Queueing Models for Cable Access Networks
Queueing Models for Cable Access Networks / by J.S.H. van Leeuwaarden. 
– Eindhoven : Technische Universiteit Eindhoven, 2005 
Proefschrift. – ISBN 90-386-0554-4 
NUR 919 
Subject headings : queueing theory, cable networks 
2000 Mathematics Subject Classification : 34A25, 35A22, 60K25, 68M20, 90B18 

Printed by Ponsen & Looijen BV 
Cover design by Paul Verspaget
Queueing Models for Cable Access Networks

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de Rector Magnificus, prof.dr.ir. C.J. van Duijn, voor een commissie aangewezen door het College voor Promoties in het openbaar te verdedigen
op maandag 13 juni 2005 om 16.00 uur

door

Johannes Simeon Hendrikus van Leeuwaarden

geboren te Eindhoven
Dit proefschrift is goedgekeurd door de promotoren:

prof.dr.ir. O.J. Boxma
en
prof.dr.ir. S.C. Borst

Copromotor:
dr. J.A.C. Resing
The effort that resulted in this thesis began back in 2002, and I owe many people gratitude for their help and encouragement since. Four people I would like to thank in particular.

First, I would like to express my gratitude to my supervisor Onno Boxma for letting me be part of his group. His great intuition, inspiring enthusiasm and good humor made the work challenging and fun at the same time. Onno guided me in the right directions and helped me out at numerous occasions.

Next, I would like to thank my advisor Jacques Resing for his encouragement and steadfast support, and the joint work covered in [P2, P3, P7, P8]. I am particularly grateful that Jacques let me join in on his idea, which eventually has led to Chapters 9 and 10 of this thesis.

I also take great pleasure in thanking Dee Denteneer from Philips Research. Dee has been the driving force behind the Pelican project: A joint project of Philips Research and EURANDOM on multi-access in cable networks. This thesis is one of the outcomes of the Pelican project. I have hugely benefitted from Dee’s ability to come up with (and solve) new and exciting problems. Our joint work has led to [P2, P3, P4, P13] and the material covered in Chapters 6-8.

Finally, I would like to thank Guido Janssen, also from Philips Research. When I knocked on his door some years ago, I did not expect this to have such a great impact on my work. The cooperation with Guido has been one big lesson for me, and has led to [P4, P5, P6, P9] and the material covered in Chapters 2-6. Particularly, Guido’s crucial idea to use Fourier sampling made it possible to derive the results in Chapters 3 and 4.

Many other people have kindly spent time sharing their knowledge with me. I wish to thank my second supervisor Sem Borst for his advice and for carefully proofreading the entire manuscript. I also thank Richard Boucherie, Herwig Bruneel and Erik Fledderus for serving on my doctoral committee.

I thank Ivo Adan for encouraging me to pursue a PhD, for helping me out at various occasions, and for joint work on [P11, P12]. I thank my office-mate Erik Winands for joint work on [P10, P12] and Monique van den Broek for joint work on [P11]. In connection with Chapter 11, I wish to thank Yiqiang Zhao and David McDonald for fruitful discussions. I thank Bart Steyaert for explaining a method I could use in [P3, P14], Koenraad Laevens for some helpful remarks on Chapter 5, and Ronald Rietman for commenting upon sections of this work.
I would like to thank Philips Research for funding my PhD position. Also, I thank all people at EURANDOM and the Stochastic Operations Research group at the Eindhoven University of Technology for creating a stimulating environment.

I owe a great debt of thanks to my friends and family. I thank Rob and Jan for their willingness to assist me during the thesis defense. A big thanks to my parents and my sister for their unconditional love. Lastly, I thank Anke for supporting me in everything but mathematics.

Johan van Leeuwaarden
May 2005


## Contents

Acknowledgments v

1 Motivation 1
   1.1 Cable access networks ........................................... 2
   1.2 Periodic scheduling .............................................. 7
   1.3 Tandem queues with shared service capacity ...................... 10
   1.4 Outline of the monograph ....................................... 13

I The discrete bulk service queue 15

2 Historical perspective and methodology 17
   2.1 Historical perspective ........................................... 18
   2.2 Generating function technique .................................. 21
   2.3 Random walk theory .............................................. 26
   2.4 Wiener-Hopf technique ........................................... 29
   2.5 Our contribution ................................................... 32
   2.6 Proof of Theorem 2.2.1 .......................................... 34
   2.7 Proof of Lemma 2.2.2 .............................................. 39

3 Fourier sampling 43
   3.1 A sketch of the approach ......................................... 44
   3.2 Generalized Szegö curves and Fourier sampling ................. 47
   3.3 Moments of the queue length .................................... 56
   3.4 Stationary queue length distribution ........................... 62
   3.5 Conclusions ........................................................... 66

4 Back to the roots 69
   4.1 Roots and Fourier series ......................................... 70
   4.2 Roots through fixed-point iteration ................................ 71
   4.3 Comparison of the two methods: Binomial case ................. 74
   4.4 On a result for a right-continuous random walk ............... 77
   4.5 Numerical results ................................................... 79
## CONTENTS

### 5 Relaxation time

- 5.1 Introduction and motivation .............................................. 84
- 5.2 Results ............................................................................. 88
- 5.3 Details for Theorem 5.2.1 .................................................. 90
- 5.4 Details for Theorem 5.2.2 .................................................. 93
- 5.5 Proof of Theorem 5.2.3 ..................................................... 96
- 5.6 Examples ......................................................................... 99
- 5.7 Results on the functions $\phi$ and $G$ ................................. 105

### 6 Moment bounds

- 6.1 Preliminaries ...................................................................... 110
- 6.2 Overview and results ....................................................... 111
- 6.3 Representations of the $\mu$-series and $\sigma^2$-series .......... 115
- 6.4 Bounds for the $\mu$-series ................................................ 117
- 6.5 Bounds for the $\sigma^2$-series ............................................. 118
- 6.6 Detailed results for the Poisson distribution ...................... 122
- 6.7 Numerical examples of series bounds ................................ 126
- 6.8 Numerical examples of moment bounds ......................... 132

### II Shared capacity models

### 7 Periodic scheduling

- 7.1 Introduction ...................................................................... 138
- 7.2 Models ............................................................................. 138
- 7.3 Stationary queue length .................................................. 139
- 7.4 Packet delay ..................................................................... 144
- 7.5 Numerical results ........................................................... 149
- 7.6 Conclusions ..................................................................... 152

### 8 Periodic scheduling with transmission delay

- 8.1 Model description and overview ....................................... 154
- 8.2 Mean queue length .......................................................... 155
- 8.3 An approximation for the correlation term ....................... 157
- 8.4 Properties ......................................................................... 159
- 8.5 Adaptive scheduling ....................................................... 160
- 8.6 Numerical evaluation ...................................................... 162
- 8.7 Proof of Theorem 8.2.2 .................................................... 167

### 9 Tandem queue with coupled processors

- 9.1 Introduction ...................................................................... 170
- 9.2 Model description and functional equation ....................... 170
- 9.3 Analysis of the kernel ..................................................... 172
- 9.4 Boundary value problem I ............................................... 176
- 9.5 Boundary value problem II ............................................. 178
## CONTENTS

9.6 Performance measures ........................................ 180  
9.7 Computational issues ........................................ 183  
9.8 Some examples ............................................... 188  

10 Two-station network with coupled processors .......... 191  
10.1 Introduction .............................................. 192  
10.2 Model description and functional equation .......... 192  
10.3 Performance measures ...................................... 194  
10.4 Preemptive priority at one of the stations .......... 196  
10.5 Analysis of the kernel ...................................... 198  
10.6 Boundary value problem .................................... 200  
10.7 Conclusions and further research ....................... 202  

11 Tail asymptotics ........................................... 205  
11.1 Introduction .............................................. 206  
11.2 Model description and functional equation .......... 207  
11.3 Priority for station 1 ...................................... 208  
11.4 Coupled processors and analytic continuation ...... 213  
11.5 Tail behavior ............................................. 220  
11.6 Alternative approaches .................................... 222  
11.7 Conclusions and further research ....................... 232  

Bibliography ....................................................... 235  
Index ............................................................... 247  
Samenvatting ....................................................... 249  
About the author ................................................ 253
Chapter 1

Motivation

Cable networks were originally designed to broadcast analogue television signals from the service provider to its users. With the help of hybrid fiber coaxial technology, most cable networks have been upgraded to provide bidirectional data transfer. The upgraded networks, referred to as cable access networks, thus allow for users to transmit signals to the service provider, which opens up a new world of interactive multimedia services with Internet browsing as most prominent application.

The upgrade of cable networks asks for ways to deal with the new situation of bidirectional data transfer. There has been a broad research effort on the description and investigation of protocols for regulating cable access networks. The work presented in this monograph is part of this effort.

This chapter serves to introduce the basic characteristics of cable access networks, and to describe how these give rise to challenging research issues. In particular, we discuss how the division of network capacity among its users leads to a two-stage process, which can be described in terms of several queueing models. In later chapters, most of these queueing models are solved analytically. From these solutions, we derive performance measures that can be used to assess the performance of cable access networks, expressed in terms of capacity and delay characteristics.

The queueing models presented in this monograph are interesting in their own right. Apart from their application in cable access networks, the models may find application in other fields. This particularly holds for the discrete bulk service queue, which is the subject of Chapters 2-6 and one of the standard models in digital communication. The queueing models covered in Chapters 7-11 incorporate characteristics of multi-access communication and resource sharing, issues that are the topic of ongoing research in fields like computer networks, radio frequency tagging, networks on chips, satellite systems, mobile telephony, and many more.
1.1 Cable access networks

The central point of the cable access network, which is connected to all users, is referred to as the head-end (HE). A service provider can transmit signals from the HE to the users over the downstream channel, and users can transmit signals from their location to the HE over the upstream channel (see Fig. 1.1). The downstream channel is used exclusively by the HE, while the upstream channel is shared by the users. Typically, the number of users connected to the same cable network ranges between 100 and 1000.

![Schematic view of a cable access network with N users.](image)

**Figure 1.1**: Schematic view of a cable access network with N users.

1.1.1 Multi-access communication and request-grant mechanism

The shared upstream channel is an example of multi-access communication, in which multiple users have access to the same communication channel. The range of applications of multi-access communication goes far beyond cable access networks and has attracted much attention from many researchers. For an overview we refer to Bertsekas & Gallager [30].

A typical problem in multi-access communication is that whenever users transmit signals simultaneously, a collision causing signal loss occurs. This is also the case for the shared upstream channel in cable access networks. A concomitant problem for cable access networks is that users are incapable of monitoring each other’s behavior and therefore cannot coordinate their transmissions themselves.

A common way to deal with this situation is to use a random access protocol, in which users transmit their message immediately without any form of coordination. Collisions might occur, so users must be informed when their message has been lost due to a collision. In case of collision, the HE sends a message to the user, and upon receipt the user will give it another try and retransmit the message. A random access protocol might work well, in particular when the load on the communication channel is low. When the load is high, collisions are more likely to occur and cause a substantial loss of capacity, which calls for a more sophisticated protocol.

To construct such a protocol, a scheduler can be installed at the HE that allocates the upstream capacity among the users. Then, each user must inform the HE about its capacity needs, after which the HE constructs a schedule and informs each user of the capacity it will receive. This information exchange between users and HE is often based on a request-grant mechanism, which can be described as follows.
Before a user can transmit its actual message, it sends a request message to the HE. This request message only contains the specifications of the actual message that a user wants to transmit. The request messages are handled by a random access protocol and could collide. However, the collided capacity remains limited, since the request messages are small. Still, the request messages require upstream capacity, but in return the HE gets the information based on which a collision-free schedule can be constructed. The HE informs the users when they can transmit their actual messages during reserved and thus collision-free time intervals.

1.1.2 Queueing theory and performance analysis

Queueing theory deals with analyzing congestion problems. Congestion may occur when users share a service system with limited capacity. Whenever the total demand to a system is more than its service capacity, the users should form some queue or waiting line. Users often decide individually when they need a certain service. Due to this uncontrolled arrival process to the system and the often varying service requirements of the users, queues may build up and dissolve over time, which leads to the formulation of stochastic models.

In the context of the cable access network, the upstream channel is the service system and the amount of data that users want to transmit is the service demand. The capacity of the upstream channel is limited, so queues will be formed. The available capacity and the way in which the users are served fully describe the service system. How the queues evolve, though, depends on both the service system and the behavior of the users.

Queues cause delay, which for some services could be problematic. That is, delay causes longer transmission times of data and therefore affects the quality of service provided to the user. Consequently, delay characteristics provide measures for the quality of the service system.

