty} \frac{\pi_{j+d}}{\pi_j} = z_1^d, \tag{3.25}$$

so that the unique positive root inside the unit circle of the characteristic equation (3.9) determines this tail behaviour completely. Thus, if $d = 1$, then the tail of the stationary distribution is asymptotically geometric in the strict sense. Although d may be larger than one, we henceforth say that, if the limit (3.25) holds, the stationary distribution has a geometric tail behaviour.

For the case $d = 1$, Tijms & Van de Coevering [1991] present an efficient numerical algorithm for the computation of the stationary distribution, which exploits the geometric tail behaviour of this distribution. Since, for our model, d is not necessarily equal to one, we adapt their algorithm to the case $d > 1$. This algorithm is in the sequel denoted by the geometric tail (GT) technique.

By the existence of the limit (3.25), we have, for j tending to infinity,

$$\pi_{j+d} \sim z_1^d \pi_j.$$

Hence, a straightforward approximation is

$$\pi_{j+d} = z_1^d \pi_j, \qquad j \geq J, \tag{3.26}$$

with J an integer for which the quotient π_{J+d}/π_J is (fairly) good approximated by z_1^d. So, it remains to compute the probabilities $\pi_0, \pi_1, \ldots, \pi_{J+d-1}$. As mentioned earlier, we omit the equilibrium equation for state $j = 0$. Then, these stationary probabilities are the unique solution of the equilibrium equations for states j, with $j = 1, 2, \ldots, J + d - 1$, and the normalisation equation, after substituting the approximation (3.26) into these equations. The normalisation equation then reads

$$\sum_{j=0}^{J-1} \pi_j + \sum_{i=1}^{d} \frac{1}{1 - z_1^d} \pi_{J+i-1} = 1. \tag{3.27}$$

In this way, we obtain an approximation for the stationary distribution.

To use this approximation, we have to compute the root z_1 of the characteristic equation (3.9) accurately. Since this is the unique root in the interval $(0, 1)$, bisection is a natural technique for the determination of this root. However, this technique may lead to numerical problems. The reason for this is that the right-hand side of the characteristic equation, after subtracting z^{T_H} from both sides, may become rather flat in the neighbourhood of the actual root z_1, so that the values in this neighbourhood may not be distinguishable from zero within the machine precision. Therefore, as suggested by Tijms & Van de Coevering [1991], we use logarithms to avoid this numerical difficulty. More precisely, we divide both sides of the characteristic equation (3.9) by z^{T_H} and then take logarithms at both sides. In this way, the resulting right-hand side is not flat anymore in the neighbourhood of the root z_1, so that the positive root z_1 can be computed accurately.

To illustrate this numerical difficulty, we consider the characteristic equation

$$z^{T_H} = e^{-\mu(1-z)},$$

which follows from $q_{T_H-h} = \beta_h = e^{-\mu}\mu^h/h!$, that is, $\{\beta_h, h = 0, 1, 2, \ldots\}$ is a Poisson distribution with mean μ. As we shall see in Section 3.5, this characteristic equation corresponds to a Markov chain describing the queue-length process considered at arrival instants in a $D/E_{T_H}/1$ queueing system. Now, let $f(z) := e^{-\mu(1-z)} - z^{T_H}$. In Figure 3.1, we depict this function $f(z)$ in the neighbourhood of the actual positive zero $z_1 = 0.107$, for the case $T_H = 50$ and $\mu = 125$. Although it may seem that the function is identically zero for $0 < z < z_1$, this function assumes values that are quite small but strictly positive. Using single machine precision ε, the computer cannot distinguish $f(z)$ from zero, if $|f(z)| < \varepsilon$. In this case, $|f(z)| < \varepsilon$ for $0 < z < 0.125$. Thus, the machine precision is insufficient to accurately compute the root z_1 by bisection within a fairly reasonable relative error of 10^{-6}. As already mentioned, this problem can be circumvented by dividing both sides of the characteristic equation (3.9) by z^{T_H} and then taking logarithms at both sides. However, we note that this particular numerical problem could also have been avoided by using double machine precision. But, even when using double machine precision, there are still many examples for which applying bisection to the characteristic equation fails. So, for safety, we shall use logarithms to compute the root z_1 by bisection.

Clearly, for the GT technique, the computational effort is low and the results are accurate if (3.26) is a good approximation for the quotient π_{j+d}/π_j for small values of J. It turns out that this algorithm performs quite well for relatively small values of J (see, for instance, Tijms & Van de Coevering [1991], and Section 3.5 and Chapter 4 of this monograph). For our numerical examples, to be presented in the subsequent chapters, J is usually smaller than $D + T_H$, which is equal to the number of (boundary) equations which have to be solved by the method of particular solutions.

Nevertheless, the theoretical foundation of the GT technique is still incomplete, so that it is unclear why it performs so well, and it is not possible to give general rules for the value of J. The appropriate value of J has to be determined experimentally, for instance, by investigating different values of J until the stationary probabilities do not alter significantly or until the quotient π_{j+d}/π_j is sufficiently close to z_1^d for $j = J, J+1, \ldots, J+d-1$. The appropriate value depends of course on the accuracy required. We shall apply this technique in subsequent chapters for determining the stationary queue-length distribution of systems with periodic service.

3.5 The $GI/E_r/1$ queueing system

As we have seen, the quality of the approximations by the GT technique depends on the rate at which the root z_1 becomes the dominating term in the linear combination (3.11). Consequently, this quality

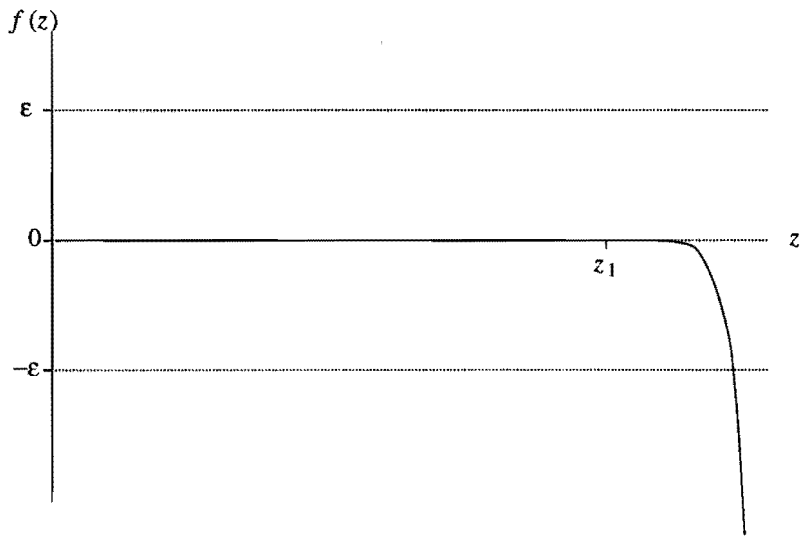

Figure 3.1: *The function* $f(z) = e^{-125(1-z)} - z^{50}$ *in the neighbourhood of the zero* z_1.

depends on how fast the quotient π_{j+d}/π_j is sufficiently well approximated by z_1^d. To get an idea of this rate, we look at a Markov chain that can be studied analytically, like the one defined in Section 3.2.

For the $GI/E_r/1$ queueing system, Adan & Zhao [1994] derive sufficient conditions on the arrival process of customers for the method of particular solutions to be successfully applicable. In this section, we use this system to investigate how fast the root z_1 dominates the linear combination (3.11). In Section 3.5.1, we describe the queueing system and determine the equilibrium equations of the Markov chain representing the queue-length process at arrival instants. The analysis of this process is briefly given in Section 3.5.2. The reasons that numerical problems can be avoided are as follows. Firstly, the determination of roots reduces to solving r fixed-point and contraction equations and, secondly, the boundary equations can be solved explicitly. In Section 3.5.3, we investigate the rate at which the quotient π_{j+1}/π_j is approximated by z_1.

3.5.1 The model and the equilibrium equations

We consider a single-server queueing system at which customers arrive with interarrival times having a common distribution function $\{A(t), t > 0\}$ with mean $1/\lambda$. The service times of customers are Erlangian distributed with r service phases and mean r/μ, and customers are served in the order of their arrival. Finally, we assume that the utilisation factor of the system $\rho := \lambda r/\mu$ is strictly less than one.

Let X_k denote the number of uncompleted service phases just prior to the arrival of the k-th customer. Then, the stochastic process $\{X_k, k = 1, 2, 3, \ldots\}$ is a discrete-time Markov chain with state space $\{0, 1, 2, \ldots\}$ and with $X_1 = n$, where n is a (possibly random) non-negative integer. The station-

ary transition probability $p_{i,j}$ from state i to state j of this chain is given by

$$p_{i,j} = \begin{cases} \sum_{n=i+r}^{\infty} \beta_n, & i = 0, 1, 2, \ldots \text{ and } j = 0, \\ \beta_{i+r-j}, & i = 0, 1, 2, \ldots \text{ and } j = 1, 2, \ldots, i+r, \\ 0, & i = 0, 1, 2, \ldots \text{ and } j \geq i+r+1, \end{cases}$$

where β_n denotes the probability that n service phases are completed during an interarrival time, that is,

$$\beta_n = \int_{t=0}^{\infty} \frac{(\mu t)^n}{n!} e^{-\mu t} \mathrm{d}A(t), \quad n = 0, 1, 2, \ldots .$$

The Markov chain is obviously aperiodic and irreducible, so that together with the assumption that $\rho < 1$, this chain is ergodic (cf. Pakes [1969]). Thus, this chain has a unique stationary distribution $\{\pi_j, j = 0, 1, 2, \ldots\}$, which is the unique solution of the equilibrium equations

$$\pi_0 = \pi_0 \sum_{n=r}^{\infty} \beta_n + \pi_1 \sum_{n=r+1}^{\infty} \beta_n + \pi_2 \sum_{n=r+2}^{\infty} \beta_n + \cdots, \tag{3.28}$$

$$\pi_j = \pi_0 \beta_{r-j} + \pi_1 \beta_{r-j+1} + \pi_2 \beta_{r-j+2} + \cdots, \qquad j = 1, 2, \ldots, r-1, \tag{3.29}$$

$$\pi_j = \pi_{j-r} \beta_0 + \pi_{j-r+1} \beta_1 + \pi_{j-r+2} \beta_2 + \cdots, \qquad j = r, r+1, r+2, \ldots, \tag{3.30}$$

and the normalisation equation. As before, the equations (3.30) are called the inner equations and the equations (3.28) and (3.29) are called the boundary equations.

Clearly, the Markov chain belongs to the class of Markov chains which is defined in Section 3.2, with, as can easily be verified, $D = 0$, $T_H = r$, and $q_h = \beta_{h+r}$ for $h = 0, 1, 2, \ldots$. Hence, we can apply the method of particular solutions of Section 3.3.1 for determining the stationary distribution, that is, expressing this distribution as a linear combination of geometric distributions.

3.5.2 Solving the equilibrium equations

To express the stationary distribution as a linear combination of geometric distributions, we substitute, along the lines of Section 3.3.1, $\pi_j = z^j$ into the inner equations (3.30) and divide these equations by z^{j-r}. So, we obtain the characteristic equation

$$z^r = \beta_0 + \beta_1 z + \beta_2 z^2 + \cdots = \tilde{A}(\mu(1-z)), \tag{3.31}$$

with $\tilde{A}(s)$ the Laplace-Stieltjes transform of the interarrival-time distribution. By Lemma 3.3.1, this equation has exactly r solutions inside the unit circle. As is shown in Adan & Zhao [1994], if both the conditions that $\tilde{A}(s)$ has no zeroes in the half plane $\mathrm{Re}(s) > 0$ and that

$$\left| -\frac{\tilde{A}'(s)}{\tilde{A}(s)} \right| \leq -\frac{\tilde{A}'(\mathrm{Re}(s))}{\tilde{A}(\mathrm{Re}(s))}, \qquad \text{for all } s \text{ with } \mathrm{Re}(s) > 0,$$

are fulfilled, then these r solutions are determined by the r fixed-point and contraction equations

$$z = \varphi_k \sqrt[r]{\tilde{A}(\mu(1-z))}, \qquad k = 0, 1, \ldots, r-1,$$

with $\varphi_k = e^{2\pi ik/r}$ and $i = \sqrt{-1}$, so that the φ_k's are the r roots of the equation $\varphi^r = 1$. Consequently, these r roots of the characteristic equation are simple. Examples of interarrival-time distributions satisfying the two conditions are deterministic, shifted exponential, mixed Erlang, Gamma, and hyperexponential interarrival times.

To complete the determination of the stationary distribution, we have to solve the boundary equations (3.28) and (3.29), and the normalisation equation. These equations are solved by linearly combining the solutions z_k^j, that is, using the form

$$\pi_j = \sum_{k=1}^{r} \lambda_k z_k^j,$$

with the λ_k's denoting complex numbers. Since the roots z_k inside the unit circle of the characteristic equation (3.31) are simple and because $D = 0$, the conditions for applying Lemma 3.3.3 are fulfilled, so that the coefficients λ_k are explicitly given by

$$\lambda_k = (1 - z_k) \frac{\prod_{\substack{j=1 \\ j \neq k}}^{r} (z_j^{-1} - 1)}{\prod_{\substack{j=1 \\ j \neq k}}^{r} (z_j^{-1} - z_k^{-1})}, \qquad k = 1, 2, \ldots, r.$$

In the next section, we consider several interarrival-time distributions to investigate the rate at which the largest root of the equation (3.31) dominates the linear combination.

3.5.3 Numerical examples

As noted before, the quality of the approximation for fixed J and the computational effort of the GT technique depend on the rate at which π_{j+1}/π_j is approximated well by z_1 and on how fast z_1 dominates the linear combination of geometric distributions (3.11). To get an idea of these two aspects, we use the $GI/E_r/1$ queueing system for several interarrival-time distributions, namely, deterministic, Erlangian, and hyperexponential interarrival times.

In the examples, we set $\mu = r$, so that the average service time is equal to 1, and we take λ equal to ρ. Hyperexponential interarrival times are with probability q exponentially distributed with parameter μ_1 and with probability $1 - q$ exponentially distributed with parameter μ_2. So, the average interarrival time is equal to $p_1/\mu_1 + (1 - p)/\mu_2$. The parameters of the hyperexponential distribution H_2 are chosen such that $p/\mu_1 = (1 - p)/\mu_2$ (that is, a hyperexponential distribution with balanced means), and that the squared coefficient of variation is equal to 2.

In Table 3.1, we list the quotient of the second-largest root (denoted by z_2) of equation (3.31) and z_1. This table shows that the quotients $|z_2|/|z_1|$ may differ considerably for the different examples. Further, these quotients may be close to one, indicating that the rate at which z_1 dominates the linear combination of geometric distributions may be rather slow.

However, more important are the results presented in Table 3.2. In this table, we list the quotient π_{j+1}/π_j for the exact values of π_j and π_{j+1}. From this table, we observe that the geometric tail behaviour seems to appear rather fast, if the quotient $|z_2|/|z_1|$ is fairly small (see, for example, the results for the $D/E_r/1$ queueing system). But, if this quotient is rather close to one, as for the $H_2/E_{50}/1$

ρ	r	$D/E_r/1$	$E_{10}/E_r/1$	$H_2/E_r/1$
0.95	5	0.381	0.453	0.724
	10	0.493	0.603	0.842
	50	0.739	0.876	0.965
0.99	5	0.365	0.441	0.723
	10	0.473	0.591	0.841
	50	0.711	0.870	0.965

Table 3.1: *The quotient of the second-largest root z_2 and the largest root z_1 in absolute value.*

queueing system, then this tail behaviour only seems to hold approximately for states at a fairly large distance from the boundary of the state space.

	$D/E_r/1$			$E_{10}/E_r/1$			$H_2/E_r/1$		
j	$r=5$	$r=10$	$r=50$	$r=5$	$r=10$	$r=50$	$r=5$	$r=10$	$r=50$
1	1.0383	1.0266	0.9852	1.0980	1.0788	1.0255	1.2348	1.1149	1.0225
2	0.9497	0.9749	0.9724	1.0149	1.0464	1.0234	1.2429	1.1172	1.0226
3	0.9023	0.9359	0.9604	0.9498	1.0153	1.0214	1.2503	1.1194	1.0227
4	0.8935	0.9111	0.9494	0.9183	0.9871	1.0193	1.2571	1.1216	1.0228
5	0.9011	0.8996	0.9393	0.9248	0.9640	1.0172	0.6294	1.1237	1.0229
10	0.9017	0.9020	0.9069	0.9339	0.9520	1.0067	0.9163	0.5771	1.0234
15	0.9017	0.9017	0.9002	0.9337	0.9498	0.9966	0.9808	1.0252	1.0239
20	0.9017	0.9017	0.9016	0.9337	0.9500	0.9881	0.9899	0.9419	1.0244
25	0.9017	0.9017	0.9018	0.9337	0.9500	0.9826	0.9907	0.9980	1.0248
50	0.9017	0.9017	0.9017	0.9337	0.9500	0.9932	0.9906	0.9952	0.5328
75	0.9017	0.9017	0.9017	0.9337	0.9500	0.9931	0.9906	0.9950	1.0058
100	0.9017	0.9017	0.9017	0.9337	0.9500	0.9931	0.9906	0.9950	0.9843
150	0.9017	0.9017	0.9017	0.9337	0.9500	0.9931	0.9906	0.9950	0.9990
∞	0.9017	0.9017	0.9017	0.9337	0.9500	0.9931	0.9906	0.9950	0.9990

Table 3.2: *The quotient π_{j+1}/π_j for $\rho = 0.95$.*

We notice that, in the $H_2/E_r/1$ queueing system, the quotient π_{r+1}/π_r differs considerably from the other quotients π_{j+1}/π_j with j near r. An intuitive explanation is that the interarrival times have a high variability. These times are often small, but incidentally they are large; customers almost arrive in batches. For the sake of argument, suppose that customers arrive in batches of size two. The customers in the first batch of a busy period see 0 and r phases, respectively. So, when π_0 is fairly large, so will be π_r. The interarrival times of the other batches in this busy period are random and large. Therefore, one may expect that, when these batches arrive, the number of uncompleted phases of the customer in service is more or less random. This suggests that the probability π_r is considerably larger than π_{r+1} due to the 'start-up' effect at the beginning of a busy period.

So, we expect that the GT technique gives accurate results for the $D/E_r/1$ queueing system, when J is at least equal to 10 (for $r = 5$ and $r = 10$) and 25 (for $r = 50$). For the $H_2/E_r/1$ queueing system, we expect that it suffices to take J equal to 25 (for $r = 5$), 50 (for $r = 10$), and 150 (for $r = 50$). These

expectations are confirmed by the results in the Tables 3.3 and 3.4. In these tables, we list the computed values of some stationary probabilities for different values of J, for the $D/E_r/1$ and $H_2/E_r/1$ queueing system with $\rho = 0.95$. We notice that smaller values of J can be used, when less accurate results are required.

		J							
r	j	5	10	15	25	50	75	100	Exact
5	0	0.1729	0.1754	0.1754	0.1754	0.1754	0.1754	0.1754	0.1754
	1	0.0681	0.0687	0.0687	0.0687	0.0687	0.0687	0.0687	0.0687
	2	0.0708	0.0713	0.0713	0.0713	0.0713	0.0713	0.0713	0.0713
	3	0.0674	0.0677	0.0677	0.0677	0.0677	0.0677	0.0677	0.0677
10	0	0.2323	0.2309	0.2307	0.2307	0.2307	0.2307	0.2307	0.2307
	1	0.0613	0.0612	0.0611	0.0611	0.0611	0.0611	0.0611	0.0611
	2	0.0629	0.0628	0.0628	0.0627	0.0627	0.0627	0.0627	0.0627
	3	0.0612	0.0612	0.0612	0.0612	0.0612	0.0612	0.0612	0.0612
50	0	0.4822	0.4365	0.4314	0.4325	0.4325	0.4325	0.4325	0.4325
	1	0.0043	0.0042	0.0041	0.0042	0.0042	0.0042	0.0042	0.0042
	2	0.0042	0.0042	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041
	3	0.0041	0.0040	0.0040	0.0040	0.0040	0.0040	0.0040	0.0040

Table 3.3: *Approximations obtained by the GT technique for the stationary probabilities π_j of the $D/E_r/1$ queueing system with $\rho = 0.95$ for different values of J.*

		J							
r	j	5	10	15	25	50	75	150	Exact
5	0	0.0280	0.0275	0.0289	0.0296	0.0296	0.0296	0.0296	0.0296
	1	0.0047	0.0046	0.0048	0.0049	0.0049	0.0049	0.0049	0.0049
	2	0.0058	0.0057	0.0059	0.0060	0.0060	0.0060	0.0060	0.0060
	3	0.0072	0.0071	0.0074	0.0075	0.0075	0.0075	0.0075	0.0075
10	0	0.0530	0.0206	0.0319	0.0288	0.0288	0.0288	0.0288	0.0288
	1	0.0033	0.0020	0.0025	0.0023	0.0023	0.0023	0.0023	0.0023
	2	0.0037	0.0022	0.0028	0.0026	0.0026	0.0026	0.0026	0.0026
	3	0.0041	0.0024	0.0031	0.0029	0.0029	0.0029	0.0029	0.0029
50	0	0.0903	0.0764	0.0641	0.0444	0.0322	0.0290	0.0282	0.0282
	1	0.0009	0.0008	0.0007	0.0006	0.0003	0.0004	0.0004	0.0004
	2	0.0009	0.0008	0.0007	0.0006	0.0003	0.0004	0.0004	0.0004
	3	0.0009	0.0008	0.0007	0.0006	0.0003	0.0004	0.0005	0.0005

Table 3.4: *Approximations obtained by the GT technique for the stationary probabilities π_j of the $H_2/E_r/1$ queueing system with $\rho = 0.95$ for different values of J.*

3.6 Conclusions

In this chapter, we imposed restrictions on a one-dimensional Markov chain such that the equilibrium equations of this chain constitute a homogeneous linear difference equation with constant coefficients. Firstly, from a certain state i onwards, the transition probability to states j, with $j \geq 1$, depend on $j - i$ only. Secondly, jump sizes to higher states are uniformly bounded by some constant.

Because of this structure of the equilibrium equations, the stationary distribution of the Markov chain is a linear combination of geometric distributions. Standard analytical techniques from the theory of difference equations can be used to obtain this linear combination. However, in Chapter 2, we saw that these techniques give numerical difficulties. Since we cannot link these difficulties to properties of the Markov chain, we used a numerical technique to approximate the stationary distribution. This numerical approach (the GT technique) exploits the geometric tail behaviour. As a result, we obtained an efficient algorithm to approximate the stationary distribution of the Markov chain and, moreover, this algorithm gives excellent approximations without too much computational effort.

In Chapter 4, we shall apply the GT technique to study the queue-length process of a broad class of queueing systems with periodic service. Furthermore, we shall study the queue-length process of some modifications of these systems in the Chapters 6 and 7 by this technique as well. Just like we saw in Chapter 2 and Section 3.5, the GT technique will give excellent approximations for the stationary queue-length distributions, and the computational effort to obtain these results will be fairly low.

4

A Numerical Technique for Queueing Systems with Periodic Service

4.1 Introduction

In Chapter 3, we presented a numerical technique for the determination of the stationary distribution of a class of one-dimensional Markov chains. We show in this chapter that this technique can be used for analysing a broad class of discrete-time queueing systems with periodic service. Hence, this enables one to evaluate a wide range of interesting real-life situations like fixed-cycle traffic lights, communication systems with periodic access schemes, and periodic production rules. Furthermore, it enables one to compare periodic service with other service policies.

We consider a single-server multi-queue system with periodic service. For this system, the time axis consists of intervals of equal length, called cycles. In a cycle, the server visits the different queues to serve customers. The order in which he visits the queues is the same for all cycles. Furthermore, the time instants within a cycle at which he starts switching from one queue to another are fixed and the same for all cycles. Switching from one queue to another may take some time. In Figure 4.1, we repeat the representation for the example in Section 1.1, with two queues and deterministic switch-over times. In this example, the server attends queue 1 for three time units to serve customers in this queue. After that, he requires two time units to switch to queue 2. Then, he attends queue 2 for four time units to serve customers in this queue. Finally, the server switches back to queue 1, which requires one time unit, after which the service policy starts over again. In this example, the time interval (0, 10) can be considered as a cycle.

For queueing systems with periodic service, the queue-length processes do not affect each other, because the switching policy of the server is rigid. As a result, the analysis of the joint queue-length process (and hence, the determination of the sojourn-time distributions) reduces to the analysis of a number of single-queue systems; one for each queue. To study the queue-length process of customers in a specific queue, we consider the system at the start of cycles. We try to apply the numerical tech-

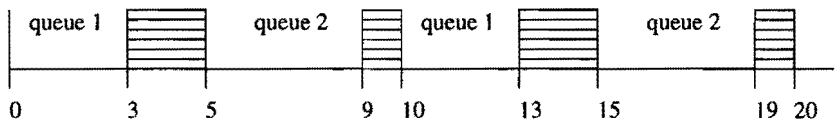

Figure 4.1: *A representation of a periodic service policy for a queueing system with two queues and deterministic switch-over times.*

nique of Chapter 3, that is, the geometric tail (GT) technique, to determine the stationary queue-length distribution at these imbedded time instants.

However, the applicability of the GT technique is restricted to the class of one-dimensional Markov chains for which the stationary distribution has an asymptotically geometric tail. Therefore, we have to impose restrictions on the queueing system, such as restrictions on the arrival process and service-time distribution of customers. Fortunately, these restrictions are not too severe, so that we are able to evaluate a broad class of queueing systems with periodic service. More precisely, this class includes queueing systems with periodic service where customers arrive according to a periodically time-dependent Bernoulli process, with service times having either an arbitrary distribution with bounded support or a distribution that is a mixture of a finite number of negative binomial distributions with the same parameter q.

After the application of the GT technique to the imbedded queue-length process, we can study the queue-length process at arbitrary (and in particular at arrival) instants. However, we are not merely interested in these processes, but also in the sojourn times of customers. Therefore, in this chapter, we pay attention to the determination of the sojourn-time distribution as well. Once the stationary imbedded queue-length distribution is found, the sojourn-time distribution can be calculated exactly.

As stated in Chapter 1, our main objective is to develop techniques for evaluating queueing systems with periodic service. Therefore, the emphasis in this chapter lies on the application of the GT technique and the determination of the sojourn-time distribution. However, we are also interested in the quality of the approximation of the GT technique and in the computational efficiency of the determination of the sojourn-time distribution.

The GT technique reduces the infinite system of equilibrium equations of a Markov chain to a finite one by imposing that the solution to the infinite system is geometric from a certain state onwards. To form an idea of the quality of the approximation of the GT technique for queueing systems with periodic service, we compare the results with those obtained by solving these equations after truncating the state space (and hence, also reducing the infinite system of equilibrium equations to a finite one). This quality turns out to be excellent, even if the geometric tail is imposed on the stationary probabilities of states fairly close to the boundary of the state space. Besides, by this comparison, we also get an idea of the benefits resulting from 'cleverly' reducing the infinite system to a finite system as compared to using brute computational force.

The queue-length process of the class of queueing systems, which we consider in this chapter, can also be studied exactly by the matrix-analytic approach introduced in Neuts [1981,1989]. This approach requires the solution of a polynomial matrix equation. As mentioned in Chapter 1, the computation of this solution is quite sensitive to the utilisation of the system, and this approach faces a 'dimensionality curse'. For the case that this polynomial matrix equation reduces to a quadratic matrix equation, Ramaswami & Lucantoni [1985] give an algorithm for the determination of the waiting-time distribution. After adapting this algorithm to the determination of the sojourn-time distribution,

we show that the procedure developed in this chapter requires less computational effort for most of the relevant cases.

The outline of this chapter is as follows. In Section 4.2, we first describe the periodic service policy in detail. After that, we introduce a broad class of arrival processes and service-time distributions. In Section 4.3, we show that the queue-length process of the corresponding class of queueing systems with periodic service can be studied by the GT technique. In Section 4.4, we present an algorithm for the computation of the sojourn-time distribution of customers. These three sections are extensions of the exposition in Van Eenige, Adan, Resing & Van der Wal [1995b].

In Section 4.5, we briefly outline the matrix-analytic approach to analyse the queue-length process of customers for a subclass of the models introduced in Section 4.2. This part is based on the discussion in Van Eenige, Resing & Van der Wal [1993]. After that, we compare the computational effort of the algorithm in Ramaswami & Lucantoni [1985] for the determination of the sojourn-time distribution with the procedure of Section 4.4. In Section 4.6, we apply the GT technique to numerical examples, and we compare the computational effort of this technique with that of a simple truncation in order to obtain the same accuracy. Finally, we give a summary and the main conclusions of this chapter in Section 4.7.

4.2 The model

In this section, we present a rich class of queueing systems with periodic service that, as will be confirmed in Section 4.3, can be analysed by the GT technique developed in Chapter 3. We describe the periodic switching pattern of the server and introduce some notation in Section 4.2.1. In Section 4.2.2, we give the class of arrival processes and service-time distributions of customers to be considered in this chapter. Finally, in Section 4.2.3, we introduce some additional notation and conventions to facilitate the analysis in subsequent sections.

4.2.1 The periodic service policy

Queueing systems with periodic service can be characterised as follows. A server attends a finite number of queues to serve customers. Switching from one queue to another may take some time. As noted before, the time axis consists of intervals of the same length, called cycles. The order in which the server visits the queues is the same for all cycles. Moreover, the time instants within a cycle at which he starts switching from one queue to another are fixed and the same for all cycles. Further, it is assumed that the arrival processes and service times of customers are independent.

So, in a periodic service policy, the time instants at which the server starts switching do not depend on the queue lengths. Furthermore, since the arrival and service processes are assumed to be independent, the queue-length processes can be studied separately. Therefore, from now on, we consider one queue only.

In a cycle, the server is alternately attending (on-periods) and not attending (off-periods) the queue under consideration. For convenience, it is assumed that a cycle begins with an off-period and ends with an on-period. An off-period and the next on-period together are called a subcycle. Let the number of subcycles in a cycle be equal to N. Recall that the time instants within a cycle at which the server starts switching are fixed and the same for all cycles. So, the length of a subcycle is constant.

As mentioned in Chapter 1, we consider these queueing systems in discrete time. Therefore, we divide each cycle into C intervals of the same length, called slots and numbered $1, 2, \ldots, C$. We assume that, in each slot of the cycle, the server is either attending the queue or not. As a result, we suppose that the on- and off-periods begin at slot boundaries. Further, the slots in an on-period and off-period are called on-slots and off-slots.

Because the switch-over times to the queue under consideration may be random, the length of the off-periods and of the on-periods may also be random (but recall that the length of an off-period and the next on-period together is constant). These switch-overs are assumed to be completed before the server departs from this queue. Let the length of the i-th subcycle be denoted by C_i, with $i = 1, 2, \ldots, N$, so that $\sum_{i=1}^{N} C_i = C$. Further, the (random) length of the on-period in the i-th subcycle is denoted by (the random variable) A_i with probability distribution $\{a_i(j), j = 0, 1, \ldots, C_i\}$, for $i = 1, 2, \ldots, N$. Hence, the length of the off-period in the i-th subcycle is equal to $C_i - A_i$. In Figure 4.2, we represent a cycle graphically, where the dotted lines indicate the random part of the length of the on- and off-periods.

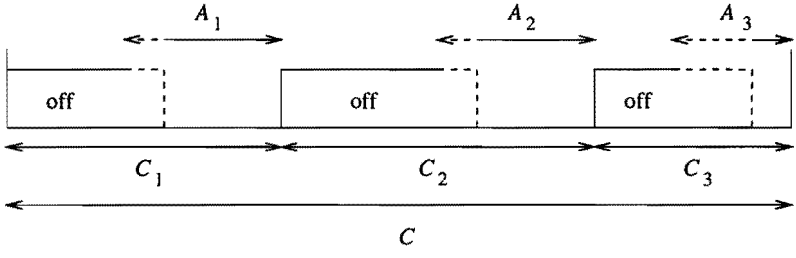

Figure 4.2: *A representation of a cycle for* $N = 3$.

4.2.2 A class of arrival processes and service-time distributions

In this section, we present the class of arrival processes and service-time distributions to be considered in this chapter. This class is broad, and it captures or approximates well most of the important arrival processes and service-time distributions considered in the literature on queueing systems with periodic service.

We assume that in slot n of the cycle, with $n = 1, 2, \ldots, C$, one customer arrives with probability p_n, where $0 \leq p_n \leq 1$. So, with probability $1 - p_n$ no customer arrives in slot n of the cycle. Arrivals in different slots are assumed to be independent. Hence, the arrival process of customers is a periodically time-dependent Bernoulli process.