The upstream channel of cable access networks regulated by a request-grant mechanism might be viewed as a two-stage tandem queue. When a user wants to transmit data, it first joins the request queue where it waits until its request gets granted. Once granted, the user moves to the data queue and waits until its data gets transmitted. The data queue is virtual in the sense that packets are not actually lined up in a queue. Instead, the users hold their packets until they are allowed to actually transmit these. The total service capacity for both queues is equal to the capacity of the upstream channel. How the upstream capacity is scheduled, i.e. divided among the two queues, will determine the delay experienced by the users at each of the two stages.

1.1.3 Scheduling the upstream capacity

Starting from the abstraction of the two-stage tandem queue we now address the issue of scheduling the upstream capacity.

A first way to schedule the upstream capacity is to give priority to one of the queues. If priority is given to the request queue, all incoming requests are handled
until no requests are left. Then, the data of granted requests is transmitted only until a first new request arrives, and users might experience a substantial delay at the data queue. If priority is given to the data queue, a user is taken into service at the data queue right after getting its request granted at the request queue. In this case, the user has no delay at the data queue but could experience substantial delay at the request queue.

Giving priority to the data queue seems reasonable. That is, when both queues are nonempty, the users in both queues benefit from serving the data queue, since all users require service there eventually. On the contrary, the users at the data queue do not benefit from serving the request queue. This intuition is further substantiated by a result from Klimov [104], who indeed proves (under simplifying assumptions) that the optimal schedule in a tandem queue, in terms of the mean delay, is to give all capacity to the last nonempty queue in line. A similar result has been obtained recently by Wang & Wolff [160]. They consider a tandem queue, where a fraction \( p \) of the service capacity goes to queue 1 and \( 1 - p \) to queue 2, when both queues are nonempty. Under work conservation and first-in-first-out (FIFO), Wang & Wolff show that sample-path wise the delay in the system of every customer increases with \( p \). This again suggests that giving priority to the data queue would be optimal in minimizing the mean total delay.

What makes cable access networks different from the above standard tandem queue settings, is transmission delay. Due to the transmission delay, it takes a while before the scheduling instructions sent by the HE reach the users. So, from the moment a user gets served at the request queue it takes a while before the user is informed by the scheduler. Therefore, the upstream capacity cannot be used immediately for transmitting the data of this user. In Fig. 1.2 this is illustrated schematically.

![Figure 1.2: Schematic view of the upstream channel of a cable network regulated by a request-grant mechanism.](image-url)

The transmission delay could influence the behavior of the system considerably. Sala et al. [141] investigated the strategy that gives priority to the data queue by simulating a cable access network regulated by a request-grant mechanism with transmission delay. The capacity not needed for serving the data queue is used for the request queue. They observe that this type of scheduling results in a cyclic behavior. Serving the request queue for a longer period due to the transmission delay allows relatively many users to get their requests granted. These users are
served from the moment that the first user arrives at the data queue, resulting in a burst of data from users with granted requests. This burst of data will lead to requests being held relatively long at the users until all data have been transmitted, again inducing a burst of requests.

Sala et al. [141] compared the priority strategy to strategies that reduce the cycles by forcing upstream capacity to the request queue, even if there is data to be transmitted. They show that these strategies, which give every fixed period some of the capacity to the request queue, lead to a smoother process and may lead to shorter delays. This scheduling effect has also been observed in other simulations of cable access networks, see e.g. Golmie et al. [80] and Pronk et al. [137].

1.1.4 Key characteristics and research goal

Let us now summarize some of the characteristics discussed earlier, and relate these to the goals we would like to pursue in this monograph. For cable access networks regulated by a request-grant mechanism, we aim at incorporating the following characteristics into our models:

Request queue and data queue. A user first sends a request message to the HE and once this request gets granted, the user is allowed to send the actual message. This leads to the following abstraction. At the moment a user generates a request message, it joins the request queue. Once the request gets granted, the user leaves the request queue and joins the data queue, where it waits until it is allowed to transmit the actual message.

Transmission delay. It takes a while before a signal has been sent from one place to another in the network. The round-trip time is defined as the time it takes to send a signal from the HE to a user and from the user back to the HE. Scheduling instructions sent from the HE to the users are therefore delayed by half the round-trip time. In terms of the abstraction of the request and data queues, it takes half the round-trip time for a user to move from the request queue to the data queue.

Centralized scheduling. At the HE, a scheduler is installed that determines the way in which the capacity of the upstream channel is divided among the users. Using the abstraction of the request and data queues, the upstream capacity is divided among these two queues. The scheduling instructions sent by the HE to the users will be delayed.

Forced capacity for the request queue. Due to specific properties of cable access networks, including the delayed scheduling instructions, it might be favorable to force upstream capacity to the request queue, so that the arrival process of granted requests at the data queue gets smoother. The amount of forced capacity can be seen as a scheduling parameter, and we aim at investigating the impact of this parameter on several performance characteristics.

The two-stage tandem queue that consists of the request and data queues is our point of departure. The request queue in fact represents the process of handling requests with a random access protocol. These types of protocols have been thoroughly studied. The best known protocol is ALOHA, see Roberts [140], in which users send their requests without any form of coordination. Since requests may
collide, it is necessary that the users obtain some form of feedback. This can be achieved by an acknowledgment by the central scheduler. Alternatively, the users could listen to the channel and detect transmission conflicts themselves. In case the users cannot quickly detect transmission conflicts, more sophisticated random access protocols are preferable to ALOHA. These protocols are based on contention trees introduced by Capetanakis [45] and Tsybakov & Mikhailov [155]. In the context of cable access networks, the contention tree works as follows. Define a request slot as a fixed amount of capacity given to the request-grant mechanism and assume that each request slot is divided into a fixed number of mini-slots. A tree is then initialized when the requests of a group of users arrive in the same mini-slot (causing a collision). This group is then recursively split by dividing the users over the mini-slots of another request slot, where this division is usually achieved by random choice. Splitting the group continues until all users get their requests granted (by arriving as the only user in a mini-slot).

Contention trees for specific application in cable access networks have been studied by Denteneer [60]. Denteneer also investigates the overall system of the request queue and data queue. In doing this, he decomposes the two-stage tandem queue into a separate request queue and data queue, and studies the request queue by the machine repair model and the data queue by the so-called delayed bulk service queue. By obtaining expressions for the mean delay in both queues, Denteneer has been able to determine the mean delay in the two-stage tandem queue. This decomposition has its limitations, though, since it cannot be generalized to higher moments of the delay. That is, to obtain delay characteristics other than the mean, one needs to consider the interaction between the two queues.

The interaction between the two queues represents one of the most challenging aspects of the performance analysis of cable access networks. Several studies on the performance of cable networks have reported results obtained by simulation [80, 137, 141], but few have addressed the issues from an analytical viewpoint. The results presented in this monograph attempt to alleviate this hiatus in the literature. Notable contributions are Denteneer [60], as explained above, and Palmowski et al. [127]. In the latter paper, a two-stage tandem queue is considered in which the service requirement of a user at the second queue is coupled to its sojourn time at the first queue. In [127] Wiener-Hopf factorization for Markov modulated random walks is applied, which hints at the mathematical challenges involved in analyzing such a two-dimensional model.

Our goal also is to analyze the two-stage tandem queue. In doing this, inspired by the simulation results in [80, 137, 141], we focus on investigating the data queue. We take two approaches. As a first approach, we model the request and data queues as a discrete-time system. We incorporate characteristics of the cable access network like periodic scheduling, forced capacity for the request queue and transmission delay. The approach taken is discussed in Sec. 1.2. As a second approach, we analyze the tandem queue with shared service capacity using the theory of boundary value problems. This is discussed in Sec. 1.3.
1.2 Periodic scheduling

Let us consider the request and data queues (see Subsec. 1.1.4) as a discrete-time packet-based system, where time is divided in slots, and each slot is equal to the time needed to transmit one data packet (or handling requests in multiple mini-slots). The centralized scheduling of the upstream capacity then comes down to deciding for each slot whether it is used to serve the request queue or the data queue.

We will present several models for the data queue, where the data queue is defined as the amount of data (in terms of numbers of packets) that belongs to actual messages for which the request message has been granted, but that are still waiting to be transmitted. Clearly, if a slot is used for handling requests (request slot), new packets can enter the queue, and if a slot is used for data transmission (data slot), a packet can leave the queue.

Due to the substantial transmission delay, scheduling decisions must be taken in advance so that they can be communicated to the users. Consequently, there is a time lag between granting a request message and transmitting the data associated with the actual message. Therefore, one is naturally led to consider periodic scheduling, for which slots are grouped together into frames composed of both request and data slots. The designation of each slot in the frame is periodically determined and broadcast to all users, and the timing is such that each user is aware of the layout of a frame before it actually starts.

1.2.1 Fixed and flexible boundary model

We consider two periodic (frame-based) scheduling strategies. The first strategy uses no information about the system’s state and constitutes a queueing model that we refer to as fixed boundary model. Each frame (defined as \( f \) consecutive slots) consists of \( c \) request slots followed by \( s = f - c \) data slots. Let the random variable \( Y_{ti} \) denote the number of arriving packets during the \( i \)th request slot of frame \( t \), and assume that the \( Y_{ti} \) are independent and identically distributed (i.i.d.) for all \( t \) and \( i \). Further assume that packets that arrive during frame \( t \) cannot depart from the queue until the beginning of frame \( t + 1 \). We then have the following evolution equation that relates the queue lengths at the beginning of two consecutive frames:

\[
X_{t+1} = (X_t - s)^+ + \sum_{i=1}^{c} Y_{ti},
\]

where \( x^+ = \max\{0, x\} \) and \( X_t \) denotes the queue length at the beginning of frame \( t \) (see Fig. 1.3). The model essentially divides the upstream capacity among the request and data queues according to fixed fractions \( c/f \) and \( s/f \). Also, (1.1) falls within the class of the classical discrete bulk service queue, which is one of the best-known discrete queueing models. In this model, a fixed number of packets is transmitted periodically, while new packets arrive to the queue according to some stochastic process. The discrete bulk service queue has been applied to model an ATM (Asynchronous Transfer Mode) switching element, see Bruneel & Wuyts [44],
and became as such one of the standard models for performance analysis of digital communication systems. We have added various new elements to the existing literature on the classical formulation of the discrete bulk service queue, which will be extensively discussed in Chapter 2 of this monograph. In the remainder of this section we discuss two modifications to (1.1), leading to two additional models for the data queue.

Clearly, if the data queue is empty at the beginning of a data slot, this capacity is lost in the fixed boundary model. Therefore, the second model considered is one that designates the unused data slots as request slots, and is referred to as the flexible boundary model, which reflects the fact that the division of a frame into request and data slots can vary from one frame to another. This leads one to consider the recursion

$$X_{t+1} = (X_t - s)^+ + \sum_{i=1}^{c+(s-X_t)^+} Y_{ti}. \quad (1.2)$$

We refer to the $c$ request slots that are scheduled at the beginning of every frame as forced request slots, and to the $(s - X_t)^+$ slots as additional request slots. As mentioned earlier, in Sala et al. [141] the flexible boundary model has been investigated through simulations, and the results suggest that inducing request slots at the beginning of each frame reduces the data queue length and thus the delay experienced by the users. Intuitively, the flexible boundary model is more efficient than the fixed boundary model, but one wants to have a clear quantitative understanding of these benefits. We will provide such understanding by analyzing the packet delay in either model.

We now comment on the model assumptions for both (1.1) and (1.2). First we comment on the independence of the $Y_{ti}$, as assumed in both models. Clearly, the correlation between the $Y_{ti}$ depends on the exact way in which the request procedure is organized. For cable access networks, the requests are usually transmitted in contention with other users and based on ALOHA or contention trees (see Bertsekas & Gallager [30]). These procedures have a considerable randomness in the

Figure 1.3: The fixed boundary model. A frame of $f$ slots consists of $c$ request slots, followed by a maximum of $s = f - c$ data slots. Packets that arrive during frame $t$ cannot depart from the queue until the beginning of frame $t + 1$. 
order in which users are actually successful: The time that a user has already been
active in the contention procedure is not very significant as to its chances of being
the next user to successfully transmit its request, see e.g. Boxma et al. [38] and
Denteneer [60]. This suggests that the independence assumption made should be a
good approximation.

It remains to comment on the transmission delay, which is inherent to cable
networks and causes one to consider frame-based scheduling, see e.g. Golmie et
al. [80, 81]. Usually, the frame length \( f \) and the number of forced request slots \( c \)
are chosen so that the capacity \( s \) is greater than the transmission delay. In this way
one ensures that a schedule for frame \( t + 1 \) can include all successful requests from
the forced request slots in frame \( t \). Specifically, this ensures that arrivals during
the forced request slots of frame \( t \) can potentially depart in frame \( t + 1 \). Note that
this implies that, in case of the flexible boundary model, we must take into account
the exact location of the additional request slots. If they are located early within
a frame, they may still be included in a schedule for the next frame. If, however,
an additional arrival slot is located at the end of frame \( t \), the corresponding request
cannot be included in a schedule for frame \( t + 1 \) and must await the schedule for frame
\( t + 2 \). In our treatment of the flexible boundary model, we have taken an optimistic
viewpoint and have assumed that all granted requests in the additional request slots
of frame \( t \) can be scheduled in frame \( t + 1 \). Hence, while the transmission delay is
the main reason for applying periodic scheduling, due to the above assumptions, it
does not play a role in the analysis of both (1.1) and (1.2). Next, we modify (1.2)
such that it does incorporate the transmission delay.

1.2.2 Periodic scheduling with large round-trip delay

We now introduce the delay parameter \( d \), and we assume that the actual message
for which the request message gets granted in frame \( t \) can only be transmitted at the
earliest in frame \( t + 1 + d \). In other words, sending the request from the user to the
HE and transmitting the acknowledgment of a granted request from the HE to the
user takes \( d \) frames. Therefore, the user is informed of the scheduling instructions
\( d \) frames after its request has been sent. This gives rise to the following model
for the data queue, referred to as delayed flexible boundary model:

\[
X_{t+1} = (X_t - s)^+ + \sum_{i=1}^{c+(s-X_{t-d})^+} Y_{t-d,i}.
\]  

(1.3)

Finding the stationary distribution of the multi-dimensional Markov chain (1.3) is
much harder than in case of the one-dimensional Markov chains (1.1) and (1.2).
We therefore use approximating techniques, heuristic arguments and simulations to
study (1.3), and the influence of \( d \) in particular.

We deduce interesting properties of the mean queue length, and we use these
to construct an adaptive scheduling strategy that designates for every frame \( t \) the
number of request slots denoted by \( c_t \). The adaptive scheduling strategy defines
\( c_t \) as a function of the forced request slots scheduled in the previous \( d \) frames, i.e.
$c_{t-1}, c_{t-2}, \ldots, c_{t-d}$, and the queue length at the beginning of frame $t$. Note that the scheduling strategy in case of the flexible boundary model ($d = 0$) only depends on $c$ and $X_t$. The adaptive scheduling strategy uses more detailed information in order to cope with the transmission delay. It is shown that the adaptive scheduling strategy leads to significant reductions in both the mean and the variance of the stationary queue length.

1.2.3 Our contribution to periodic scheduling

The fixed and flexible boundary model are examples of queueing models with periodic service. Van Eenige [63] gives a broad overview of the work done on queueing models with periodic service and provides applications to traffic light queues and logistic systems. We contribute to this field by providing a detailed analysis for the evolution equations (1.1)-(1.3), presented in Chapters 7 and 8 of this monograph. What distinguishes (1.2) and (1.3) from most models in the literature is that the arrival process depends on the queue length process, which considerably complicates the analysis.

For the fixed boundary model (1.1) we show that the probability generating function of the stationary queue length follows from the solution of the classical discrete bulk service queue. We next derive, using a more advanced technique, the probability generating function of the packet delay. From these transform solutions, the entire probability distributions can be obtained, as well as explicit expressions for more specific performance characteristics like the mean and variance. For the flexible boundary model (1.2) we obtain similar results, although the derivation gets slightly more complicated. For both models we investigate the impact of the forced arrival slots $c$, in relation with other settings like the frame length and type of arrival process. For the delayed flexible boundary model (1.3) we derive bounds and approximations to investigate the influence of $c$ and $d$ on the mean and variance of the stationary queue length.

1.3 Tandem queues with shared service capacity

We now leave the discrete-time assumption and model the request and data queues as a continuous-time two-stage tandem queue for which the total service capacity should be divided among the two queues.

Although the fixed and flexible boundary models describe dependence between the two queues, these models are relatively easy to analyze. The main reasons for this are the fact that the two-dimensional system of the request and data queues is reduced to a one-dimensional model for the data queue by treating the request queue as a black box, and the fact that the transmission delay is partially ignored. The delayed flexible boundary model could not be solved explicitly. For the continuous-time models we aim at solving the two-dimensional system, where we keep track of both the request and data queues.
1.3 Tandem queues with shared service capacity

1.3.1 Coupled processors

Without loss of generality we assume that the service capacity of the upstream channel equals one unit of work per time unit. Then, whenever both the request queue and data queue are nonempty, this capacity should be divided: a proportion $p$ of the capacity is given to the request queue, and $1 - p$ to the data queue.

Let us now translate the discrete-time models introduced in Sec. 1.2 into their continuous-time counterparts. In the fixed boundary model, the capacity of the two queues is divided according to fixed fractions $p = c/f$ and $1 - p = s/f$, irrespective of whether one of the queues is empty. In the flexible boundary model, the unused capacity of the data queue is used for the request queue. So, whenever both queues are nonempty the service capacity is still divided according to $p = c/f$ and $1 - p = s/f$, but when the data queue is empty, $p$ is increased from $c/f$ to 1. We will refer to this scheduling discipline as partial coupling. Under partial coupling, the service capacity of the request queue depends on the workload of the data queue, and this interdependence between the queues severely complicates the analysis.

A natural extension of partial coupling is then full coupling, where not only the capacity of the request queue is increased from $c/f$ to 1 when the data queue is empty, but the capacity of the data queue is also increased from $s/f$ to 1 when the request queue is empty. Both partial and full coupling guarantee a minimum rate $p = c/f$ for the request queue and $1 - p = s/f$ for the data queue whenever there is work to be done at the queue in question. However, contrary to partial coupling, full coupling is work-conserving in the sense that the service (upstream) capacity is always fully used, irrespective of one of the queues being empty or not.

A service discipline that changes the service rates whenever one of the queues is empty is known in the queueing literature as coupled processors. If the coupled processors discipline is work-conserving, it reduces to full coupling. Full coupling is better known as generalized processor sharing (GPS). GPS is a popular scheduling discipline in modern communication networks, since it provides a way to achieve service differentiation among different types of traffic classes. For an overview of the literature on GPS we refer to Borst et al. [36], and the references therein. In the remainder of this monograph we will refer to GPS/full coupling as coupled processors.

1.3.2 Boundary value problems

When we assume that users arrive to the request queue according to a Poisson process, and that they require exponential service times at both queues, no coupling results in a tandem queue of two independent $M/M/1$ queues. Since this is a standard Jackson network, the stationary joint queue length distribution possesses a pleasant product form, see p. 193.

This does not hold for partial and full coupling. These service disciplines give rise to two-dimensional Markov processes that can be solved using the theory of boundary value problems. This is because the joint queue length can be modelled as a random walk on the lattice in the first quadrant, and belongs as such to the class of nearest-
neighbor random walks (only transitions to immediate neighbors may occur). A pioneering study of these types of random walks is the one of Malyshev [114], whose technique was introduced to queueing theory by Fayolle & Iasnogorodski [67]. They analyzed two parallel queues with coupled processors, each queue having Poisson arrivals and exponential service times. They showed that the functional equation for the probability generating function of the joint queue length distribution can be transformed to a Riemann-Hilbert boundary value problem. Cohen & Boxma [54] have presented a systematic and detailed study of the technique of reducing a two-dimensional functional equation of a random walk or queueing model to a boundary value problem, and discuss in detail the numerical issues involved. In particular, the analytic solution to the boundary value problem requires the determination of some conformal mapping, which can be accomplished via the solution of singular integral equations. In most cases, this requires a numerical approach (see Cohen & Boxma [54], Part IV).

Blanc [33] has investigated the transient behavior of the ordinary two-station tandem queue (so without coupled processors). In his analysis, Blanc transforms the functional equation for the probability generating function of the joint queue length distribution into a Riemann-Hilbert boundary value problem, using the same technique as introduced by Fayolle & Iasnogorodski [67]. For the two-stage tandem queue with coupled processors, Resing & Örmeci [139] made a similar transformation. Other applications of the theory of boundary value problems to queueing models can be found in Blanc [31], Coffman et al. [49], Cohen [53], Cohen & Boxma [54], Fayolle et al. [68], Fayolle et al. [69], De Klein [100], Mikou [117], Nauta [122], and references therein.

### 1.3.3 Our contribution to tandem queues with shared service capacity

For the two-stage tandem queue with coupled processors we show that the problem of finding the generating function of the joint stationary queue length distribution can be reduced to two different Riemann-Hilbert boundary value problems. We discuss the similarities and differences between the two boundary value problems, and relate them to the computational aspects of obtaining performance measures like the mean queue length and the fraction of time a queue is empty. Our detailed account of the numerical issues that arise when implementing a formal solution to a Riemann-Hilbert boundary value problem, is illustrative and may serve as an example for other types of queues that can be solved using the same technique. For the two-stage tandem queue with partial coupling we will show that the problem of finding the bivariate generating function of the joint stationary queue length distribution can be reduced to a Riemann-Hilbert boundary value problem of a slightly different type. The solution to this boundary value problem is more involved than the one for the coupled processors discipline. We indicate how the solution to the model with partial coupling can be obtained, but we do not discuss all details.

Next, we present a more general model of a two-station network with coupled processors. After receiving service at a station, a user either joins the queue of the same station, joins the queue of the other station, or leaves the system, each with
1.4 Outline of the monograph

We have described some of the key characteristics of cable access networks regulated by a request-grant mechanism in Sec. 1.1, and translated these into several discrete-time and continuous-time queueing models in Secs. 1.2 and 1.3, respectively. Each of these models will be addressed in Part II of this thesis. Part I is devoted entirely to the discrete bulk service queue, which came up in the formulation of the fixed boundary model defined by (1.1). As mentioned earlier, the discrete bulk service queue is a classical model in queueing theory, and its range of applications goes far beyond the scope of cable access networks. As such, Part I will be presented in general terms and can be read separately from the models considered in Part II that were inspired by cable access networks. Also, a detailed outline of Part I (which covers Chapters 2-6) will be given in Chapter 2. We will use some of the results obtained on the discrete bulk service queue for the analysis of the models in Part II. The outline of Part I is further specified in Sec. 2.5.

Throughout Part II of this monograph, we focus on deriving characteristics of the stationary queue length distribution or the stationary delay distribution for the models introduced in Secs. 1.2 and 1.3.

In Chapter 7, we consider the fixed and flexible boundary models. For both models we obtain expressions for the pgf of the stationary queue length and stationary packet delay. We investigate the impact of the forced request slots on various
14 Motivation

performance characteristics.

In Chapter 8, we study the delayed flexible boundary model. The models developed in Chapter 7 are based on the assumption that a packet that arrives in frame $t$ can be transmitted, at the earliest, in frame $t + 1$. Instead, we now assume that this packet can only be transmitted from frame $t + d$ on. We derive an exact expression for the mean stationary queue length at the beginning of a frame and present bounds for this expression. We further investigate several scheduling strategies using simulation.

In Chapter 9 we give a treatment of the two-stage tandem queue with coupled processors. We will show that the problem of finding the pgf of the joint stationary queue length distribution can be reduced to a Riemann-Hilbert boundary value problem. Starting from the solution of the boundary value problem, we consider the issues that arise when calculating performance measures like the mean queue length and the fraction of time a station is empty. We further briefly discuss the two-stage tandem queue with partial coupling.

In Chapter 10 we present the two-station network with coupled processors. For an open queueing network with two single-server stations, Poisson arrival streams, exponential service times and probabilistic routing, we will show that a similar approach can be taken as for the two-stage tandem queue.

In Chapter 11 we present asymptotic expressions for the tail distribution of the stationary queue length in the tandem queue with coupled processors. In particular, we perform an analytic continuation of the pgf of the joint stationary queue length distribution and determine its dominant singularities, from which the asymptotic expressions follow.