The service times of customers are measured in numbers of slots. For a customer arriving in slot n of the cycle, with $n = 1, 2, \ldots, C$, the probability generating function $F_n(z)$ of his service time is

$$F_n(z) = \sum_{i=1}^{B_n} b_n(i) \left(\frac{1-\beta}{1-\beta z} \right)^i z^i, \qquad 0 \leq \beta < 1,$$

where B_n is a non-negative constant and $\{b_n(i), i = 1, 2, \ldots, B_n\}$ a probability distribution. In other words, the service time of this customer is i slots, if $\beta = 0$, or is distributed as the sum of i independent

geometric distributions with parameter β, if $0 < \beta < 1$. For convenience, in the sequel, we say that the service of a customer arriving in slot n consists of i service phases with probability $b_n(i)$. Such a phase is called deterministic, if $\beta = 0$, and geometric, if $0 < \beta < 1$.

Clearly, the class of arrival processes is broad, and it captures many important arrival processes. Further, the class of service-time distributions consists of probability distributions with bounded support and of finite mixtures of negative binomial distributions with the same parameter q. Moreover, any arbitrary probability distribution can be approximated as accurately as desired by a probability distribution with bounded support (although one may have to take B_n quite large). For the case that probability distributions are Poisson mixtures, we can approximate them arbitrarily close by a mixture of negative binomial distributions (cf. Steutel & Van Eenige [1996]). Thus, most of the arrival processes and service-time distributions considered in the literature on queueing systems with periodic service (see Section 1.2) are captured or approximated well by this class of arrival processes and service-time distributions. Hence, we conclude that this class is rich and important.

For the class of arrival processes and service-time distributions mentioned, the stochastic process describing the number of service phases in the queue at the start of cycles can be analysed by the GT technique, as we shall formally show in Section 4.3. The reason for this is that this process fulfils the condition that the queue-length process at the start of cycles can be described by a Markov chain for which the equilibrium equations constitute a homogeneous linear difference equation with constant coefficients. That is, this process satisfies the following three conditions (see Section 3.2)

(*i*) the process can be described by a one-dimensional Markov chain,

(*ii*) the transition probabilities from state i to state j of this chain, with $j \geq 1$, depend on i and j only through their difference $j - i$ from a certain state i onwards,

(*iii*) the maximal jump size (that is, the maximal difference $j - i$) is uniformly bounded from above by some positive constant.

For the arrival process of customers, the first condition implies that this process has to be independent and identical for all cycles, so that no supplementary state descriptors are necessary. Further, this arrival process may not depend on the queue length, if this queue length exceeds some constant, because otherwise the second condition is violated. Finally, the number of arrivals in a cycle has to be limited, since otherwise the jump sizes are not uniformly bounded. Notice that these implications allow some generalisations of the aforementioned class of arrival processes. For example, we may allow dependencies between arrivals in different slots in the same cycle, and dependencies between the arrival process and the queue length, if this queue length is smaller than some constant. However, we do not consider these generalisations in order to facilitate the exposition.

For the service-time distributions of customers, condition (*i*) implies that all service phases have to be stochastically identical and that the number of on-slots needed to handle one such phase has to satisfy the lack-of-memory property. Otherwise, we have to use supplementary state descriptors. To avoid violation of condition (*ii*), the number of service phases arriving in a slot may depend on the queue length upon arrival, but only if this queue length is not larger than some constant. The implication of condition (*iii*) is that the number of service phases arriving in a slot is uniformly bounded by some constant. So, the class of service-time distributions could have been extended to the case that the number of service phases arriving in a slot depends on the queue length upon arrival as long as this

queue length is smaller than some constant. We do not consider this extension for convenience. However, at the end of Section 4.3, we shall make a small exception, because there we shall briefly discuss the queue-length process at a fixed-cycle traffic-light queue. For traffic-light queues, there is a natural dependence between the service time and the queue length (recall the traffic-light queue assumption, that is, the TLQ assumption, as defined in Section 1.2). We emphasise that the three conditions prohibit that some customers have deterministic service phases and that other customers have geometric service phases, or that the service times have a general discrete phase-type distribution as defined in Neuts [1981].

So, the class of arrival processes and service-time distributions as introduced at the beginning of this section is broad. Furthermore, this class captures almost all possible arrival processes and service-time distributions that satisfy the condition mentioned that the queue-length process at the start of cycles can be described by a Markov chain for which the equilibrium equations constitute a homogeneous linear difference equation with constant coefficients. Hence, the GT technique is applicable to a rich class of queueing systems with periodic service. To facilitate the analysis in subsequent sections, we introduce some notation and conventions.

4.2.3 Additional notation and conventions

Let $A_{\min}$ and $A_{\max}$ denote the minimal and maximal number of on-slots in a cycle, that is,

$$A_{\min} := \sum_{i=1}^{N} \min\{j | a_i(j) > 0, j = 0, 1, \ldots, C_i\}, \tag{4.1}$$

$$A_{\max} := \sum_{i=1}^{N} \max\{j | a_i(j) > 0, j = 0, 1, \ldots, C_i\}. \tag{4.2}$$

For notational convenience, we set $B_n := 0$ and $b_n(0) := 1$, if $p_n = 0$, with $n = 1, 2, \ldots, C$. Furthermore, when appropriate, we consider deterministic service phases as geometric service phases with $\beta = 0$. Further, let $B_{\min}$ and $B_{\max}$ denote the minimal and maximal number of service phases arriving in a cycle, that is,

$$B_{\min} := \sum_{n=1}^{C} \min\{i | b_n(i) > 0, i = 0, 1, \ldots, B_n\} \quad \text{and} \quad B_{\max} := \sum_{n=1}^{C} B_n. \tag{4.3}$$

Customers are served in the order of their arrival, and the number of service phases of the arriving customer is assumed to be known upon arrival. The arrival process and service times are supposed to be independent. Moreover, the service of a customer that is interrupted (due to an off-period) is resumed where it was interrupted.

Finally, customer arrivals, and the start and completion of a service phase occur at slot boundaries. Hence, services start and are completed at slot boundaries as well. For convenience, we assume that the completions of service phases occur *just before* slot boundaries, and that customer arrivals and the start of service phases occur *just after* slot boundaries. A customer arriving at the boundary between slot $n-1$ and slot n in the cycle (that is, the n-th slot boundary) is said to arrive in slot n, and a customer for whom the service is completed at the n-th slot boundary is said to depart in slot $n-1$ (see Figure 4.3). Further, if the server is idle upon a customer arrival, then he starts servicing this customer immediately.

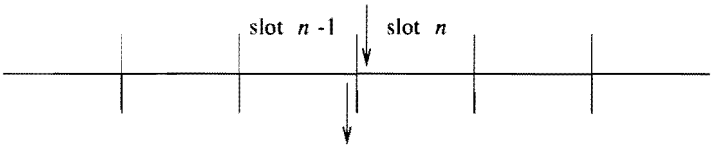

Figure 4.3: *A customer departure in slot $n-1$, and a customer arrival in slot n.*

4.3 The queue-length process

In this section, we show that the queue-length process of customers of the model introduced in Section 4.2 can indeed be treated by the GT technique. Furthermore, we exploit the structure of this model to conclude that more equilibrium equations have the constant structure than is suggested in Section 3.2. This observation can be utilised by the GT technique. Finally, we illustrate that this technique can also be applied, if there is some dependence between the service time of a customer and the queue length at his arrival. This illustration is based upon and motivated by the study of fixed-cycle traffic-light queues, where cars arriving in a green period at an empty queue experience no delay at all. Before presenting the outline of this section, we introduce a Markov chain, which is the main object of study in this section.

To analyse the queue-length process of customers, we consider the system at the first slot boundary of cycles, that is, the slot boundaries between two consecutive cycles (see Figure 4.4). Hence, these imbedded time instants are just after a possible service phase completion (and just after an on-period), but just before a possible arrival (and just before the start of an off-period). Then, the number of uncompleted service phases at these imbedded instants constitutes a homogeneous discrete-time Markov chain.

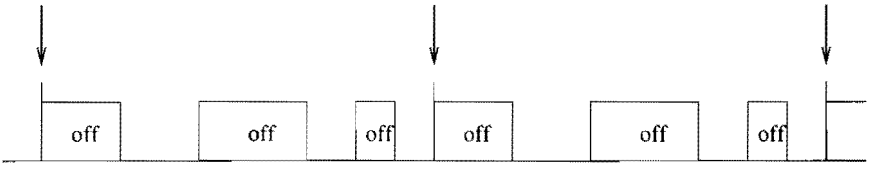

Figure 4.4: *The imbedded time instants for the case $N = 3$.*

We assume that the state space of this Markov chain consists of the non-negative integers (by some small and straightforward adaptations, the analysis of this section can be applied when this state space consists of multiples of some integer, which may be the case for $\beta = 0$). Further, we assume that this chain is irreducible and aperiodic. Finally, it is assumed that the average number of slots of work arriving per cycle is strictly less than the average service capacity per cycle, that is,

$$\sum_{n=1}^{C} \sum_{i=1}^{B_n} p_n b_n(i) \frac{i}{1-\beta} < \mathrm{E}\{A\},$$

where $\mathrm{E}\{A\}$ denotes the average total length of the on-periods in a cycle, that is,

$$\mathrm{E}\{A\} := \sum_{i=1}^{N} \mathrm{E}\{A_i\}.$$

Under these assumptions, the Markov chain is ergodic (cf. Pakes [1969]), so that it has a unique stationary distribution $\{\pi_j, j = 0, 1, 2, \ldots\}$.

For the queueing system in Chapter 2, with $p_n = p$, $B_n = 1$, $b_n(1) = 1$, for all $n = 1, 2, \ldots, C$, and $N = 1$ and $A_1 = A$, the latter assumption simply reduces to $Cp < A$.

We notice that, by assuming that the Markov chain is irreducible, we have implicitly assumed for deterministic service phases that $B_{\max} > A_{\min}$. In this way, we exclude trivialities. Otherwise, the state space of this chain would be finite.

Further, once the stationary distribution $\{\pi_j, j = 0, 1, 2, \ldots\}$ is obtained, we can straightforwardly determine the stationary imbedded queue-length distributions at other slot boundaries in the cycle. Given the stationary queue-length distribution at the n-th slot boundary, we can compute this distribution for the $(n+1)$-st slot boundary by considering the one-slot transition probabilities, for $n = 2, 3, \ldots, C$. However, since these transition probabilities may depend on the length of the on-period of the subcycle to which slot n belongs, we first have to condition on this length. For $n = 1, 2, \ldots, C$, the stationary imbedded queue-length distribution at the n-th slot boundary is denoted by $\{\pi_{j,n,0}, j = 0, 1, 2, \ldots\}$, if slot n is an off-slot, and by $\{\pi_{j,n,1}, j = 0, 1, 2, \ldots\}$, if slot n is an on-slot.

This section is organised as follows. In Section 4.3.1, we show that the stationary distribution of the Markov chain can be determined by the GT technique. Furthermore, we utilise the structure of this chain to characterise its transition probabilities. From this characterisation, it follows that the number of boundary equations is smaller than the number suggested in Section 3.2. In Section 4.3.2, we show, for a specific example, that the GT technique can also be applied, if there is some dependence between the service time of a customer and the queue length upon his arrival. This specific example is motivated by studies of fixed-cycle traffic-light queues.

Before presenting the results, we remark that these results are valid for the general model introduced in Section 4.2. They may, however, be improved for special cases of the model.

4.3.1 Validation of the applicability of the GT technique

In the first part of this section, we show that the Markov chain defined in Section 4.2.2 belongs to the class of Markov chains defined in Section 3.2. Thereafter, we characterise the transition probabilities of this chain. From this characterisation, we conclude that the number of boundary equations is smaller than the number of boundary equations (3.6) and (3.7) in Section 3.2.

The results in this section depend on whether the service phases are deterministic or geometric. The reason for this is that, for geometric service phases, no service phase may be completed in an on-slot, even if there is one such phase at the start of this slot, whereas for deterministic service phases, exactly one such phase (if any) is handled in each on-slot.

As in Chapter 3, we use the non-negative integers D, T_L, and T_H. For completeness, D denotes the minimal integer for which the transition probabilities $p_{i,j}$ of the Markov chain are equal to q_{j-i}, for $i \geq D$ and $j \geq 1$. The integers T_L and T_H indicate the largest possible jump out of state i to a lower and a higher state, respectively, for $i \geq D$.

In the next lemma, we show that the Markov chain belongs to the class of Markov chains introduced in Section 3.2 by specifying the values of the quantities D, T_L, and T_H, where

$$\delta_\beta := \begin{cases} 1, & \text{if } \beta = 0, \\ 0, & \text{if } 0 < \beta < 1. \end{cases}$$

Lemma 4.3.1 *For the Markov chain of Section 4.2, we have*

(*i*) $D \leq A_{\max}$,

(*ii*) $T_L = A_{\max} - B_{\min}$,

(*iii*) $T_H = B_{\max} - \delta_\beta A_{\min}$.

Proof. We begin with proving that $D \leq A_{\max}$. The number of on-slots in a cycle is at most $A_{\max}$, so that the server cannot become idle in a cycle if the number of uncompleted service phases i at the start of this cycle is at least $A_{\max}$. Of course, the number of uncompleted service phases j at the start of the next cycle is simply equal to i minus the number of service phases completed in the cycle plus the number of service phases arriving in the cycle. But, if $i \geq A_{\max}$, then this transition from state i to state j does not depend on the slots in which these phases arrive and in which they are completed. So, we have $D \leq A_{\max}$.

The lowest state that can be reached by a single transition occurs if the maximal number of service phases is completed in a cycle and if the minimal number of service phases arrives in this cycle, so that we have $T_L = A_{\max} - B_{\min}$.

Finally, the result $T_H = B_{\max}$, for $0 < \beta < 1$ (that is, for geometric service phases), immediately follows from the observation that the maximal transition occurs if no service phase is completed in a cycle at all and if the maximal number of service phases arrives in the same cycle. The result $T_H = B_{\max} - A_{\min}$, for $\beta = 0$ (that is, for deterministic service phases), is obtained in a similar way by including that, for $i \geq D$, at least $A_{\min}$ phases are completed in a cycle. □

So, we have proved that the Markov chain belongs to the class of Markov chains of Section 3.2. As a result, the transition probabilities $p_{i,j}$ of this chain depend on i and j only through their difference $j - i$, for $i \geq D$ and $j \geq 1$, and they may be nonzero for $j - i = -T_L, -T_L + 1, \ldots, T_H$. Consequently, when defining $q_h := p_{i,i+h}$, for $i \geq D$ and $h = -T_L, -T_L + 1, \ldots, T_H$, the equilibrium equations of the Markov chain can be partitioned as follows (cf. the equations (3.6), (3.7), and (3.8))

$$\pi_0 = \pi_0 p_{0,0} + \pi_1 p_{1,0} + \cdots + \pi_{\max\{D-1,T_L\}} p_{\max\{D-1,T_L\},0}, \tag{4.4}$$

$$\pi_j = \pi_0 p_{0,j} + \cdots + \pi_{D-1} p_{D-1,j} + \pi_D q_{j-D} + \cdots + \pi_{j+T_L} q_{-T_L}, \qquad 1 \leq j < D + T_H, \tag{4.5}$$

$$\pi_j = \pi_{j-T_H} q_{T_H} + \pi_{j-(T_H-1)} q_{T_H-1} + \cdots + \pi_{j+T_L} q_{-T_L}, \qquad j \geq D + T_H. \tag{4.6}$$

In the remainder of this section, we assume that $D = A_{\max}$. However, D may be less than $A_{\max}$, as we saw for the Markov chain in Chapter 2, where D was equal to one, whereas $A_{\max}$ was equal to A.

Since the Markov chain under study belongs to the class of Markov chains of Section 3.2, it follows that the stationary distribution of this chain can be determined from the equilibrium equations (4.4), (4.5), and (4.6) by the GT technique. In order to actually solve these equilibrium equations, it remains to determine the probabilities q_h and the transition probabilities $p_{i,j}$, with $i < A_{\max}$, appearing in these equations.

To determine the probability distribution $\{q_h, h = -T_L, -T_L + 1, \ldots, T_H\}$, we introduce its shifted probability generating function $Q(z)$. Furthermore, we define the probability generating functions $\alpha_i(z)$ and $\beta_n(z)$ of the length of the on-period in the i-th subcycle of the cycle and of the number of

service phases arriving in slot n of the cycle, respectively. More precisely,

$$Q(z) := z^{T_L} \sum_{h=-T_L}^{T_H} q_h z^h, \tag{4.7}$$

$$\alpha_i(z) := \sum_{j=0}^{C_i} a_i(j) z^j, \qquad i = 1, 2, \ldots, N, \tag{4.8}$$

$$\beta_n(z) := 1 - p_n + p_n \sum_{i=1}^{B_n} b_n(i) z^i, \qquad n = 1, 2, \ldots, C. \tag{4.9}$$

The reason for the use of the shifted probability generating function is merely to avoid negative powers.

Lemma 4.3.2 *The shifted probability generating function $Q(z)$ of the probability distribution $\{q_h, h = -T_L, -T_L + 1, \ldots, T_H\}$ satisfies*

$$Q(z) = z^{T_L} \prod_{i=1}^{N} \alpha_i(\beta + (1-\beta)/z) \prod_{n=1}^{C} \beta_n(z).$$

Proof. For $h = -T_L, -T_L + 1, \ldots, T_H$, the probability q_h is actually the conditional probability that the net increase of the number of uncompleted service phases in a cycle is equal to h phases, given that the initial number of service phases at the start of the cycle prevents the server from becoming idle in this cycle. Let the random variable H denote this net increase, and the random variables X and Y_i the number of service phases arriving in a cycle and the number of service phases completed in the i-th subcycle of the cycle (given that the server does not become idle) with $i = 1, 2, \ldots, N$, respectively. Then, we have obviously the following relation between these random variables

$$H = X - \sum_{i=1}^{N} Y_i.$$

From this relation, it immediately follows that

$$Q(z) = z^{T_L} \mathrm{E}\{z^H\} = z^{T_L} \mathrm{E}\{z^{X - \sum_{i=1}^{N} Y_i}\} = z^{T_L} \mathrm{E}\{z^X\} \prod_{i=1}^{N} \mathrm{E}\{(1/z)^{Y_i}\}, \tag{4.10}$$

where we use that the arrival process and service process are independent. Clearly, the probability generating function of the random variable X satisfies

$$\mathrm{E}\{z^X\} = \prod_{n=1}^{C} \beta_n(z), \tag{4.11}$$

so that it remains to determine $\mathrm{E}\{(1/z)^{Y_i}\}$, for $i = 1, 2, \ldots, N$.

To determine $\mathrm{E}\{(1/z)^{Y_i}\}$, with $i = 1, 2, \ldots, N$, let the random variable $K_{i,n}$ denote the number of service phases completed in the n-th on-slot of the i-th subcycle, for $n = 1, 2, \ldots, A_i$. These random variables are independent and have a Bernoulli distribution with parameter $1 - \beta$. Since

$$Y_i = \sum_{n=1}^{A_i} K_{i,n}, \qquad i = 1, 2, \ldots, N,$$

and because the random variable A_i has probability generating function $\alpha_i(s)$, we have

$$E\{s^{Y_i}\} = \alpha_i(\beta + (1-\beta)s) \qquad i = 1, 2, \ldots, N. \tag{4.12}$$

Substituting the relations (4.11) and (4.12) into equation (4.10) completes the proof. □

From the generating function $Q(z)$, we can determine the probabilities q_h by equating the coefficients which belong to the same power, for $h = -T_L, -T_L + 1, \ldots, T_H$. So, it remains to characterise the transition probabilities $p_{i,j}$ of the Markov chain, for $i < A_{\max}$.

If the number of uncompleted service phases i at the start of the cycle is smaller than $A_{\max}$, then a transition to state j may depend on the slots in which service phases arrive and are completed. The reason for this is that the server may be idle in the cycle. Hence, the transition probabilities $p_{i,j}$ of the Markov chain, for $i < A_{\max}$, are in general not equal to q_{j-i}. However, for some $i < A_{\max}$ and j, the transition probabilities $p_{i,j}$ appear to be equal to q_{j-i}. The implication of this result is that, as we shall see, more equilibrium equations have the constant structure (4.6) than is suggested in Section 3.2. Before giving the lemma presenting this result, we introduce an integer K, illustrate the importance of this integer by an example, and mention a subtlety that may occur for the case of deterministic service phases.

Consider the last possible on-slot in a cycle in which a customer may arrive (that is, the last possible on-slot n in a cycle with $p_n > 0$). If slot n lies in the N-th subcycle, then the integer K denotes the number of possible on-slots in the cycle, before slot n, in which no customer can arrive. Otherwise, the integer K denotes the number of possible on-slots, which lie in the first $N-1$ subcycles, in which no customer can arrive. In Figure 4.5, we give three examples of cycles with on-periods of fixed duration. For the first example, the last on-slot of the cycle in which a customer may arrive is slot 5, whereas it is slot 4 for the second example. Nevertheless, for both examples, the number of on-slots, before this on-slot (that is, slot 5 and slot 4, respectively), in which no customer can arrive is equal to one (namely, slot 3). So, since slot 4 and slot 5 lie in the last (and only) subcycle of the cycle, we have $K = 1$ for both cases. For the third example, N is equal to two and the last on-slot in which a customer can arrive is slot 4. Since this slot lies in the first of the two subcycles and because $p_5 = 0$, we have $K = 1$.

For geometric service phases, the next lemma states that, for $K \leq i < A_{\max}$ and $j \geq B_{\max}$, the transition probabilities $p_{i,j}$ are equal to q_{j-i} as well. The transition probability from state i to state j, with $i < K$ and $j \geq B_{\max}$, is not equal to q_{j-i}, because the server may be idle. To illustrate this latter remark (and so, the importance of the integer K), consider the following example.

Consider a cycle consisting of one on- and off-period, and let the length of this cycle be equal to three slots. The length of the on-period is supposed to be constant and equal to two slots. In the first two slots of the cycle, no customer can arrive. In the last slot of the cycle a customer can arrive, whose service consists of at most three phases. So, we have $K = 1$ and $B_{\max} = 3$ (see also Figure 4.6). Suppose that there is no uncompleted service phase at the start of the cycle. Then, if in the last slot of the cycle the maximal number of service phases arrives and if none of these arriving phases is completed in this slot, then the number of uncompleted service phases at the start of the next cycle is equal to 3. The probability that this transition occurs is equal to $\beta b_3(3)$, whereas $q_3 = \beta^2 b_3(3)$.

For deterministic service phases, the reason behind the introduction of the integer K is similar as for geometric service phases. For these service phases, the result presented in the next lemma also deals with the subtlety that there are on-slots n with both $p_n > 0$ and $b_n(1) = 1$ (that is, on-slots in which an arriving customer has always a service time equal to one slot). To illuminate this subtlety, we use the example corresponding to Figure 4.7, where $K = 0$ and the length of the on-period is constant. If the

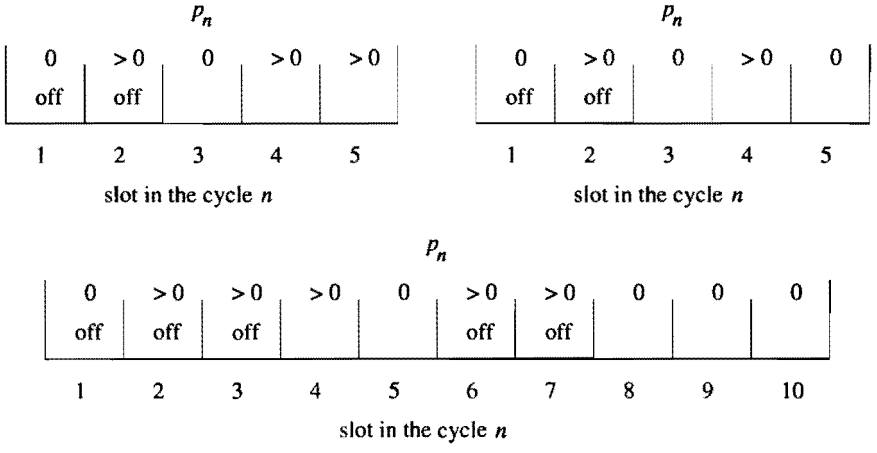

Figure 4.5: *Three cycles for which K is equal to one.*

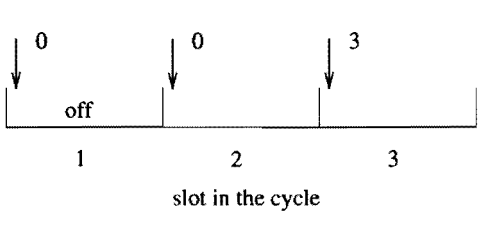

Figure 4.6: *The maximal number of service phases arriving in the slots of a cycle.*

queue is empty at the boundary between slot 1 and slot 2, then it is still empty at the next slot boundary, irrespective of whether a customer arrives in slot 2 or not. Further, if the queue is empty at the start of the cycle, then exactly three service phases have to arrive in slot 3 in order to have two service phases at the start of the next cycle. However, these three phases may not arrive as one phase in slot 2 and two phases in slot 3. This indicates that the transition depends on the arrival pattern of the phases and not on the number of arriving service phases only. So, this transition cannot be equal to q_2. To deal with this subtlety, we consider transitions from states i, with $K+1 \leq i < A_{\max}$ instead of $K \leq i < A_{\max}$, to states j, with $j \geq B_{\max} - \max\{0, A_{\min} - K\} + 1$ instead of $j \geq B_{\max} - \max\{0, A_{\min} - K\}$ (notice that K may be larger than $A_{\min}$).

Lemma 4.3.3

For geometric service phases, we have

$$p_{i,j} = q_{j-i}, \quad K \leq i < A_{\max} \text{ and } j \geq B_{\max}.$$

For deterministic service phases, we have

$$p_{i,j} = q_{j-i}, \quad K+1 \leq i < A_{\max} \text{ and } j \geq B_{\max} - \max\{0, A_{\min} - K\} + 1.$$

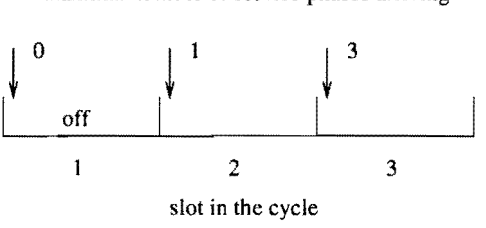

Figure 4.7: *The maximal number of service phases arriving in the slots of a cycle.*

Proof. For both geometric ($0 < \beta < 1$) and deterministic ($\beta = 0$) service phases, the proof of the present lemma consists of two parts. We first prove that a transition from state i to state j (with i and j satisfying the above conditions) implies that the server cannot become idle in the cycle. If the server may not become idle in the cycle, then this may imply that service phases must arrive in certain slots of the cycle. In that case, the transition probability $p_{i,j}$ is not equal to q_{j-i}. However, in the second part of the proof, we show that the initial number of uncompleted service phases i prevents the server from becoming idle, even if the service phases arrive as late as possible in the cycle. Combining these two parts yields the desired result.

Let i and j satisfy the above conditions, and assume that the total length of the on-periods is equal to A slots with $i < A \leq A_{\max}$. If $A \leq i$, then the proof of the present lemma is trivial, because the server cannot become idle in the cycle, so that a transition from state i to state j depends on the number of arriving service phases only.

To prove the first part for $0 < \beta < 1$, suppose that the server is idle in the n-th on-slot. Clearly, $n \geq i + 1$, and in at most K out of the first n on-slots no customer can arrive. Since $n - K \geq i + 1 - K \geq 1$, there is at least one on-slot m with $m \leq n$ and $p_m > 0$. Hence, at most $B_{\max} - 1$ service phases can arrive after on-slot n in the cycle, so that a transition to state $j \geq B_{\max}$ is not possible.

For the second part of the proof for $0 < \beta < 1$, let $j = B_{\max} + k$, with $k = 0, 1, \ldots, i$ (notice that $p_{i,j} = 0$ for $j > B_{\max} + i$). Then, a transition from state i to state $B_{\max} + k$ indicates that at most $i - k$ service phases may be completed in the cycle; otherwise a transition to state $B_{\max} + k$ is not possible, because at most $B_{\max}$ phases arrive in a cycle. If less than i phases are completed in the cycle, then it is clear that the server cannot have been idle, irrespective of when the phases arrived. If i service phases are completed in the cycle (so, $k = 0$) then, in order to make a transition to state j, $B_{\max}$ phases must have arrived in this cycle. Hence, in this case, the server is certainly not idle in the on-slots n with $p_n > 0$. Further, since $i \geq K$, he is also not idle in the first K on-slots m with $p_m = 0$. Finally, because the server may not complete any of the arriving phases, he does not become idle between the last on-slot n, with $p_n > 0$, and the start of the next cycle. Thus, we have proved the second part, so that combining this part with the first part completes the proof for the case $0 < \beta < 1$.

To prove the first part for $\beta = 0$, we assume that K from the i initial service phases are handled in the first K on-slots m with $p_m = 0$, so that the server is not idle in these slots. Furthermore, we remove these K on-slots from the cycle, so that the resulting cycle has the same off-periods, but has now only $A - K$ on-slots. This resulting cycle is called the new cycle in order to distinguish it from the original cycle. Notice that, by the definition of K, it is still possible that this new cycle has on-slots m with $p_m = 0$. However, between the start of any of these on-slots and the start of the next cycle, no customer can arrive. Hence, if the server is idle in any of these on-slots, then the queue is empty at the

start of the next cycle, so that a transition to state j is not possible.

To complete the first part for $\beta = 0$, suppose that the server is idle in the n-th on-slot of the new cycle, with $p_n > 0$. We recall that, in the new cycle, there are no on-slots m, with $p_m = 0$, between two arbitrary on-slots in which customers may arrive. Hence, the number of service phases at the start of the next cycle is maximal if the maximal number of phases arrives in each of the on-slots after on-slot n. We remark that, if the first on-slots m, after on-slot n, have $b_m(1) = 1$, the number of service phases at both the start and the end of these slots is zero, irrespective of whether the service phase arrives or not. For convenience, however, we assume that in these on-slots the service phase arrives. Obviously, the maximal number of phases arriving after on-slot n is at most $B_{\max} - n$. If the maximal number of phases arrives, then $A - K - n$ of these phases or all arriving phases are completed in this new cycle, whichever is smaller. So, the maximal number of phases at the start of the next cycle is at most

$$\max\{0, B_{\max} - n - (A - K - n)\} < B_{\max} - \max\{0, A_{\min} - K\} + 1.$$

Thus, for $\beta = 0$, we conclude that the server cannot be idle in the cycle, if a transition from state i to state j occurs.

Next, we prove the second part for $\beta = 0$. If a transition from state i to state j occurs, then exactly $j - i + A$ service phases arrive in the (original) cycle. The reason for this is that, from the first part of the proof of the present lemma, the server cannot become idle if a transition to state j occurs. Suppose that these $j - i + A$ service phases arrive as late as possible in the cycle. Since

$$j - i + A \geq B_{\max} - i - \max\{0, A_{\min} - K\} + A + 1,$$

no service phase may arrive in at most the first $i + \max\{0, A_{\min} - K\} - A - 1$ on-slots m with $p_m > 0$. This implies that no service phase may arrive in at most the first $i + \max\{0, A_{\min} - K\} - (A - K) - 1$ on-slots, because there are at most K on-slots in the cycle, before the first on-slot n with $p_n > 0$, in which no service phase can arrive. Since the initial number of service phases is equal to i and

$$i \geq i + \max\{0, A_{\min} - K\} - (A - K) - 1,$$

the server cannot become idle in the cycle, even if the $j - i + A$ service phases arrive as late as possible in the cycle. Combining this part with the first part proves the result of the present lemma for the case $\beta = 0$. □

So, we have now characterised the probabilities q_h in Lemma 4.3.2, for $h = -T_L, -T_L + 1, \ldots, T_H$, and the transition probabilities $p_{i,j}$ in Lemma 4.3.3, for $K \leq i < A_{\max}$ and $j \geq B_{\max}$ (if $0 < \beta < 1$) and for $K + 1 \leq i < A_{\max}$ and $j \geq B_{\max} - \max\{0, A_{\min} - K\} + 1$ (if $\beta = 0$). As a result of this characterisation, we have that $p_{i,j} = q_{j-i}$ for $j \geq B_{\max} + K$ (if $0 < \beta < 1$) and for $j \geq B_{\max} - \max\{0, A_{\min} - K\} + K + 1$ (if $\beta = 0$). So, the number of boundary equations is equal to $B_{\max} + K$ and $B_{\max} - \max\{0, A_{\min} - K\} + K + 1$, respectively, instead of $B_{\max} + T_H$ as is suggested by the equations (4.4) and (4.5). This circumstance can be exploited by the GT technique.

Finally, to apply the GT technique, we have to determine the transition probabilities $p_{i,j}$, with $i < A_{\max}$, that are not contained within Lemma 4.3.3. Unfortunately, these transition probabilities generally depend on the slots in which service phases arrive and are completed. Hence, in general, it is hard to characterise these probabilities explicitly. We can, however, determine them recursively by the on-slot transition probabilities (after conditioning on the lengths of the on-periods).

4.3.2 The fixed-cycle traffic-light queue

In the above analysis of queueing systems with periodic service, we did not allow any dependence between the service times of a customer and the queue length upon his arrival. This assumption can be relaxed such that the GT technique is still applicable. In this section, we illustrate this for a specific case, namely, the case that the service times of customers are equal to zero, if they arrive in an on-period at an empty queue. This case is motivated by studies of fixed-cycle traffic-light queues, where this dependence arises quite naturally (see also Section 1.2).