1.4.1 Literature summary

We now give an overview of the reports and papers upon which this thesis is largely built. Concerning Chapter 2, Sec. 2.4 stems from the paper Janssen & Van Leeuwaarden [P6], and Sec. 2.6 is based on the paper Adan et al. [P12]. Chapters 3, 4 and 5 are based on the papers Janssen & Van Leeuwaarden [P5, P6, P9]. Chapter 6 is mainly based on Denteneer et al. [P4], while some initial material was presented in the master thesis Van Leeuwaarden [P1] and the conference paper Denteneer et al. [P2]. Chapter 7 is based on work in Van Leeuwaarden [P1], which is also covered in the paper Van Leeuwaarden et al. [P3]. The material in Chapter 8 is partly based on the conference paper Denteneer & Van Leeuwaarden [P13]. The patentability of the scheduling algorithm described in Chapter 8 is currently being investigated. Chapter 9 is based on the paper Van Leeuwaarden & Resing [P7]. Chapter 10 is based on a preliminary version of Van Leeuwaarden & Resing [P8] and the material in Chapter 11 has not yet been published.
Part I

The discrete bulk service queue
Throughout Part I of this monograph we focus on deriving characteristics of the stationary queue length distribution for the discrete bulk service queue. This model has a deeply rooted place in queueing theory and appeared throughout the twentieth century in a variety of applications. The work done on the discrete bulk service queue runs to a large extent parallel to the maturing of queueing theory as a branch of mathematics. We therefore give an extensive description of the historical perspective in which the discrete bulk service queue can be placed. Next, we give a detailed account of the methodology that can be applied to solve for the stationary queue length distribution. The methodology can be roughly categorized into three techniques: The generating function technique, random walk theory, and the Wiener-Hopf technique. Depending on the technique used, characteristics of the stationary distribution can be expressed in terms of either the roots of some equation, or infinite series that involve convolutions of some probability distribution.

The three techniques cover the existing methodology to a large extent, both from the analytical and computational viewpoint. We will discuss each of the techniques, which facilitates us to give a precise formulation of the contributions that we have made to the existing literature. The historical overview is given in Sec. 2.1. We then present the generating function technique in Sec. 2.2, random walk theory in Sec. 2.3, and the Wiener-Hopf technique in Sec. 2.4. We end this chapter in Sec. 2.5 with a description of our contributions to the discrete bulk service queue. We relate our contributions to the three techniques and give an overview of the remaining chapters of Part I of this monograph.
2.1 Historical perspective

The first, somewhat disguised, appearance of the discrete bulk service queue was in the theory of telephone exchanges, going by the name $M/D/s$ queue. This model was introduced in the 1920’s by Erlang (see [40]), who is considered to be the founding father of queueing theory. At a telephone exchange with $s$ available channels, calls arrive according to a Poisson process. Each call occupies a channel for a constant time (holding time). Let $X_n$ denote the number of calls (both waiting and in service just after the $n$th holding time). Then, the following relation holds:

$$X_{n+1} = (X_n - s)^+ + A_n,$$  \hspace{1cm} (2.1)

where $x^+ = \max\{0, x\}$ and $A_n$ denotes the number of newly arriving calls during the $n$th holding time. It should be noted that due to the assumption of constant holding times, the calls which are in progress at the end of the $n$th holding time must have started during this holding time. Also, the calls which terminate during the $n$th holding time must have started before the beginning of this holding time.

The random variables $A_n$, $n = 0, 1, \ldots$ are assumed to be i.i.d. according to a random variable $A$ that has a Poisson distribution. Under the assumption that $EA < s$, the stationary distribution of the Markov chain defined by (2.1) exists. Denote by $X$ a random variable that has the same distribution as the stationary queue length.

Erlang obtained expressions for both the first moment and the distribution function of the stationary waiting time for values of $s = 1, 2, 3$. A first formal proof has been derived by Crommelin [55] in 1932, although this had already been indicated by Erlang. Crommelin used the generating function technique, which was remarkable at such an early stage, to obtain the pgf of $X$ expressed in terms of the $s$ roots on and within the unit circle of $z^s = \exp(\lambda(z - 1))$. From this pgf, Crommelin could obtain the distribution function of the stationary waiting time. At about the same time, Pollaczek treated the $M/D/s$ queue in a series of papers, generalizing it to the $M/G/s$ queue. Pollaczek’s work [128] was difficult to read, since he relied on rather complicated analysis, so Crommelin [56] gave an exposition of Pollaczek’s theory for the $M/D/s$ queue and found his own results in agreement with those of Pollaczek. Both methods lead to a solution in terms of infinite series that involve convolutions of the Poisson distribution. It is noteworthy that, after a lull in the literature of more than sixty years, Franx [76] came up recently with alternative expressions for the stationary waiting time distribution in the $M/D/s$ queue.

The infinite series-type result was generalized by Pollaczek [129]. In his derivation of the stationary waiting time distribution for the $G/G/1$ queue, Pollaczek obtained an identity, which was some years later obtained independently and by a different method by Spitzer [146]. Pollaczek again used complicated analysis, whereas Spitzer gave an elegant combinatorial proof. This is probably the reason why the result goes down in history as Spitzer’s identity, despite the efforts of Syski [149], who pointed out the equivalence of the two results. For a detailed treatment of Spitzer’s identity, we refer to Sec. 2.3.
2.1 Historical perspective

2.1.1 From telephony to digital data transfer

Recursion (2.1) that describes the queue length process in the \( M/D/s \) queue fits into the framework of bulk service queues. In this type of queues, at each epoch of service, a number of customers is taken from the queue. The bulk service queue originates from the work of Bailey [27] in 1954. Bailey modelled the situation where a doctor is prepared to see a maximum of no more than \( s \) patients per clinic session. The new patients who arrive during the clinic session join the queue right after the session ends. Bailey assumed that patients arrive according to a Poisson process, and in case of deterministic visiting times (Bailey allows for generally distributed visiting times), the recursive relation (2.1) would hold. Note that both the \( M/D/s \) queue and Bailey’s bulk service queue are continuous-time models that can be described in terms of discrete random variables by assuming Poisson arrivals and considering the queue at specific (embedded) points in time.

The first real discrete-time bulk service queue was introduced by Boudreau et al. [37] in 1962. They modelled the situation of a helicopter leaving a station every twenty minutes carrying a maximum of \( s \) passengers. Passengers that arrive between subsequent departures join the queue just after the next departure instant, again leading to (2.1), except now \( A \) can be any discrete random variable (with \( E(A) < s \)), instead of just Poisson. This generalization does not increase the complexity much, and so the method applied by Boudreau et al. [37] is almost identical to that of Bailey.

Up till the mid 1970’s, applications of bulk service queues were scarce. The most notable exception is the problem of estimating delays at traffic lights that alternate between periods of red and green (yellow is disregarded) of fixed length. For this traffic problem, bulk service queueing theory has been used to develop closed-form approximations for the expected delay (see e.g. Darroch [59], McNeill [116], Miller [119], Newell [125] and Webster [161]).

The real resurrection of the interest in the bulk service queue came in the mid 1970’s with the emergence of computer applications and digital data transfer. During the last decades of the twentieth century, discrete-time models have been applied to model digital communication systems such as multiplexers and packet switches. In this field, the discrete bulk service queue plays a key role due to its wide range of applications, among which the Asynchronous Transfer Mode (ATM) switching element (see Bruneel & Kim [43] and the references therein). For this model, time is divided into slots of fixed length, and again (2.1) holds with \( X_n \) the queue content (in terms of packets) at the beginning of slot \( n \), \( A_n \) the number of new packets that arrive during slot \( n \), and \( s \) the maximum number of packets that can be served during one slot. Besides the discrete bulk service queue, there are many other types of bulk queueing models, for which we refer to Baghi & Templeton [26], Bruneel & Kim [43], Cohen [51], Chaudhry & Templeton [48] and Powell [133].
2.1.2 Methodology

Deriving expressions for Laplace-Stieltjes transforms or pgf’s that contain roots of some equation has become a classic procedure in queueing theory. When applying the generating function technique, as introduced by Crommelin [55], the consideration of roots is often inevitable. Initially, the need for roots was considered to be a slur on the transform solutions, since the determination of the roots could be numerically troublesome and the roots themselves have no probabilistic interpretation. However, due to advanced numerical algorithms and increased computational power, root-finding has become more or less straightforward. In Chaudhry et al. [46] it is demonstrated that root-finding in queueing theory is well-structured, in the sense that the roots are distinct for most models and that their location is well-predictable, so that numerical problems are not likely to occur.

In case of the discrete bulk service queue, there is at least one alternative to root-finding. Using the recursive relation (2.1), the distribution of $X_{n+1}$ follows from the convolution of the distribution of $(X_n - s)^+$ and the distribution of $A$. Since discrete convolutions are not so hard to compute (see e.g. Ackroyd [21]), one could iteration (2.1) to obtain the transient queue length distributions which eventually will tend to the stationary distribution for increasing values of $n$. This idea of iterating (2.1) can be made more rigorous using random walk (or fluctuation) theory.

Many of the results from random walk theory are important for queueing theory. In particular, the waiting-time process in the $G/G/1$ queue where customers are served in order of arrival can be viewed as a random walk with a reflecting barrier at zero. The evolution equation that relates the waiting times of two subsequent customers is nowadays referred to as Lindley’s equation and given by

$$W_{n+1} = (W_n + B_n - C_n)^+, \quad n = 0, 1, \ldots, \quad (2.2)$$

where $W_n$ denotes the waiting time of the $n$th arriving customer, $B_n$ denotes the service time of the $n$th arriving customer, and $C_n$ denotes the interarrival time between the $n$th and $(n+1)$st arriving customer. Lindley [113] showed that, due to the $\max\{0, \cdot\}$ operator, finding the stationary waiting-time distribution requires the solution of a Wiener-Hopf type integral equation. With these observations, Lindley opened up a new field of research in which the Wiener-Hopf technique (see e.g. Smith [145], De Smit [144]) and other methods from random walk theory were used to study queueing models. For many types of queues, the Wiener-Hopf technique leads to an explicit factorization in terms of the roots of some characteristic equation. For an overview of the results from random walk theory that play a role in queueing theory we refer to Cohen [51], Sec. I.6.6, and Asmussen [25], Chapter 8. Perhaps the most famous result is the earlier-mentioned Spitzer’s identity which, among other things, expresses the Laplace transform of the stationary waiting-time distribution in terms of an infinite series that involves convolutions of some given probability distribution (see Sec. 2.3 for a detailed treatment).

It is quite common that for a particular queueing model, one or more of the processes of interest may be described in terms of a Lindley equation. In fact, (2.1) is a Lindley equation as well. This means that the methods developed to solve
Lindley’s equation for the general case become also available for the discrete bulk service queue. Equation (2.1) allows for a Wiener-Hopf factorization, which results in the same solution for the pgf of the stationary queue length as obtained with the generating function technique. Again, the solution requires the roots of some characteristic equation.

We have mentioned three techniques that can be applied to deal with the discrete bulk service queue: The generating function technique, random walk theory and the Wiener-Hopf technique. All three techniques can be applied to solve for the stationary regime and result in the pgf of the stationary queue length, denoted by $X(z)$. The generating function technique is the most traditional method and leads to an expression for $X(z)$ that includes the roots on and inside the unit circle of some equation. Random walk theory comes into the picture when one observes that the queue length process is a random walk with a reflecting barrier at zero. Spitzer’s identity then yields an expression for $X(z)$ in terms of infinite series that involve convolutions of the probability distribution of $A$. The Wiener-Hopf technique allows for two solutions, $X(z)$ in terms of roots as obtained by the generating function technique, and $X(z)$ in terms of infinite series as obtained from random walk theory. In that respect we might say that the Wiener-Hopf technique can be considered as the broadest approach. However, its application is far from straightforward and requires more advanced mathematics than is needed for the generating function technique and random walk theory. Therefore, we first present the latter two techniques, and then derive the same results with the Wiener-Hopf technique. Although the three techniques each have a broad range of applications, we present them, for reasons of clarity, in the context of the discrete bulk service queue.

### 2.2 Generating function technique

The discrete bulk service queue is defined by the recursion

$$X_{n+1} = (X_n - s)^+ + A_n. \tag{2.3}$$

Here, time is assumed to be slotted, $X_n$ denotes the queue length at the beginning of slot $n$, $A_n$ denotes the number of new packets that arrive during slot $n$, and $s$ denotes the maximum number of packets that can be transmitted in one slot. Packets that arrive to the queue in slot $n$ can be transmitted at the earliest from the beginning of slot $n + 1$. This is no restrictive assumption, since studying the queue $X_{n+1} = (X_n + A_n - s)^+$ is equivalent, see Sec. 2.3.

We denote for a non-negative discrete random variable $Y$ its mean by $\mathbb{E}Y$ or $\mu_Y$, its variance by $\sigma_Y^2$ and $\mathbb{P}(Y = j)$ by $y_j$. Furthermore, we denote the pgf of $Y$ by $Y(z)$, i.e.

$$Y(z) = \sum_{j=0}^{\infty} y_j z^j, \tag{2.4}$$

which is known to be analytic for $|z| < 1$ and continuous for $|z| \leq 1$. The numbers of new packets that arrive per slot are assumed to be i.i.d. according to a discrete
random variable $A$ with $a_j = \mathbb{P}(A = j)$ and pgf $A(z)$. We assume that $a_0 > 0$, which involves no essential limitation: If $a_0$ were zero, we would replace the distribution \( \{a_i\}_{i \geq 0} \) by \( \{a_i'\}_{i \geq 0} \) where $a_i' = a_{i+m}$, $a_m$ being the first non-zero entry of \( \{a_i\}_{i \geq 0} \), and a corresponding decrease in the maximum number of packets transmitted per slot according to $s' = s - m$.

Assume that $\mu_A < s$. Then, the stationary queue length distribution exists (see e.g. Bruneel & Kim [43]). Let $X$ denote the random variable following the stationary distribution of the Markov chain defined by (2.3), with

$$x_j = \mathbb{P}(X = j) = \lim_{n \to \infty} \mathbb{P}(X_n = j), \quad j = 0, 1, 2, \ldots.$$  \hfill (2.5)

The stationary queue length distribution satisfies the balance equations

$$x_k = \sum_{j=s}^{s+k} x_j a_{k-j+s} + \sum_{j=0}^{s-1} x_j a_k, \quad k = 0, 1, 2, \ldots.$$  \hfill (2.6)

Multiplying both sides of the above expression with $z^k$ and summing over all values of $k$ yields

$$X(z) = \sum_{k=0}^{\infty} x_k z^k$$

$$= \sum_{k=0}^{\infty} \sum_{j=s}^{s+k} x_j a_{k-j+s} z^k + \sum_{j=0}^{s-1} x_j a_k z^k$$

$$= z^{-s} \sum_{j=s}^{\infty} x_j z^j \sum_{k=j-s}^{\infty} a_{k-j+s} z^{k-j+s} + \sum_{j=0}^{s-1} x_j \sum_{k=0}^{s-1} a_k z^k$$

$$= z^{-s} X(z) A(z) - z^{-s} \sum_{j=0}^{s-1} x_j z^j A(z) + \sum_{j=0}^{s-1} x_j A(z).$$  \hfill (2.7)

Rewriting (2.7) results in the following expression for $X(z)$ (see e.g. Bruneel & Kim [43])

$$X(z) = \frac{A(z) \sum_{j=0}^{s-1} x_j (z^s - z^j)}{z^s - A(z)}, \quad |z| \leq 1.$$  \hfill (2.8)

The expression (2.8) is of indeterminate form, but the $s$ unknowns $x_0, \ldots, x_{s-1}$ can be determined by consideration of the zeros of the denominator in (2.8) that lie on or within the unit circle (see e.g. Bailey [27], Zhao & Campbell [165]).

We can prove the following result:

**Theorem 2.2.1** Under the condition that $\mu_A < s$, it holds that the function $z^s = A(z)$ has $s$ roots on or within the unit circle.

**Proof** See Sec. 2.6. \hfill \square
The \( s \) roots of \( z^s = A(z) \) in \(|z| \leq 1\) are denoted by \( z_0 = 1, z_1, \ldots, z_{s-1} \). For the ease of presentation we assume that these roots are distinct, but the theory presented below can be easily extended to the case in which there are multiple roots, see Remark 2.2.3.

Since the function \( X(z) \) is finite on and inside the unit circle, the numerator of the right-hand side of (2.8) needs to be zero for each of the roots, see Remark 2.2.3. The ease of presentation we assume that these roots are distinct, but the theory should vanish at the exact points where the denominator of the right-hand side of (2.8) vanishes. This gives the following \( s \) equations

\[
\sum_{j=0}^{s-1} x_j (z_k^s - z_j^s) = 0, \quad k = 0, 1, \ldots, s-1. \tag{2.9}
\]

For \( z_0 = 1 \), the above equation has a trivial solution, but the normalization condition \( X(1) = 1 \) provides an additional equation. Using l'Hôpital's rule, this equation is found to be

\[
s - \mu_A = \sum_{j=0}^{s-1} x_j (s - j), \tag{2.10}
\]

where both sides represent the average unused service capacity.

The system of equations can be written in matrix form \( A x = b \), where \( x \) denotes the column vector \((x_0, x_1, \ldots, x_{s-1})^T\), and \( b \) the column vector with all entries zero except for the first entry which is equal to \( s - \mu_A \). The matrix \( A \) is given by

\[
A = \begin{pmatrix}
  s & s - 1 & \ldots & 1 \\
  z_1^s - 1 & z_1^s - z_1 & \ldots & z_1^s - z_1^{s-1} \\
  z_2^s - 1 & z_2^s - z_2 & \ldots & z_2^s - z_2^{s-1} \\
  \vdots & \vdots & \ddots & \vdots \\
  z_{s-1}^s - 1 & z_{s-1}^s - z_{s-1} & \ldots & z_{s-1}^s - z_{s-1}^{s-1}
\end{pmatrix}. \tag{2.11}
\]

For this system of \( s \) equations to have a unique solution, all \( s \) equations should be linearly independent. Denote the determinant of a matrix \( C \) as \(|C|\). For the case that the roots \( z_0 = 1, z_1, \ldots, z_{s-1} \) are distinct Bailey [27] has shown that \(|A| = |V|\), where \( V \) is some Vandermonde matrix. In that case, \( A \) is non-singular and a unique solution \( x_0, x_1, \ldots, x_{s-1} \) exists. Using some additional arguments, we can derive explicit expressions for the \( x_j \) as given in the following lemma:

**Lemma 2.2.2** If the roots \( z_0 = 1, z_1, \ldots, z_{s-1} \) are distinct, the set of equations (2.9) together with the normalization condition (2.10) constitute a system of \( s \) linearly independent equations. The unique solution is given by

\[
x_j = (-1)^{j+2} (s - \mu_Y) \frac{S_{s-j} + S_{s-j-1}}{\prod_{k=1}^{s-1} (z_k - 1)}, \quad j = 0, 1, \ldots, s - 1, \tag{2.12}
\]

where \( S_j \) denotes the elementary symmetric function of degree \( j \), having as variables \( z_1, \ldots, z_{s-1} \), i.e.

\[
S_j = \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq s-1} z_{i_1} z_{i_2} \cdots z_{i_j}. \tag{2.13}
\]
Proof See Sec. 2.7. \qed

Remark 2.2.3 If one (or more) of the roots $z^s = A(z)$ in $|z| \leq 1$ has multiplicity higher than 1, an expression like (2.12) for the $x_j$ cannot be derived. However, for the pgf $X(z)$ to be finite on and inside the unit circle, the numerator of (2.8) should still have the same zeros as the denominator of (2.8), and with the same multiplicity. For $z_0 = 1$ it can be verified that this root has multiplicity 1, and we have argued before that this root places no restriction on the probabilities $x_0, \ldots, x_{s-1}$ whatsoever. For all other roots, the fact that the numerator of (2.8) should vanish does yield a restriction on $x_0, \ldots, x_{s-1}$. Assume, for example, that $z_1$ has multiplicity 2. Then $z_1$ should be a double root of the numerator of (2.8), yielding next to (2.9),

$$\sum_{j=0}^{s-1} x_j(sz_1^{s-1} - jz_1^{j-1}) = 0,$$

as an additional restriction on $x_0, \ldots, x_{s-1}$. In a similar way, whatever the multiplicity of the roots would be, we can construct $s - 1$ equations. Together with the normalization equation (2.10) this gives $s$ equations for $s$ unknowns. Since the Markov chain has a unique stationary distribution, we know that this system of equations has a unique solution.

So we can determine the probabilities $x_0, \ldots, x_{s-1}$ either explicitly through (2.12), or implicitly through a system of linear equations as described in Remark 2.2.3. From these probabilities, the entire probability distribution can be found. That is, from matching coefficients at both sides of

$$(z^s - A(z))X(z) = A(z) \sum_{j=0}^{s-1} x_j(z^s - z^j)$$

we find that

$$x_j = \frac{1}{a_0} \left( x_{j-s} - a_{j-s} \sum_{n=0}^{s-1} x_n - \sum_{n=0}^{j-s-1} x_{s+n}a_{j-s-n} \right), \quad j \geq s.$$

2.2.1 Roots on and inside the unit circle

We can go a step further and eliminate $x_0, \ldots, x_{s-1}$ from (2.8). Write

$$\sum_{j=0}^{s-1} x_j(z^s - z^j) = \gamma_1 (z - 1) \prod_{k=1}^{s-1} (z - z_k),$$

where the constant $\gamma_1$ can be determined from differentiating both sides of (2.17) with respect to $z$, and using the normalization condition (2.10). This gives

$$\gamma_1 = \frac{s - \mu_A}{\prod_{k=1}^{s-1} (1 - z_k)},$$
2.2 Generating function technique

and so

\[ \sum_{j=0}^{s-1} x_j (z^s - z^j) = (s - \mu_A) (z - 1) \prod_{k=1}^{s-1} \frac{z - z_k}{1 - z_k}. \]  

(2.19)

Together with (2.8) this yields the following result:

**Theorem 2.2.4** The pgf of the stationary queue length distribution is given by

\[ X(z) = \frac{A(z)(z - 1)(s - \mu_A)}{z^s - A(z)} \prod_{k=1}^{s-1} \frac{z - z_k}{1 - z_k}, \quad |z| \leq 1. \]  

(2.20)

Explicit expressions for the mean \( \mu_X \) and variance \( \sigma_X^2 \) of the stationary queue length can be obtained by evaluating derivatives of \( X(z) \) at \( z = 1 \), i.e.

\[ \mu_X = \frac{\sigma_X^2}{2(s - \mu_A)} + \frac{1}{2} \mu_A - \frac{1}{2} (s - 1) + \sum_{k=1}^{s-1} \frac{1}{1 - z_k}, \]  

(2.21)

\[ \sigma_X^2 = \sigma_A^2 + \frac{A''(1) - s(s - 1)(s - 2)}{3(s - \mu_A)} + \frac{A''(1) - s(s - 1)}{2(s - \mu_A)} \]

\[ + \left( \frac{A''(1) - s(s - 1)}{2(s - \mu_A)} \right)^2 - \sum_{k=1}^{s-1} \frac{z_k}{(1 - z_k)^2}. \]  

(2.22)

### 2.2.2 Roots outside the unit circle

When \( A \) has finite support, i.e. \( A \leq m \), we know that \( A(z) \) is a polynomial of degree \( m \). It then immediately follows that \( z^s = A(z) \) has \( m - s \) roots outside the unit circle, to be denoted by \( z_{s}, z_{s+1}, \ldots, z_{m-1} \), and so we can write (with \( m > s \))

\[ \sum_{j=0}^{s-1} x_j (z^s - z^j) = \frac{\gamma_2}{z^s - A(z)} \prod_{k=0}^{s-1} (z - z_k) \prod_{k=s}^{m-1} (z - z_k), \]  

(2.23)

where \( \gamma_2 \) is a constant. From the normalization condition \( X(1) = 1 \) it follows that \( \gamma_2 = \prod_{k=s}^{m-1} (1 - z_k) \), and so we arrive at

**Theorem 2.2.5** The pgf of the stationary queue length distribution is given by

\[ X(z) = A(z) \prod_{k=s}^{m-1} \frac{1 - z_k}{z - z_k}, \quad |z| \leq 1. \]  

(2.24)
From (2.24) we obtain (see e.g. Chaudhry & Kim [47], Zhao & Campbell [165])

$$
\mu_X = \mu_A + \sum_{k=s}^{m-1} \frac{1}{z_k - 1}, \quad (2.25)
$$

$$
\sigma_X^2 = \sigma_A^2 + \sum_{k=s}^{m-1} \frac{1}{(z_k - 1)^2} + \sum_{k=s}^{m-1} \frac{1}{z_k - 1}. \quad (2.26)
$$

Using partial-fraction expansion (see e.g. Henrici [90]), inverting $X(z)$ is a simple exercise. Write

$$
\prod_{k=s}^{m-1} \frac{1 - z_k}{z - z_k} = \sum_{i=s}^{m-1} \frac{r_i}{z - z_i}, \quad (2.27)
$$

where

$$
r_i = \lim_{z \to z_i} \frac{(1 - z_k)}{\prod_{k=s}^{m-1} (z_i - z_k)}, \quad i = s, \ldots, m - 1. \quad (2.28)
$$

Then rewrite the right-hand side of (2.27) as

$$
\sum_{i=s}^{m-1} \frac{r_i}{z - z_i} = \sum_{n=0}^{\infty} \sum_{i=s}^{m-1} \left( \frac{r_i}{z_i} \right) \left( \frac{1}{z_i} \right)^n z^n. \quad (2.29)
$$

Hence, the probability distribution $\{x_j\}_{j=0}^{\infty}$ is given by

$$
x_j = -\sum_{n=0}^{\infty} \sum_{i=s}^{m-1} \left( \frac{r_i}{z_i} \right) \left( \frac{1}{z_i} \right)^n a_n, \quad j = 0, 1, 2, \ldots. \quad (2.30)
$$

**Remark 2.2.6** For $j$ large enough, the sum on the right-hand side of (2.30) is dominated by the pole of $X(z)$ with the smallest modulus, to be denoted by $\hat{z}$. This pole can be shown to be the unique root of $z^s = A(z)$ contained in the interval $(1, \infty)$, see e.g. Tijms [153]. Omitting all fractions in (2.30) other than the one that corresponds to $\hat{z}$ gives the following approximation for the tail probabilities:

$$
x_j \approx -\sum_{n=0}^{\infty} \left( \frac{1}{\hat{z}} \right)^n a_n, \quad j \to \infty. \quad (2.31)
$$

### 2.3 Random walk theory

Most results from random walk theory that are important for queueing theory have been presented in the context of the waiting time of a customer in the $G/G/1$
2.3 Random walk theory

queue, see e.g. Asmussen [25], Chapter 10. We first show that the discrete bulk service queue may be viewed as a special type of \( G/G/1 \) queue. Then we invoke a result for the \( G/G/1 \) queue known as Spitzer’s identity that leads to an alternative expression for the pgf of the stationary queue length in the discrete bulk service queue.