With this adaptation, Lemma 4.3.1 is still valid. The reason for this is that the server cannot become idle in the cycle, if the number of service phases at the start of this cycle is at least $A_{\max}$. As a result, all customers arriving in this cycle have service times that are stochastically identical to those defined in Section 4.2. Hence, the arguments in the proof of Lemma 4.3.1 can be applied directly, so that this lemma holds.

The fact that Lemma 4.3.1 is valid implies that the Markov chain describing the imbedded queue-length process can be analysed by the GT technique. Furthermore, by this lemma, the shifted probability generating function $Q(z)$ of the probabilities $q_{j-i} = p_{i,j}$, for $i \geq D$ and $j \geq 1$, is given by Lemma 4.3.2.

Finally, Lemma 4.3.3 is valid as well, because the transitions in this lemma prevent the server from becoming idle, even if customers arrive as late as possible in the cycle. So, we conclude that the equilibrium equations for states j, with $j \geq B_{\max} + K$ (if $0 < \beta < 1$) or $j \geq B_{\max} - \max\{0, A_{\min} - K\} + K + 1$ (if $\beta = 0$), are identical to those for the queueing system without the dependence between the service time and queue length. We again notice that the transition probabilities, which are not captured by Lemma 4.3.2 and Lemma 4.3.3, can be determined recursively. Furthermore, these transition probabilities are in general not equal to those in Section 4.3.1, because of the imposed dependence.

4.4 The sojourn-time distribution

In this section, we present a procedure for the computation of the sojourn-time distribution of an arbitrary customer arriving at the system in statistical equilibrium. In contrast to the study of the queue-length process, this computation is exact (given the stationary imbedded queue-length distribution). The sojourn time of a customer is defined as the length of the time interval between his arrival and his departure, and is measured in numbers of slots. Notice that the sojourn time of a customer is not affected by customers who arrive after him, so that we can disregard any subsequent arrival.

We compute the probability that the sojourn time S of an arbitrary customer is at most equal to s slots, with $s = 1, 2, 3, \ldots$. To compute this probability, we first condition on the slot of arrival (and whether this slot is an on- or off-slot) and determine the number of service phases immediately after this arrival. Because the number of on-slots between the slot of arrival and the s-th slot after this arrival is not necessarily fixed, we also condition on the number of on-slots between these two slots. Then, conditional on this number of on-slots, we derive the probability that the number of service phases just after the arrival is completed within this number of on-slots, so that the sojourn time of the customer is at most equal to s slots.

So, we first condition on the slot of arrival n, with $n = 1, 2, \ldots, C$, and on whether slot n is an off-slot ($\delta_n := 0$) or an on-slot ($\delta_n := 1$). For $n = 1, 2, \ldots, C$, let the function $i(n)$ denote the subcycle in which slot n lies, and define $\overline{C}_k := \sum_{j=1}^{k} C_j$, for $k = 1, 2, \ldots, N$. Then, the probability $\psi_{n,0}$ that an

arbitrary customer arrives in slot n and that this slot is an off-slot is equal to

$$\psi_{n,0} = \frac{p_n}{\sum_{k=1}^{C} p_k} \sum_{h=0}^{\overline{C}_{i(n)}-n} a_{i(n)}(h), \tag{4.13}$$

and the probability $\psi_{n,1}$ that the slot of arrival is slot n and that this slot is an on-slot is equal to

$$\psi_{n,1} = \frac{p_n}{\sum_{k=1}^{C} p_k} \sum_{h=\overline{C}_{i(n)}-n+1}^{C_{i(n)}} a_{i(n)}(h). \tag{4.14}$$

The number of service phases just after the arrival in slot n is equal to the sum of the number of service phases upon arrival and the number of service phases of the arriving customer. By the Bernoulli-arrivals-see-time-average property (BASTA property, cf. Halfin [1983]), the number of service phases upon arrival has probability distribution $\{\pi_{j,n,\delta_n},\ j = 0, 1, 2, \ldots\}$. So, the conditional probability that the number of service phases W_{n,δ_n} immediately after the arrival in slot n is equal to w slots is given by

$$\Pr\{W_{n,\delta_n} = w\} = \sum_{j=0}^{w-1} \pi_{j,n,\delta_n} b_n(w-j), \qquad w = 1, 2, 3, \ldots, \tag{4.15}$$

where, of course, $b_n(j) = 0$ for $j > B_n$.

We are interested in the probability that the sojourn time of an arbitrary customer is at most s slots, with $s = 1, 2, 3, \ldots$. This indicates that we have to determine the probability that w service phases are completed within the first s slots (including slot n) after the arrival. Since the number of on-slots within the first s slots after the arrival is not necessarily deterministic, we also condition on this number of on-slots. To determine the probability generating function $H_{n,s,\delta_n}(z)$ of this number of on-slots, we make a distinction between the following three cases. Firstly, we consider the case that slot n and the s-th slot after (and including) slot n lie in the same subcycle, secondly, the case that these slots lie in the same cycle (but in different subcycles), and thirdly the case that these slots lie in different cycles. Notice that, in the first two cases, the s-th slot after the arrival instant is slot $n + s - 1$, because we assumed that customers arrive just after slot boundaries and depart just before slot boundaries (see Figure 4.8).

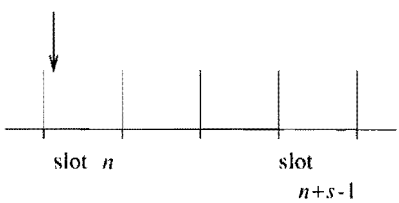

Figure 4.8: *A customer arriving in slot slot n, and the latest slot $n + s - 1$ in which this customer may depart such that his sojourn time is at most s slots, for the case $s = 4$.*

Case 1. *Slot n and slot $n + s - 1$ lie in the same subcycle.*
If, in this case, slot n is an on-slot, then clearly, slot $n + s - 1$ is an on-slot as well, so that

$$H_{n,s,1}(z) = z^s.$$

Suppose that slot n is an off-slot. Then, there are no on-slots between slot n and slot $n+s-1$, if the length of the on-period in the $i(n)$-th subcycle is less than or equal to $\overline{C}_{i(n)} - (n+s-1)$, and there are j on-slots, if this length is equal to $\overline{C}_{i(n)} - (n+s-1) + j$, with $j = 1, 2, \ldots, s-1$. Hence,

$$H_{n,s,0}(z) = \sum_{h=0}^{\overline{C}_{i(n)}-(n+s-1)} \frac{a_{i(n)}(h)}{\sum_{k=0}^{\overline{C}_{i(n)}-n} a_{i(n)}(k)} + \sum_{j=1}^{s-1} \frac{a_{i(n)}(\overline{C}_{i(n)} - (n+s-1)+j)}{\sum_{k=0}^{\overline{C}_{i(n)}-n} a_{i(n)}(k)} z^j.$$

Case 2. *Slot n and slot $n+s-1$ lie in the same cycle, but in difference subcycles.*
In this case, the computation of $H_{n,s,\delta_n}(z)$ is divided into three parts. Firstly, we determine the number of on-slots between the start of slot n and the start of subcycle $i(n)+1$. Secondly, we compute the number of on-slots in the subcycles $i(n)+1, i(n)+2, \ldots, i(n+s-1)-1$ together. Thirdly, we determine the number of on-slots between the start of subcycle $i(n+s-1)$ and the end of slot $n+s-1$.

By similar arguments as in case 1, the probability generating function $P_{1,n,s}(z)$ of the number of on-slots between the start of slot n and the start of subcycle $i(n)+1$ is

$$P_{1,n,s}(z) = \begin{cases} \displaystyle\sum_{h=0}^{\overline{C}_{i(n)}-n} \frac{a_{i(n)}(h)}{\sum_{k=0}^{\overline{C}_{i(n)}-n} a_{i(n)}(k)} z^h, & \text{if slot } n \text{ is an off-slot,} \\ z^{\overline{C}_{i(n)}-n+1}, & \text{if slot } n \text{ is an on-slot.} \end{cases}$$

The probability generating function $P_{2,n,s}(z)$ of the total number of on-slots in the subcycles $i(n)+1, i(n)+2, \ldots, i(n+s-1)-1$ is obviously

$$P_{2,n,s}(z) = \prod_{j=i(n)+1}^{i(n+s-1)-1} \alpha_j(z).$$

Further, using similar arguments as in case 1, the probability generating function $P_{3,n,s}(z)$ of the number of on-slots between the start of subcycle $i(n+s-1)$ and the end of slot $n+s-1$ is

$$P_{3,n,s}(z) = \sum_{h=0}^{\overline{C}_{i(n+s-1)}-(n+s-1)} a_{i(n+s-1)}(h) + \sum_{h=1}^{(n+s-1)-\overline{C}_{i(n+s-1)-1}} a_{i(n+s-1)}(\overline{C}_{i(n+s-1)} - (n+s-1)+h) z^h.$$

Combining these results, we have $H_{n,s,\delta_n}(z) = P_{1,n,s}(z) P_{2,n,s}(z) P_{3,n,s}(z)$.

Case 3. *Slot n and the s-th slot after the start of slot n lie in different cycles.*
For convenience, we denote the s-th slot after the start of slot n by $n+s-1$, and the subcycle of the cycle in which this slot lies $i(n+s-1)$. In this case, the determination of the probability generating function $H_{n,s,\delta_n}(z)$ consists of three parts. Firstly, we determine the probability distribution of the number of on-slots between the arrival instant and the start of the first cycle after the arrival. Secondly, we use a recursive relation to incorporate the number of on-slots between the start of the first cycle after the arrival and the start of the cycle in which slot $n+s-1$ lies. Finally, we determine the number of on-slots between the start of the cycle in which slot $n+s-1$ lies and slot $n+s-1$. In Figure 4.9, we represent these three parts.

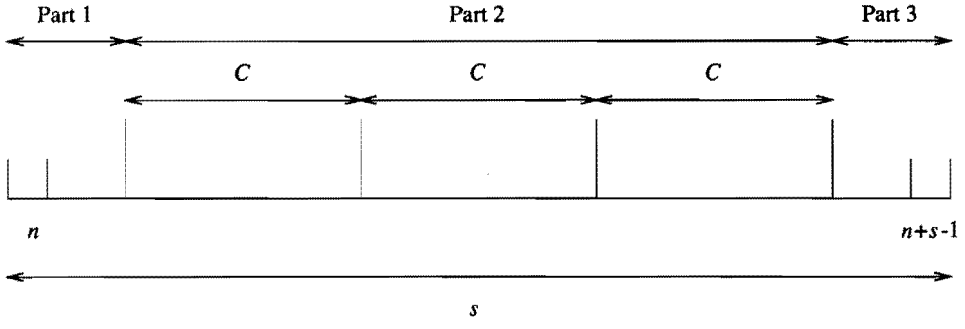

Figure 4.9: *A representation of the three parts to determine the number of on-slots between the slot of arrival n and slot n + s − 1.*

Using the arguments in case 1, the probability generating function $G_{1,n}(z)$ of the number of on-slots in the first part (that is, part 1) is given by

$$G_{1,n}(z) = \begin{cases} \displaystyle\sum_{h=0}^{\overline{C}_{i(n)}-n} \frac{a_{i(n)}(h)}{\sum_{j=0}^{\overline{C}_{i(n)}-n} a_{i(n)}(j)} z^h \prod_{k=i(n)+1}^{N} \alpha_k(z), & \text{if slot } n \text{ is an off-slot,} \\ z^{\overline{C}_{i(n)}-n+1} \displaystyle\prod_{k=i(n)+1}^{N} \alpha_k(z), & \text{if slot } n \text{ is an on-slot.} \end{cases} \tag{4.16}$$

For the second part, let $G_{k,n}(z)$ denote the probability generating function of the number of on-slots between the slot of arrival n and the start of the k-th cycle after the arrival, with $k = 1, 2, 3, \ldots$. Then, it is easily verified that the following recursive relation holds

$$G_{k+1,n}(z) = G_{k,n}(z) \prod_{j=1}^{N} \alpha_j(z), \qquad k = 1, 2, 3, \ldots . \tag{4.17}$$

Finally, for part 3, let the $\bar{s}$-th slot in subcycle $i(n+s-1)$ correspond to slot $n+s-1$. Then, the probability generating function $\hat{G}_{n,s}(z)$ of the number of on-slots between the start of the cycle in which slot $n+s-1$ lies and slot $n+s-1$ is given by

$$\hat{G}_{n,s}(z) = \left(\prod_{j=1}^{i(n+s-1)-1} \alpha_j(z) \right) \left(\sum_{k=0}^{C_{i(n+s-1)}-\bar{s}} a_{i(n+s-1)}(k) + \sum_{k=1}^{\bar{s}} a_{i(n+s-1)}(C_{i(n+s-1)} - \bar{s} + k) z^k \right). \tag{4.18}$$

Combining the relations (4.16), (4.17), and (4.18) yields the probability generating function $H_{n,s,\delta_n}(z)$ for case 3.

From $H_{n,s,\delta_n}(z)$, we can now determine the probability distribution of the number of on-slots between slot n and the s-th slot after the start of slot n. Let $\{h_{n,s,\delta_n}(j), j = 0, 1, \ldots, s\}$ denote this probability distribution. Then, the conditional probability $\omega_{n,\delta_n}(w, s)$ that w service phases are completed in s

slots or less, given the slot of arrival n and δ_n, equals

$$\omega_{n,\delta_n}(w,s) = \sum_{j=w}^{s} \binom{j-1}{w-1} (1-\beta)^w \beta^{j-w} h_{n,s,\delta_n}(j), \quad w = 1, 2, \ldots, s; \; s = 1, 2, 3, \ldots. \tag{4.19}$$

Combining the above steps yields the following procedure for the computation of the sojourn-time distribution.

An algorithm for the computation of the sojourn-time distribution

Step 1. Initialisation. Set $G_1(z) := 1$ and $n := 1$, and compute $G_k(z)$, for $k = 1, 2, \ldots, \lceil s/C \rceil$, using the recursive relation

$$G_{k+1}(z) = G_k(z) \prod_{j=1}^{N} \alpha_j(z).$$

Step 2. Iteration.

(*i*) For $w = 1, 2, \ldots, s$, compute the probabilities $\Pr\{W_{n,0} = w\}$ and $\Pr\{W_{n,1} = w\}$ that the number of service phases just after the arrival in slot n equals w, given that this slot is an off-slot or on-slot, according to equation (4.15).

(*ii*) Compute the probability distribution $\{h_{n,s,\delta_n}(j), j = 0, 1, 2, \ldots, s\}$ of the number of on-slots between slot n and the s-th slot after the start of slot n in the way described above (using the generating functions $G_k(z)$ as computed in Step 1). For $w = 1, 2, \ldots, s$, compute the probabilities $\omega_{n,0}(w, s)$ and $\omega_{n,1}(w, s)$ using (4.19).

(*iii*) Compute the probabilities $\Pr\{S_{n,0} \le s\}$ and $\Pr\{S_{n,1} \le s\}$ that the sojourn time of the customer arriving in slot n, given that this slot is an off-slot or on-slot, is at most s slots

$$\Pr\{S_{n,0} \le s\} = \sum_{w=1}^{s} \Pr\{W_{n,0} = w\} \omega_{n,0}(w, s),$$

$$\Pr\{S_{n,1} \le s\} = \sum_{w=1}^{s} \Pr\{W_{n,1} = w\} \omega_{n,1}(w, s).$$

Step 3. If $n < C$, then set $n = n + 1$ and execute Step 2. Otherwise, compute the probability that the sojourn time of an arbitrary customer is at most s slots (using (4.13) and (4.14))

$$\Pr\{S \le s\} = \sum_{n=1}^{C} (\psi_{n,0} \Pr\{S_{n,0} \le s\} + \psi_{n,1} \Pr\{S_{n,1} \le s\}).$$

To conclude this section, we determine the time complexity of this algorithm. The computations in Step 1 have to be executed only once. The time complexity for the computation of the generating functions $G_k(z)$ is $O(\lceil s/C \rceil sC)$. Part (*i*) of Step 2 requires a number of operations that is proportional to sB^*, with $B^* := \max\{B_n, n = 1, 2, \ldots, C\}$. Part (*ii*) of this step has time complexity $O(s^2)$. Finally, the number of operations to execute part (*iii*) of Step 2 is proportional to s. From these arguments, we conclude that the total time complexity of Step 2 is $O(s^2)$. Let r be equal to C plus the number of slots in a cycle that can be an off-slot in one cycle and an on-slot in another. Then, Step 2 has to be executed r times, so that the total number of operations for this algorithm is proportional to rs^2.

4.5 A matrix-analytic approach

For the queueing system introduced in Section 4.2, an exact technique for analysing the queue-length process is the matrix-analytic approach as introduced in Neuts [1981,1989]. To apply the matrix-analytic approach for analysing Markov chains, these chains have to fulfil the following two conditions. Firstly, the state space of these chains has to be two dimensional, and it may be denumerably large in one dimension only (the first dimension, say). Secondly, for all i, a transition from state (i, j) to state (i', j') is only possible if either $i' \leq i + 1$ or $i' \geq i - 1$. For this latter condition, transitions are said to be skip-free to the right or to the left with respect to the first dimension. The first condition indicates that the first dimension should represent the queue length. The second condition suggests that we have to consider the queue length at slot boundaries, because otherwise the corresponding Markov chains may not be skip-free to the right or to the left. So, these chains are periodic with period C. Yet, different Markov chains describing the queue-length process can be defined that can be analysed by the matrix-analytic approach.

The matrix-analytic approach consists of solving a polynomial matrix equation first, and then several systems of linear equations. The degree of this polynomial matrix equation and the number of equations to be solved depend on the Markov chain defined. In this section, we focus on the Markov chain describing the queue-length process in terms of the number of customers. For this case, the degree of the polynomial equation is minimal and equal to two. Hence, we can use an adapted version of the algorithm of Ramaswami & Lucantoni [1985] to determine the sojourn-time distribution of customers. Since the algorithm in Section 4.4 is exact, once the stationary imbedded queue-length distribution has been found, we can compare the computational effort of these algorithms.

This section is organised as follows. The Markov chain to be studied is defined and analysed by the matrix-analytic approach in Section 4.5.1, because Ramaswami & Lucantoni [1985] use notation and results from this analysis. For this chain, they exploit its structure to determine the waiting-time distribution of a customer. In Section 4.5.2, we adapt this algorithm to determine the sojourn-time distribution of a customer and show that the procedure in Section 4.4 is more efficient from a computational point of view than this algorithm for most of the important cases.

4.5.1 The queue-length process

For convenience, we make in this section the additional assumption that the service times are independent and identically distributed. More precisely, for $n = 1, 2, \ldots, C$, we assume that $B_n = B$ and $b_n(i) = b(i)$, with $i = 1, 2, \ldots, B$. In this section, we first describe the Markov chain to study the queue-length process. After that, we analyse this chain by the matrix-analytic approach. Actually, we apply the matrix-geometric approach of Neuts [1981].

Let X_t denote the number of customers in the queue under consideration, including a possible customer in service, at the t-th slot boundary, for $t = 1, 2, 3, \ldots$. Further, let Y_t denote the triple consisting of, firstly, the slot in the cycle that starts at the t-th slot boundary, secondly, whether this slot is an on-slot or not, and finally, the residual number of uncompleted service phases of the customer in service at the t-th slot boundary. Notice that the state space of Y_t is finite, so that this state space is essentially one dimensional. More precisely, if r denotes the number of slots in the cycle C plus the number of slots that can be an on-slot in one cycle and an off-slot in another cycle, then Y_t can assume rB different states. Therefore, we consider the state space of Y_t as being one dimensional.

The stochastic process $\{(X_t, Y_t), t = 1, 2, 3, \ldots\}$ is a two-dimensional discrete-time Markov chain with state space $\mathcal{S}$. This chain is periodic with period C, and as before it is assumed to be irreducible. Because, in Section 4.3, we assumed that the average number of slots of work arriving per cycle is strictly less than the average service capacity per cycle, this Markov chain is positive recurrent (cf. Pakes [1969]), so that it has a unique stationary distribution.

To determine this stationary distribution by the matrix-analytic approach, we partition the state space $\mathcal{S}$ into levels. Level i, for $i = 0, 1, 2, \ldots$, is defined as the set $\mathcal{S}_i$ of all states for which the number of customers is equal to i, that is,

$$\mathcal{S}_i := \{(i, j) | (i, j) \in \mathcal{S}\}, \qquad i = 0, 1, 2, \ldots .$$

We order the states at these levels lexicographically. Using this partition of the state space, the transition matrix $\boldsymbol{P}$ of the Markov chain has the following block tri-diagonal form

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{B}_0 & \boldsymbol{B}_1 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{B}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \cdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where the matrices $\boldsymbol{B}_0$, $\boldsymbol{B}_1$, and $\boldsymbol{B}_2$ denote the matrices consisting of the transition probabilities from states at level 0 to states at level 0, from states at level 0 to states at level 1, and from states at level 1 to states at level 0, respectively. For $i = 1, 2, 3, \ldots$, the square matrices $\boldsymbol{A}_0$, $\boldsymbol{A}_1$, and $\boldsymbol{A}_2$ denote the matrices consisting of the transition probabilities from states at level i to states at level $i + 1$, to states at level i, and to states at level $i - 1$, respectively. The precise entries of these matrices are not relevant for the exposition in this section. Markov chains for which the transition matrix has a block tri-diagonal form are also known as quasi-birth-death processes.

Now, we can apply the matrix-analytic approach to determine the stationary distribution of the Markov chain. For $i = 1, 2, 3, \ldots$, let π_i, denote the row vector consisting of the stationary probabilities $\pi_{i,j}$, for the states $(i, j) \in \mathcal{S}_i$. The set of row vectors $\{\pi_i, i = 0, 1, 2, \ldots\}$ is the unique solution of the equilibrium equations

$$\pi_0 = \pi_0 \boldsymbol{B}_0 + \pi_1 \boldsymbol{B}_2, \tag{4.20}$$

$$\pi_1 = \pi_0 \boldsymbol{B}_1 + \pi_1 \boldsymbol{A}_1 + \pi_2 \boldsymbol{A}_2, \tag{4.21}$$

$$\pi_i = \pi_{i-1} \boldsymbol{A}_0 + \pi_i \boldsymbol{A}_1 + \pi_{i+1} \boldsymbol{A}_2, \qquad i = 2, 3, 4, \ldots ,$$

and the normalisation equation. Further, define the matrix $\boldsymbol{A}$ as

$$\boldsymbol{A} := \boldsymbol{A}_0 + \boldsymbol{A}_1 + \boldsymbol{A}_2. \tag{4.22}$$

Clearly, the matrix $\boldsymbol{A}$ is a stochastic matrix. We assume that this matrix corresponds to a transition matrix of a finite, irreducible and periodic Markov chain. Then, by Theorem 1.3.2 in Neuts [1981], we have

$$\pi_i = \pi_{i-1} \boldsymbol{R} = \pi_1 \boldsymbol{R}^{i-1}, \qquad i = 1, 2, 3, \ldots ,$$

where the matrix $\boldsymbol{R}$ is the non-negative solution of the quadratic matrix equation

$$\boldsymbol{A}_0 + \boldsymbol{R}\boldsymbol{A}_1 + \boldsymbol{R}^2\boldsymbol{A}_2 = \boldsymbol{R},$$

with $\boldsymbol{R} \leq \boldsymbol{R}$, for any other non-negative solution $\boldsymbol{R}$ of this matrix equation. The remaining vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ can be obtained from the matrix equations (4.20) and (4.21), and the normalisation equation. When $\boldsymbol{e}$ denotes the column vector with all its entries equal to one, the normalisation equations reads

$$\boldsymbol{\pi}_0\boldsymbol{e} + \boldsymbol{\pi}_1(\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{e} = 1,$$

since the spectral radius of $\boldsymbol{R}$ is positive and strictly less than one (cf. Theorem 1.2.1 and Lemma 1.3.5 in Neuts [1981]). The stationary distribution is said to have a modified matrix-geometric form. Next, we use the algorithm of Ramaswami & Lucantoni [1985] to determine the sojourn-time distribution of customers.

4.5.2 The sojourn-time distribution

For the continuous-time $GI/Ph/1$ queueing system, Ramaswami & Lucantoni [1985] present an algorithm for the computation of the waiting-time distribution. Their algorithm can also be used to determine the waiting-time distribution for quasi-birth-death processes by exploiting the similarity in the structure of these processes and of the $GI/Ph/1$ queueing system. In this section, we first adapt this algorithm for computing the sojourn-time distribution of a customer in statistical equilibrium. After that, we compare the computational effort of this algorithm with that of Section 4.4.2. We remark that both these algorithms are exact, provided that the stationary queue-length distribution has been found.

The sojourn time of a customer corresponds to the time until absorption in the Markov chain with transition matrix $\boldsymbol{P}$, in which all subsequent arrivals are ignored, that is, with transition matrix

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{B}_2 & \boldsymbol{A}_1 + \boldsymbol{A}_0 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{A}_2 & \boldsymbol{A}_1 + \boldsymbol{A}_0 & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_2 & \boldsymbol{A}_1 + \boldsymbol{A}_0 & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_2 & \boldsymbol{A}_1 + \boldsymbol{A}_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

and with initial probability vector

$$(\boldsymbol{0}, \tilde{\boldsymbol{\pi}}_1, \tilde{\boldsymbol{\pi}}_2, \tilde{\boldsymbol{\pi}}_3, \ldots).$$

The row vector $\tilde{\boldsymbol{\pi}}_i$ consists of the probabilities $\tilde{\pi}_{i,j}$ that the state of the system immediately after the arrival is state $(i, j) \in \mathcal{S}$. We first determine the initial probability vector, and then the sojourn-time distribution.

The probability that an arbitrary customer arrives in slot n of the cycle is equal to

$$\frac{p_n}{\sum_{k=1}^{C} p_k}, \qquad n = 1, 2, \ldots, C.$$

We condition on the slot of arrival, and recall that $\sum_{i=0}^{\infty}\sum_{j=1}^{rB} \pi_{i,j} = 1$ and that the period of the Markov chain, with transition matrix $\boldsymbol{P}$, is C. Then, by the BASTA property, the state of this chain upon arrival

is $(i, j) \in \mathcal{S}_i$ with probability $\pi_{i,j}C$. Hence, we have

$$\tilde{\pi}_{i,j} = \pi_{i-1,j}C\frac{p_n}{\sum_{k=1}^{C} p_k}, \qquad i = 2, 3, 4, \ldots, \tag{4.23}$$

where j corresponds (among other things) to slot n. From (4.23), we see that there is a square matrix $\boldsymbol{C}$, with nonzero entries at the main diagonal only, such that

$$\tilde{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_{i-1}\boldsymbol{C}, \quad i = 2, 3, 4, \ldots.$$

For $\tilde{\boldsymbol{\pi}}_1$, we can obtain a similar expression as (4.23). But, since level zero may have less states than level one, we present this relation as

$$\tilde{\boldsymbol{\pi}}_1 = \boldsymbol{\pi}_0\tilde{\boldsymbol{B}}_1\boldsymbol{C},$$

where the matrix $\tilde{\boldsymbol{B}}_1$ maps the states at level 0 to states at level 1 just after the arrival.

Now, the initial probability vector of the Markov chain with transition probability matrix $\boldsymbol{P}$ is given by

$$(\boldsymbol{0}, \boldsymbol{\pi}_0\tilde{\boldsymbol{B}}_1\boldsymbol{C}, \boldsymbol{\pi}_1\boldsymbol{C}, \boldsymbol{\pi}_1\boldsymbol{R}\boldsymbol{C}, \boldsymbol{\pi}_1\boldsymbol{R}^2\boldsymbol{C}, \ldots).$$

Before we determine the sojourn-time distribution, we introduce some additional notation to facilitate the exposition.

Consider the finite Markov chain consisting of rB states, with transition probability matrix $\boldsymbol{A}$ as defined by (4.22). Let, for $n = 0, 1, 2, \ldots$ and $m = 0, 1, \ldots, n$, the square matrix $\boldsymbol{K}_m^{(n)}$ of order rB have as its (i, j) entry the conditional probability that, given that this finite chain starts in state i and that it has made n transitions, m of these transitions correspond to customer departures and that the state of this chain after these n transitions is j. Then, we have clearly

$$\boldsymbol{K}_0^{(n)} = (\boldsymbol{A}_1 + \boldsymbol{A}_0)^n, \quad n = 0, 1, 2, \ldots, \quad \text{and} \quad \boldsymbol{K}_1^{(1)} = \boldsymbol{A}_2, \tag{4.24}$$

and the recursive relation

$$\boldsymbol{K}_m^{(n+1)} = \boldsymbol{K}_{m-1}^{(n)}\boldsymbol{A}_2 + \boldsymbol{K}_m^{(n)}(\boldsymbol{A}_1 + \boldsymbol{A}_0), \qquad n = 1, 2, 3, \ldots \text{ and } m = 1, 2, \ldots, n+1. \tag{4.25}$$

Further, notice that

$$\boldsymbol{K}_m^{(n)} = \boldsymbol{0}, \qquad m > n. \tag{4.26}$$

Let $\Pr\{S > s\}$ denote the probability that the sojourn time of the customer is larger than s slots, with $s = 1, 2, 3, \ldots$. Then, this probability can be related to the matrices $\boldsymbol{K}_m^{(n)}$ (and hence, to the finite Markov chain with transition probability matrix $\boldsymbol{A}$) as follows. Suppose that the state of the system just after the arrival is state $(k, i) \in \mathcal{S}$, which corresponds to state i of the finite Markov chain. If this chain makes s transitions (starting in state i) and at most $k - 1$ of these transitions correspond to a customer departure, then the sojourn time of the customer is larger than s slots. The probability that at most $k - 1$ customer departures correspond to these s transitions (starting in state i) is given by the i-th entry in

$$\sum_{m=0}^{k-1} \boldsymbol{K}_m^{(s)}\boldsymbol{e}.$$

The probability of state (k, i) just after the arrival is given by $\tilde{\pi}_{k,i}$, so that we have

$$\Pr\{S > s\} = \sum_{k=1}^{\infty} \sum_{n=0}^{k-1} \tilde{\pi}_k K_n^{(s)} e, \qquad s = 1, 2, 3, \ldots . \tag{4.27}$$

When we use the approximations of $\tilde{\pi}_k$ in Section 4.3, we have an alternative way of computing the sojourn-time distribution. Calculating the sojourn-time distribution in this way, most of the computational effort is used for computing the matrices $K_n^{(s)}$, for $n = 0, 1, 2, \ldots$. Using the relations (4.24) and (4.25), we multiply $s + k$ square matrices of order rB for computing these matrices. The time complexity of multiplying two square matrices of order rB is $O((rB)^3)$. So, the total time complexity for computing the matrices $K_n^{(s)}$, for $n = 0, 1, 2, \ldots$, is $O((s + k)(rB)^3)$. The number of operations for computing the sojourn-time distribution by the algorithm in Section 4.4 is proportional to rs^2. Hence, as long as the magnitude of s is at most proportional to $r^2 B^3$, the procedure in Section 4.4 is at least as efficient as the one in this section.

If we use the probabilities $\tilde{\pi}_{i,j}$ as given by (4.23), then relation (4.27) can be computed in a recursive fashion, as we show in the next lemma.