The discrete bulk service queue is closely related to the discrete \( D/G/1 \) queue. The latter refers to a single server queue at which customers arrive with discrete and deterministic interarrival times, are served on a first-come-first-served basis and have service requirements that are i.i.d. according to a discrete random variable \( A \). The waiting time of the \( n \)th customer, denoted by \( W_n \), then satisfies (see e.g. Servi [142])

\[
W_{n+1} = (W_n + A_n - s)^+, \quad n = 0, 1, \ldots 
\]  
(2.32)

Here, \( A_n \) denotes the service time of customer \( n \) and the integer \( s \) denotes the interarrival time between two consecutive customers. When \( \mathbb{E}A < s \), the stationary waiting time denoted by \( W \) exists (see e.g. Servi [142]). By comparing (2.32) and (2.3), it is immediately clear that the stationary regimes of the discrete bulk service queue and the discrete \( D/G/1 \) queue are related as

\[
X(z) = A(z)W(z).
\]

Hence, a solution for \( W(z) \) yields the solution for \( X(z) \) and vice versa.

From the evolution equation (2.32) it can be seen that the distribution of \( W_{n+1} \) follows from the convolution of the distribution of \( W_n \) and that of \( A_n - s \), corrected for the \( \max\{0, \cdot\} \) operator. Hence, by iterating on (2.32) one can obtain transient characteristics of the model. This idea of iterating can be made more rigorous using random walk theory. When we set \( W_0 \) equal to zero, the following result is known as Spitzer’s identity (see Spitzer [146], p. 207):

**Theorem 2.3.1** (Spitzer’s identity) For \( 0 \leq t < 1 \) and \( |z| \leq 1 \),

\[
\sum_{n=0}^{\infty} t^n \mathbb{E}z^{W_n} = \exp \left\{ \sum_{l=1}^{\infty} t^{l-1} \mathbb{E}z^{S_l^+} \right\},
\]  
(2.33)

where \( S_l = \sum_{i=1}^{l} (A_i - s) \), \( A_i \) i.i.d. as \( A \).

From (2.33) the distribution of the stationary waiting time \( W \) can be obtained. When we write (2.33) as

\[
(1 - t) \sum_{n=0}^{\infty} t^n \mathbb{E}z^{W_n} = \exp \left\{ \sum_{l=1}^{\infty} t^{l-1}(\mathbb{E}z^{S_l^+} - 1) \right\},
\]  
(2.34)

it follows from Abel’s theorem (see Spitzer [146], p. 207, Cohen [51], p. 650), that \( W(z) \) is given by

\[
W(z) = \lim_{t \uparrow 1} (1 - t) \sum_{n=0}^{\infty} t^n \mathbb{E}z^{W_n} = \exp \left\{ \sum_{l=1}^{\infty} t^{l-1}(\mathbb{E}z^{S_l^+} - 1) \right\}.
\]  
(2.35)
Now we return to the discrete bulk service queue. The pgf of the stationary queue length is given by
\[ X(z) = A(z)W(z) \] with \( W(z) \) as in (2.35), which gives the following result.

**Theorem 2.3.2** The pgf of the stationary queue length distribution is given by
\[ X(z) = A(z) \exp \left\{ \sum_{l=1}^{\infty} \frac{1}{l} \mathbb{E}(z^{S^+_l} - 1) \right\}, \quad |z| \leq 1, \]
(2.36)

where \( S_l = \sum_{i=1}^{l} (A_i - s) \), \( A_i \) i.i.d. according to \( A \).

The mean and variance of the stationary queue length follow from taking derivatives of (2.35). Note that
\[ W'(1) = \sum_{l=1}^{\infty} \frac{1}{l} \mathbb{E}(S^+_l z^{S^+_l - 1}) W(z) \bigg|_{z=1} = \sum_{l=1}^{\infty} \frac{1}{l} \mathbb{E}(S^+_l), \]
(2.37)
\[ W''(1) = \sum_{l=1}^{\infty} \sum_{k=1}^{\infty} \frac{1}{l} \mathbb{E}(S^+_l) \frac{1}{k} \mathbb{E}(S^+_k) + \sum_{l=1}^{\infty} \frac{1}{l} \mathbb{E}(S^+_l (S^+_l - 1)). \]
(2.38)

Denoting by \( A^{*l} \) the random variable that follows the \( l \)-fold convolution of the distribution of \( A \), this gives after some rewriting
\[ \mu_X = \mu_A + \sum_{l=1}^{\infty} \frac{1}{l} \sum_{j=ls}^{\infty} (j - ls) \mathbb{P}(A^{*l} = j), \]
(2.39)
\[ \sigma^2_X = \sigma^2_A + \sum_{l=1}^{\infty} \frac{1}{l} \sum_{j=ls}^{\infty} (j - ls)^2 \mathbb{P}(A^{*l} = j), \]
(2.40)
which are root-free expressions for \( \mu_X \) and \( \sigma^2_X \), and alternative expressions for (2.21)-(2.22) and (2.25)-(2.26).

Moreover, introducing the short-hand notation \( C_{z^j} [f(z)] \) for the coefficient of \( z^j \) in \( f(z) \), the following result follows from (2.36):

**Lemma 2.3.3** The stationary queue length distribution is given by (for \( j = 0, 1, \ldots \))
\[ x_j = \mathbb{P}(W = 0) \sum_{k=0}^{l} a_k C_{z^{j-l}} \left[ \exp \left\{ \sum_{l=1}^{\infty} \sum_{i=1}^{\infty} \frac{1}{l} \mathbb{P}(A^{*l} = ls + i) z^i \right\} \right], \]
(2.41)
and
\[ \mathbb{P}(W = 0) = \exp \left\{ - \sum_{l=1}^{\infty} \sum_{i=ls+1}^{\infty} \frac{1}{l} \mathbb{P}(A^{*l} = i) \right\}. \]
(2.42)

Expression (2.41) provides for each \( x_j \) a root-free representation, as an alternative for (2.12), (2.16) and (2.30) that do depend on the roots of \( z^s = A(z) \). For determining the coefficients \( C_{z^j} \) in (2.41), the following property can be used:
Property 2.3.4 For \( K(z) = \sum_{j=0}^{\infty} k_j z^j \) and \( M(z) = \sum_{j=0}^{\infty} m_j z^j \) with \( K(z) = \exp\{M(z)\} \), the coefficients \( k_j \) follow recursively from the coefficients \( m_j \) (and vice versa) according to

\[
k_0 = \exp(m_0); \quad k_j = \frac{1}{j} \sum_{n=1}^{j} nm_n k_{j-n}, \quad j = 1, 2, \ldots
\]  

(2.43)

Proof The proof consists of computing the \( k_j \)'s successively by equating coefficients in \( K'(z) = M'(z)K(z) \). 

\[\square\]

Remark 2.3.5 Several authors (Konheim [105], Murata & Miyahara [120], Stadje [147]) have suggested to approximate the \( G/G/1 \) queue by its discrete counterpart. This can be done as follows. Denote by \( B_n \) the service time of customer \( n \) and by \( C_n \) the interarrival time between customer \( n \) and \( n+1 \). Choose \( B_n \) and \( C_n \) i.i.d. according to discrete random variables \( B \) and \( C \), respectively. Moreover, assume \( C \leq s \). Then \( W_n \) satisfies

\[
W_{n+1} = (W_n + B_n - C_n)^+ = (W_n + A_n - s)^+, \quad n = 0, 1, \ldots
\]  

(2.44)

with \( A_n \) assumed i.i.d. as \( A = B - C + s \). The discrete approximation to the \( G/G/1 \) queue then fits into the framework of the \( D/G/1 \) queue.

2.4 Wiener-Hopf technique

The Wiener-Hopf technique stems from mathematical physics, and found its way to the field of applied probability through Smith [145], Kemperman [97] and Cohen [50, 51] (also see Regterschot [138]). Perhaps the most famous application of the Wiener-Hopf technique is in the context of random walks, see e.g. Cohen [51] or Asmussen [25], due to the fact that the Wiener-Hopf technique is a powerful tool for the analysis of Markov processes whose evolution equation contains the \( \max\{0, \cdot\} \) operator, as in the case of the discrete bulk service queue. We will apply the Wiener-Hopf technique to obtain alternative derivations of the expressions for \( X(z) \) given by (2.20), (2.24) and (2.36). In addition to this application, the Wiener-Hopf technique is also frequently applied in the analysis of the trajectories of random walks (see e.g. Asmussen [25]).

Let us first describe the role of the \( \max\{0, \cdot\} \) operator. From recursion (2.3) we have

\[
\mathbb{E}(z^{X_{n+1}}) = \mathbb{E}(z^{A_n}1\{X_n \leq s\}) + \mathbb{E}(z^{X_n+A_n-s}1\{X_n > s\})
\]

\[
= \mathbb{P}(X_n \leq s)\mathbb{E}(z^{A_n}) + \mathbb{E}(z^{X_n+A_n-s}) - \mathbb{E}(z^{X_n+A_n-s}1\{X_n \leq s\}),
\]  

(2.45)

where \( 1\{x\} = 1 \) if \( x \) is true and 0 otherwise. Letting \( n \to \infty \) and observing that \( X_n \) and \( A_n \) are independent then yields

\[
\xi_+(z)(1 - z^{-s}A(z)) = \xi_-(z),
\]  

(2.46)
where $\xi_+(z) = X(z)/A(z)$ and $\xi_-(z) = P(X \leq s) - E(z^{X-s}1\{X \leq s\})$. Observe that $\xi_+$ (respectively $\xi_-$) is analytic and bounded in $|z| < 1$ (respectively $|z| > 1$), and both $\xi_+, \xi_-$ are continuous up to $|z| = 1$.

In order to find an explicit expression for $\xi_+(z)$ we need to factorize the function $1 - z^{-s}A(z)$. In more general terms, we need to factorize a function $1 - Y(z)$, where $Y(z)$ is the pgf of a random variable $Y$ for which it holds that $EY < 0$ (in the case of the discrete bulk service queue $Y = A - s$, i.e. $Y(z) = z^{-s}A(z)$). Such a factorization is known as the Wiener-Hopf factorization. The treatment of this factorization in terms of a characteristic function prevails in the literature, but we present the theory here for $Y(z)$ being a pgf (Bayer [28] does this also). Furthermore, it is common practice to present a factorization of the bivariate function $1 - rY(z)$, $0 \leq r < 1$, but since we are interested in the stationary distribution only (and not in the transient distribution), we will stick to the analysis of the univariate function $1 - Y(z)$. The Wiener-Hopf factorization identity then reads (see Asmussen [25], p. 228, Prabhu [135], p. 22):

**Theorem 2.4.1** (Wiener-Hopf factorization identity) *The following decomposition exists:*

$$1 - Y(z) = \phi_+(z)\phi_-(z), \quad |z| = 1,$$

*(2.47)*

where $\phi_+$ (respectively $\phi_-$) is analytic and bounded in $|z| < 1$ (respectively $|z| > 1$), and both $\phi_+, \phi_-$ are continuous up to $|z| = 1$.

Hence, once we know the functions $\phi_+, \phi_-$ we can write (2.46) as

$$\xi_+(z)\phi_+(z) = \frac{\xi_-(z)}{\phi_-(z)},$$

*(2.48)*

where the left-hand side (respectively right-hand side) of (2.48) represents a function that is analytic and bounded in $|z| < 1$ (respectively $|z| > 1$), and both sides of (2.48) are functions continuous up to $|z| = 1$. Therefore, their analytic continuation contains no singularities in the entire complex plane. Liouville’s theorem then says

**Theorem 2.4.2** (Liouville) *Let $f(z)$ be analytic for all values of $z$ and let $|f(z)| < K$ for all values of $z$, where $K$ is a constant (so that $|f(z)|$ is bounded as $|z| \to \infty$). Then $f(z)$ is seen to be constant.*

Whence upon using Liouville’s theorem the left-hand side of (2.48) is constant, and since $\xi_+(1) = 1$, we obtain

$$\xi_+(z) = \frac{\phi_+(1)}{\phi_+(z)}.$$  

*(2.49)*

With the machinery described above, we present alternative proofs of Thms. 2.2.4, 2.2.5 and 2.3.2, where we rely on three different factorizations of the function $1 - Y(z)$.

**Alternative proof of Thm. 2.3.2.** Start from the basic identity

$$1 - z = \exp\{\ln(1 - z)\} = \exp\left\{-\sum_{l=1}^{\infty} \frac{1}{l} z^l \right\}, \quad |z| \leq 1, \quad z \neq 1.$$  

*(2.50)*
We have denoted \( \sum_{i=1}^{L} (A_i - s) \) by \( S_l \) for which it holds that \( \mathbb{E}(z^{S_l}) = (z^{-s}A(z))^l \) and \( |z^{-s}A(z)| < 1 \) for \( |z| = 1 \). Hence, we can write (for \( |z| = 1 \))

\[
1 - z^{-s}A(z) = \exp \left\{ - \sum_{l=1}^{\infty} \frac{1}{l} (z^{-s}A(z))^l \right\} = \phi_+(z)\phi_-(z),
\]

where

\[
\phi_+(z) = \exp \left\{ - \sum_{l=1}^{\infty} \frac{1}{l} \mathbb{E}(z^{S_l} \mathbf{1}\{S_l > 0\}) \right\},
\]

\[
\phi_-(z) = \exp \left\{ - \sum_{l=1}^{\infty} \frac{1}{l} \mathbb{E}(z^{S_l} \mathbf{1}\{S_l \leq 0\}) \right\}.
\]

Observe that

\[
\phi_+(1) = \lim_{z \to 1} \frac{z^s - A(z)}{(z-1)\prod_{k=1}^{s-1}(z-z_k)} = \frac{s - \mu_A}{\prod_{k=1}^{s-1}(1-z_k)},
\]

which by (2.49) completes the proof. \( \square \)

The type of Wiener-Hopf factorization as outlined above can be applied for the more general \( G/G/1 \) queue in a similar fashion. For a subclass of queues, the Wiener-Hopf technique allows for an explicit factorization that relies on the consideration of the roots of some equation (see e.g. Asmussen [25], Cohen [51], Kleinrock [101]).

For the discrete bulk service queue, the knowledge on the roots of \( z^s = A(z) \) given by Thm. 2.2.1 can be applied to prove the following two previously derived results:

**Alternative proof of Thm. 2.2.4.** We construct an explicit factorization of \( 1 - z^{-s}A(z) \) by choosing

\[
\phi_+(z) = \frac{z^s - A(z)}{\prod_{k=0}^{s-1}(z-z_k)}, \quad \phi_-(z) = \prod_{k=0}^{s-1} \left( \frac{1}{z^s} \right).
\]

With

\[
\phi_+(1) = \lim_{z \to 1} \frac{z^s - A(z)}{(z-1)\prod_{k=1}^{s-1}(z-z_k)} = \frac{s - \mu_A}{\prod_{k=1}^{s-1}(1-z_k)},
\]

this completes the proof. \( \square \)

**Alternative proof of Thm. 2.2.5.** In case \( A \leq m \), we construct an explicit factorization of \( 1 - z^{-s}A(z) \) by choosing

\[
\phi_+(z) = \gamma \prod_{k=s}^{m-1} (z-z_k), \quad \phi_-(z) = \prod_{k=0}^{s-1} \left( \frac{1}{z^s} \right),
\]

with \( \gamma \) a constant. We have that

\[
\phi_+(1) = \gamma \prod_{k=s}^{m-1} (1-z_k),
\]

which completes the proof. \( \square \)
2.5 Our contribution

The generating function technique is widely applied in the discrete-time modeling of communication systems. The required procedures of numerical root-finding are known to a broad community of engineers, including electrical engineers and computer scientists. On the contrary, the Wiener-Hopf technique and random walk theory, while often of great interest to theoreticians, are less popular among practitioners, probably due to the advanced mathematical techniques involved and the lack of clearly described computational schemes for determining certain performance characteristics. For the discrete bulk service queue, this separation between theoretical and practical results is far less clear-cut. The results obtained by all three methods in fact might complement each other well, since clear descriptions of the computational schemes are available.

Depending on the method used, one can obtain a transform solution of the stationary queue length distribution either in terms of the roots of $z^* A(z)$ or in terms of infinite series that involve convolutions of the distribution of $A$. The solution in terms of roots can be obtained using either the generating function technique or the Wiener-Hopf technique. The solution in terms of the infinite series follows from random walk theory (Spitzer’s identity), and can also be obtained using the Wiener-Hopf technique. All four options have been demonstrated in this chapter. To make a rough distinction, two courses can be followed in obtaining characteristics of the stationary queue length distribution of the discrete bulk service queue: roots or infinite series. That is, for the mean, variance and probability distribution we have the expressions in terms of roots and the expressions in terms of infinite series. From a practical viewpoint, both courses have their difficulties: Roots need to be determined and infinite series need to be truncated.

In Fig. 2.1 we give an overview of the methods that can be applied to deal with the discrete bulk service queue, where we distinguish between the existing methods described in the previous sections, and the contributions made in this thesis.

First of all, in Chapter 3, we will derive the expressions in terms of infinite series from the expressions in terms of roots. For that, we first make the observation that the roots can be considered to be located on a specific curve, referred to as generalized Szegő curve. By adopting a special parametrization of this Szegő curve, the relevant roots occur as equidistant samples of a $2\pi$-periodic function whose Fourier coefficients can be determined analytically. This makes that we can write the expressions in terms of roots as Fourier aliasing series with terms given in analytic form, eventually leading to expressions in terms of infinite series. We demonstrate this procedure for the mean and variance of the queue length, as well as for the entire queue length distribution.

In Chapter 4 we present for a large class of distributions analytic representations of the roots, based on the Fourier series representation presented in Chapter 3. We show that the resulting computational scheme is easy to implement and numerically stable. We also discuss a second method to determine the roots based on applying successive substitutions to a fixed-point equation. We outline under which condi-
2.5 Our contribution

The numbers refer to the chapters in which the topic is addressed, where * means that this is our contribution.

In Chapter 5 we further investigate the infinite series that involve convolutions of the probability distribution of $A$. For practical purposes, the infinite series should be truncated, and we therefore seek for some means to characterize the speed with which these series converge. Such a characterization is related to the notion of relaxation time in queueing theory, a generic term for the time required for transient system characteristics to tend to their stationary values. We derive an asymptotic expression for the relaxation time in a purely analytical way, mostly relying on the saddle point method. We present a simple and useful upper bound which may serve as a stopping criterion for the level at which one can truncate the infinite series in case the load on the system is not very high. A more detailed upper bound is then developed that continues to be sharp for high loads.

Finally, in Chapter 6 we present sharp bounds for the mean and variance of the stationary queue length. The bounds are in closed form and hold for a general arrival process. Considerable attention is paid to the case of Poisson arrivals, for which we prove additional properties of the expressions that involve roots, leading to even sharper bounds. The Poisson case serves as a pilot study for a broader range of distributions. We present many numerical examples and compare the new bounds with existing ones in the literature.
2.6 Proof of Theorem 2.2.1

This section is based on Adan et al. [P12]. Rouché’s theorem is the standard tool for proving Thm. 2.2.1, since it is typically used to determine regions of the complex plane in which there may be zeros of a given analytic function. We focus on the zeros of the function $z^s - A(z)$ (i.e. the roots of $z^s = A(z)$) on or within the unit circle.

Rouché’s theorem is a direct consequence of the argument principle and the scope of application of Rouché’s theorem goes well beyond the field of queueing theory. While the verification of the conditions needed to apply Rouché’s theorem can become rather difficult, in queueing theory this is usually straightforward. For most queueing applications, the region of interest is typically the unit disk, and the ingredient that makes Rouché’s theorem work is oftentimes the stability condition. This is why Rouché’s theorem is a popular and standardized tool in queueing theory.

However, in order to apply Rouché’s theorem it is required that $A(z)$ has a radius of convergence larger than 1 (see Lemma 2.6.2), which is not true in general. A pgf obeys all the rules of power series with non-negative coefficients, and since $A(1) = 1$ the radius of convergence of a pgf is at least 1. The shoe thus pinches for those pgf’s for which the radius of convergence is exactly 1 (some examples are given at the end of Subsec. 2.6.2).

In papers like Bruneel [42], Darroch [59], Powell & Humblet [134], Servi [142] and Zhao & Campbell [165], the assumption is made that $A(z)$ has a radius of convergence larger than 1, so that Rouché’s theorem can be applied. In this chapter, we impose no such restriction on $A(z)$. Instead of excluding those functions $A(z)$ with radius of convergence 1, we present a proof of Thm. 2.2.1 that does not rely on Rouché’s theorem and holds for general $A(z)$.

Several other authors proved such a generalization. Abolnikov & Dukhovny [19] apply the so-called generalized principle of the argument (that was proved by Gakhov et al. [79] in 1973) to give a proof for general $A(z)$. Klimenok [103] extended this result to a larger class of functions (so including $z^s - A(z)$), again using the generalized principle of the argument. An alternative approach to deal with general $A(z)$ was presented by Boudreau et al. [37]. Under the condition that all zeros in the unit disk are distinct, they were able to apply the implicit function theorem to prove the existence of the zeros. However, examples can be constructed for which there are multiple zeros, and so this approach does not cover the issue in full generality. The key idea of Boudreau et al. is to study the parameterized function $z^s - tA(z)$, $0 \leq t < 1$, and then letting $t$ tend to one. The same idea, without making the assumption of distinct zeros, has been used by Gail et al. [78] for a larger class of zeros, including $z^s - A(z)$. We present an elementary proof of the existence of the zeros for general $A(z)$ using the classical argument principle and truncation of $A(z)$.

In Subsec. 2.6.1 we first describe the classical application of Rouché’s theorem. In Subsec. 2.6.2 we give our proof for general $A(z)$. 
2.6 Proof of Theorem 2.2.1

2.6.1 Classical setting

Let us first state Rouché’s theorem (see e.g. Titchmarsh [154]):

**Theorem 2.6.1** (Rouché) Let the bounded region $D$ have as its boundary a simple closed contour $C$. Let $f(z)$ and $g(z)$ be analytic both in $D$ and on $C$. Assume that $|f(z)| < |g(z)|$ on $C$. Then $f(z) - g(z)$ has in $D$ the same number of zeros as $g(z)$, all zeros counted according to their multiplicity.

When $A(z)$ has a radius of convergence larger than one, we can prove the following result concerning the number of zeros on and within the unit circle of $z^s - A(z)$ by using Rouché’s theorem:

**Lemma 2.6.2** Let $A(z)$ be a pgf that is analytic in $|z| \leq 1 + \nu$, $\nu > 0$. Assume that $A'(1) < s$, $s \in \mathbb{N}$. Then the function $z^s - A(z)$ has exactly $s$ zeros in $|z| \leq 1$.

**Proof** Define the functions $f(z) := A(z)$, $g(z) := z^s$. Because $f(1) = g(1)$ and $f'(1) = A'(1) < s = g'(1)$, we have, for sufficiently small $\epsilon > 0$,

$$f(1 + \epsilon) = f(1 + \epsilon) < g(1 + \epsilon). \quad (2.59)$$

Consider all $z$ with $|z| = 1 + \epsilon$. By the triangle inequality and (2.59) we have that

$$|f(z)| = \sum_{j=0}^{\infty} a_j |z|^j = f(1 + \epsilon) < g(1 + \epsilon) = |g(z)|, \quad (2.60)$$

and hence $|f(z)| < |g(z)|$. Because both $f(z)$ and $g(z)$ are analytic for $|z| \leq 1 + \epsilon$, Rouché’s theorem tells us that $g(z)$ and $f(z) - g(z)$ have the same number of zeros in $|z| \leq 1 + \epsilon$. Letting $\epsilon$ tend to zero yields the proof. \qed

2.6.2 New setting

Before we present our main result, we first prove a result on the number and location of zeros of $z^s - A(z)$ on the unit circle. We define the period $p$ of a series $\sum_{j=-\infty}^{\infty} b_j z^j$ as the largest integer for which $b_j = 0$ whenever $j$ is not divisible by $p$.