Theorem 4.5.1 *The sojourn-time distribution can be computed by*

$$\Pr\{S > s\} = (\pi_0 \tilde{B}_1 + \pi_1 (I - R)^{-1}) C (A_1 + A_0)^s e + \pi_1 C (I - R)^{-1} H_s e, \qquad s = 1, 2, 3, \ldots ,$$

where

$$H_1 = A_2, \quad \text{and} \quad H_{n+1} = H_n (A_1 + A_0) + K_0^{(n)} A_2 + R H_n A_2, \quad n = 1, 2, 3, \ldots .$$

Proof. The proof of this lemma is similar to the one in in Ramaswami & Lucantoni [1985]. From (4.27), we have, for $s = 1, 2, 3 \ldots$,

$$\Pr\{S > s\} = \sum_{k=1}^{\infty} \sum_{m=0}^{k-1} \tilde{\pi}_k K_m^{(s)} e.$$

We partition the sum over m into $m = 0$ and $m = 1, 2, \ldots, k - 1$, and we use relation (4.24) to obtain

$$\begin{aligned}
\Pr\{S > s\} &= \sum_{k=1}^{\infty} \tilde{\pi}_k (A_1 + A_0)^s e + \sum_{k=2}^{\infty} \sum_{m=1}^{k-1} \tilde{\pi}_k K_m^{(s)} e \\
&= (\pi_0 \tilde{B}_1 + \sum_{k=1}^{\infty} \pi_1 R^{k-1}) C (A_1 + A_0)^s e + \sum_{k=2}^{\infty} \sum_{m=1}^{k-1} \pi_{k-1} C K_m^{(s)} e \\
&= (\pi_0 \tilde{B}_1 + \pi_1 (I - R)^{-1}) C (A_1 + A_0)^s e + \sum_{k=1}^{\infty} \sum_{m=1}^{k} \pi_k C K_m^{(s)} e.
\end{aligned}$$

So, it remains to rewrite the last part of the right-hand side of this equation. Interchanging the summons over k and m, and using relation (4.26) yield

$$\begin{aligned}
\sum_{k=1}^{\infty} \sum_{m=1}^{k} \pi_k C K_m^{(s)} e &= \sum_{m=1}^{s} \sum_{k=m}^{\infty} \pi_k C K_m^{(s)} e = \pi_1 C \sum_{m=1}^{s} \sum_{k=m}^{\infty} R^{k-1} K_m^{(s)} e \\
&= \pi_1 C (I - R)^{-1} \sum_{m=1}^{s} R^{m-1} K_m^{(s)} e.
\end{aligned}$$

Define

$$H_s := \sum_{m=1}^{s} R^{m-1} K_m^{(s)}.$$

Then, by relation (4.25), we obtain the recursive relation, for $n = 0, 1, 2, \ldots$,

$$\begin{aligned} H_{n+1} &= \sum_{m=1}^{n+1} R^{m-1} K_m^{(n+1)} = \sum_{m=1}^{n+1} R^{m-1} (K_m^{(n)} (A_1 + A_0) + K_{m-1}^{(n)} A_2) \\ &= \sum_{m=1}^{n} R^{m-1} K_m^{(n)} (A_1 + A_0) + K_0^{(n)} A_2 + R \sum_{m=2}^{n+1} R^{m-2} K_{m-1}^{(n)} A_2 \\ &= H_n (A_1 + A_0) + K_0^{(n)} A_2 + R H_n A_2. \end{aligned}$$

Finally, observing that $H_1 = K_1^{(1)} = A_2$, completes the proof. □

4.6 Numerical examples

In this section, we present some numerical examples. These examples first of all illustrate that the GT technique gives excellent results for the approximation of the stationary imbedded queue-length distribution, even if the geometric tail behaviour is imposed on the stationary probabilities of states fairly close to the boundary of the state space. Further, these examples show that, to obtain the same accuracy, the computational effort of the GT technique is much lower than that of simple truncation.

As an example, we consider a cycle that consists of one subcycle and assume that the length A of the on-periods is constant. Customers are supposed to arrive according to a homogeneous Bernoulli process with parameter p. For the service of customers, we consider the following three cases: it consists of three deterministic service phases, of one geometric service phase requiring on average three on-slots, and of two geometric service phases, each of them requiring on average one and a half on-slots. Notice that these service times correspond to service times that are deterministic, geometric, and negative binomial, respectively, and that the average service times are equal to three on-slots. These three cases are denoted by Det, Geo, and NBin. In the examples, we examine for the cycle length C the cases $C = 60, 120, 180$ and for A the cases $A = 0.25C, 0.50C, 0.75C$.

We begin with illustrating that the GT technique gives excellent results when the geometric tail behaviour is imposed on the stationary probabilities of states fairly close to the boundary of the state space. In Table 4.1, we display the average number of service phases in statistical equilibrium at the first slot boundary of a cycle, as obtained by the application of the GT technique for different values of the threshold J. In Table 4.2, we list these results for the standard deviation of this number of service phases. In the columns Exact, we display the 'exact' results as obtained by truncating the states $j > T$ of the Markov chain, with T sufficiently large. For all the examples, we have set the effective utilisation factor ρ equal to 0.95, that is, the arrival intensity p is chosen such that $3(C/A)p = 0.95$. As these tables show, the GT technique gives good results for all examples, even if J is fairly small; using $J = 40$ gives results that are sufficiently accurate for most of the examples.

Next, we compare the computational efficiency of the GT technique with that of simple truncation. For this purpose, we list in Table 4.3 the minimal thresholds J_E and J_{std} (the minimal thresholds T_E and T_{std}) such that the average and the standard deviation of the number of service phases, respectively,

	C	A	Exact	$J=20$	$J=40$	$J=60$	$J=80$	$J=100$
Det	60	15	23.337	23.346	23.337	23.337	23.337	23.337
		30	20.391	20.403	20.391	20.391	20.391	20.391
		45	18.433	18.379	18.433	18.433	18.433	18.433
	120	30	22.035	22.044	22.035	22.035	22.035	22.035
		60	18.759	18.608	18.761	18.759	18.579	18.759
		90	16.805	16.406	16.811	16.805	16.806	16.805
	180	45	21.062	20.994	21.062	21.062	21.062	21.062
		90	17.561	17.141	17.578	17.560	17.561	17.561
		135	15.596	14.870	15.622	15.596	15.596	15.597
Geo	60	15	14.006	14.006	14.006	14.006	14.006	14.006
		30	12.949	12.949	12.949	12.949	12.949	12.949
		45	12.334	12.333	12.334	12.334	12.334	12.334
	120	30	13.430	13.430	13.430	13.430	13.430	13.430
		60	12.217	12.216	12.217	12.217	12.217	12.217
		90	11.615	11.618	11.615	11.615	11.615	11.615
	180	45	12.991	12.990	12.991	12.991	12.991	12.991
		90	11.660	11.662	11.660	11.660	11.660	11.660
		135	11.052	11.068	11.052	11.052	11.052	11.052
NBin	60	15	18.688	18.688	18.688	18.688	18.688	18.688
		30	16.661	16.665	16.661	16.661	16.661	16.661
		45	15.369	15.381	15.368	15.369	15.369	15.369
	120	30	17.730	17.735	17.730	17.730	17.730	17.730
		60	15.461	15.486	15.461	15.461	15.461	15.461
		90	14.176	14.152	14.175	14.176	14.176	14.176
	180	45	17.013	17.033	17.013	17.013	17.013	17.013
		90	14.570	14.567	14.569	14.571	14.570	14.570
		135	13.274	13.124	13.277	13.274	13.274	13.274

Table 4.1: *The exact values and the approximations of the average number of service phases at the start of a cycle for the examples, for different values of the threshold J and with $\rho = 0.95$.*

have a six-decimal accuracy when applying the GT technique (when using simple truncation). As we see from this table, the computational effort of the GT technique is much lower than that for simple truncation. So, it is more advantageous from a numerical point of view to use the GT technique than truncation. Further, the number of equations to be solved increases dramatically for simple truncation when the utilisation of the system is increased. For the GT technique, however, the thresholds increase only slowly, when the utilisation increases. We remark that these observations are in compliance with those in Tijms & Van de Coevering [1991].

Finally, we list in Table 4.4 the number of boundary equations for the different configurations. We remark that this number exploits the additional structure imposed, compared to the general model described in Section 4.2. Comparing this number of boundary equations with the thresholds for the GT technique in Table 4.1 and Table 4.2, we observe that excellent results are obtained when J is taken considerably smaller than the number of boundary equations.

	C	A	Exact	$J = 20$	$J = 40$	$J = 60$	$J = 80$	$J = 100$
Det	60	15	26.858	26.860	26.858	26.858	26.858	26.858
		30	24.299	24.307	24.299	24.299	24.299	24.299
		45	21.826	21.806	21.827	21.826	21.826	21.826
	120	30	26.676	26.682	26.676	26.676	26.676	26.676
		60	23.977	23.911	23.964	23.977	23.977	23.977
		90	21.406	21.188	21.414	21.406	21.406	21.406
	180	45	26.506	26.482	26.506	26.506	26.506	26.506
		90	23.674	23.451	23.700	23.676	23.674	23.674
		135	21.007	20.531	21.034	21.007	21.007	21.007
Geo	60	15	15.466	15.466	15.466	15.466	15.466	15.466
		30	14.633	14.633	14.633	14.633	14.633	14.633
		45	13.832	13.832	13.832	13.832	13.832	13.832
	120	30	15.408	15.408	15.408	15.408	15.408	15.408
		60	14.528	14.527	14.528	14.528	14.528	14.528
		90	13.698	13.700	13.698	13.698	13.698	13.698
	180	45	15.353	15.353	15.353	15.353	15.353	15.353
		90	14.426	14.427	14.426	14.426	14.426	14.426
		135	13.565	13.576	13.565	13.565	13.565	13.565
NBin	60	15	21.176	21.176	21.176	21.176	21.176	21.176
		30	19.485	19.480	19.485	19.485	19.485	19.485
		45	17.856	17.862	17.855	17.856	17.856	17.856
	120	30	21.057	21.059	21.057	21.057	21.057	21.057
		60	19.271	19.286	19.271	19.271	19.271	19.271
		90	17.578	17.567	17.577	17.578	17.578	17.578
	180	45	20.943	20.952	20.943	20.943	20.943	20.943
		90	19.068	19.070	19.067	19.068	19.068	19.068
		135	17.310	17.214	17.312	17.310	17.310	17.310

Table 4.2: *The exact values and the approximations of the standard deviation of the number of service phases at the start of a cycle, for different values of the threshold J and with $\rho = 0.95$.*

			Deterministic				Geometric				Negative Binomial			
ρ	C	A	J_E	J_{std}	T_E	T_{std}	J_E	J_{std}	T_E	T_{std}	J_E	J_{std}	T_E	T_{std}
0.75	60	15	40	40	110	110	20	20	50	60	30	30	70	80
		30	40	50	80	100	20	20	50	60	30	30	70	70
		45	40	40	80	90	30	30	60	60	30	40	60	70
	120	30	40	50	90	130	20	20	50	60	40	40	80	80
		60	50	50	80	90	30	30	50	60	40	40	60	70
		90	50	60	70	80	30	30	50	50	40	40	80	70
	180	45	50	60	90	100	30	30	50	60	30	40	70	90
		90	60	60	90	90	30	30	60	50	40	40	70	80
		135	50	60	70	90	30	30	50	50	40	50	60	60
0.90	60	15	50	40	260	250	20	20	150	170	30	30	210	230
		30	50	50	240	230	30	30	140	180	40	40	190	270
		45	60	60	240	250	30	30	140	140	50	50	170	200
	120	30	50	50	260	260	30	30	150	200	40	40	210	200
		60	60	60	230	230	30	40	140	170	50	60	200	200
		90	70	90	230	230	30	30	130	150	70	60	160	190
	180	45	60	60	250	280	30	30	140	170	40	50	210	220
		90	70	90	220	220	30	50	150	160	60	70	180	200
		135	90	110	210	220	40	50	140	140	60	80	160	180
0.95	60	15	40	40	480	530	20	20	300	310	30	30	370	430
		30	60	50	440	490	30	30	250	280	30	30	340	390
		45	50	60	430	450	30	30	250	280	40	50	330	350
	120	30	50	50	480	530	20	30	280	300	50	40	370	450
		60	60	60	450	550	30	30	260	350	50	60	380	380
		90	70	70	400	460	30	30	260	280	50	70	330	340
	180	45	60	70	490	530	30	30	270	310	50	40	370	420
		90	80	90	470	480	30	30	260	280	50	60	350	440
		135	90	90	390	450	40	40	280	270	60	60	300	350

Table 4.3: *The thresholds J_E and J_{std} for the GT technique and the thresholds T_E and T_{std} for the simple truncation in order to compute the average and the standard deviation of the number of service phases in six-decimal accuracy for the examples with $\rho = 0.95$.*

	$C = 60$			$C = 120$			$C = 180$		
	A			A			A		
	15	30	45	30	60	90	45	90	135
Det	165	150	135	330	300	270	495	450	405
Geo	60	60	60	120	120	120	180	180	180
NBin	120	120	120	240	240	240	360	360	360

Table 4.4: *The number of boundary equations for the examples.*

4.7 Conclusions

In this chapter, we have shown that the GT technique is applicable to a broad and important class of arrival processes and service time distributions for discrete-time queueing systems with periodic service. This class contains almost all of the possible arrival processes and service-time distributions that fulfil the condition that the equilibrium equations of the Markov chain, describing the queue-length process at the start of cycles, constitute a homogeneous linear difference equation with constant coefficients. Further, numerical examples show that this technique yields excellent results, even if the geometric tail behaviour is imposed on the stationary probabilities of states fairly close to the boundary of the state space. Moreover, these results show that the GT technique is much less sensitive to the utilisation of the system than other techniques like simple truncation.

Further, we have developed an algorithm for determining of the sojourn-time distribution. Given the stationary imbedded queue-length distribution, this algorithm computes the sojourn-time distribution exactly. Moreover, for most practically relevant cases, this algorithm is more efficient than the algorithm of Ramaswami & Lucantoni [1985] to compute the sojourn-time distribution.

5

A Moment-Iteration Technique for Queueing Systems with Periodic Service

5.1 Introduction

In Chapter 4, we applied the GT technique to study the queue-length process in queueing systems with periodic service. This technique makes a detailed use of the service-time distribution of customers. In practice, however, one usually has only (approximate) knowledge about the first two moments of this distribution. So, it makes sense to develop a technique that uses only this information.

In this chapter, we develop such a technique for queueing systems with periodic service. This technique approximates the stationary queue-length distributions at the slot boundaries in the cycle. Furthermore, it can deal with a larger class of service-time distributions than the class defined in Section 4.2. More specifically, the service time of a customer has a general discrete probability distribution that, as in the previous chapter, may depend on the slot of arrival. Once the stationary queue-length distributions are approximated, we use the algorithm of Section 4.4 to compute the sojourn-time distribution of customers.

The main idea for this approximation technique results from the circumstance that the relation between the queue length at consecutive slot boundaries has a form similar to Lindley's equation for the $GI/G/1$ queueing system (see, for example, Grimmett & Stirzaker [1992]). As mentioned in Chapter 2, De Kok [1989] develops an efficient moment-iteration algorithm for approximating the limiting solution to this equation. Furthermore, this algorithm only uses the first two moments of the interarrival-time and of the service-time distributions of customers. The iteration technique developed in this chapter is an adapted version of this algorithm, and it is called the moment-iteration (MI) technique.

To the first two moments of the interarrival-time and of the service-time distributions, De Kok [1989] fits probability distributions that are easy to evaluate numerically. More precisely, he uses the well-known procedure for continuous distributions as described in Tijms [1986] (we give this procedure in the next section). Since the random variables involved in the MI technique are discrete, it is

more natural to fit discrete distributions instead of continuous ones. In the literature, several procedures for fitting discrete distributions on the first two moments of a non-negative random variable can be found. However, some of these procedures do not capture all possible combinations of these moments, whereas others are not that useful for applying in the MI technique. Therefore, in this chapter, we also develop a novel procedure for fitting discrete distributions by matching the first two moments. The distributions involved capture all possible combinations of the first two moments and they are convenient from a numerical point of view. This procedure is a discrete analogue to the one in Tijms [1986].

We remark that the MI technique can also be used to study the transient behaviour of the queue-length process, but we do not consider this behaviour.

This chapter, which is largely based on Adan, Van Eenige & Resing [1995], is organised as follows. In Section 5.2, we describe the moment-iteration method of De Kok [1989] for the continuous-time $GI/G/1$ queueing system. The queueing system with periodic service is briefly described in Section 5.3. In Section 5.4, we first show that the relation between queue-lengths at consecutive slot boundaries has a structure similar to Lindley's equation. After that, we present an adapted version of the method in Section 5.2 to approximate the stationary queue-length distributions at slot boundaries. The procedure for fitting discrete distributions on the first two moments of a non-negative random variable is presented in Section 5.5. In Section 5.6, we use the approximations of the stationary queue-length distributions to compute the sojourn-time distribution by the algorithm in Section 4.4. To illustrate the performance of the MI technique, we present some numerical examples in Section 5.7. Finally, Section 5.8 gives a summary of this chapter.

5.2 A moment-iteration method for the $GI/G/1$ queueing system

This section shows the moment-iteration method of De Kok [1989] for determining the limiting solution of Lindley's equation for the continuous-time $GI/G/1$ queueing system. Later in this chapter, this will facilitate the step to applying (an adapted version of) this method to the evaluation of queueing systems with periodic service. In fact, this step will appear to be quite natural.

We first give a description of the $GI/G/1$ queueing system. After that, we give Lindley's equation for the waiting-time process of customers in this system, and finally, we approximate the limiting solution of this equation by the moment-iteration method of De Kok [1989]. The idea of iterating such equations has already been employed by, for instance, Bagchi & Templeton [1972], Kleinrock [1976], Ackroyd [1980], and Powell [1986].

Consider a single-server queueing system at which customers arrive at the epochs $t_1, t_2, t_3, \ldots$. For $k = 1, 2, 3, \ldots$, the interarrival times $A_k = t_{k+1} - t_k$ are independent and identically distributed random variables with distribution function $\{A(t), t > 0\}$ and finite mean α. In order to use the fitting procedure in Tijms [1986] for the algorithm of De Kok [1989], it is assumed that the second moment of these times is also finite. The service times of customers are independent and identically distributed random variables $B_1, B_2, B_3, \ldots$ with distribution function $\{B(t), t > 0\}$ and finite mean β. For reasons to be stated later, we assume that the second moment of these times is finite as well. All interarrival times and service times are assumed to be independent, and customers are served in the order of their arrival. Finally, we assume that the queue is stable, that is, we assume that $\beta < \alpha$.

Let the random variable W_k denote the waiting time of the k-th customer, for $k = 1, 2, 3, \ldots$. Sup-

pose that the first customer arrives at $t_1 = 0$ at an empty system, so that $W_1 = 0$. Then, it is easily seen that the following relation holds

$$W_{k+1} = \max\{0, W_k + B_k - A_k\}, \qquad k = 1, 2, 3, \ldots . \tag{5.1}$$

This equation is known as Lindley's equation (cf. Grimmett & Stirzaker [1992]).

Because of the stability assumption, the limiting solution of this equation, for k tending to infinity, exists (see Lindley [1952]). In general, however, it is hard to determine this solution. Therefore, De Kok [1989] develops a method to approximate two important waiting-time characteristics in statistical equilibrium, namely, the probability of waiting and the average waiting time. These quantities are well defined, since we have assumed that the queue is stable, and that the first moment of the interarrival times and the first two moments of the service times are all finite (see, for instance, Asmussen [1987]). This method uses only the first two moments of the interarrival-time and of service-time distributions, and, starting with an initial solution, it iterates Lindley's equation. Furthermore, it involves the fitting of continuous distributions on the first two moments of non-negative random variables. Before presenting this method, we give a well-known fitting procedure.

A common way to fit a continuous distribution to the mean m and the coefficient of variation c (that is, the quotient of the standard deviation and the mean) of a given non-negative random variable is the following (see, for example, Tijms [1986]).

A procedure for fitting a continuous distribution on the first two moments

$0 < c < 1$: Fit a mixture of two Erlang distributions with the same parameter μ. More specifically, fit a distribution with probability density function

$$f(t) = q\frac{\mu(\mu t)^{n-1}}{(n-1)!}e^{-\mu t} + (1-q)\frac{\mu(\mu t)^{n}}{(n)!}e^{-\mu t},$$

where $1/(n+1) \leq c^2 < 1/n$, for certain $n = 1, 2, 3, \ldots$ and

$$q = \frac{(n+1)c^2 - \sqrt{(n+1)(1+c^2) - (n+1)^2c^2}}{1+c^2}, \quad \text{and} \quad \mu = \frac{k+1-q}{m}.$$

$c \geq 1$: Fit a hyperexponential distribution with balanced means. This distribution has probability density function

$$g(t) = q\mu_1 e^{-\mu_1 t} + (1-q)\mu_2 e^{-\mu_2 t},$$

with

$$q = \frac{1}{2}\left(1 + \sqrt{\frac{c^2-1}{c^2+1}}\right), \quad \mu_1 = \frac{2q}{m}, \quad \text{and} \quad \mu_2 = \frac{2(1-q)}{m}.$$

It is easily verified that the parameters involved in this procedure are unique. Moreover, we notice that these distributions can be evaluated well for numerical purposes. Let the generic random variables A and B have distribution function $\{A(t), t > 0\}$ and $\{B(t), t > 0\}$, respectively. Further, W denotes the limiting solution of equation (5.1). The moment-iteration method of De Kok [1989] can be described as follows, where $\mathrm{E}\{W_k\}$, $\mathrm{E}\{W_k^2\}$, and σ_k should be read as the *approximations* for the first two moments and the standard deviation of W_k.

A moment-iteration algorithm for the $GI/G/1$ queueing system

Step 1. Initialisation: Set $\mathrm{E}\{W_1\} = \mathrm{E}\{W_1^2\} := 0$, so that $\sigma_1 = 0$, and set $k := 1$. Define $F_A(\cdot)$ as the distribution obtained by matching the first two moments of the random variable A according to the above procedure.

Step 2. Iteration: Compute the first two moments of $W_k + B$, using the approximations for the first two moments of W_k. Define $F_k(\cdot)$ as the distribution obtained by matching the first two moments of $W_k + B$ according to the above procedure. Compute the approximations for the first two moments and the standard deviation of W_{k+1} using equation (5.1), that is,

$$\begin{aligned} \mathrm{E}\{W_{k+1}\} &= \int_{t=0}^{\infty}\int_{x=t}^{\infty}(x-t)\,\mathrm{d}F_k(x)\,\mathrm{d}F_A(t), \\ \mathrm{E}\{W_{k+1}^2\} &= \int_{t=0}^{\infty}\int_{x=t}^{\infty}(x-t)^2\,\mathrm{d}F_k(x)\,\mathrm{d}F_A(t), \end{aligned}$$

and $\sigma_{k+1} = \sqrt{\mathrm{E}\{W_{k+1}^2\} - (\mathrm{E}\{W_{k+1}\})^2}$, respectively.

Step 3. Stopping criterion: If both $|\mathrm{E}\{W_{k+1}\} - \mathrm{E}\{W_k\}|$ and $|\sigma_{k+1} - \sigma_k|$ are small enough, then execute Step 4. Otherwise, set $k := k + 1$ and repeat Step 2.

Step 4. Approximation: The mean and standard deviation of W are approximated by $\mathrm{E}\{W_{k+1}\}$ and σ_{k+1}, and $\Pr\{W > 0\}$ is approximated by

$$\int_{t=0}^{\infty}(1 - F_k(t))\,\mathrm{d}F_A(t).$$

So, the moment-iteration algorithm only uses the first two moments of the interarrival-time and of the service-time distribution of customers. In De Kok [1989], it is conjectured that this method always terminates. Numerical examples in the same paper show good performance of this method.

5.3 The model

In this chapter, we consider almost the same queueing system as in Chapter 4. The difference is that we extend the class of service-time distributions of Section 4.2. For completeness, we first repeat some notation and conventions. After that, we give the extension of this class of service-time distributions.

As before, we consider one queue only. The time periods during which the server is attending and not attending this queue are called on-periods and off-periods. In a periodic service policy, the time axis consists of time intervals of equal length, called cycles. The number of on- and off-periods in a cycle is the same for all cycles. Moreover, the time instants within a cycle at which off-periods begin are fixed and the same for all cycles. For convenience, we assume that a cycle begins with an off-period. The number of off-periods (and of on-periods) in a cycle is denoted by N. One off-period and the next on-period together are called a subcycle, and slots in on-periods and in off-periods are called on-slots and off-slots, respectively. Further, we again use

C = the number of slots in a cycle,

C_i = the fixed length of the i-th subcycle, $i = 1, 2, \ldots, N$,

A_i = the random variable of the length of the i-th on-period in the cycle, $i = 1, 2, \ldots, N$.

The probability distribution of A_i is denoted by $\{a_i(j), j = 0, 1, \ldots, C_i\}$, for $i = 1, 2, \ldots, N$.

Customers are supposed to arrive according to a periodically time-dependent Bernoulli process. More precisely, with probability p_n exactly one customer arrives in slot n of the cycle and with probability $1 - p_n$ no customer, with $n = 1, 2, \ldots, C$, and the arrivals in different slots are independent. The service times are measured in numbers of (on-)slots of work. This number of slots of work for the customer arriving in slot n is denoted by the positive random variable B_n, where $n = 1, 2, \ldots, C$. We only use the first two moments of these random variables and assume that these moments are finite. So, this extends the class of service-time distributions in Section 4.2. Further, we assume that the service times are known upon arrival and that these times are independent. Finally, we assume that the average number of slots of work arriving in a cycle is smaller than the average service capacity of the server per cycle.

It turns out to be convenient to define the non-negative random variable $\overline{B}_n$ as the number of slots of work arriving in slot n, that is, $\overline{B}_n := K_n B_n$, with K_n and B_n independent and

$$K_n := \begin{cases} 0, & \text{with probability } 1 - p_n, \\ 1, & \text{with probability } p_n. \end{cases}$$

Further, we again assume that arrivals and start of services occur just after slot boundaries, and that service completions occur just before slot boundaries (recall Figure 4.3). Finally, the service of a customer who arrives in an on-period and finds the server being idle, starts immediately.

5.4 The queue-length process

In this section, we develop a moment-iteration technique (or MI technique, for short) for studying the queue-length process. We first consider this process at the start of cycles, and show that this imbedded process can in general not be described by a relation that has a structure similar to Lindley's equation. After that, we demonstrate that the relation between the queue lengths at successive slot boundaries does have this structure. Because of the structure of this relation, we use (an adapted version of) the method in De Kok [1989] for approximating the stationary queue-length distributions at the slot boundaries in the cycle. Throughout this section, we implicitly assume that we have a procedure for fitting tractable *discrete* distributions on the first two moments of a non-negative random variable. In Section 5.5, we shall present such a procedure.

In general, the queue length, measured in numbers of customers or slots of work, can increase during on-periods. In that case, the number of customers served or the number of slots of work handled in a cycle does not only depend on the number of customers or slots of work arrived in this cycle, but also on the slots of arrival. This indicates that the queue-length process at the start of cycles can in general not be described by relations that have the same structure as Lindley's equation (we recall that the queueing system in Chapter 2 is an exception).

However, if we consider the queue lengths at the start of consecutive slots, then the queue-length process can be expressed by relations that have this structure. As an illustration of this circumstance, in Section 5.4.1, we first consider the case that $N = 1$ (that is, one on- and off-period in a cycle) and that the length of the on-periods is deterministic. After that, we describe the small adaptations in order to apply this technique when $N > 1$ and when the lengths of the on-periods are random. In Section 5.4.2, we formalise the ideas of Section 5.4.1 and present the MI technique for approximating the stationary queue-length distributions at the slot boundaries in a cycle.

5.4.1 The queue-length process at slot boundaries

Consider the case that $N = 1$ and that the length of the on-periods is constant. For convenience, let A indicate the length of the on-period in a cycle instead of A_1. In this case, for all cycles, the first $C - A$ slots are off-slots and the last A slots are on-slots. For $k = 1, 2, 3, \ldots$ and $n = 1, 2, \ldots, C$, let $X_{k,n}$ be the (random) number of slots of work in the system at the n-th slot boundary in the k-th cycle. Then, it is clear that, for $k = 1, 2, 3, \ldots$,

$$X_{k,n+1} = \begin{cases} X_{k,n} + \overline{B}_n, & n = 1, 2, \ldots, C - A, \\ \max\{0, X_{k,n} + \overline{B}_n - 1\}, & n = C - A + 1, C - A + 2, \ldots, C, \end{cases} \tag{5.2}$$

where $X_{k,C+1}$ should be read as $X_{k+1,1}$. From the assumptions in Section 5.3, it follows that, for k tending to infinity and n fixed, the probability distribution of $X_{k,n}$ converges to the stationary queue-length distribution at the n-th slot boundary in the cycle.

For k and n both fixed, relation (5.2) has a form similar to Lindley's equation (equation (5.1)). So, once the approximations for the first two moments of $X_{k,n}$ are given, we can apply Step 2 of the algorithm in Section 5.2 to approximate these moments for $X_{k,n+1}$. This suggests to adapt this algorithm, for approximating the first two moments of the stationary queue-length distributions at the slot boundaries in the cycle, as follows. Keep k fixed, and suppose that the approximations for the first two moments of $X_{k,1}$ are given. Then, in a similar way as in Step 2, we use relation (5.2) successively for approximating the first two moments and the standard deviation $\sigma_{k,n+1}$ of $X_{k,n+1}$, for $n = 1, 2, \ldots, C$. In Step 3, we now compute the difference between the approximations for $\mathrm{E}\{X_{k,C+1}\}$ and $\mathrm{E}\{X_{k-1,C+1}\}$ and that between the approximations for $\sigma_{k,C+1}$ and $\sigma_{k-1,C+1}$. If these differences are both small, then we have an approximation for the average and standard deviation of the stationary queue-length distribution at the start of cycles. Otherwise, we set $k := k + 1$, and we repeat the iteration step. To sum up, Step 2 and Step 3 are altered into

Step 2. Iteration: For $n = 1, 2, \ldots, C$,

- (*i*) compute the first two moments of $X_{k,n} + \overline{B}_n$, using the approximations for the first two moments of $X_{k,n}$; fit a probability distribution $F_{k,n}(\cdot)$ to the approximate first two moments of $X_{k,n} + \overline{B}_n$;
- (*ii*) approximate the first two moments and the standard deviation of $X_{k,n+1}$, using relation (5.2) and $F_{k,n}(\cdot)$ as an approximation for the probability distribution of $X_{k,n} + \overline{B}_n$.

Step 3. Stopping criterion: If the approximations for the differences $|\mathrm{E}\{X_{k,C+1}\} - \mathrm{E}\{X_{k-1,C+1}\}|$ and $|\sigma_{k,C+1} - \sigma_{k-1,C+1}|$, are both small enough, then stop. Otherwise, set $X_{k+1,1} := X_{k,C+1}$ and $k := k + 1$, and repeat Step 2.

Given the approximations for the first two moments of the stationary queue-length distribution at the start of a cycle, we use the iteration step to compute these two quantities for the other slot boundaries in the cycle. Using the fitting procedure to be presented in Section 5.5, we obtain an approximation for the stationary queue-length distributions at the slot boundaries in the cycle.

Notice that this algorithm only uses the first two moments of B_n. In Section 5.4.3, we present a refinement of this algorithm. This refinement requires the knowledge of or an approximation for the probability that the queue length does not increase during a slot (that is, the probability $\Pr\{\overline{B}_n = 0\} =$

$1 - p_n$, if slot n is an off-slot, and the probability $\Pr\{\overline{B}_n \leq 1\} = 1 - p_n + p_n \Pr\{B_n = 1\}$, if slot n is an on-slot). Furthermore, we refine Step 2 in the sense that the first two moments of $X_{k,C-A+1}$ are approximated using the relation

$$X_{k,C-A+1} = X_{k,1} + \sum_{n=1}^{C-A} \overline{B}_n,$$

instead of iterating relation (5.2) for $n = 1, 2, \ldots, C - A$. Before presenting the moment-iteration algorithm, we describe the small adaptations of this algorithm in case that the lengths of the on-periods are random and in case that $N > 1$.

On-periods of random duration

When the length of the on-periods is random (that is, when A is random), then it is not known in advance what slots are on-slots and what slots are off-slots. However, if we condition on the length of the on-period in a cycle, then it is known what slots are on-slots and what are off-slots. Hence, conditional on this length, we can use the procedure described for on-periods of fixed duration. By using this procedure for all possible values of A, we obtain for each possible length of the on-period an approximation for the first two conditional moments of the number of slots of work at the start of the next cycle. Unconditioning then yields an approximation for the first two moments of the number of slots of work at the start of this cycle.

More than one on- and off-period in a cycle

In the case $N > 1$, that is, the case that a cycle contains more than one subcycle, we apply the procedure described above to each subcycle. More precisely, suppose that we have the approximations for the first two moments of $X_{k,1}$, for some fixed k with $k = 1, 2, 3, \ldots$. Then, we use this technique to compute the approximations for the first two (unconditional) moments of the number of slots of work at the start of the second subcycle. Given these approximations, we use this procedure to the second subcycle for approximating the first two (unconditional) moments of the number of slots work at the start of the third subcycle. Repeating this procedure for the remaining subcycles in the cycle, we obtain approximations for the first two (unconditional) moments of $X_{k+1,1}$.

5.4.2 The MI technique

In Section 5.4.1, we sketched the main ideas of the MI technique. Based on these ideas, we formalise this technique for approximating the stationary queue-length distributions at the slot boundaries in the cycle. For convenience, we present the MI technique for the case that $N = 1$ and that on-periods are of fixed duration only. Using the adaptations mentioned in Section 5.4.1, this technique can be applied to the case that the lengths of the on-periods are random and to the case that $N > 1$.

Let Y be a non-negative discrete random variable, and let Z have the same distribution as Y conditional on $Y > 0$. In the MI technique, we fit a distribution on the first two moments of $Z - 1$ instead of Y; in the sequel we write $(Y|Y > 0) - 1$ instead of $Z - 1$. In this way, we reduce the impact of the probability $\Pr\{Y = 0\}$ on the fitted distribution. Since this refinement explicitly uses that the random variables involved are discrete, it is important to have a procedure for fitting *discrete* distributions. As mentioned earlier, such a procedure is developed in Section 5.5. Further, we refine the MI technique in the sense that we do not iterate the relation (5.2) for approximating the first two moments of $X_{k,C-A+1}$,

but we use, for all k, the relation

$$X_{k,C-A+1} = X_{k,1} + \sum_{j=1}^{C-A} \overline{B}_j. \tag{5.3}$$

Finally, as in Chapter 2, we do not fit a distribution to the first two conditional moments of $X_{k,n} + \overline{B}_n$, but to the first two conditional moments of $X_{k,n}$ (that is, we fit a distribution to the first two moments of $(X_{k,n}|X_{k,n} > 0) - 1$ instead of to the first two moments of $(X_{k,n} + \overline{B}_n|X_{k,n} + \overline{B}_n > 0) - 1$).

A consequence of these refinements is that we have to know the probability $\Pr\{\overline{B}_n \leq 1\} = 1 - p_n + p_n \Pr\{B_n = 1\}$ or an approximation for this probability, for all on-slots n in the cycle, that is, for $n = C - A + 1, C - A + 2, \ldots, C$. To clarify this, let n be an on-slot, and define (in distribution)

$$Y_{k,n} \stackrel{d}{=} (X_{k,n}|X_{k,n} > 0) - 1, \qquad k = 0, 1, 2, \ldots \text{ and } n = 1, 2 \ldots, C. \tag{5.4}$$

In order to compute an approximation for $\Pr\{X_{k,n+1} > 0\}$, for some fixed k, we use relation (5.2). As mentioned, in the refined MI technique, we fit a distribution on the (approximate) first two moments of $(X_{k,n}|X_{k,n} > 0) - 1$. In order to compute the probability $\Pr\{X_{k,n+1} > 0\}$, we have to compute

$$\begin{aligned} \Pr\{X_{k+1,n} > 0\} &= \Pr\{X_{k,n} + \overline{B}_n - 1 > 0\} \\ &= \Pr\{X_{k,n} = 0\}\Pr\{\overline{B}_n > 1\} + \Pr\{X_{k,n} > 0\}\Pr\{(X_{k,n}|X_{k,n} > 0) + \overline{B}_n > 1\} \\ &= \Pr\{X_{k,n} = 0\}\Pr\{\overline{B}_n > 1\} + \Pr\{X_{k,n} > 0\}(1 - (1 - p_n)\Pr\{Y_{k,n} = 0\}), \end{aligned} \tag{5.5}$$

so that we need information about $\Pr\{\overline{B}_n \leq 1\}$.

The MI technique for the case $N = 1$ and on-periods of fixed duration is described as follows.

The MI technique for the stationary queue-length distributions

Step 1. Initialisation: Set $\mathrm{E}\{X_{1,1}\} = \mathrm{E}\{X_{1,1}^2\} := 0$, so that $\sigma_{1,1} = 0$, and $\Pr\{X_{1,1} = 0\} = 1$. Set $k := 1$ and $i := 1$.

Step 2. Iteration: Approximate the first two moments of $X_{k,C-A+1}$ and the probability $\Pr\{X_{k,C-A+1} > 0\}$ from relation (5.3), using the approximations for the first two moments of $X_{k,1}$.
For $n = C - A + 1, C - A + 2, \ldots, C$,

(*i*) compute the approximations for the first two moments of $Y_{k,n}$ (using (5.4)); fit a discrete distribution to $Y_{k,n}$ by matching these moments in order to approximate $\Pr\{Y_{k,n} = 0\}$;

(*ii*) compute the approximations for the first two moments of $X_{k,n+1}$ using (5.2), and the approximation for the probability $\Pr\{X_{k,n+1} > 0\}$ using (5.5).

Step 3. Stopping criterion: Compute the approximation for the standard deviation $\sigma_{k,C+1}$ of $X_{k,C+1}$. If the approximations for the differences $|\mathrm{E}\{X_{k,C+1}\} - \mathrm{E}\{X_{k-1,C+1}\}|$ and $|\sigma_{k,C+1} - \sigma_{k-1,C+1}|$ are both small enough, then execute Step 4. Otherwise, set $k := k + 1$ and $X_{k,1} := X_{k-1,C+1}$, and execute Step 2.

Step 4. Approximation: The first two moments of the stationary imbedded queue-length distribution are approximated by $\mathrm{E}\{X_{k,C+1}\}$ and $\mathrm{E}\{X_{k,C+1}^2\}$, and the stationary probability that the queue

is not empty by $\Pr\{X_{k,C+1} > 0\}$. Starting with these approximations, we approximate these quantities for the other slot boundaries n in the cycle; for $n = 2, 3, \ldots, C - A + 1$ using (5.3), with $C - A$ substituted by $n - 1$, and for $n = C - A + 2, C - A + 3, \ldots, C$ using the iteration step.

Using a procedure for fitting discrete distributions on the first two moments, we obtain the approximations of the stationary queue-length distributions at the slot boundaries in the cycle. In Section 5.7, we shall illustrate the performance of this moment-iteration algorithm. But, as noted before, we present in the next section a useful procedure for fitting discrete distributions on the first two moments of non-negative random variables.

Numerical examples indicate that this MI technique always terminates. However, we have not been able yet to prove this.

5.5 Fitting discrete distributions on the first two moments

In the literature, several methods are presented to approximate a discrete distribution by matching moments. Here, we mention some of these methods. Drew [1968] gives an informal approach. More specifically, he suggests to fit a binomial distribution, a Poisson distribution, and a negative binomial distribution, if the coefficient of variation is appreciably smaller than one, equal to one, and appreciably larger than one, respectively. In Ord [1972], a classification is presented of discrete distributions satisfying a difference equation that was already studied by Pearson [1895]. This class of distributions is the discrete analogue to the Pearson system of (continuous) distributions. Based on this classification, Ord fits a discrete distribution on the first three moments. Powell [1986] uses the shifted negative binomial distribution to match the first two moments of a random variable, because of its ability to incorporate a wide range of coefficients of variation. Unlike the previous three approaches, Brahimi & Worthington [1991] do not use classical distributions. They derive a system of three non-linear equations to match the first two moments. In this way, they obtain a class of discrete distributions with bounded support having these first two moments.

The methods of Drew [1968] and Powell [1986] do not capture all possible combinations of the first two moments of discrete non-negative random variables. The method of Ord [1972] requires the first *three* moments, and the approach of Brahimi & Worthington [1991] is not that useful for applying in the MI technique. Therefore, in this section, we construct a novel procedure for fitting discrete distributions on the first two moments. This procedure is a discrete analogue to the one in Section 5.2. It turns out that this construction is not trivial. First of all, unlike the continuous case, for some pairs of non-negative numbers (m, c) there does not exist a discrete random variable concentrated on the non-negative integers with mean m and coefficient of variation c. Secondly, the discrete analogues to the Erlang distribution and hyperexponential distribution (that is, the negative binomial distribution and a mixture of two geometric distributions, respectively) are not sufficient to fit all possible pairs of non-negative numbers (m, c).

Consider an arbitrary pair of non-negative and real numbers (m, c). Before we come to the issue of how to fit a discrete distribution with mean m and coefficient of variation c, we first answer the question what combinations (m, c) are possible for discrete distributions concentrated on the non-negative integers. For continuous distributions on the non-negative real numbers, all combinations of a positive first moment and positive coefficient of variation are possible. For discrete distributions, however, this

turns out to be not the case as we show in the next lemma, where $[m]$ denotes the largest integer not exceeding m.

Theorem 5.5.1 *For a pair of non-negative and real numbers (m, c), there exists a random variable on the non-negative integers with mean m and coefficient of variation c, if, and only if,*

$$c^2 \geq \frac{(k+1-m)(m-k)}{m^2}, \tag{5.6}$$

with $k = [m]$.

To prove Theorem 5.5.1, we use the result presented in the next lemma, where $\stackrel{d}{=}$ denotes equality in distribution.

Lemma 5.5.1 *Let X be a random variable on the non-negative integers with mean μ, where $0 < \mu \leq 1$, and coefficient of variation c_X. Further, let Y be a random variable on $\{0, 1\}$ with mean μ and coefficient of variation c_Y. Then,*

$$c_X^2 \geq c_Y^2,$$

with equality if, and only if, $X \stackrel{d}{=} Y$.

Proof. We first notice that

$$\Pr\{Y = 1\} = 1 - \Pr\{Y = 0\} = \mu.$$

Since X is concentrated on the non-negative integers, we have

$$\mathrm{E}\{X^2\} \geq \mathrm{E}\{X\} = \mathrm{E}\{Y\},$$

with equality if, and only if, $X \stackrel{d}{=} Y$. Hence,

$$\mathrm{var}(X) = \mathrm{E}\{X^2\} - (\mathrm{E}\{X\})^2 \geq \mathrm{E}\{X\} - (\mathrm{E}\{X\})^2 = \mu(1-\mu) = \mathrm{var}(Y),$$

with equality if, and only if, $X \stackrel{d}{=} Y$. Finally, since $\mathrm{E}\{X\} = \mathrm{E}\{Y\}$, we obtain

$$c_X^2 = \frac{\mathrm{var}(X)}{(\mathrm{E}\{X\})^2} \geq \frac{\mathrm{var}(Y)}{(\mathrm{E}\{X\})^2} = c_Y^2,$$

which completes the proof. □

Proof of Theorem 5.5.1. We first prove that the inequality (5.6) is a necessary condition. Define $k := [m]$, and let $Y_{k,1}$ be a random variable on $\{k, k+1\}$ with probability distribution

$$\Pr\{Y_{k,1} = k+1\} = 1 - \Pr\{Y_{k,1} = k\} = m - k.$$

As can easily be verified,

$$\mathrm{E}\{Y_{k,1}\} = m \quad \text{and} \quad c_{Y_{k,1}}^2 = \frac{(k+1-m)(m-k)}{m^2}.$$

The implication of Lemma 5.5.1 is that a discrete random variable Y assuming values on two consecutive non-negative integers has the smallest coefficient of variation of all discrete random variables

on the non-negative integers with the same mean as Y. This implies that, for all discrete non-negative random variables X with the same mean as $Y_{k,1}$,

$$c_X^2 \geq c_{Y_{k,1}}^2.$$

Hence, the inequality (5.6) is a necessary condition. To prove that this inequality is a sufficient condition as well, we need to show that, for any pair (m, c) satisfying (5.6), there exists a random variable X with $\mathrm{E}\{X\} = m$ and $c_X = c$. We do this as follows.

Let $k = [m]$ and, for $n = 1, 2, 3, \ldots$, let $Y_{k,n}$ be a random variable on $\{k, k+n\}$ with probability distribution

$$\Pr\{Y_{k,n} = k+n\} = 1 - \Pr\{Y_{k,n} = k\} = \frac{m-k}{n}.$$

So, the mean and coefficient of variation of $Y_{k,n}$ are

$$\mathrm{E}\{Y_{k,n}\} = m \quad \text{and} \quad c_{Y_{k,n}}^2 = \frac{(k+n-m)(m-k)}{m^2},$$

respectively. It is easily verified that, for fixed k, $c_{Y_{k,n}}^2$ is strictly increasing in n and that $c_{Y_{k,n}}^2$ tends to infinity as n tends to infinity. Thus, for a random variable with mean m and squared coefficient of variation c^2 satisfying inequality (5.6), there exists an n with

$$c_{Y_{k,n}}^2 \leq c^2 \leq c_{Y_{k,n+1}}^2.$$

Furthermore, it is readily verified that the random variable X defined by

$$X = pY_{k,n} + (1-p)Y_{k,n+1},$$

with

$$p = \frac{c_{Y_{k,n+1}}^2 - c_X^2}{c_{Y_{k,n+1}}^2 - c_{Y_{k,n}}^2},$$

has mean m and squared coefficient of variation c^2. Hence, the inequality (5.6) is also a sufficient condition, so that the proof is complete. □

In Figure 5.1, the 'impossible' regions in the (m, c)-plane are shaded.

To construct a discrete analogue to the method described in Section 5.2, it does not suffice to use the discrete analogues to a mixture of two Erlang distributions and to a hyperexponential distribution. The reason for this is that these two discrete analogues do not cover the unshaded area in Figure 5.1. It turns out that this area can be covered by four classes of distributions. Before presenting the discrete fitting procedure, we introduce some notations.

Geo(p) denotes a random variable with probability distribution $\{(1-p)p^i, i = 0, 1, 2, \ldots\}$, with $0 < p < 1$, and NB(n, p) the sum of n independent Geo(p) random variables; finally, Bin(n, p) denotes a random variable having a binomial distribution with n the number of trials and p the success probability. In the method for fitting discrete distributions, the parameter

$$\theta := c^2 - 1/m$$

plays an important part, where m and c denote the mean and coefficient of variation, respectively, of a random variable. Our method for fitting discrete distributions is presented in the next theorem. This method is in a sense a formalisation and generalisation of the informal approach used in Drew [1968].

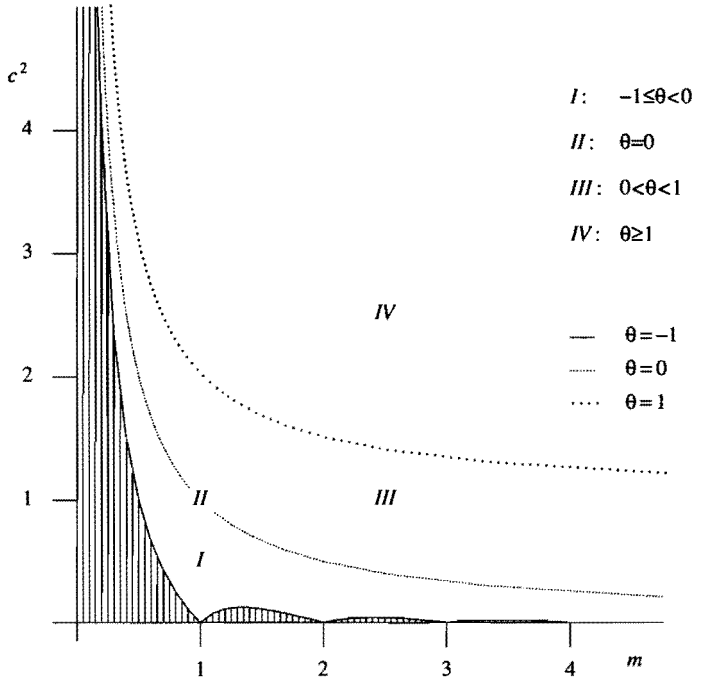

Figure 5.1: *The shaded regions denote the 'impossible' regions for a discrete random variable on the non-negative integers, and the other regions are the four regions for θ indicating which distribution is used to match the first two moments of this random variable.*

Theorem 5.5.2 *Let Z be a random variable on the non-negative integers with mean m and coefficient of variation c. Then, the discrete random variable Y matches the first two moments of Z, if Y is chosen as follows:*

1. *If* $-1/n \le \theta < -1/(n+1)$, *for certain* $n = 1, 2, 3, \ldots$, *then*

$$Y = \begin{cases} Bin(n, p), & \text{with probability } q, \\ Bin(n+1, p), & \text{with probability } 1-q, \end{cases}$$

where

$$q = \frac{1 + \theta(n+1) + \sqrt{-\theta n(n+1) - n}}{1+\theta} \quad \text{and} \quad p = \frac{m}{n+1-q}.$$

2. *If* $\theta = 0$, *then Y has a Poisson distribution with mean m.*

3. *If* $1/(n+1) \le \theta < 1/n$, *for certain* $n = 1, 2, 3, \ldots$, *then*

$$Y = \begin{cases} NB(n, p), & \text{with probability } q, \\ NB(n+1, p), & \text{with probability } 1-q, \end{cases}$$

where

$$q = \frac{(n+1)\theta - \sqrt{(n+1)(1-\theta n)}}{1+\theta} \quad \textit{and} \quad p = \frac{m}{n+1-q+m}.$$

4. *If* $\theta \geq 1$, *then*

$$Y = \begin{cases} Geo(p_1), & \textit{with probability } q_1, \\ Geo(p_2), & \textit{with probability } q_2, \end{cases}$$

where

$$p_1 = \frac{m(1+\theta+\sqrt{\theta^2-1})}{2+m(1+\theta+\sqrt{\theta^2-1})}, \quad q_1 = \frac{1}{1+\theta+\sqrt{\theta^2-1}},$$
$$p_2 = \frac{m(1+\theta-\sqrt{\theta^2-1})}{2+m(1+\theta-\sqrt{\theta^2-1})}, \quad q_2 = \frac{1}{1+\theta-\sqrt{\theta^2-1}}.$$

Proof. From condition (5.6) and the fact that k is the unique integer satisfying $k \leq m < k+1$, we have

$$c^2 - \frac{1}{m} \geq \frac{(k+1-m)(m-k)}{m^2} - \frac{1}{m} \geq -1.$$

Thus, we have to consider the case that $\theta \geq -1$ only. Further, it is straightforward to check that the given distributions indeed have the same mean and coefficient of variation as Z. □

It can be verified that $p \leq 1$ for case 1 of Theorem 5.5.2 is a consequence of the inequality (5.6). Further, if $\theta > 0$ (that is, cases 3 and 4 of Theorem 5.5.2), then it is also possible to fit an NB(n, q) distribution with real-valued n. An advantage of the solution proposed in Theorem 5.5.2 is that the fitted distributions allow a simple interpretation in terms of sums or mixtures of geometric distributions. Finally, we note that by the Continuity Theorem (cf. Feller [1968]) the Bin(n, p) and the NB(n, q) converge to the Poisson distribution with mean λ for $p = \lambda/n$ and $q = n/(n+\lambda)$, respectively, and letting n tend to infinity.

In the Section 5.7, we investigate the performance of the MI technique, using the fitting procedure of Theorem 5.5.2. But first, we compute the sojourn-time distribution of customers.

5.6 The sojourn-time distribution

Given the number of slots of work just after the arrival of a customer, we developed in Section 4.4 an algorithm for computing the sojourn-time distribution of this customer exactly. The number of slots of work just after an arrival in statistical equilibrium is equal to the number of slots of work upon arrival and the number of slots of work of this customer. This latter number of slots has the same probability distribution as B_n, if the customer arrives in slot n of the cycle, with $n = 1, 2, \ldots, C$. By the Bernoulli-arrivals-see-time-average property (cf. Halfin [1983]), the number of slots of work upon arrival in slot n equals the number of slots of work at the boundary between slot $n-1$ and slot n. For this number of slots, we use the MI technique for approximating its first two moments (possibly after conditioning on the length of the subcycle of arrival). Thus, we have an approximation for the first two moments of the total number of slots of work just after the arrival. Using the fitting procedure in Section 5.5, we approximate the probability distribution of the number of slots of work just after the arrival. Given this approximation, we use the algorithm in Section 4.4 for computing the sojourn-time distribution.

5.7 Numerical examples

In this section, we investigate the performance of the MI technique for approximating the stationary queue-length distributions at the slot boundaries in the cycle. Further, we examine the results for the sojourn-time distribution when these approximations are used in the algorithm of Section 4.4. More specifically, for some numerical examples, we compare the (practically) exact values and the approximations of several important performance measures as obtained by the techniques of Chapter 4 and this chapter. Firstly, we consider the probability $\Pr\{Q_1 > 0\}$ that the queue is not empty at the first slot boundary of a cycle and the average number of slots of work $\mathrm{E}\{Q_1\}$ at this time instant. Secondly, the average number of slots of work $\mathrm{E}\{L\}$ in the system in an arbitrary slot and the average sojourn time $\mathrm{E}\{S\}$ of an arbitrary arriving customer is considered. Finally, we investigate the tail probabilities of the sojourn-time distribution of an arbitrary arriving customer. For the MI technique, we use the fit procedure as exposed in Section 5.6.

To investigate the performance of the MI technique, we use the same examples as in Section 4.6. So, we have a cycle consisting of one subcycle and we assume that the length of the off-periods and of the on-periods are both constant. Customers are supposed to arrive according to a homogeneous Bernoulli process with parameter p. The average service times of these customers are assumed to be equal to three on-slots. For these times, we consider three cases, namely, the cases that they are deterministic, geometric, and negative binomial of order two. Further, we examine the cases that the length C of a cycle is $C = 60, 120, 180$ and that the length A of the on-periods is $A = 0.25C, 0.50C, 0.75C$.

In Table 5.1, we display the 'exact' (Ex) and approximative (MI) values of $\Pr\{Q_1 > 0\}$, $\mathrm{E}\{Q_1\}$, $\mathrm{E}\{L\}$, and $\mathrm{E}\{S\}$ for deterministic service times. The 'exact' values are those computed when the states $j > T$ are truncated, with T sufficiently large. For geometric and negative binomial service times, these results are listed in Table 5.2 and Table 5.3, respectively. In these tables, ρ denotes the effective utilisation of the system, that is, $\rho = 3p(C/A)$.

From these tables, we first of all observe that the MI technique generally gives good approximations for $\mathrm{E}\{Q_1\}$. Usually, the relative error is within 7%. When the utilisation is low, the relative error may be large, but the absolute errors are small; in these cases, the average number of slots of work corresponds to less than one customer in the system on average. For the approximations of $\Pr\{Q_1 > 0\}$, we can make similar remarks.

Further, these tables show that the results for $\mathrm{E}\{L\}$ and $\mathrm{E}\{S\}$ are all within 6% of the exact values; for most of the cases, this error is even much smaller. So, the MI technique approximates the average number of customers in the system quite accurate.

For the tail probabilities of the sojourn-time distribution, we list in Table 5.4 the quantiles of order α, for $\alpha = 0.70, 0.80, 0.90, 0.95$, in case of deterministic service times. As we see, using the approximations obtained by the MI technique, the shape of the tail of the resulting sojourn-time distribution is fairly well described. The corresponding results for the case of geometrically and negative binomially distributed service times are similar, and these results are, therefore, omitted.

The examples considered indicate that the MI technique gives good approximations, compared to the information used (that is, the first two moments of the service-time distributions and the probability that a service time equals one slot).

			$\Pr\{Q_1 > 0\}$		$\mathrm{E}\{Q_1\}$		$\mathrm{E}\{L\}$		$\mathrm{E}\{S\}$	
ρ	C	A	Ex	MI	Ex	MI	Ex	MI	Ex	MI
0.75	60	15	0.371	0.400	2.20	2.31	6.43	6.53	36.03	37.65
		30	0.397	0.407	1.73	1.91	7.01	7.16	20.19	20.62
		45	0.504	0.515	1.88	2.00	5.49	5.56	11.01	11.17
	120	30	0.288	0.315	1.57	1.84	9.74	9.99	53.70	55.42
		60	0.341	0.340	1.19	1.51	11.08	11.33	31.05	31.66
		90	0.477	0.493	1.53	1.60	7.70	7.74	14.94	15.04
	180	45	0.241	0.264	1.20	1.50	13.29	13.57	72.65	74.24
		90	0.316	0.311	0.94	1.36	15.38	15.71	42.50	43.23
		135	0.468	0.475	1.39	1.40	10.04	10.04	19.09	19.11
0.90	60	15	0.683	0.720	9.81	9.90	15.02	15.01	68.50	70.64
		30	0.681	0.706	8.26	8.27	15.12	15.13	35.09	35.63
		45	0.742	0.749	7.60	7.62	12.75	12.77	20.15	20.20
	120	30	0.612	0.642	8.70	8.91	18.84	19.04	85.49	88.96
		60	0.615	0.641	6.97	7.04	20.08	20.14	46.12	46.82
		90	0.701	0.701	6.41	6.61	15.73	15.88	24.55	24.81
	180	45	0.563	0.596	7.91	8.25	22.97	23.31	103.85	108.29
		90	0.572	0.592	6.08	6.12	25.41	25.44	57.97	58.59
		135	0.676	0.668	5.61	6.00	19.01	19.30	29.41	29.81
0.95	60	15	0.828	0.864	23.33	23.14	28.87	28.68	123.20	124.78
		30	0.824	0.854	20.39	20.01	27.81	27.45	60.05	59.99
		45	0.858	0.875	18.43	18.06	24.18	23.86	35.18	34.87
	120	30	0.784	0.818	22.03	20.01	32.85	32.61	139.92	141.87
		60	0.780	0.816	18.76	18.31	33.04	32.61	71.06	71.38
		90	0.829	0.847	16.80	16.33	27.41	26.00	39.72	39.33
	180	45	0.752	0.784	21.06	20.97	37.14	37.06	157.99	161.51
		90	0.748	0.789	17.56	16.94	38.69	38.10	82.94	83.35
		135	0.808	0.824	15.60	15.04	31.00	30.52	44.76	44.31

Table 5.1: *Numerical results for deterministic service times.*

			$\Pr\{Q_1 > 0\}$		$E\{Q_1\}$		$E\{L\}$		$E\{S\}$	
ρ	C	A	Ex	MI	Ex	MI	Ex	MI	Ex	MI
0.75	60	15	0.478	0.485	4.87	4.84	9.12	9.09	48.65	48.72
		30	0.468	0.475	4.19	4.16	9.58	9.56	25.54	25.62
		45	0.539	0.539	4.42	4.36	8.23	8.20	14.63	14.58
	120	30	0.387	0.393	3.85	3.97	12.08	12.19	64.40	65.43
		60	0.399	0.398	3.19	3.27	13.27	13.34	35.40	35.53
		90	0.503	0.497	3.66	3.83	10.21	10.30	18.15	18.27
	180	45	0.331	0.340	3.18	3.40	15.36	15.56	81.91	83.25
		90	0.362	0.355	2.60	2.80	17.30	17.47	46.14	46.38
		135	0.486	0.481	3.25	3.65	12.38	12.61	22.02	22.37
0.90	60	15	0.755	0.774	18.49	18.28	23.70	23.51	105.32	104.66
		30	0.739	0.763	16.71	16.32	23.62	23.26	52.48	52.00
		45	0.777	0.796	16.15	15.72	21.41	21.08	31.71	31.37
	120	30	0.694	0.711	16.95	16.77	27.12	26.94	120.49	120.62
		60	0.682	0.710	14.86	14.41	28.07	27.66	62.38	62.39
		90	0.741	0.756	14.43	13.81	23.98	23.50	35.52	35.02
	180	45	0.650	0.666	15.82	15.80	30.92	30.90	137.38	139.18
		90	0.642	0.670	13.52	12.99	32.99	32.51	73.31	73.51
		135	0.717	0.724	13.17	12.55	26.90	26.42	39.85	39.26
0.95	60	15	0.871	0.889	42.02	41.67	47.56	47.22	193.70	197.08
		30	0.860	0.887	38.85	38.02	46.29	45.51	97.44	96.03
		45	0.880	0.902	37.00	36.29	42.80	42.18	60.07	59.34
	120	30	0.835	0.859	40.29	39.58	51.12	50.42	208.10	210.62
		60	0.824	0.857	36.65	35.39	50.99	49.79	107.32	105.66
		90	0.857	0.884	34.85	33.29	45.57	44.22	63.96	62.44
	180	45	0.808	0.833	38.97	38.22	55.08	54.34	224.05	227.12
		90	0.798	0.837	34.98	33.43	56.18	54.71	118.25	116.87
		135	0.840	0.868	33.16	31.14	48.75	47.00	68.42	66.55

Table 5.2: *Numerical results for geometric service times.*

			$\Pr\{Q_1 > 0\}$		$E\{Q_1\}$		$E\{L\}$		$E\{S\}$	
ρ	C	A	Ex	MI	Ex	MI	Ex	MI	Ex	MI
0.75	60	15	0.429	0.431	2.87	2.90	7.33	7.14	40.11	40.46
		30	0.427	0.429	2.33	2.38	7.99	7.68	21.62	21.73
		45	0.539	0.539	2.49	2.56	6.59	6.20	11.91	11.98
	120	30	0.517	0.512	2.14	2.30	10.50	10.47	56.88	57.86
		60	0.361	0.355	1.65	1.82	11.92	11.73	32.12	32.39
		90	0.485	0.482	2.02	2.32	8.68	8.48	15.71	15.99
	180	45	0.277	0.288	1.66	1.87	13.96	13.97	75.24	76.32
		90	0.330	0.321	1.31	1.56	16.14	16.03	43.35	43.74
		135	0.473	0.476	1.82	2.26	10.96	10.85	19.78	20.22
0.90	60	15	0.725	0.741	12.00	11.88	17.63	17.09	79.14	79.03
		30	0.709	0.732	10.36	10.14	17.82	17.03	39.67	39.56
		45	0.756	0.772	9.72	9.44	15.64	14.68	23.05	22.81
	120	30	0.653	0.669	10.76	10.77	21.21	20.91	95.39	96.65
		60	0.643	0.670	8.90	8.60	22.53	21.77	50.31	50.46
		90	0.716	0.721	8.36	8.08	18.38	17.53	27.26	27.00
	180	45	0.604	0.625	9.86	9.99	25.21	25.06	113.19	115.90
		90	0.600	0.622	7.88	7.47	27.71	26.88	61.86	61.92
		135	0.690	0.687	7.42	7.25	21.52	20.79	31.96	31.74
0.95	60	15	0.854	0.876	28.03	27.66	34.28	33.20	143.64	142.84
		30	0.842	0.870	24.99	24.33	33.30	31.79	69.69	68.67
		45	0.868	0.890	23.05	22.36	29.86	28.21	41.43	40.71
	120	30	0.811	0.835	26.60	26.11	37.87	36.94	159.64	159.09
		60	0.799	0.837	23.19	22.21	38.14	36.56	80.34	79.44
		90	0.840	0.866	21.26	20.10	32.73	30.90	45.76	44.66
	180	45	0.779	0.804	25.52	25.10	41.99	41.20	177.14	177.93
		90	0.769	0.812	21.86	20.57	43.58	41.78	91.92	91.20
		135	0.820	0.844	19.91	18.44	36.13	34.09	50.63	49.22

Table 5.3: *Numerical results for negative binomial service times.*

			Quantiles of order							
			0.70		0.80		0.90		0.95	
ρ	C	A	Ex	MI	Ex	MI	Ex	MI	Ex	MI
0.75	60	15	45	47	50	54	66	70	79	83
		30	28	28	32	32	35	37	42	43
		45	15	15	18	18	21	21	24	24
	120	30	73	75	83	85	93	96	105	113
		60	45	46	53	54	60	61	63	66
		90	22	22	27	28	33	33	35	36
	180	45	102	104	117	118	131	133	138	142
		90	63	64	75	76	86	87	92	93
		135	29	29	37	37	45	45	48	49
0.90	60	15	81	87	102	107	139	139	176	170
		30	41	43	51	54	68	70	85	84
		45	24	25	30	31	40	40	50	49
	120	30	101	108	123	132	160	167	197	199
		60	59	60	66	69	84	86	100	106
		90	32	33	37	39	46	48	57	58
	180	45	129	134	147	156	186	199	222	234
		90	78	79	88	90	101	104	119	125
		135	41	41	47	48	54	58	64	68
0.95	60	15	146	154	190	192	265	252	340	310
		30	71	74	91	91	125	121	159	148
		45	42	43	54	54	74	71	95	88
	120	30	165	174	209	213	284	277	359	337
		60	84	87	104	108	138	140	172	168
		90	47	48	59	60	80	79	100	99
	180	45	187	199	230	239	305	304	381	370
		90	99	102	120	121	154	159	188	192
		135	54	56	66	67	87	87	107	109

Table 5.4: *Quantiles of the sojourn-time distribution for deterministic service times.*

5.8 Conclusions

In this chapter, we have developed an approximative moment-iteration technique (MI technique) to study the queue-length process in discrete-time queueing systems with periodic service. Unlike the GT technique in Chapter 4, this technique uses only the first two moments of the service-time distributions.

The MI technique is based on the circumstance that the queue-lengths at consecutive slot boundaries can be described by a relation that has a structure similar to Lindley's equation for the $GI/G/1$ queueing system. Using an adapted version of the moment-iteration method of De Kok [1989] and a novel procedure for fitting discrete distributions by matching moments, we approximate the stationary queue-length distributions. Numerical examples show that the MI technique gives very good approximations for the performance measures of interest. Only when the average queue lengths are very small, the relative errors may be larger, but in these cases the absolute errors are small.

6

Make to Stock and Overtime in Queueing Systems with Periodic Service

6.1 Introduction

In Chapter 1, we mentioned that the techniques of the Chapters 4 and 5 are applicable to modifications of discrete-time queueing systems with periodic service. These modifications are motivated by the study of periodic production rules. Of all possible modifications, we discuss three important ones. In Chapter 7, we shall consider the case that there are customers arriving randomly, and that other customers arrive according to a periodic pattern, having priority for service over the randomly arriving customers. The aim of this chapter is to show that the possibility of making products to stock and of working overtime can both be analysed by the techniques of the Chapters 4 and 5. This enables one to evaluate the effect of these possibilities on the performance of queueing systems with periodic service.

Because of the motivation, we say that customers place orders at the server. In practice, once the size of an order is given, the service time of this order is usually fairly well known. So, it seems reasonable to consider only deterministic service phases. Such a service phase is called a product.

Make to stock

The server can only make products to stock in on-periods. Within these periods, products are made to stock if there are no orders for these products and if the number of these products in stock is less than the maximal number of them allowed in stock. When a product is taken from stock, we can interpret this as if the server orders a product at himself. Because of this interpretation, the resulting queueing system features the main characteristics of an inventory system governed by a base-stock policy. In particular, this queueing system features the main characteristics of an $(S-1, S)$ inventory system (see, for instance, Hadley & Whitin [1963]), with S the maximal number of products allowed in stock. Consequently, this queueing system is closely related to inventory systems with spare parts and

repairable items (see, for example, Sherbrooke [1968] and Nahmias [1981]). Further, this queueing system is related to a system studied in Federgruen & Katalan [1995]. Federgruen & Katalan study a cyclic polling system in which the server uses a base-stock policy for each queue. The main difference between our model and the other models is the following. In our model the server may not always increase the stock to the base-stock level before he switches to the next queue (due to the periodic service policy), whereas in the other models the server always increases the stock to the base-stock level before switching.

Overtime

The server can work overtime at the end of on-periods only. The decision whether to work overtime or not is taken according to a threshold policy. More precisely, if at the end of an on-period the number of ordered products in the corresponding queue exceeds a certain threshold, then the server works a specific amount of overtime. We assume that working overtime does not affect the future switch-over instants of the server. For example, working overtime represents work executed outside the normal working hours, work executed by another server, hiring some additional service capacity, or buying products from another firm. The use of overtime is applied, for instance, in Goodwin, Jr., Elvers & Goodwin [1978] for a general job shop environment, in Scudder [1985] and Scudder & Chua [1987] for a repair shop environment, and in Dellaert [1988] for a firm making a variety of products to order. Furthermore, overtime can be regarded as a temporary expansion of the service capacity, so that the service rate is speeded up temporary.

This chapter, which is largely based on Van Eenige, Adan, Resing & Van der Wal [1995a], is organised as follows. In Section 6.2, we discuss queueing systems with periodic service and make to stock. Queueing systems with periodic service and overtime are studied in Section 6.3. Finally, we give a summary of this chapter in Section 6.4.

6.2 Queueing systems with periodic service and make to stock

In this section, we first specify the possibility of making products to stock. After that, in Section 6.2.2, we analyse the queue-length process and compute the sojourn-time distribution of customers by the techniques in Chapter 4. In Section 6.2.3, we first apply a moment-iteration technique similar to the one in Chapter 5 to analyse the queue-length process. Then, given the approximation for the stationary queue-length distributions at arrival instants, we use the technique of Chapter 4 to calculate the sojourn-time distribution. Finally, in Section 6.2.4, some numerical examples are given.

6.2.1 The model

We consider the queueing system with periodic service as introduced in Chapter 4. For convenience, we treat the case $N = 1$ only, that is, the case that a cycle consists of one on- and off-period (the analysis can easily be extended to the case $N > 1$). As noted earlier, the service phases (which are called products in this chapter) are deterministic.

As already mentioned, products can only be made to stock in on-periods. The decision whether to make a product to stock or not is taken immediately after possible arrival instants in on-slots. Recall

that we assumed that arrivals occur just after slot boundaries and that products are completed just before slot boundaries (see Figure 4.3). This decision depends on whether the queue is empty or not, and on the number of products already in stock. More precisely, the server makes a product to stock in an on-slot if, at the decision epoch, the queue is empty and the number of products in stock is smaller than V, with V a non-negative integer. So, V denotes the maximal number of products allowed in stock. We remark that V may depend on the slot in the cycle, but, for clarity of presentation, we assume that V is the same for each slot.

Customers who find their requested products in stock are served immediately upon arrival, irrespective of whether they arrive in an on- or off-period. A customer who can be served only partially from stock has a reduced service time. Further, it is clear that the case $V = 0$ corresponds to the standard queueing system with periodic service as studied in the Chapters 4 and 5.

Finally, for completeness, we repeat some notation. $A_{\min}$ and $A_{\max}$ denote the minimal and maximal number of on-slots in a cycle, respectively (see also definitions (4.1) and (4.2) with $N = 1$). The minimal and maximal number of products ordered in a cycle are denoted by $B_{\min}$ and $B_{\max}$, respectively (recall definition (4.3)). Further, we again define the integer K, for which we implicitly use that $N = 1$. Consider the last possible on-slot n in a cycle with $p_n > 0$. Then, K denotes the number of possible on-slots in the cycle, before slot n, in which no customer can arrive.

6.2.2 The GT technique

In this section, we use the techniques of Chapter 4 to study queueing systems with periodic service and make to stock. In Section 6.2.2.1, we show that the queue-length process at the start of cycles is stochastically identical to the imbedded queue-length process studied in Chapter 4 after a transformation. This indicates that we can use the results of Section 4.3 directly. In Section 6.2.2.2, we shall see that the sojourn time of a customer (that is, the delivery time of an ordered product) can be computed exactly. This computation is almost the same as in Section 4.4. The difference is that, with the possibility of making products to stock, the sojourn time of a customer may be equal to zero, namely, when he is completely delivered from stock.

6.2.2.1 The queue-length process

To analyse the queue-length process of customers, we look at the system at the first slot boundary of cycles. If the queue is empty at the start of the k-th cycle, then X_k denotes the difference between V and the number of products in stock, for $k = 1, 2, 3, \ldots$. If the queue is not empty at this instant, then X_k denotes the number of products in the queue plus V. So, for example, state 0 denotes an empty queue and the maximal number of products in stock, state V denotes both an empty queue and no stock, and state $V + 3$ indicates no stock and 3 ordered products in the queue.

The stochastic process $\{X_k, k = 1, 2, 3, \ldots\}$ is a discrete-time Markov chain, and this process is also called the imbedded queue-length process. We assume that $B_{\max} > A_{\min}$, so that the state space of this chain is denumerable, and suppose that this state space is the set $\{0, 1, 2, \ldots\}$ (by some small adaptations the analysis can be used when the state space consists of multiples of some integer only). Further, it is assumed that the Markov chain is irreducible and aperiodic. Finally, as in Chapter 4, we suppose that the average number of products ordered in a cycle is strictly less than the average service capacity of the server per cycle. Under these assumptions, the Markov chain has a unique stationary distribution (cf. Pakes [1969]), which is the unique solution of the equilibrium equations

and the normalisation equation.

These equilibrium equations are identical to those of the Markov chain defined in Section 4.3. This can be verified by considering the difference between V and the number of products in stock as additionally ordered products. So, in this case, X_k denotes the number of ordered products including the additionally ordered ones. For instance, in this case, state 0 corresponds to an empty queue, and state V indicates that there are V ordered products in the queue (that is, the V additionally ordered products). As a result, the queue-length processes in this section and in Section 4.3 are stochastically identical. Hence, the Markov chains of these sections are the same, so that the equilibrium equations of both chains are the same. Consequently, the results of Section 4.3 can readily be applied. For completeness, we restate these results without proofs.

Recall that D denotes the minimal integer for which the transition probabilities $p_{i,j}$ of the Markov chain are equal to q_{j-i}, for $i \geq D$ and $j \geq 1$. Further, T_L and T_H indicate the largest possible jump out of state i to a lower and higher state, respectively. The values of these quantities are specified in the next lemma.

Lemma 6.2.1 *For the Markov chain describing the imbedded queue-length process of the queueing system with periodic service and make to stock, we have*

(*i*) $D \leq A_{\max}$,

(*ii*) $T_L = A_{\max} - B_{\min}$,

(*iii*) $T_H = B_{\max} - A_{\min}$.

Let $Q(z)$ denote the shifted probability generating function of the probability distribution $\{q_h, h = -T_L, -T_L + 1, \ldots, T_H\}$ (see also definition (4.7)). Further, $\alpha(z)$ and $\beta_n(z)$ denote the probability generating function of the length of the on-periods and of the number of products ordered in slot n, respectively (see also definitions (4.8), with $N = 1$ and omitting the subscript i, and (4.9)).

Lemma 6.2.2 *The shifted probability generating function $Q(z)$ of the probability distribution $\{q_h, h = -T_L, -T_L + 1, \ldots, T_H\}$ satisfies*

$$Q(z) = z^{T_L}\alpha(1/z)\prod_{n=1}^{C}\beta_n(z).$$

Finally, we recall that some transition probabilities $p_{i,j}$ with $i < D$ are equal to q_{j-i} as well.

Lemma 6.2.3 *For $K + 1 \leq i < D$ and $j \geq B_{\max} - \max\{0, A_{\min} - K\} + 1$, we have $p_{i,j} = q_{j-i}$.*

As in Chapter 4, the transition probabilities $p_{i,j}$, with $i < D$, that are not captured by Lemma 6.2.3 are determined recursively by the one-slot transition probabilities.

6.2.2.2 The sojourn-time distribution

As in the Chapters 4 and 5, the sojourn time of a customer denotes the length of the time interval (measured in numbers of slots) between his arrival instant and his departure. Notice that, with the possibility of making products to stock, the sojourn time of a customer is not affected by any subsequent arrival.

In this section, we show that the sojourn-time distribution in statistical equilibrium can be computed in almost the same way as in Chapter 4.

For $n = 1, 2, \ldots, C$, let Y_n denote the number of ordered products in the queue just after an arrival in slot n in statistical equilibrium. If $Y_n = 0$, then this customer is completely delivered from stock, so that his sojourn time is equal to zero. Otherwise, this customer stays in the system until the server has made Y_n products. Furthermore, in this case, his sojourn time is stochastically identical to the sojourn time of a customer arriving in slot n of the model without make to stock of Chapter 4, who has to stay Y_n on-slots in the system. Hence, once the number of ordered products in the queue immediately after his arrival is known, we can use the algorithm of Section 4.4 to compute his sojourn time. So, it remains to calculate the probability distribution of this number of ordered products.

The number of ordered products immediately after a customer arrival is zero, if, upon arrival, the number of products in stock is at least equal to the number of products he ordered. Otherwise, this number of ordered products is equal to the number of products in the queue upon arrival plus the number of products he ordered. To compute the queue-length distribution at arrival instants in statistical equilibrium, we use the Bernoulli-arrivals-see-time-average property (BASTA property, see Halfin [1983]). By the BASTA property, this distribution is equal to the stationary queue-length distribution at the boundary between slot $n-1$ and slot n, for the customer arriving in slot n, with $n = 1, 2, \ldots, C$. Hence, by using the on-slot transition probabilities, this distribution can be obtained from the stationary imbedded queue-length distribution studied in Section 6.2.2.1. The convolution of this distribution and the distribution of the number of products ordered by the customer is the distribution of Y_n.

6.2.3 The MI technique

From Section 6.2.2.1, we know that the Markov chains describing the imbedded queue-length process for the queueing system with make to stock and for the system without make to stock are the same. In fact, the periodic Markov chains describing the queue-length process at slot boundaries for these two systems are the same. So, except for the interpretation of the state of the system, we can directly use the algorithm of Section 5.4 to approximate the stationary queue-length distributions at slot boundaries.

Further, from Section 6.2.2.2, we know that, given the number of ordered products immediately after a customer arrival, the sojourn time of this customer can be computed using the algorithm in Section 4.4. So, for computing the sojourn-time distribution of a customer arriving in slot n of the cycle, with $n = 1, 2, \ldots, C$, it remains to calculate the probability distribution of the number of ordered products immediately after his arrival. By the same arguments as in Section 6.2.2.2, this distribution is the convolution of the stationary queue-length distribution at the n-th slot boundary and the service-time distribution of this customer. The first two moments of the stationary queue-length distribution at the n-th slot boundary can be approximated by the MI technique. From these approximations, the approximate first two moments of the distribution of the number of ordered products just after the arrival are easily computed. Finally, by fitting a discrete distribution to these moments using Theorem 5.5.2, we obtain an approximation for the complete distribution.

6.2.4 Numerical examples

The GT technique is used to approximate the same system of equations as in Chapter 4. In that chapter, we already saw that this technique gives excellent approximations for the solution, that is, for the (complete) stationary imbedded queue-length distribution. Therefore, we do not present numerical

examples for the investigation of the quality of these approximations.

As an illustration of the quality of the approximations obtained by the MI technique, we use the following example. Consider a cycle, with length $C = 120$, that consists of one subcycle. For the constant length A of the on-periods, we consider the cases $A = 30, 60, 90$. Customers place orders according to a Bernoulli process with parameter p, and each customer orders three products. For the effective utilisation ρ of the system, that is, $\rho := 3pC/A$, we consider the cases $\rho = 0.75, 0.90, 0.95$.

For different values of V, in Table 6.1, we present the 'exact' values (Ex) and the moment-iteration approximation (MI) of the average sojourn time $E\{S\}$. In Table 6.2, we display the 'exact' values and the MI approximations for the quantiles of the sojourn-time distribution. The 'exact' values are computed by the GT technique, when imposing the geometric tail behaviour on the stationary probabilities of states sufficiently far from the boundary of the state space.

		V							
		0		3		6		9	
ρ	A	Ex	MI	Ex	MI	Ex	MI	Ex	MI
0.75	30	53.70	55.42	40.12	43.34	28.89	32.63	19.95	23.64
	60	31.05	31.66	24.36	25.79	18.80	20.50	14.11	15.87
	90	14.94	15.04	10.71	11.41	7.63	8.30	5.31	5.94
0.90	30	85.49	88.96	73.03	76.32	61.57	64.82	51.23	54.52
	60	46.12	46.82	39.90	40.29	34.20	34.58	28.96	29.54
	90	24.55	24.81	20.48	20.88	16.93	17.47	13.82	14.52
0.95	30	139.92	141.87	127.73	129.68	116.06	118.00	105.01	106.94
	60	71.06	71.38	64.96	65.02	59.14	59.06	53.58	53.48
	90	39.72	39.33	35.68	35.29	31.92	31.58	28.42	28.17

Table 6.1: *Approximations (MI) and the exact values (Ex) for the average sojourn time* $E\{S\}$.

From Table 6.1, we see that the quality of the approximations varies considerably for $\rho = 0.75$ and $V = 3, 6, 9$; the relative errors are between 6% and 18.5%. The main reason for these errors is that the MI technique does not approximate the probability that a customer is delivered from stock accurately, that is, the probability that the sojourn time is zero. More precisely, for the case $\rho = 0.75$, this probability is relatively large, so that an error in the approximation of this probability has a relatively large effect on the approximation for the average sojourn time. If the utilisation of the system is high, then the probability that a customer is delivered from stock is rather small, so that a fairly poor approximation of this probability does not affect the average sojourn time too much. As we see, the relative errors are all within 6.5% of the exact values for the cases $\rho = 0.90$ and $\rho = 0.95$; for most of these examples, these errors are even much lower.

Table 6.2 shows that the MI technique describes the shape of the tail of the sojourn-time distribution well.

6.3 Queueing systems with periodic service and overtime

In this section, we focus on the second modification of queueing systems with periodic service, namely, the possibility of working overtime. This queueing system is described in Section 6.3.1. In Section

			Quantiles of order							
			0.70		0.80		0.90		0.95	
ρ	A	V	Ex	MI	Ex	MI	Ex	MI	Ex	MI
0.75	30	0	73	75	83	85	93	96	105	113
		3	58	62	70	73	83	88	93	99
		6	44	48	56	61	71	77	82	90
		9	29	34	42	48	59	66	71	80
	60	0	45	46	53	54	60	61	63	66
		3	38	39	45	47	54	56	59	62
		6	30	32	38	40	47	50	53	57
		9	22	24	30	33	40	43	47	51
	90	0	22	22	27	28	33	33	35	36
		3	17	17	22	23	28	29	31	32
		6	12	12	17	18	24	24	28	28
		9	6	6	12	12	19	19	23	24
0.90	30	0	101	108	123	132	160	167	197	199
		3	89	96	111	120	148	156	185	188
		6	78	85	99	107	136	145	173	176
		9	66	73	87	94	124	132	161	165
	60	0	59	60	66	69	84	86	100	106
		3	53	54	61	63	78	80	94	100
		6	46	47	55	58	71	74	88	93
		9	40	41	49	52	65	68	82	86
	90	0	32	33	37	39	46	48	57	58
		3	28	29	33	35	42	44	53	54
		6	24	24	29	31	38	41	49	50
		9	20	20	25	27	34	37	45	46
0.95	30	0	165	174	209	213	284	277	359	337
		3	153	162	197	201	272	266	347	325
		6	141	150	185	189	260	254	335	313
		9	129	139	173	177	248	243	323	301
	60	0	84	87	104	108	138	140	172	168
		3	78	80	98	101	132	134	166	162
		6	72	75	92	95	126	128	160	157
		9	66	69	86	88	120	122	154	151
	90	0	47	48	59	60	80	79	100	99
		3	43	45	55	56	76	75	96	94
		6	39	41	51	52	72	71	92	90
		9	35	37	47	48	68	67	88	86

Table 6.2: *Quantiles of the sojourn-time distribution.*

6.3.2, we study this queueing system by the techniques of Chapter 4. In Section 6.3.3, we first approximate the stationary queue-length distributions at slot boundaries by the MI technique. After that, these approximations are used to compute the sojourn-time distribution. Finally, in Section 6.3.4, we present some numerical examples.

6.3.1 The model

We consider the same model as introduced in Chapter 4. As in Section 6.2, we consider the case that $N = 1$ only (by some small adaptations the analysis can be used for analysing the case $N > 1$). Again, the service phases (which are called products) are deterministic.

The decision whether to work overtime or not is taken at the end of each on-period. That is, this decision is taken just before the first slot boundary of each cycle, but just after a possible completion of a product. This decision depends on the number of ordered products in the queue at these time instants. More precisely, if, at the decision epoch, this number of products is at least equal to the lower bound L, then the server works OT slots in overtime, where the constants L and OT are non-negative integers. As mentioned in Section 6.1, we assume that working overtime is organised such that it neither lengthens the on-periods, nor does it affect the future switch-over instants of the server. One might think of work executed outside the normal working hours or work that is put out to contract. Clearly, if $L = OT = 0$, then this queueing system corresponds to the queueing system studied in the Chapters 4 and 5. Finally, we again use the quantities $A_{\min}$, $A_{\max}$, $B_{\min}$, $B_{\max}$, and K as defined in Section 6.2.1.

6.3.2 The GT technique

The organisation of this section is equivalent to that of Section 6.2.2. More specifically, we use the GT technique to study the queue-length process in Section 6.3.2.1. In Section 6.3.2.2, we compute the sojourn-time distribution of an arbitrary customer exactly, given the stationary queue-length distributions at the slot boundaries in the cycle. However, this computation differs from the one in Section 4.4 in the sense that, for the possibility of working overtime, we have to keep track of the queue length as long as the customer is in the system.

6.3.2.1 The queue-length process

As usual, we consider the system at the first slot boundary of cycles for studying the queue-length process. Note that we look at the system immediately after a possible period of working overtime. Then, the stochastic process of the number of ordered products in the queue at these instants constitutes a discrete-time Markov chain. This process will also be called the imbedded queue-length process.

We assume that $B_{\max} > A_{\min} + OT$, so that the state space of the Markov chain is denumerable, and suppose that this state space is the set $\{0, 1, 2, \ldots\}$ (as before, after some adaptations, this analysis can be used for analysing the case that this state space consists of multiples of some integer). Furthermore, we suppose that this chain is irreducible and aperiodic. In addition, if the average number of products ordered in a cycle is strictly less than the service capacity including overtime per cycle, that is,

$$\sum_{n=1}^{C} \sum_{j=1}^{B_n} j p_n b_n(j) < \sum_{j=0}^{C} (j + OT) a(j),$$

with $\{a(j), j = 0, 1, \ldots, C\}$ the probability distribution of the length of the on-period. Then, the Markov chain has a unique stationary distribution (cf. Pakes [1969]). This stationary distribution is the unique solution of the equilibrium equations and the normalisation equation.

To show that the equilibrium equations of the Markov chain can be solved by the GT technique, we utilise the following observation that connects the queueing system with overtime to one without overtime. Suppose that the initial number of ordered products at the start of a cycle prevents the server from being idle in this cycle. In addition, assume that this number of ordered products also ensures that the server works the maximal number of slots overtime in this cycle. Then, it is not difficult to verify that the corresponding transition probabilities are equal to those of the Markov chain describing the imbedded queue-length process of the following system without overtime. Consider the model as introduced in Chapter 4 with $N = 1$ and adapt a cycle as follows. The on-period in each cycle is lengthened with OT slots, so that its length becomes $A + OT$, and in the last OT slots of this new on-period, no customer can arrive.

If the number of ordered products at the start of a cycle is at least $A_{\max}$, then the server will not be idle in the on-period. If we add $L + OT$ to this number of products, then the server works the maximal number of slots overtime as well. From these observations, it is easily verified that the transition probabilities $p_{i,j}$ of the Markov chain are equal to q_{j-i}, for $i \geq A_{\max} + L + OT$. Hence, we have $D \leq A_{\max} + L + OT$.

Consider the above queueing system without overtime. Suppose that the number of ordered products at the start of a cycle ensures that the server does not idle and that he works the maximal number of slots overtime. Since at least $B_{\min}$ products are ordered and because the server makes at most $A_{\max} + OT$ products in a cycle, the largest jump out of state i to a lower state is $A_{\max} + OT - B_{\min}$. So, we have $T_L = A_{\max} + OT - B_{\min}$. Finally, since at most $B_{\max}$ products are ordered and because the server makes at least $A_{\min} + OT$ products in a cycle, we have $T_H = B_{\max} - A_{\min} - OT$. These results are summarised in the next lemma.

Lemma 6.3.1 *For the Markov chain describing the imbedded queue-length process of the queueing system with periodic service and overtime, we have*

(*i*) $D \leq A_{\max} + L + OT$,

(*ii*) $T_L = A_{\max} + OT - B_{\min}$,

(*iii*) $T_H = B_{\max} - A_{\min} - OT$.

To characterise the probability distribution $\{q_h, h = -T_L, -T_L + 1, \ldots, T_H\}$, we utilise the aforementioned link between queueing systems with and without overtime. This link implies that we can use the result of Lemma 4.3.2 directly, if the server does not idle in a cycle and if he has to work the maximal number of slots overtime in this cycle. Let $Q(z)$ denote the shifted probability generating function of the probability distribution $\{q_h, h = -T_L, -T_L + 1, \ldots, T_H\}$ as defined in (4.7), and $\beta_n(z)$ the probability generating function of the number of service phases arriving in slot n of the cycle, for $n = 1, 2, \ldots, C$ (see definition (4.9)). Further, let $\alpha^*(z)$ denote the probability generating function of the length of an on-period plus the overtime period, that is,

$$\alpha^*(z) = z^{OT} \sum_{j=0}^{C} a(j) z^j = z^{OT} \alpha(z),$$

with $\alpha(z)$ as defined in (4.8) with $N = 1$ and omitting the subscript i. So, $\alpha^*(z)$ can be interpreted as the length of the on-period in the above queueing system without overtime that corresponds to the system with overtime. Then, the probability distribution $\{q_h, h = -T_L, -T_L + 1, \ldots, T_H\}$ can be determined from its shifted probability generating function $Q(z)$, which is presented in the following lemma.

Lemma 6.3.2 *The shifted probability generating function $Q(z)$ of the probability distribution $\{q_h, h = -T_L, -T_L + 1, \ldots, T_H\}$ satisfies*

$$Q(z) = z^{T_L}\alpha^*(1/z)\prod_{n=1}^{C}\beta_n(z).$$

For $i < D$, the transition probabilities $p_{i,j}$ of the Markov chain may in general not be equal to q_{j-i}, because the server may idle or may not work overtime in a cycle. As in Chapter 4 and Section 6.2.2.3, however, some of these transition probabilities are still equal to q_{j-i}.

Lemma 6.3.3 *For $K + L + 1 \leq i < D$ and $j \geq L + B_{\max} - \max\{0, A_{\min} + OT - K\} + 1$, we have $p_{i,j} = q_{j-i}$.*

Proof. This lemma can be proved along the lines of Lemma 4.3.3. The difference is that we also have to check that the number of ordered products at the end of the last slot in a cycle ensures that the server has to work the maximal number of slots overtime. □

To apply the GT technique, we need the transition probabilities $p_{i,j}$, with $i < D$, that are not contained within Lemma 6.3.3. As before, these probabilities depend in general on the slots in which customers arrive and depart. So, it is in general hard to characterise these probabilities explicitly. As before, these probabilities can be determined from the one-slot transition probabilities.

6.3.2.2 The sojourn-time distribution

In order to compute the sojourn time of a customer, we have to define how overtime is incorporated. In this section, we assume that work executed in overtime takes no time; in overtime, the service rate is supposed to be infinity. We give the main ideas for the computation of the sojourn-time distribution of the customer arriving in slot n, with $n = 1, 2, \ldots, C$. For convenience, we consider the case that the length of the on-periods is constant and denoted by A; the ideas are easily adapted to the case that this length is random.

The starting point for the computation of the sojourn-time distribution is to calculate the number of ordered products W_n immediately after the arrival of the customer in slot n in statistical equilibrium, with $n = 1, 2, \ldots, C$. By the BASTA property, the probability distribution of W_n is the convolution of the stationary queue-length distribution at the n-th slot boundary and the service-time distribution of the customer. As usual, we compute the stationary queue-length distribution at the n-th slot boundary from the stationary imbedded queue-length distribution studied in Section 6.3.2.1, by using the one-slot transition probabilities.

By conditioning on W_n and the number of products ordered after the arrival, the probabilities that the customer is served at the end of slot m, with $m = n, n + 1, \ldots, C$, are easy to compute; in these cases, the customer is served in the cycle of arrival. Next, we derive recurrence relations for computing the sojourn-time distribution, when the customer is not served in the cycle of arrival.

For $i, k = 1, 2, 3, \ldots$ and $j = 0, 1, 2, \ldots$, let $p_n(i, j, k)$ be the probability that, at the start of the k-th cycle after the arrival in slot n, i from the W_n ordered products are still in the queue and j products

are ordered after the arrival. Before computing these probabilities, we illustrate how they are used for the calculation of the sojourn-time distribution.

Suppose that, at the start of the k-th cycle after the arrival of the customer in slot n, i from the W_n ordered products are in the queue. For $1 \leq i \leq A$ and k fixed, it is clear that this customer leaves the system in the i-th on-slot of the k-th cycle after his arrival. The probability of this event is $\sum_{j=0}^{\infty} p_n(i, j, k)$. For $A < i \leq A + OT$ and k fixed, the customer only leaves the system at the end of the k-th cycle after the arrival if the server works overtime. In other words, when j products are ordered between slot n and the start of the k-th cycle after this slot, the customer leaves the system at this instant if $i - A + j + M \geq L$, where M denotes the number of products ordered in a cycle. Hence, the probability of this event is given by

$$\sum_{m=B_{\min}}^{B_{\max}} r(m) p_n(i, j, k) \delta_{i-A+j+m},$$

with $\{r(m), m = B_{\min}, B_{\min} + 1, \ldots, B_{\max}\}$ the probability distribution of M and

$$\delta_m = \begin{cases} 0, & \text{if } m < L, \\ 1, & \text{if } m \geq L. \end{cases}$$

Thus, once the probabilities $p_n(i, j, k)$ are known for $1 \leq i \leq A + OT$, it is straightforward to compute the sojourn-time distribution of the customer. So, it remains to compute these probabilities.

Let $\{r_n(m), m = 0, 1, \ldots, B_{\max}\}$ be the probability distribution of the number of products ordered in the remainder of the cycle after slot n. The probabilities $p_n(i, j, k)$ can be computed in a recursive fashion as follows. We condition on the number of products ordered after slot n and note that the server works overtime when the number of ordered products at the end of the cycle is at least L. Then, it is easily verified that, for $i = 1, 2, 3, \ldots$ and $j = 0, 1, \ldots, B_{\max}$,

$$p_n(i, j, 1) = (1 - \delta_{i+j}) \Pr\{W_n = i + A\} r_n(j) + \Pr\{W_n = i + A + OT\} r_n(j) \delta_{i+OT+j},$$

and for $i, k = 1, 2, 3, \ldots$ and $j = 0, 1, 2, \ldots$

$$p_n(i, j, k+1) = \sum_{m=B_{\min}}^{B_{\max}} r(m) \left((1 - \delta_{i+j}) p_n(i + A, j - m, k) + p_n(i + A + OT, j - m, k) \delta_{i+OT+j} \right),$$

where $p_n(i, h, k) := 0$ for $h < 0$. So, we can compute the probabilities $p_n(i, j, k)$ in a recursive fashion, and hence use them to compute the sojourn-time distribution of customers. We remark that the computational effort of this procedure is higher than of the procedure in Section 4.4, because we have to keep track of the queue length as long as the customer is in the system.

6.3.3 The MI technique

In this section, we use an adapted version of the MI technique of Chapter 5 for studying the queue-length process for the system with overtime. We consider overtime as an expansion of the on-period. Given the approximations of the stationary queue-length distributions at arrival instants, we use the same technique as in Section 6.3.2.2 to compute the sojourn-time distribution of customers. For clarity of presentation, in this section, we only discuss the case that the length of the on-periods is constant.

Before we apply the MI technique to study the queue-length process, we repeat some notation. Let $X_{k,n}$ denote the number of ordered products at the n-th slot boundary in the k-th cycle, for $k =$

$1, 2, 3, \ldots$ and $n = 1, 2, \ldots, C$. Further, $\overline{B}_n$ denotes the number of products ordered in slot n of the cycle.

For $n = 1, 2, \ldots, C-1$, the server does not work overtime at the end of these slots, so that the recurrence equation (5.2) holds. Hence, we can use the iteration step of the algorithm in Section 5.4 to iterate this equation for $n = 1, 2, \ldots, C-1$.

At the end of slot C, the server may have to work overtime. Hence, relation (5.2) does not hold for $n = C$. Therefore, we adapt the iteration step for slot C.

The decision whether to work overtime or not is taken at the end of slot C (just after a possible completion of a product). This decision depends on the number of ordered products in the queue at this time instant. Let $Y_{k,C}$ denote the number of ordered products at this instant in the k-th cycle, for $k = 1, 2, 3, \ldots$. Then, it is easily seen that

$$Y_{k,C} = \max\{0, X_{k,C} + \overline{B}_n - 1\},$$

so that we can use the iteration step in Section 5.4 for approximating the first two moments of $Y_{k,C}$. To these first two moments, we fit a discrete distribution for approximating the probability $\Pr\{Y_{k,C} \geq L\}$.

Let Z_k denote the number of ordered products after overtime in the k-th cycle, conditional on the event that the server worked overtime. Then,

$$Z_k = \max\{0, (Y_{k,C}|Y_{k,C} \geq L) - OT\}.$$

Using this equation, we can approximate the first two moments of Z_k. The approximations for the first two moments of $X_{k+1,1}$ are then given by

$$\begin{aligned} \mathrm{E}\{X_{k+1,1}\} &= \mathrm{E}\{Y_{k,C}|Y_{k,C} < L\}\Pr\{Y_{k,C} < L\} + \mathrm{E}\{Z_k\}\Pr\{Y_{k,C} \geq L\}, \\ \mathrm{E}\{X^2_{k+1,1}\} &= \mathrm{E}\{Y^2_{k,C}|Y_{k,C} < L\}\Pr\{Y_{k,C} < L\} + \mathrm{E}\{Z^2_k\}\Pr\{Y_{k,C} \geq L\}. \end{aligned}$$

Adapting the iteration step for $n = C$, in this way, we obtain an MI technique for the queueing system with overtime.

To compute the sojourn-time distribution, we determine as usual the number of ordered products immediately after the arrival first, and after that, investigate when these products are made.

The number of ordered products just after the arrival consists of the number of products upon arrival and the number of products ordered by the arriving customer. As before, the probability distribution of the number of products upon arrival in statistical equilibrium is equal to the stationary queue-length distribution at the slot boundary just before the arrival. This distribution can be approximated by the adapted MI technique described above. Consequently, we can approximate the probability distribution of the number of ordered products just after the arrival.

Given the approximation of the number of ordered products just after a customer arrival, we determine the sojourn-time distribution of this customer in a similar way as in Section 6.3.2.2. The difference is that we use here the MI technique to approximate the probabilities $p_n(i, j, k)$.

6.3.4 Numerical examples

The GT technique basically approximates the same system of equations as in Chapter 4. In that chapter, it was already shown that this technique gives excellent approximations. Therefore, we do not present numerical examples to demonstrate the accuracy of the approximations.

As an illustration of the quality of the approximations obtained by the MI technique, we consider the same examples as in Section 6.2.2.3, except that we now allow overtime instead of make to stock. In these examples, we have set $L := 0$.

In Table 6.3, we display the 'exact' values (Ex) and the approximations obtained by the MI technique (MI) for the average sojourn time E$\{S\}$. The 'exact' values are computed by the GT technique. The 'exact' values and MI approximations for the quantiles of the sojourn-time distribution are listed in 6.4.

		OT							
		0		3		6		9	
ρ	A	Ex	MI	Ex	MI	Ex	MI	Ex	MI
0.75	30	53.70	55.42	47.72	49.40	46.06	47.00	45.41	45.92
	60	31.05	31.66	28.70	29.58	28.18	28.64	28.00	28.19
	90	14.94	15.04	13.45	13.53	13.11	13.07	13.02	12.99
0.90	30	85.49	88.96	58.00	61.49	50.93	53.24	48.46	49.91
	60	46.12	46.82	37.05	38.27	33.73	34.86	32.26	33.22
	90	24.55	24.81	19.86	20.58	17.89	18.49	17.00	17.37
0.95	30	139.92	141.87	66.43	70.93	53.83	57.06	49.92	51.94
	60	71.06	71.38	45.20	46.56	37.63	39.15	34.57	36.02
	90	39.72	39.33	26.67	27.20	21.70	22.53	19.48	20.29

Table 6.3: *The average sojourn time E$\{S\}$.*

From Table 6.3, we see that the approximations of the average sojourn time are very good, since the relative errors are smaller than 7%. Furthermore, Table 6.4 demonstrates that the MI technique gives a good insight in the tail of the sojourn-time distribution. So, we may conclude that the MI technique performs very well.

6.4 Conclusions

In this chapter, we considered two modifications of queueing systems with periodic service. These modifications were the possibilities of making products to stock and of working overtime. By linking the corresponding queueing systems to a queueing system without either of these opportunities, we could use the techniques of the Chapters 4 and 5 to analyse the modifications. These techniques enable one the evaluate the effect of making products to stock and of working overtime on, for example, the sojourn-time distribution of customers.

			Quantiles of order							
			0.70		0.80		0.90		0.95	
ρ	A	V	Ex	MI	Ex	MI	Ex	MI	Ex	MI
0.75	30	0	73	75	83	85	93	96	105	113
		3	67	69	77	79	87	89	92	95
		6	66	67	75	75	85	85	90	91
		9	65	66	75	75	84	85	89	90
	60	0	45	46	53	54	60	61	63	66
		3	43	44	50	51	58	59	61	63
		6	42	43	49	50	57	58	61	61
		9	42	42	49	50	57	57	61	61
	90	0	22	22	27	27	33	33	35	36
		3	20	20	25	25	30	31	33	33
		6	19	19	24	24	30	30	32	33
		9	19	19	24	24	30	30	32	32
0.90	30	0	101	108	123	132	160	167	197	199
		3	76	80	86	90	100	109	121	131
		6	70	72	79	82	89	91	93	103
		9	68	69	77	79	86	88	91	93
	60	0	59	60	66	69	84	86	100	106
		3	50	52	57	59	64	68	77	81
		6	47	49	54	55	60	63	64	70
		9	46	47	52	54	59	61	62	66
	90	0	32	33	37	39	46	48	57	58
		3	27	28	31	33	37	40	45	47
		6	25	26	29	30	33	36	38	41
		9	24	25	28	29	32	34	35	38
0.95	30	0	165	174	209	213	284	277	359	337
		3	83	88	93	101	121	131	145	154
		6	72	76	82	85	91	97	107	116
		9	69	71	78	80	87	90	92	97
	60	0	84	87	104	108	138	140	172	168
		3	57	59	64	68	83	86	101	105
		6	51	52	57	59	63	69	78	82
		9	48	50	54	56	60	63	64	71
	90	0	47	48	59	60	80	79	100	99
		3	33	35	39	42	52	53	65	65
		6	29	30	32	35	40	43	50	51
		9	27	28	30	32	34	38	41	44

Table 6.4: *Quantiles of the sojourn-time distribution.*

7

Regular and Incidental Customers in Queueing Systems with Periodic Service

7.1 Introduction

In Chapter 6, we considered two modifications of discrete-time queueing systems with periodic service, namely, the possibility of making products to stock and of working overtime. These possibilities were studied by the techniques of the Chapters 4 and 5. In this chapter, we focus on a third modification that can be studied by these techniques. In this modification, it is assumed that there are customers arriving in a periodic way and that other customers arrive randomly. Moreover, it is assumed that the periodically arriving customers have priority for service over the other customers. Like the modifications in Chapter 6, this modification is mainly motivated by the study of a periodic production rule at production centres. To give a more detailed motivation, consider the following situation.

A production centre has regular and incidental customers, who both represent firms. Incidental customers place their orders in a highly irregular way, and the size of their orders may vary considerably. Each regular customer has a fairly steady production process and is able to place orders according to a rather steady pattern. However, these steady order patterns do in general not coincide with the start of the on-periods for these orders. When, for instance, the production centre makes good agreements with these regular customers, these customers may adjust their order patterns to the production rule. As a result, the centre faces a rather smooth arrival process of orders from regular customers. Moreover, the total size of these orders may be nearly deterministic. When the centre gives the regular customers priority for service over incidental customers, it does not have to use much additional capacity to guarantee that these customers are served before the end of the on-period of their arrival. Hence, in this case, the orders can practically always be handled in the on-period of arrival, so that the regular customers do not have to keep much safety stock.

In this chapter, we consider a discrete-time queueing system with periodic service. Some customers (henceforth called regular customers) arrive in a periodic way at the start of on-periods. Other

customers (in the sequel called incidental customers) arrive according to a random pattern. Further, the regular customers have preemptive priority for service over the incidental customers and are served before the start of the next off-period (by working overtime, for example). The aim of this chapter is to demonstrate that the techniques of the Chapters 4 and 5 can be used for studying this system. These techniques enable one to evaluate, for instance, the effect of splitting the arrival process of customers.

The remainder of this chapter is organised as follows. In Section 7.2, we describe the model. In Section 7.3, we study the additional capacity required in order to serve the regular customers before the start of the first off-period after their arrival. In Section 7.4, we use the techniques of the Chapters 4 and 5 for determining the stationary queue-length distribution and the sojourn-time distribution of incidental customers. We present a numerical example in Section 7.5. Finally, a summary of this chapter is given in Section 7.6.

7.2 The model

For convenience, we consider a cycle as defined in Chapter 4 with one on- and off-period. Each cycle consists of C slots, and the probability distribution of the length A of the on-periods has probability distribution $\{a(j), j = 0, 1, \ldots, C\}$.

Regular customers are supposed to arrive just after the first slot boundary of cycles instead of at the start of on-periods as mentioned in Section 7.1; the arrival instants of regular customers may otherwise be different for each cycle, because the lengths of the on-periods are not necessarily deterministic. The service times of these customers are measured in numbers of slots of work. The total service time B_R of regular customers arriving at the start of a cycle has probability distribution $\{b_R(j), j = 0, 1, 2, \ldots\}$. Furthermore, regular customers who are not served before the start of the first cycle after their arrival are all assumed to be served in overtime; the server works at infinite speed or work is put out to contract at the end of on-periods. It is important to note that there is no work of regular customers left at the start of cycles.

The arrival process and service times of incidental customers are supposed to be equal to those in Chapter 4. More precisely, customers are assumed to arrive according to a periodically time-dependent Bernoulli process and the service times have either a discrete distribution with bounded support or a distribution that is a mixture of a finite number of negative binomial distributions with the same parameter β. Incidental customers arriving in the first slot of the cycle are supposed to arrive just after regular customers. For the other slots in the cycle, we make the usual assumptions with respect to the arrival instants of incidental customers, and the start and completion of service phases (see also Figure 4.3).

As mentioned in Section 7.1, the regular customers have preemptive priority for service over the incidental customers as follows. At the start of each on-period, the server begins servicing the regular customers until either all regular customers are served or the next cycle begins, whatever occurs first. In the latter case, the remaining regular customers are assumed to be served by other means such as overtime or other firms. In the former case, the remaining on-period in the present cycle is used to serve incidental customers.

7.3 Regular customers

From the service policy, we know that there are no regular customers at the start of cycles. So, the queue-length process and the sojourn-time distribution of these customers are no interesting objects of study. However, the number of slots of work of regular customers that is not handled in the cycle of arrival, but that is handled by other means as overtime, is interesting to investigate from the server's point of view. In this section, we focus on this performance measure.

Let the random variable E denote the extended service capacity required (in terms of slots) in order to serve the regular customers in their cycle of arrival. Then, we have

$$E = \max\{0, B_R - A\}.$$

From this equation, the probability that the server requires no additional capacity is

$$\Pr\{E = 0\} = \sum_{j=0}^{C} \Pr\{B_R \leq j\} a_i(j) = \sum_{j=0}^{C} \sum_{k=0}^{j} b_R(k) a(j), \tag{7.1}$$

and the probability that the required additional service capacity equals k slots is

$$\Pr\{E = k\} = \sum_{j=0}^{C} b_R(j+k) a(j), \qquad k = 1, 2, 3, \ldots. \tag{7.2}$$

When only the first two moments of B_R, we may use the fitting procedure in Chapter 5 for approximating the probabilities (7.1) and (7.2).

7.4 Incidental customers

In this section, we determine the stationary queue-length and sojourn-time distribution of incidental customers. In fact, we show that the techniques of the Chapters 4 and 5 can readily be applied after a redefinition of the on- and off-periods.

Incidental customers are served during those parts of the on-periods in which no regular customers are in service. Due to the arrival process and the preemptive priority of regular customers, the time intervals during which these customers are served, effectively lengthen the off-periods faced by incidental customers. This circumstance suggests to consider an off-period for incidental customers as the original off-period plus the part of the on-period used for servicing regular customers. Accordingly, the on-period for incidental customers is the part of the original on-period that is not used for servicing regular customers. Let the random variable A^* denote the length of the on-period for incidental customers in the cycle. For the case that the length of the off-periods is fixed, a representation of the off- and on-period in a cycle as faced by incidental customers is given in Figure 7.1. The dotted lines denote the part used for servicing regular customers.

Now, we clearly have

$$A^* = \max\{0, A - B_R\},$$

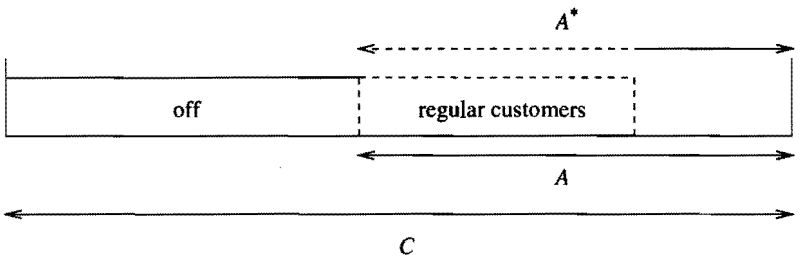

Figure 7.1: *A representation of a cycle as faced by incidental customers.*

Consequently, the probability distribution of A^* is given by

$$\Pr\{A^* = k\} = \begin{cases} \sum_{j=0}^{C} \sum_{l=j}^{\infty} b_R(l)a(j), & k = 0, \\ \sum_{j=k}^{C} b_R(j-k)a(j), & k = 1, 2, \ldots, C. \end{cases}$$

By the redefinition of the on- and off-periods, we may use the results and techniques of the Chapters 4 and Chapter 5 directly, when reading A^* instead of A.

7.5 A numerical example

As an illustration of the effect of splitting the arrival process of customers, we use the following example. Consider a production centre that produces raw materials, like plastic granules, to order for both incidental and regular customers. As regular customers, one might think of manufacturers of plastic toys, dashboards of cars, and plastic bags. For convenience, the regular customers are called firms.

A firm makes products to order as well. At the firm, the average number of products ordered per week is 18. The time to make such a product has a mean and a standard deviation of two hours. For each hour of production, the firm uses a fixed amount of raw materials. Its order policy for raw materials is the following. If, at the end of a week, the number of production hours since the last order exceeds 400, then the firm places a new order at the production centre for raw materials. The amount of raw materials that is used in 400 hours production, is produced in 200 minutes by the centre.

The number of hours of work arriving at the firm has mean 36 and standard deviation 144. Since the coefficient of variation of this number is small, we use a normal distribution to approximate the probability distribution of this number of hours work. Suppose that there are 40 firms placing orders in this way at the production centre. Using simulation, the total size of the orders placed by these firms at the start of each week has a mean of 12 hours and a standard deviation of 6.2 hours.

The incidental customers place orders for raw materials at the production centre according to a Bernoulli process with rate 0.075. Such an order requires on average 2 hours production, with a standard deviation of 3 hours.

The production centre uses a production cycle of two weeks, in which it can work 80 hours. The last 40 hours in a cycle are used to manufacture the raw materials. We consider an hour as a slot and assume

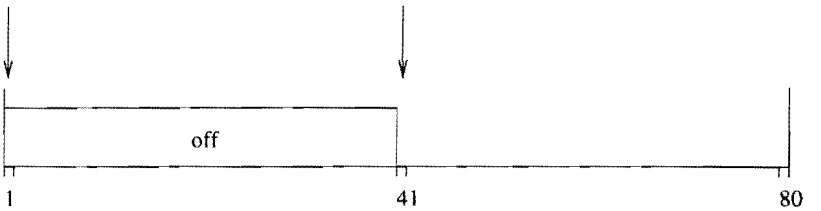

Figure 7.2: *A cycle and the instants at which firms place their orders.*

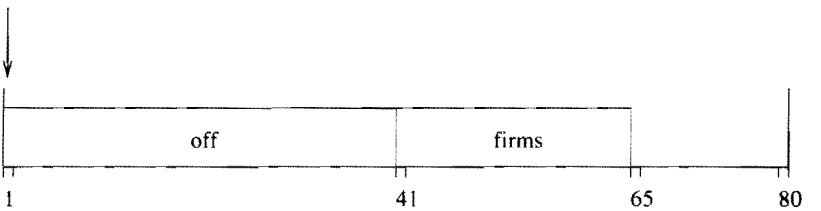

Figure 7.3: *A cycle as faced by incidental customers*

that firms place orders in the first and 41-st slot of the cycle, just before a possible incidental customer (see Figure 7.2). Further, we assume that the centre serves the firms and the incidental customers in the order of their arrival; no priority for service is giving to either of them.

Using the fit procedure of Theorem 5.5.2 and the MI technique, the approximation of the average sojourn time of incidental customers is 61.3 hours. The approximation for the quantiles of the sojourn-time distribution for incidental customers are given in the first row of Table 7.1.

	Quantiles of order			
	0.70	0.80	0.90	0.95
No split	76	94	122	151
Split	77	104	150	197

Table 7.1: *Quantiles for the sojourn-time distribution.*

Now, suppose that, by making good agreements, all firms place their orders every five cycles (that is, every 10 weeks). Moreover, the production centre indicates which firm orders at the start of which cycle. Then, this centre produces raw materials for 8 customers in each cycle, and delivers them at the end of the cycle. In each cycle, the total demand of these firms has an average of 24 production hours with a standard deviation of 0.44. So, the part of the production capacity claimed by firms is almost deterministic. Neglecting the variability in this part, we determine the sojourn-time distribution for incidental customers, by considering this part as an extension of the off-period. In Figure 7.3, we represent the cycle as faced by incidental customers.

The average sojourn time of incidental customers is now 70.5 hours approximately. In the second row of Table 7.1, we list the quantiles of the sojourn-time distribution of these customers. As we see, the average sojourn time increases with about 15%. Furthermore, the quantiles for the sojourn-time

distribution also increase. However, the degradation in performance for the incidental customers may be regarded as small. For instance, roughly 95% of these customers is delivered within 5 weeks instead of 4 weeks.

7.6 Conclusions

In this chapter, we demonstrate that a third modification of queueing systems with periodic service can be analysed by the techniques of the Chapters 4 and 5. For this modification, some customers arrive randomly and other customers arrive according to a periodic pattern. The latter customers arrive at the start of cycles only and they have preemptive priority for service over the randomly arriving customers. Furthermore, the periodically arriving customers are served in the cycle of their arrival, for instance, by overtime.

For the periodically arriving customers, we investigated the additional capacity required in order to serve them before the start of the first off-period after their arrival. For the incidental customers, we redefined the on- and off-periods as faced by them. After this redefinition, we could use the results and techniques of the Chapters 4 and 5 directly for evaluating the stationary queue-length distributions at slot boundaries and their sojourn-time distribution.

8

Conclusions and Suggestions for Future Research

In this monograph, we studied single-server multi-queue systems with periodic service. For these systems, the time axis consists of intervals of equal length, called cycles. In a cycle the server visits the different queues to serve customers. The order in which he visits the queues is the same for all cycles. Moreover, the time instants within a cycle at which he starts switching from one queue to another are fixed and the same for all cycles. Further, switching from one queue to another may take some time. These systems are used to model, for example, fixed-cycle traffic lights at intersections, communication and computer systems with periodic access schemes, and periodic production rules.

Under the common assumption that the arrival and service processes of customers are independent, the analysis of the joint queue-length process reduces to analysing the queue-length process for each queue separately. In spite of this reduction, there are not many useful techniques found in the literature for studying these processes. On one hand, classical analytical techniques such as the use of generating functions and Laplace-Stieltjes transforms may lead to analytical and numerical problems. On the other hand, some numerical techniques seem to be applicable to a rather limited class of queueing systems with periodic service, whereas other numerical approaches are focussed on deriving approximations for the average queue length and sojourn time only; information about averages is often insufficient for evaluating queueing systems.

The objective of this monograph was the development of techniques for analysing queueing systems with periodic service; in particular techniques for computing the queue-length and sojourn-time distributions of customers in statistical equilibrium. Considering these systems in discrete time and looking at one of the queues only, we have described two techniques for determining the stationary queue-length distribution for this queue. Once this distribution has been found, the sojourn-time distribution of customers in this queue could be computed exactly.

As an illustration of the two numerical techniques for determining the queue-length distribution, we considered in Chapter 2 a specific queueing system with periodic service. For this system, the queue-length process is described by a one-dimensional Markov chain. The equilibrium equations of

this chain constitute a homogeneous linear difference equation with constant coefficients, so that the stationary queue-length distribution is a linear combination of powers. Before presenting the two numerical techniques, we demonstrated the numerical difficulties that can occur when classical analytical techniques are used for determining this linear combination.

The first numerical technique (GT technique) discussed in Chapter 2, exploits the fact that the tail of the stationary queue-length distribution is asymptotically geometric. The unique parameter describing this tail behaviour can be computed easily and accurately. By imposing this behaviour on the stationary probabilities from a certain state onwards, we reduce the denumerable system of equilibrium equations to a finite system. Solving this finite system and the normalisation equation yields an approximation of the stationary queue-length distribution. This technique is based on Tijms & Van de Coevering [1991]. As numerical examples for the specific queueing system showed, the GT technique gives excellent results, even if the tail behaviour is imposed on states fairly close to the boundary of the state space.

The second numerical technique in Chapter 2 utilises the fact that the Markov chain corresponds to the waiting-time process in a special discrete-time $D/G/1$ queueing system. This process can also be represented by Lindley's equation. The limiting solution of this equation equals the stationary queue-length distribution. To compute this solution for the continuous-time case, De Kok [1989] developed an efficient moment-iteration algorithm that only uses the first two moments of the random variables involved. To these two moments, he fits a (continuous) probability distribution. Since we consider the queueing system in discrete time, it is more natural to use discrete distributions. Using a (novel) procedure for fitting discrete distributions on the first two moments, we applied the algorithm of De Kok [1989] to approximate the stationary queue-length distribution. Numerical examples showed that the approximations are excellent.

The GT technique in Chapter 2 exploited the structure of the solution to the equilibrium equations of a Markov chain that constitute a homogeneous linear difference equation with constant coefficients. In Chapter 3, we first investigated what class of Markov chains has equilibrium equations constituting such a difference equation. For these chains, we showed that transitions from state i to state j may depend on $j-i$ only, from a certain state i onwards, and that the upward jumps have to be uniformly bounded by some constant. Markov chains with this structure have a stationary distribution that is a linear combination of a finite number of powers, possibly except for a finite number of states at the boundary of the state space. Unlike the chain in Chapter 2, the tail behaviour of this linear combination may be described by more than one parameter. Therefore, we adapted the GT technique in Chapter 3 to these cases. Numerical results showed that this technique is efficient and that it gives accurate results without too much computational effort.

In Chapter 4, we applied the GT technique for determining the stationary queue-length distributions in queueing systems with periodic service. In order to use this technique, we had to impose restrictions on the arrival and service processes of customers. Fortunately, these restrictions were not too severe. For example, the GT technique is applicable when customers arrive according to a periodically time-dependent Bernoulli process and when the service times have an arbitrary discrete distribution with bounded support or a distribution that is a mixture of a finite number of negative binomial distributions with the same parameter q. Given the stationary queue-length distribution, we developed in Chapter 4 also an efficient algorithm for computing the sojourn-time distribution of customers. Numerical examples showed that the GT technique gives excellent results, and that it is much less sensitive to the utilisation of the system than, for example, simple truncation techniques. Furthermore, these examples demonstrated that the GT technique is much more advantageous from a computational point of view than brute computational force.

The GT technique used the complete service-time distributions of customers. However, in practice, one usually has only (approximate) information about the first two moments of these distributions. In Chapter 5, we presented the MI technique, which only uses this limited information. As a result, a larger class of service-time distributions could be dealt with than in Chapter 4. This technique resulted from the circumstance that the queue length at consecutive discrete time instants could be described by a relation that has a structure similar to Lindley's equation for the $GI/G/1$ queueing system. As mentioned, De Kok [1989] developed an efficient moment-iteration method in order to compute the limiting solution to this equation for the continuous-time case, using the first two moments of the interarrival-time and service-time distributions only. To these first two moments, he fitted a continuous probability distribution. The MI technique is an adapted version of this method. Since the random variables involved in this technique are discrete, it is more natural to fit discrete distributions. For fitting discrete distributions, several approaches can be found in the literature. However, some of these approaches do not capture all possible combinations of the first two moments, whereas others are not that useful for the MI technique. Therefore, we developed in Chapter 5 a novel procedure for fitting discrete distributions on the first two moments. Given the results as obtained by the MI technique, we used the algorithm in Chapter 4 to compute the sojourn-time distributions of customers. Numerical examples demonstrated that the MI technique, with this fit procedure, gives good results for the stationary queue-length and sojourn-time distributions. Only when the average queue lengths or sojourn times are very small, the relative errors may be larger, but in these cases the absolute errors are small.

The techniques developed in the Chapters 4 and 5 can be used to investigate different queueing systems with periodic service and to compare the performance of a periodic service policy with other service policies. Further, these techniques can also be applied to modifications of these systems. We considered three modifications that were motivated by the study of periodic production rules at production centres.

In Chapter 6, we considered the case that a production centre may make a limited number of products to stock and the case that the centre may work a limited amount of overtime. The analysis of these modifications was based on the circumstance that the corresponding queue-length processes could be related to those for a specific queueing system with periodic service but without either of these modifications. After some small adaptations, we computed the sojourn-time distribution of customers by the algorithm in Chapter 4.

In Chapter 7, we considered the case that some customers (regular customers, say) adapt their order pattern at a production centre to the periodic production rule. Furthermore, these customers have preemptive priority for service over the other customers, and they are served before the server leaves the queue (by overtime, for instance). By considering the time periods during which the server serves these regular customers as extensions of the switch-over times, the performance measures for the other customers could be computed by the same techniques as in the Chapters 4 and 5.

In this monograph, we presented the GT technique and the MI technique for evaluating the queue-length process in queueing systems with periodic service. Both techniques give approximations for the stationary queue-length distributions at the slot boundaries in the cycle. Numerous examples demonstrate that the approximations are accurate. However, in the literature, not much attention is focussed on theoretical results that give error bounds for these approximations or that give lower and upper bounds for the quantities of interest. For the GT technique, the theoretical results may be used to choose an appropriate value for J instead of the execution of several experiments for determining this

value. For the MI technique, these results may give insight in the accuracy of these approximations when no exact results are available. The derivation of such bounds may be an interesting topic for future research.

Further, numerical examples indicate that the MI technique always terminates. However, we have not been able to prove this yet, so that it stays open for future research.

For the queueing systems considered in this monograph, the waiting room was implicitly assumed to be infinite. In some applications, such as in communication and production systems, the waiting room may be finite. In this case, the Markov chain describing the queue-length process is finite as well. The finite system of equilibrium equations can then be solved directly. However, many of these equations may have the constant structure that is exploited by the GT technique. A topic for future research may be to investigate whether the GT technique (possibly after some adaptations in order to deal with the boundary equations at both sides of the state space) can be applied to finite Markov chains or not. Another way of approximating the stationary distribution of this finite Markov chain is to use the MI technique. In this method, we fit discrete distributions to the first two moments of the random variables involved. In this case, however, some of the variables, like the queue length at slot boundaries, may assume only a finite number of values. Then, a fitting procedure is needed that uses distributions with bounded support. Distributions with a bounded support play a role in many other queueing systems as well; for instance in queueing systems with bounded waiting time (cf. Cohen [1982] and Tijms [1986]). Therefore, the development of a practical analogue to the one in Chapter 5 for fitting distributions with bounded support seems to be a challenging and important topic for future research.

Bibliography

ACKROYD, M.H. [1980], Computing the waiting-time distribution for the $GI/G/1$ queue by signal-processing methods, *IEEE Transactions on Communications* **28**, pp. 52–58.

ACKROYD, M.H. [1984], Stationary and cyclostationary finite buffer behaviour computation via Levinson's method, *AT&T Bell Laboratories Technical Journal* **63**, pp. 2159–2170.

ACKROYD, M.H. [1985], Numerical computation of delays in clocked schedules, *AT&T Technical Journal* **64**, pp. 617–631.

ADAN, I.J.B.F. [1991], *A Compensation Approach for Queueing Problems*, Ph.D. thesis, Eindhoven University of Technology.

ADAN, I., M. VAN EENIGE, AND J. RESING [1995], Fitting discrete distributions on the first two moments, *Probability in the Engineering and Informational Sciences* **9**, pp. 623–632.

ADAN, I.J.B.F., J. WESSELS, AND W.H.M. ZIJM [1993], A compensation approach for two-dimensional Markov processes, *Advances in Applied Probability* **25**, pp. 783–817.

ADAN, I.J.B.F., AND Y. ZHAO [1994], *Analyzing $GI/E_r/1$ queues*, Memorandum COSOR 94-37, Department of Mathematics and Computing Science, Eindhoven University of Technology, (to appear in *Operations Research Letters*).

ALFA, A.S., AND M.F. NEUTS [1995], Modelling vehicular traffic using the discrete time Markovian arrival process, *Transportation Science* **29**, pp. 109–117.

ANDERSON, R.R., G.J. FOSCHINI, AND B. GOPINATH [1979], A queuing model for a hybrid data multiplexer, *The Bell System Technical Journal* **58**, pp. 279–300.

ASMUSSEN, S. [1987], *Applied Probability and Queues*, John Wiley & Sons, Chichester.

ATHREYA, K.B., AND P.E. NEY [1972], *Branching Processes*, Springer-Verlag, Berlin.

BAGCHI, T.P., AND J.G.C. TEMPLETON [1972], *Numerical Methods in Markov Chains and Bulk Queues*, in: M. Beckmann, G. Goos, and H.P. Künzi (eds.), *Lecture Notes in Economics and Mathematical Systems* 72, Springer-Verlag, Heidelberg.

BAILEY, N.T.J. [1954a], A continuous time treatment of a simple queue using generating functions, *Journal of the Royal Statistical Society (Series B)* **16**, pp. 288–291.

BAILEY, N.T.J. [1954b], On queueing processes with bulk service, *Journal of the Royal Statistical Society (Series B)* **16**, pp. 80–87.

BECKMANN, M., C.B. MCGUIRE, AND C.B. WINSTEN [1956], *Studies in the Economics of Transportation*, Yale University Press, New Haven.

BELLMAN, R. [1970], *Introduction to Matrix Analysis* (second ed.), McGraw-Hill, London.

BORST, S.C., O.J. BOXMA, J.H.A. HARINK, AND G.B. HUITEMA [1994], Optimization of fixed time polling schemes, *Telecommunication Systems* **3**, pp. 31–59.

BRAHIMI, M., AND D.J. WORTHINGTON [1991], The finite capacity multi-server queue with inhomogeneous arrival rate and discrete service time distribution and its application to continuous service time problems, *European Journal of Operational Research* **50**, pp. 310–324.

BRUNEEL, H. [1986], Message delay in TDMA channels with contiguous output, *IEEE Transactions on Communications* **34**, pp. 681–684.

BUCKLEY, D.J., AND R.C. WHEELER [1964], Some results for fixed-time traffic signals, *Journal of the Royal Statistical Society (Series B)* **26**, pp. 133–140.

CHU, W.W., AND A.G. KONHEIM [1972], On the analysis and modeling of a class of computer communication systems, *IEEE Transactions on Communications* **20**, pp. 645–660.

COHEN, J.W. [1982], *The Single Server Queue* (second ed.), North-Holland, Amsterdam.

COHEN, J.W. [1994], On a class of two-dimensional nearest-neighbour random walks, in: J. Gani (ed.), *Studies in Applied Probability (Journal of Applied Probability, Special Volume 31A)*, pp. 207–237.

CONOLLY, B.W. [1958a], A difference equation technique applied to the simple queue, *Journal of the Royal Statistical Society (Series B)* **20**, pp. 165–167.

CONOLLY, B.W. [1958b], A difference equation technique applied to the simple queue with arbitrary arrival interval distribution, *Journal of the Royal Statistical Society (Series B)* **20**, pp. 168–175.

COWAN, R. [1981], An analysis of the fixed-cycle traffic-light problem, *Journal of Applied Probability* **18**, pp. 672–683.

DARROCH, J.N. [1964], On the traffic-light queue, *Annals of Mathematical Statistics* **35**, pp. 380–388.

DELLAERT, N.P. [1988], *Production to Order*, Ph.D. thesis, Eindhoven University of Technology.

DE MORAES, L.F.M., AND I. RUBIN [1984], Message delays for a TDMA scheme under a nonpreemptive priority discipline, *IEEE Transactions on Communications* **32**, pp. 583–588.

DOSHI, B.T. [1986], Queueing systems with vacations: A survey, *Queueing Systems* **1**, pp. 29–66.

DOSHI, B.T. [1990], Single server queues with vacations, in: H. Takagi (ed.), *Stochastic Analysis of Computer and Communications Systems*, Elsevier Science Publishers, North-Holland, Amsterdam, pp. 217–265.

DOWNTON, F. [1955], Waiting time in bulk service queues, *Journal of the Royal Statistical Society (Series B)* **17**, pp. 256–261.

DOWNTON, F. [1956], On limiting distributions arising in bulk service queues, *Journal of the Royal Statistical Society (Series B)* **18**, pp. 265–274.

DREW, D.R. [1968], *Traffic Flow Theory and Control*, McGraw-Hill, London.

EENIGE, M.J.A. VAN, I.J.B.F. ADAN, J.A.C. RESING, AND J. VAN DER WAL [1995a], *Periodic service with working overtime and producing to stock in a multi-product production center*, Memorandum COSOR 95-12, Department of Mathematics and Computing Science, Eindhoven University of Technology.

EENIGE, M.J.A. VAN, I.J.B.F. ADAN, J.A.C. RESING, AND J. VAN DER WAL [1995b], *Periodic versus exhaustive service in a multi-product production center*, Memorandum COSOR 95-01, Department of Mathematics and Computing Science, Eindhoven University of Technology.

EENIGE, M.J.A. VAN, J.A.C. RESING, AND J. VAN DER WAL [1993], *A matrix-geometric analysis of queueing systems with periodic service interruptions*, Memorandum COSOR 93-32, Department of Mathematics and Computing Science, Eindhoven University of Technology.

FEDERGRUEN, A., AND L. GREEN [1986], Queueing systems with service interruptions, *Operations Research* **34**, pp. 752–768.

FEDERGRUEN, A., AND Z. KATALAN [1995], *The stochastic economic lot scheduling problem: Cyclical base-stock policies with idle times*, Working paper, Graduate School of Business, Columbia University, New York, (to appear in *Management Science*).

FELLER, W. [1968], *An Introduction to Probability Theory and Its Applications*, Volume I, (third ed.), John Wiley & Sons, New York.

FELLER, W. [1971], *An Introduction to Probability Theory and Its Applications*, Volume II, (second ed.), John Wiley & Sons, New York.

FISCHER, M.J. [1977a], Analysis and design of loop service systems via a diffusion approximation, *Operations Research* **25**, pp. 269–278.

FISCHER, M.J. [1977b], An approximation to queueing systems with interruptions, *Management Science* **24**, pp. 338–344.

FOSTER, F.G. [1953], On the stochastic matrices associated with certain queuing processes, *Annals of Mathematical Statistics* **24**, pp. 355–360.

FREDERICKS, A.A. [1979], Analysis of a class of schedules for computer systems with real time applications, in: M. Arató, A. Butrimenko, and E. Gelenbe (eds.), *Performance of Computer Systems*, North-Holland, Amsterdam, pp. 201–216.

FREDERICKS, A.A. [1982], A class of approximations for the waiting time distribution in a $GI/G/1$ queueing system, *The Bell System Technical Journal* **61**, pp. 295–325.

FREDERICKS, A.A., B.L. FARRELL, AND D.F. DEMAIO [1985], Approximate analysis of a generalized clocked schedule, *AT&T Technical Journal* **64**, pp. 597–615.

FUHRMANN, S.W., AND R.B. COOPER [1985], Stochastic decompositions in the $M/G/1$ queue with generalized vacations, *Operations Research* **33**, pp. 1117–1129.

GIFFIN, W.C. [1975], *Transform Techniques for Probability Modeling*, Academic Press, New York.

GOODWIN, J.C., JR., D. ELVERS, AND J.S. GOODWIN [1978], Overtime usage in a job shop environment, *OMEGA International Journal of Management Science* **6**, pp. 493–500.

GRIMMETT, G.R., AND D.R. STIRZAKER [1992], *Probability and Random Processes* (second ed.), Oxford University Press, New York.

GROSS, D., AND C.M. HARRIS [1974], *Fundamentals of Queueing Theory*, John Wiley & Sons, New York.

HADLEY, G., AND T.M. WHITIN [1963], *Analysis of Inventory Systems*, Prentice-Hall, Englewood-Cliffs.

HAIGHT, F.A. [1959], Overflow at a traffic light, *Biometrika* **46**, pp. 420–424.

HALFIN, S. [1983], Batch delays versus customer delays, *The Bell System Technical Journal* **62**, pp. 2011–2015.

HEIDEMANN, D. [1994], Queue length and delay distributions at traffic signals, *Transportation Research B* **28**, pp. 377–389.

HENRICI, P. [1968], *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley & Sons, New York.

HOUTUM, G.J.J.A.N. VAN [1995], *New Approaches for Multi-Dimensional Queueing Systems*, Ph.D. thesis, Eindhoven University of Technology.

JURY, E.I. [1964], *Theory and Application of the z-Transform Method*, John Wiley & Sons, New York.

KAPLAN, M. [1983], A single-server queue with cyclostationary arrivals and arithmetic service, *Operations Research* **31**, pp. 184–205.

KEILSON, J., AND L.D. SERVI [1990], A distributed Poisson approximation for preempt-resume clocked schedules, *IEEE Transactions on Communications* **38**, pp. 192–198.

KLEINROCK, L. [1975], *Queueing Systems, Volume I: Theory*, John Wiley & Sons, New York.

KLEINROCK, L. [1976], *Queueing Systems, Volume II: Computer Applications*, John Wiley & Sons, New York.

KO, K.T., AND B.R. DAVIS [1984], Delay analysis for a TDMA channel with contiguous output and

Poisson message arrival, *IEEE Transactions on Communications* **32**, pp. 707–709.

KOK, A.G. DE [1989], A moment-iteration method for approximating the waiting-time characteristics of the $GI/G/1$ queue, *Probability in the Engineering and Informational Sciences* **3**, pp. 273–287.

KOSOVYCH, O.S. [1978], Fixed assignment access technique, *IEEE Transactions on Communications* **26**, pp. 1370–1376.

LAM, S.S. [1977], Delay analysis of a time division multiple access (TDMA) channel, *IEEE Transactions on Communications* **25**, pp. 1489–1494.

LEVINSON, N. [1946], The Wiener RMS (root mean square) error criterion in filter design and prediction, *Journal of Mathematics and Physics* **25**, pp. 261–278.

LINDLEY, D.V. [1952], The theory of queues with a single server, *Proceedings of the Cambridge Philosophical Society* **48**, pp. 277–289.

LUCHAK, G. [1958], The continuous time solution of the equations of the single channel queue with a general class of service-time distributions by the method of generating functions, *Journal of the Royal Statistical Society (Series B)* **20**, pp. 176–181.

MCNEIL, D.R. [1968], A solution to the fixed-cycle traffic light problem for compound Poisson arrivals, *Journal of Applied Probability* **5**, pp. 624–635.

MEISSL, P. [1962], Stochastisches Modell einer festzeitgesteuerten Rot-Grün-Signalanlage, *Mathematik Technik Wirtschaft* **9**, pp. 11–18, 65–69.

MEISSL, P. [1963], Zufallsmodell einer Signalanlage mit mehrspurigem Stauraum, *Mathematik Technik Wirtschaft* **10**, pp. 1–4, 63–68.

MILLER, A.J. [1963], Settings for fixed-cycle traffic signals, *Operational Research Quarterly* **14**, pp. 373–386.

MINE, H., AND K. OHNO [1971], Traffic light queues as a generalization to queueing theory, *Journal of Applied Probability* **8**, pp. 480–493.

MORSE, P.M. [1958], *Queues, Inventories and Maintenance*, John Wiley & Sons, New York.

NAHMIAS, S. [1981], Managing repairable items inventory systems: A review, in: L.B. Schwarz (ed.), *Multi-Level Production/Inventory Control Systems: Theory and Practice*, TIMS Studies in the Management Sciences 16, North-Holland, Amsterdam, pp. 253–277.

NEUTS, M.F. [1981], *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, The Johns Hopkins University Press, Baltimore.

NEUTS, M.F. [1989], *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, Inc., New York.

NEWELL, G.F. [1956], Statistical analysis of the flow of highway traffic through a signalized intersection, *Quarterly of Applied Mathematics* **13**, p. 353–369.

NEWELL, G.F. [1960], Queues for a fixed-cycle traffic light, *Annals of Mathematical Statistics* **31**, pp. 589–597.

NEWELL, G.F. [1965], Approximation methods for queues with application to the fixed-cycle traffic light, *SIAM Review* **7**, pp. 223–240.

OHNO, K. [1978], Computational algorithm for a fixed cycle traffic signal and new approximate expressions for average delay, *Transportation Science* **12**, pp. 29–47.

ORD, J.K. [1972], *Families of Frequency Distributions*, Griffin, London.

OTT, T.J. [1984], On the $M/G/1$ queue with additional inputs, *Journal of Applied Probability* **21**, pp. 129–142.

OTT, T.J. [1987a], On the stationary waiting time distribution in the $GI/G/1$ queue, I: Transform

methods and almost-phase-type distributions, *Advances in Applied Probability* **19**, pp. 240–265.

OTT, T.J. [1987b], The single-server queue with independent GI/G and M/G input streams, *Advances in Applied Probability* **19**, pp. 266–286.

PAKES, A.G. [1969], Some conditions for ergodicity and recurrence of Markov chains, *Operations Research* **17**, pp. 1058–1061.

PEARSON, K. [1895], Contributions to the mathematical theory of evolution-II: Skew variation in homogeneous material, *Philosophical Transactions of the Royal Society of London (Series A)* **186**, pp. 343–414.

POWELL, W.B. [1986], Approximate, closed form moment formulas for bulk arrival, bulk service queues, *Transportation Science* **20**, pp. 13–23.

PRABHU, N.U. [1965], *Queues and Inventories: A Study of Their Basic Stochastic Processes*, John Wiley & Sons, New York.

PRESS, W.H., B.P. FLANNERY, S.A. TEUKOLSKY, AND W.T. VETTERLING [1986], *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge.

RAMASWAMI, V., AND D.M. LUCANTONI [1985], Stationary waiting time distribution in queues with phase type service and in quasi-birth-and-death processes, *Stochastic Models* **1**, pp. 125–136.

ROBILLARD, P., AND P. NAOR [1968], On queueing with regular service interruptions, *Revue Française d'Informatique et de Recherche Operationelle* **2**, pp. 11–29.

ROBINSON, E.A. [1967], *Multichannel Time Series Analysis with Digital Computer Programs*, Holden-Day, San Francisco.

ROMANOVSKY, V.I. [1970], *Discrete Markov Chains*, Wolters-Noordhoff, Groningen.

RUBIN, I. [1979a], Access-control disciplines for multi-access communication channels: Reservation and TDMA schemes, *IEEE Transactions on Information Theory* **25**, pp. 516–536.

RUBIN, I. [1979b], Message delays in FDMA and TDMA communication channels, *IEEE Transactions on Communications* **27**, pp. 769–777.

RUBIN, I., AND Z. ZHANG [1988], Message delay analysis for TDMA schemes using contiguous-slot assignments, *Proceedings of the IEEE International Conference on Communications 1988*, IEEE, New York, pp. 418–422.

ŞAHIN, I., AND U.N. BHAT [1971], A stochastic system with scheduled secondary inputs, *Operations Research* **19**, pp. 436–446.

SCHASSBERGER, R. [1974], A broad analysis of single server priority queues with two independent input streams, one of them Poisson, *Advances in Applied Probability* **6**, pp. 666–688.

SCUDDER, G.D. [1985], An evaluation of overtime policies for a repair shop, *Journal of Operations Management* **6**, pp. 87–98.

SCUDDER, G.D., AND R.C. CHUA [1987], Determining overtime policies for a repair shop, *OMEGA International Journal of Management Science* **15**, pp. 197–206.

SENGUPTA, B. [1990], A queue with service interruptions in an alternating random environment, *Operations Research* **38**, pp. 308–318.

SHANTHIKUMAR, J.G. [1988], On stochastic decomposition in $M/G/1$ type queues with generalized server vacations, *Operations Research* **36**, pp. 566–569.

SHERBROOKE, C.C. [1968], METRIC: A multi-echelon technique for recoverable item control, *Operations Research* **16**, pp. 122–141.

SMIT, J.H.A. DE [1971], The transient behaviour of the queue at a fixed cycle traffic light, *Transportation Research* **5**, pp. 1–14.

STEUTEL, F.W., AND M.J.A. VAN EENIGE [1996], *Note on the approximation of distributions on $\mathbb{Z}_+$ by mixtures of negative binomial distributions*, Memorandum COSOR 96-13, Department of Mathematics and Computing Science, Eindhoven University of Technology.

STEYAERT, B., AND H. BRUNEEL [1991], A general analysis of the packet delay in TDMA channels with contiguous-slot assignments, *Proceedings of the IEEE International Conference on Communications 1991*, IEEE, New York, pp. 1539–1543.

TAKÁCS, L. [1962], *Introduction to the Theory of Queues*, Oxford University Press, New York.

TIJMS, H.C. [1986], *Stochastic Modelling and Analysis: A Computational Approach*, John Wiley & Sons, New York.

TIJMS, H.C., AND M.C.T. VAN DE COEVERING [1991], A simple numerical approach for infinite-state Markov chains, *Probability in the Engineering and Informational Sciences* **5**, pp. 85–295.

TITCHMARSH, E.C. [1960], *The Theory of Functions* (second ed.), Oxford University Press, London.

VALDEZ-FLORES, C., AND R.M. FELDMAN [1989], A survey of preventive maintenance models for stochastically deteriorating single unit systems, *Naval Research Logistics* **36**, pp. 419–446.

WARDROP, J.G. [1952], Some theoretical aspects of road traffic research, *Proceedings of the Institution of Civil Engineers II* **1**, pp. 325–362.

WEBSTER, F.V. [1958], *Traffic signal settings*, Road Research Technical Paper No. 39, Road Research Laboratory, Department of Scientific and Industrial Research, Her Majesty's Stationery Office, London.

WINSTEN, C.B. [1959], Geometric distributions in the theory of queues, *Journal of the Royal Statistical Society (Series B)* **21**, pp. 1–35.

ZHANG, Z.G., R.G. VICKSON, AND M.J.A. VAN EENIGE [1995], *Optimal two-threshold policies in an $M/G/1$ queue with two vacation types*, Memorandum COSOR 95-36, Department of Mathematics and Computing Science, Eindhoven University of Technology, (to appear in *Performance Evaluation*).

Samenvatting

Dit proefschrift is gewijd aan het ontwikkelen van technieken voor het analyseren van wachtrijsystemen met periodieke bediening. Voor deze (en andere) wachtrijsystemen zijn rijlengten en verblijftijden van klanten de meest belangrijke prestatiematen. Vrijwel alle andere prestatiematen kunnen hiervan worden afgeleid. Daarom besteden we in het bijzonder aandacht aan technieken om de rijlengte- en verblijftijdverdelingen van klanten te bepalen.

Een wachtrijsysteem met periodieke bediening bestaat uit een aantal wachtrijen waar klanten arriveren. Klanten verlangen bediening van een gemeenschappelijke bediende, die hen helpt volgens een periodieke bedieningsdiscipline. Dit betekent dat de tijdas uit intervallen van gelijke lengte bestaat, welke cycli worden genoemd. Voor iedere cyclus is de volgorde waarin de bediende de rijen bezoekt hetzelfde. De momenten binnen een cyclus waarop hij van rij verandert liggen vast en zijn voor alle cycli gelijk. Het veranderen van rij door de bediende vergt eventueel tijd. Wachtrijsystemen met periodieke bediening worden gebruikt om onder andere verkeerslichtmodellen, communicatie- en computersystemen en produktiesystemen te beschrijven en te analyseren.

Hoewel deze wachtrijsystemen uit meerdere wachtrijen bestaan, kan het rijlengteproces van iedere rij afzonderlijk worden bestudeerd. Ondanks deze decompositie zijn er in de literatuur weinig geschikte technieken om deze systemen te analyseren. Zo leveren klassieke technieken, als het gebruik van genererende functies en Laplace-Stieltjes getransformeerden, analytische en numerieke problemen op. Verder lijken vele numerieke aanpakken slechts toepasbaar op zeer beperkte klassen van deze systemen of geven zij benaderingen voor alleen de gemiddelde rijlengte en verblijftijd; informatie omtrent gemiddelden is vaak onvoldoende om wachtrijsystemen te evalueren.

Het doel van dit proefschrift is om technieken te ontwikkelen voor het evalueren van wachtrijsystemen met periodieke bediening; in het bijzonder technieken voor het bepalen van de rijlengte- en verblijftijdverdelingen in de evenwichtssituatie. Hiertoe beschouwen we deze systemen in discrete tijd en bekijken we één van de wachtrijen. Voor het bepalen van de rijlengteverdeling van deze rij ontwikkelen we twee numerieke technieken. Als de rijlengteverdeling bekend is, dan kan de verblijftijdverdeling van klanten in deze rij exact worden berekend.

Ter illustratie van de basisideeën beschouwen we in hoofdstuk 2 een specifiek wachtrijsysteem met periodieke bediening. Voor dit systeem kan het rijlengteproces worden beschreven door een ééndimensionale Markov keten. De evenwichtsvergelijkingen van deze keten vormen een homogene, lineaire differentievergelijking met constante coëfficiënten, zodat de oplossing van deze vergelijkingen een lineaire combinatie van machten is. Als eerste laten we in hoofdstuk 2 zien dat het gebruik van standaard technieken om deze oplossing te berekenen kan leiden tot numerieke problemen. In het algemeen kunnen we deze problemen niet relateren aan eigenschappen van het aankomst- en bedieningsproces van klanten. Dit is voor ons een belangrijke reden om numerieke technieken te gebruiken.

De eerste numerieke techniek in hoofdstuk 2 benut het staartgedrag van de evenwichtsverdeling van de Markov keten, zoals is voorgesteld in Tijms & Van de Coevering [1991]. Dit staartgedrag is asymptotisch geometrisch en wordt bepaald door één parameter, welke eenvoudig en nauwkeurig kan worden berekend. Deze techniek, die we de GT techniek zullen noemen, legt dit geometrische staart-

gedrag op aan de evenwichtskansen vanaf een zekere toestand. Hierdoor houden we een eindig stelsel evenwichtsvergelijkingen over. Het oplossen van dit stelsel en de normalisatievergelijking levert vervolgens een benadering voor de evenwichtsverdeling op. Numerieke voorbeelden laten zien dat de GT techniek efficiënt is en nauwkeurige resultaten geeft.

De tweede numerieke techniek in hoofdstuk 2 is gebaseerd op het feit dat de voornoemde Markov keten correspondeert met het wachttijdproces van klanten in een speciaal discrete tijd $D/G/1$ wachtrijsysteem. Dit wachttijdproces kan ook worden beschreven door Lindley's vergelijking. De limietoplossing van deze vergelijking is gelijk aan de stationaire verdeling van de Markov keten. Voor het continue tijd $D/G/1$ wachtrijsysteem, gebruikt De Kok [1989] een momenten-iteratie methode om de limietoplossing van Lindley's vergelijking te benaderen. Deze methode gebruikt alleen de eerste twee momenten van de bedieningsduurverdelingen. Op deze twee momenten fit hij continue kansverdelingen. Een aangepaste versie van de methode van De Kok [1989] is de tweede numerieke techniek in hoofdstuk 2 (voorts aangeduid met MI techniek). Maar, omdat wij een discrete tijd model hebben, gebruiken we een (nieuwe) methode voor het fitten van *discrete* verdelingen op de eerste twee momenten van kansvariabelen. Numerieke voorbeelden laten zien dat de MI techniek zeer goede benaderingen geeft.

De GT techniek toegepast in hoofdstuk 2 is gebaseerd op het benutten van het geometrische staartgedrag van de evenwichtsverdeling van Markov ketens. Als de evenwichtsvergelijkingen van een Markov keten een homogene, lineaire differentievergelijking met constante coeffcienten vormen, dan heeft de evenwichtsverdeling dit staartgedrag; de oplossing van deze vergelijking is namelijk een lineaire combinatie van machten (behalve wellicht voor een eindig aantal toestanden aan de rand van de toestandsruimte). In hoofdstuk 3 beantwoorden we de vraag: Welke klasse van één-dimensionale Markov ketens heeft evenwichtsvergelijkingen met deze structuur? Het staartgedrag van de evenwichtsverdeling van ketens uit deze klasse kan door meer dan één parameter worden bepaald. In dit hoofdstuk passen we de GT techniek uit hoofdstuk 2 aan deze gevallen aan. Numerieke voorbeelden laten zien dat de GT techniek op efficiënt wijze zeer nauwkeurige resultaten geeft.

In hoofdstuk 4 passen we de GT techniek toe om de stationaire rijlengteverdeling in wachtrijsystemen met periodieke bediening te bepalen. Om deze techniek te gebruiken, dienen we restricties aan het aankomst- en bedieningsproces van klanten op te leggen. Deze restricties zijn niet erg beperkend. Zo mogen klanten arriveren volgens een periodiek tijdsafhankelijk Bernoulli proces en mogen bedieningsduren een willekeurige discrete verdeling hebben met een eindige drager of zijn verdeeld als een mengsel van een eindig aantal negatief binomiale verdelingen met dezelfde parameter q. Numerieke resultaten laten zien dat de stationaire rijlengteverdeling betrekkelijk snel het geometrische staartgedrag vertoont, zodat de GT techniek al goede resultaten geeft wanneer dit staartgedrag wordt opgelegd aan relatief lage toestanden. Daarnaast is deze techniek veel efficiënter en minder gevoelig voor de bezettingsgraad van het systeem dan bijvoorbeeld standaard truncatiemodellen. In hoofdstuk 4 presenteren we ook een efficiënt algoritme voor het berekenen van de verblijftijdverdeling van klanten. Gegeven de rijlengteverdeling, zijn de resultaten van dit algoritme exact.

De GT techniek gebruikt de gehele bedieningsduurverdeling. In de praktijk echter is vaak niet meer informatie beschikbaar dan de eerste twee momenten van de bedieningsduurverdeling. In hoofdstuk 5 presenteren we de MI techniek die alleen deze informatie gebruikt. Hierdoor kunnen we ook een grotere klasse van bedieningsduurverdelingen beschouwen dan in hoofdstuk 4. Deze techniek is een aangepaste versie van de eerder aangehaalde methode in De Kok [1989] en is gebaseerd op het feit dat de relatie tussen de rijlengte op twee opeenvolgende tijdstippen eenzelfde structuur heeft als Lindley's vergelijking voor een $GI/G/1$ wachtrijsysteem. De MI techniek itereert deze relatie, waarbij alleen

de eerste twee momenten van de betrokken kansvariabelen worden gebruikt. Op deze kansvariabelen fitten we een discrete verdeling. Verschillende procedures voor het fitten van discrete verdelingen zijn in de literatuur te vinden. Sommige van deze technieken omvatten niet alle mogelijke combinaties van de eerste twee momenten, daar waar andere technieken niet bruikbaar lijken voor de MI techniek. Daarom ontwikkelen we in hoofdstuk 5 ook een nieuwe procedure voor het fitten van discrete verdelingen op de eerste twee momenten van niet-negatieve kansvariabelen. Numerieke resultaten geven aan dat de MI techniek goede benaderingen geeft voor de gemiddelde rijlengte en verblijftijd. Wanneer de gemiddelde rijlengten of verblijftijden heel klein zijn, kunnen de relatieve fouten groter zijn, maar in deze gevallen zijn de absolute fouten klein. De benadering verkregen met de MI techniek wordt gebruikt om de verblijftijdverdeling te berekenen met het algoritme in hoofdstuk 4. De resultaten geven aan dat de vorm van de verdeling goed wordt benaderd.

Met de technieken uit de hoofdstukken 4 en 5 kunnen we nu wachtrijsystemen met periodieke bediening door te rekenen en de prestatiematen vergelijken met die van systemen met een andere bedieningsdiscipline. De GT en MI techniek zijn ook toepasbaar op modificaties van wachtrijsystemen met periodieke bediening. In de hoofdstukken 6 en 7 bekijken we drie modificaties die zijn gebaseerd op toepassingen in produktiesystemen.

In hoofdstuk 6 behandelen we de mogelijkheid tot het houden van een beperkte voorraad produkten en de mogelijkheid om een beperkte tijd over te werken. In deze systemen is het mogelijk om de rijlengteprocessen te relateren aan de rijlengteprocessen in de systemen van hoofdstuk 4 en 5. Hieruit volgt dat de GT en MI techniek vrijwel direct toepasbaar zijn. Bovendien kan het algoritme uit hoofdstuk 4, na kleine aanpassingen, worden gebruikt om de verblijftijdverdeling van klanten te berekenen.

In hoofdstuk 7 veronderstellen we dat bepaalde klanten (zeg, vaste klanten) volgens een periodiek proces arriveren welke overeenkomt met de tijdstippen waarop de bediende naar de betreffende rij gaat. Bovendien hebben deze klanten prioriteit over de andere klanten en worden zij bediend voordat de bediende de rij weer verlaat (door middel van overwerken bijvoorbeeld). Door de tijdsperioden gedurende welke de bediende vaste klanten helpt te beschouwen als extra omschakeltijden, is voor het rijlengteproces van de andere klanten een directe vertaling mogelijk naar de systemen bestudeerd in de hoofdstukken 4 en 5.

Curriculum Vitae

De schrijver van dit proefschrift werd op 10 april 1969 geboren te Voorburg. Van 1981 tot 1987 bezocht hij het Alfrink College te Zoetermeer. Na het behalen van het V.W.O.-diploma aldaar, begon hij aan de studie econometrie aan de Erasmus Universiteit te Rotterdam. In januari 1992 studeerde hij af in de richting bedrijfseconometrie. Het afstudeerwerk betrof de worst-case analyse van twee-dimensionale on-line bin packing algoritmen. Een deel van dit afstudeerwerk werd gedurende de maanden oktober-december 1991 verricht aan de Attila Joszef Universiteit te Szeged (Hongarije) in het kader van een TEMPUS-project. Bij het afstudeerwerk werd hij begeleid door dr. J.B.G. Frenk.

Van februari 1992 tot februari 1996 is de schrijver verbonden geweest als assistent in opleiding aan de faculteit Wiskunde en Informatica van de Technische Universiteit te Eindhoven. Het onderzoek dat hij gedurende die periode heeft verricht onder begeleiding van dr.ir. J. van der Wal en prof.dr. J. Wessels heeft geleid tot de totstandkoming van dit proefschrift.

Stellingen

behorende bij het proefschrift

Queueing Systems with Periodic Service

van

Michel van Eenige

I

Beschouw het algoritme in Coppersmith & Raghavan [1989] voor het twee-dimensionale on-line bin packing probleem met rechthoekige objecten. Veronderstel dat voor ieder object a de hoogte $h(a)$ en breedte $b(a)$ voldoen aan $0 < h(a) \leq 1/r$ en $0 < b(a) \leq 1/r$, waarbij r een constante is uit de verzameling $\{2, 3, 4, \ldots\}$. Objecten met $h(a) < b(a)$ en objecten met $h(a) \geq b(a)$ worden in verschillende bins geplaatst. Dit algoritme rondt de hoogte (als $h(a) < b(a)$) of de breedte (als $h(a) \geq b(a)$) van objecten naar boven af naar de kleinste waarde uit de verzameling

$$\{\tfrac{1}{2}, \tfrac{1}{4}, \ldots, (\tfrac{1}{2})^k, \ldots\} \cup \{\tfrac{1}{3}, \tfrac{1}{6}, \ldots, \tfrac{1}{3}(\tfrac{1}{2})^{k-1}, \ldots\}$$

en plaatst hen in stroken met deze hoogte of breedte. Door de speciale afmeting kunnen deze stroken efficiënt in bins van 1 bij 1 worden geplaatst. Als de objecten volgens de First-Fit heuristiek in stroken worden geplaatst, dan is de asymptotische worst-case ratio van dit algoritme gelijk aan $(3/2)(r+1)/r$. Hierbij wordt de factor $3/2$ veroorzaakt door de maximaal mogelijke afrondfout en is $(r+1)/r$ de asymptotische worst-case ratio van First-Fit.

COPPERSMITH, D., AND P. RAGHAVAN [1989], Multidimensional on-line bin packing: Algorithms and worst-case analysis, *Operations Research Letters* **8**, pp. 17-20.

CSIRIK, J., J.B.G. FRENK, AND M. LABBÉ [1993], Two-dimensional rectangle packing: On-line methods and results, *Discrete Applied Mathematics* **45**, pp. 197-204.

EENIGE, M.J.A. VAN [1992], Worst-case analysis of two-dimensional on-line bin packing algorithms, Master's thesis, Econometric Institute, Erasmus University Rotterdam.

II

Beschouw een continue-tijd Markov-keten met toestandsruimte $\{(i, j), i = 0, 1, \ldots; j = 1, 2, \ldots, J_i\}$ waarbij J_i constanten zijn die afhangen van i. Definieer de verzameling $\{(i, j), j = 1, 2, \ldots, J_i\}$ als *level* i voor $i = 0, 1, 2, \ldots$ en verdeel de toestandsruimte in deze *levels*. Veronderstel dat de generator van de Markov-keten de blok-tri-diagonale structuur

$$\begin{pmatrix} \boldsymbol{A}_{0,0} & \boldsymbol{A}_{0,1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \boldsymbol{A}_{1,0} & \boldsymbol{A}_{1,1} & \boldsymbol{A}_{1,2} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \boldsymbol{A}_{2,1} & \boldsymbol{A}_{2,2} & \boldsymbol{A}_{2,3} & \mathbf{0} & \cdots \\ \vdots & & \ddots & \ddots & \ddots & \end{pmatrix}$$

heeft. Stel bovendien dat de matrices $\boldsymbol{A}_{i,i-1}$ te schrijven zijn als het product van een kolomvector $\boldsymbol{x}_i$ en een rijvector $\boldsymbol{y}_i$ met $\boldsymbol{y}_i\boldsymbol{e} = 1$ voor $i = 2, 3, 4, \ldots$, waarbij $\boldsymbol{e}$ de kolomvector met alle elementen gelijk aan 1 is. Het rijlengteproces van het wachtrijsysteem in Van Eenige & Van der Wal [1995] geeft een voorbeeld van zo een Markov-keten.

Als de Markov-keten ergodisch is, kan de rijvector $\boldsymbol{\pi}_i$ bestaande uit de evenwichtskansen van toestanden op level i worden geschreven als

$$\boldsymbol{\pi}_{i+1} = \boldsymbol{\pi}_i \boldsymbol{A}_{i,i+1}(-(\boldsymbol{A}_{i+1,i+1} + \boldsymbol{A}_{i+1,i+2}\boldsymbol{e}\boldsymbol{y}_{i+2})^{-1}), \qquad i = 0, 1, 2, \ldots.$$

EENIGE, M.J.A. VAN, AND J. VAN DER WAL [1995], A non-homogeneous $M/Ph/1$ queueing system subject to breakdowns, in: U. Derigs, A. Bachem, and A. Drexl (eds.), *Operations Research Proceedings 1994*, Springer-Verlag, Berlin, pp. 221-226.

III

Beschouw een $M/G/1$-wachtrijsysteem met twee typen vakanties en laat n en N niet-negatieve constanten zijn met $n \leq N$. Zodra het systeem leeg is, gaat de bediende met een type 1 vakantie. Wanneer de bediende terugkomt van een vakantie, telt hij het aantal klanten in de rij. Als het aantal klanten i kleiner is dan n, gaat de bediende weer met een type 1 vakantie. Als $n \leq i < N$, gaat de bediende met een type 2 vakantie. Als bij terugkomst het aantal klanten i tenminste N bedraagt, begint de bediende met het bedienen van klanten totdat het systeem weer leeg is. De duur van type 1 en type 2 vakanties zijn exponentieel verdeeld met parameterwaarde ν_1, respectievelijk ν_2. Voor het bepalen van het gemiddelde aantal klanten in het systeem is het gebruik van de PASTA-eigenschap (zie Wolff [1982]) gecombineerd met een mean-value analyse eenvoudiger en inzichtelijker dan analyses die gebaseerd zijn op 'busy periods' in een $M/G/1$-wachtrij.

WOLFF, R.W. [1982], Poisson arrivals see time averages, *Operations Research* **30**, pp. 223-231.

ZHANG, Z.G., R.G. VICKSON, AND M.J.A. VAN EENIGE [1995], Optimal two-threshold policies in an $M/G/1$ queueing system with two vacation types, Memorandum COSOR 95-36, Department of Mathematics and Computing Science, Eindhoven University of Technology (to appear in *Performance Evaluation*).

IV

Beschouw het model beschreven in stelling III met $v_1 < v_2$. Als de wachtkosten van klanten evenredig zijn met de tijd, de kosten voor het opstarten van een 'busy period' constant zijn en de opbrengsten gedurende beide vakanties evenredig zijn met de duur, dan kunnen de optimale waarden van (n, N) in een eindig aantal stappen worden bepaald.

ZHANG, Z.G., R.G. VICKSON, AND M.J.A. VAN EENIGE [1995], Optimal two-threshold policies in an $M/G/1$ queueing system with two vacation types, Memorandum COSOR 95-36, Department of Mathematics and Computing Science, Eindhoven University of Technology (to appear in *Performance Evaluation*).

V

Voor niet-negatieve getallen m en c, met $0 < m \leq N$, is er een kansvariabele X op $\{0, 1, \ldots, N\}$ met verwachting m en variatiecoëfficiënt c, dan en slechts dan als

$$\frac{(\lfloor m \rfloor + 1 - m)(m - \lfloor m \rfloor)}{m^2} \leq c^2 \leq \frac{N}{m} - 1,$$

met $\lfloor m \rfloor$ het grootste gehele getal dat niet groter is dan m.

VI

Zij X een kansvariabele op $\{0, 1, \ldots, N\}$ met verwachting $m > 0$ en variatiecoëfficiënt c. Definieer $\theta := c^2 - 1/m$ en zij $\mathrm{Bin}(n, p)$ een kansvariabele met een binomiale verdeling waarbij n het aantal pogingen is en p de kans op succes. Dan heeft de kansvariabele Y op $\{0, 1, \ldots, N\}$ dezelfde verwachting en variatiecoëfficiënt als X, wanneer Y als volgt wordt gekozen:

1. Als $-1/n \leq \theta < -1/(n+1)$ voor zekere $n = 1, 2, \ldots, N-1$, dan

$$Y = \begin{cases} Bin(n, p), & \text{met kans } q, \\ Bin(n+1, p), & \text{met kans } 1 - q, \end{cases}$$

waarbij

$$q = \frac{1 + \theta(n+1) + \sqrt{-\theta n(n+1) - n}}{1 + \theta} \quad \text{en} \quad p = \frac{m}{n + 1 - q}.$$

2. Als $-1/N \leq \theta \leq (N-1)/m - 1$, dan

$$Y = \begin{cases} Bin(N, p_1), & \text{met kans } m/N, \\ Bin(N, p_2), & \text{met kans } 1 - m/N, \end{cases}$$

waarbij

$$p_1 = \frac{m}{N} + \frac{1}{N}\sqrt{\frac{Nmc^2(N-m)}{N-1} - \frac{(N-m)^2}{N-1}} \quad \text{en} \quad p_2 = \frac{m(1-p_1)}{N-m}.$$

VII

Een kansverdeling $\{p_k, k = 0, 1, 2, \ldots\}$ op $\mathbb{Z}_+$ kan willekeurig dicht worden benaderd door mengsel van negatief binomiale verdelingen, dan en slechts dan als $\{p_k, k = 0, 1, 2, \ldots\}$ een Poisson mengse is, d.w.z. dan en slechts dan als de kansgenererende functie P van $\{p_k, k = 0, 1, 2, \ldots\}$ voldoet aan

$$P(z) = \int_{[0,\infty)} e^{-x(1-z)} \mathrm{d}F(x),$$

met F een verdelingsfunctie op $\mathbb{R}_+$.

STEUTEL, F.W., AND M.J.A. VAN EENIGE [1996], Note on the approximation of distributions on $\mathbb{Z}_+$ by mixtures of negative binomial distributions, Memorandum COSOR 96-13, Department of Mathematics and Computing Science, Eindhoven University of Technology.

VIII

De opkomst en het verval van de Republiek der Verenigde Nederlanden als wereldmacht lopen voo een belangrijk deel parallel aan de opkomst en het verval van deze Republiek als zeemacht.

IX

In de voorwaarden en regels die van toepassing zijn op de Seizoen Club Card van een organisatie voo betaald voetbal staat te lezen dat:'Uitlenen van de kaart aan iemand anders is uiteraard toegestaan Hiermee lijken de KNVB en deze organisatie naast een belangrijk doel van deze kaart te schieten.

X

De zorgvuldigheid waarmee sommige wiskundigen eigenschappen toekennen aan wiskundige func ties staat soms in schril contrast met de onzorgvuldigheid waarmee zij eigenschappen toeschrijven aa supporters van bepaalde voetbalverenigingen.