**Lemma 2.6.3** Let $A(z)$ be a pgf of some nonnegative discrete random variable with $A(0) > 0$. Assume $A(z)$ is differentiable at $z = 1$ and $A'(1) < s$, where $s$ is a positive integer. If $z^s - A(z)$ has period $p$, then $z^s - A(z)$ has exactly $p$ zeros on the unit circle given by the $p$-th roots of unity $\tau_k = \exp(2\pi ik/p)$, $k = 0, 1, \ldots, p - 1$. In each of these zeros, the derivative of $z^s - A(z)$ does not vanish.

**Proof** Obviously, any zero $\xi$ of $z^s - A(z)$ with $|\xi| = 1$ is simple, since $|A'(\xi)| \leq A'(|\xi|) = A'(1) < s$ and, thus, $s\xi - A'(\xi) \neq 0$. Furthermore, for any $z$ with $|z| = 1$, $|A(z)| = A(1)$ iff $z^k = 1$ whenever $a_k > 0$. This easily follows from the fact that $|a_0 + a_k z^k| < a_0 + a_k$ if $z^k \neq 1$. So, for $z$ with $|z| = 1$ and $A(z) - z^s = 0$ it
follows that $z^k = 1$ for all $k$ with $a_k > 0$, and $z^s = 1$. This implies that $z^p = 1$, which completes the proof. □

Note that the requirement $a_0 = A(0) > 0$ involves no essential limitation: If $a_0$ were zero we would replace the distribution $\{a_i\}_{i\geq 0}$ by $\{a'_i\}_{i\geq 0}$ where $a'_i = a_{i+m}$, $a_m$ being the first non-zero entry of $\{a_i\}_{i\geq 0}$, and a corresponding decrease in $s$ according to $s' = s - m$.

We are now in a position to give the main result:

**Theorem 2.6.4** Let $A(z)$ be a pgf of some nonnegative discrete random variable with $A(0) > 0$. Assume $A(z)$ is differentiable at $z = 1$ and $A'(1) < s$, where $s$ is a positive integer. Also, let $z^s - A(z)$ have period $p$. Then the function $z^s - A(z)$ has $p$ zeros on the unit circle given by $\tau_k = \exp(2\pi ik/p)$, $k = 0, 1, \ldots, p - 1$ and exactly $s - p$ zeros in $|z| < 1$.

**Proof** Lemma 2.6.3 tells us that $F(z) = z^s - A(z)$ has $p$ equidistant zeros on the unit circle, and so it remains to prove that this function has exactly $s - p$ zeros within the unit circle. Thereto, define, for $N \in \mathbb{N}$, the truncated pgf

$$A_N(z) = \sum_{j=0}^{N-1} a_j z^j + \sum_{j=N}^{\infty} a_j z^N,$$

(2.61)

where $N$ is a multiple of $p$. Then $F_N(z) = z^s - A_N(z)$ has obviously $s$ zeros in $z \in D = \{z \in \mathbb{C} : |z| \leq 1\}$, since $A_N(z)$ is a polynomial satisfying $A'_N(1) < s$, and Lemma 2.6.2 thus applies. By Lemma 2.6.3 we know that $F_N(z)$ has $p$ simple and equidistant zeros on the unit circle. We further have that

$$|A(z) - A_N(z)| \leq 2 \sum_{j=N}^{\infty} a_j, \quad |z| \leq 1,$$

(2.62)

and

$$|A'(z) - A'_N(z)| \leq 2 \sum_{j=N}^{\infty} j a_j, \quad |z| \leq 1.$$

(2.63)

Thus, $A_N(z)$ and $A'_N(z)$ converge uniformly to $A(z)$ and $A'(z)$ on $z \in D$, respectively. Moreover, if $G : D \to \mathbb{C}$ is continuous, then $G(A_N(z))$ is uniformly convergent to $G(A(z))$ on $z \in D$.

Let $z$ on $C = \{z \in \mathbb{C} : |z| = 1\}$. If for all $n \in \mathbb{N}$ there is a $z_n \in D$ with $0 < |z - z_n| < 1/n$ and $F(z_n) = 0$, then $F(z) = 0$ and

$$F'(z) = \lim_{n \to \infty} \frac{F(z_n) - F(z)}{z_n - z} = 0.$$

(2.64)

However, this is impossible by Lemma 2.6.3. Hence, there is an $\eta > 0$ such that $F(\xi) \neq 0$ for all $\xi \in D(z, \eta) := \{\xi \in D : 0 < |\xi - z| < \eta\}$. Since $C$ is compact, it can be covered by finitely many $D(z, \eta)$'s. Hence, there is a $0 < r < 1$ such that $F(z)$ has no zeros in $r \leq |z| < 1$. 

Historical perspective and methodology
2.6 Proof of Theorem 2.2.1

Now we prove that for large $N$ the function $F_N(z)$, as the function $F(z)$, has no zeros in $r \leq |z| < 1$. Thereto, we show that there is an $\epsilon > 0$ and $M \in \mathbb{N}$ such that $F_N(z) \neq 0$ for all $N \geq M$ and $0 < |z - \tau_k| < \epsilon$, $k = 0, 1, \ldots, p - 1$. Because $F'(z)$ is continuous and $F'_N(z)$ converges uniformly to $F'(z)$ on $z \in D$, there are $\epsilon > 0$ and $M \in \mathbb{N}$ such that (for $k = 0, 1, \ldots, p - 1$)

\[ |F'_N(z) - F'(#2c1)| < \delta < |F'(\tau_k)|, \quad 0 < |z - \tau_k| < \epsilon, \quad N \geq M. \quad (2.65) \]

Furthermore, we have (for $k = 0, 1, \ldots, p - 1$)

\[ |F_N(z) - F'(\tau_k)(z - \tau_k)| = \left| \int_{[\tau_k, z]} (F'_N(s) - F'(\tau_k))ds \right|. \quad (2.66) \]

where the integration is carried out along the straight line that connects $\tau_k$ and $z$. Hence, for $0 < |z - \tau_k| < \epsilon$ and $N \geq M$, we obtain (for $k = 0, 1, \ldots, p - 1$)

\[ \left| \int_{[\tau_k, z]} (F'_N(s) - F'(\tau_k))ds \right| \leq |z - \tau_k| \max_{s \in [\tau_k, z]} |F'_N(s) - F'(\tau_k)| < |z - \tau_k| \delta. \quad (2.67) \]

So, it follows that for $0 < |z - \tau_k| < \epsilon$ and $N \geq M$ (for $k = 0, 1, \ldots, p - 1$)

\[ |F_N(z)| = |F_N(z) - F'(\tau_k)(z - \tau_k) + F'(\tau_k)(z - \tau_k)| \quad (2.68) \]
\[ \geq |F'(\tau_k)||z - \tau_k| - |F_N(z) - F'(\tau_k)(z - \tau_k)| \quad (2.69) \]
\[ > (|F'(\tau_k)| - \delta)|z - \tau_k| > 0. \quad (2.70) \]

Since $F_N(z)$ converges uniformly to $F(z)$ and $F(z) \neq 0$ on the compact set (see Fig. 2.2)

\[ E = \{ z \in \mathbb{C} : r \leq |z| \leq 1 \} \setminus \bigcup_{k=0}^{p-1} D(\tau_k, \epsilon), \quad (2.71) \]
there exists an $K \in \mathbb{N}$ such that $F_N(z) \neq 0$ for all $N \geq K$ and $z \in \mathbb{C}$ with $r \leq |z| < 1$. Hence, for all $N \geq K$ the number of zeros of $F_N(z)$ with $|z| < r$ is equal to $s - p$. This number can be expressed by the argument principle (see e.g. Titchmarsh [154]) as follows

$$s - p = \frac{1}{2\pi i} \oint_{|z|=r} \frac{F'_N(z)}{F_N(z)} dz.$$  \hspace{2cm} (2.72)

The integrand converges uniformly to $F'(z)/F(z)$, and thus

$$\frac{1}{2\pi i} \oint_{|z|=r} \frac{F'(z)}{F(z)} dz = \lim_{N \to \infty} \frac{1}{2\pi i} \oint_{|z|=r} \frac{F'_N(z)}{F_N(z)} dz = s - p.$$  \hspace{2cm} (2.73)

Hence, the number of zeros of $F(z)$ with $|z| < r$ is also $s - p$. This completes the proof. □

Obviously, Thm. 2.6.4 proves Thm. 2.2.1. Due to Thm. 2.6.4, the $A(z)$ with a radius of convergence of 1 do not have to be excluded from the analysis of the zeros of $z^a - A(z)$. This further means that these pgf’s can be incorporated in the general formulation of the solution to the queueing models of interest. The $A(z)$ that have radius of convergence 1 are typically those associated with heavy-tailed random variables. Some examples are given below.

(i) The discrete Pareto distribution (e.g. Johnson et al. [94]), defined by

$$a_j = c \frac{1}{j^{p+1}}, \quad j = 1, 2, \ldots,$$  \hspace{2cm} (2.74)

with

$$c = \left( \sum_{j=1}^{\infty} a_j \right)^{-1} = \zeta(p+1)^{-1},$$  \hspace{2cm} (2.75)

where $\zeta(\cdot)$ is called the Riemann zeta function and $p > 1$. For $k < p$, the $k$th moment $\mu_k$ of the discrete Pareto distribution is given by

$$\mu_k = \frac{\zeta(p - k + 1)}{\zeta(p + 1)},$$  \hspace{2cm} (2.76)

whereas for $k \geq p$ the moments are infinite. The discrete Pareto distribution is also known as the Zipf or Riemann zeta distribution.

(ii) The discrete standard lognormal distribution, defined by

$$a_j = ce^{-\frac{(\log j)^2}{2}}, \quad j = 1, 2, \ldots,$$  \hspace{2cm} (2.77)

where $c$ is a normalization constant.

(iii) The discrete distribution, related to the continuous Weibull distribution, defined by

$$a_j = cp^{-\sqrt{j}}, \quad j = 0, 1, \ldots,$$  \hspace{2cm} (2.78)

where $p > 1$ and $c$ is a normalization constant.
2.7 Proof of Lemma 2.2.2

(iv) The Haight’s zeta distribution (see e.g. Johnson et al. [94]), defined by (with \( p > 1 \))

\[
a_j = \frac{1}{(2j-1)^p} - \frac{1}{(2j+1)^p}, \quad j = 1, 2, \ldots \quad (2.79)
\]

2.7 Proof of Lemma 2.2.2

By Cramer’s rule we have that \( x_j = |A_{j+1}|/|A|, j = 0, 1, \ldots, s - 1 \), where \( A_{j+1} \) is the matrix \( A \) except for the \( j + 1 \)-st column being replaced by \( b \). Since \( |A| = |A^T| \), we find

\[
|A| = \begin{vmatrix}
  s & z_1^s - 1 & \ldots & z_{s-1}^s - 1 \\
  s - 1 & z_1^s - z_1 & \ldots & z_{s-1}^s - z_{s-1} \\
  \vdots & \vdots & \ddots & \vdots \\
  1 & z_1^{s-1} - z_{s-1} & \ldots & z_{s-1}^{s-1} - z_{s-1}^s \\
\end{vmatrix}
\]

= \begin{vmatrix}
  1 & z_1 - 1 & \ldots & z_{s-1} - 1 \\
  1 & z_1(z_1 - 1) & \ldots & z_{s-1}(z_{s-1} - 1) \\
  \vdots & \vdots & \ddots & \vdots \\
  1 & z_1^{s-1}(z_1 - 1) & \ldots & z_{s-1}^{s-1}(z_{s-1} - 1) \\
\end{vmatrix},
\]

where the last equality follows by subtracting row \( r + 1 \) from row \( r \) for each \( r = 1, 2, \ldots, s - 1 \). Dividing each column \( k + 1 \) by \( z_k - 1 \) for \( k = 1, \ldots, s - 1 \) yields the following result

\[
|A| = \prod_{k=1}^{s-1} (z_k - 1) \begin{vmatrix}
  1 & 1 & \ldots & 1 \\
  1 & z_1 & \ldots & z_{s-1} \\
  \vdots & \vdots & \ddots & \vdots \\
  1 & z_1^{s-1} & \ldots & z_{s-1}^{s-1} \\
\end{vmatrix}
= \prod_{k=1}^{s-1} (z_k - 1) \prod_{0 \leq n < k \leq s-1} (z_k - z_n), \quad (2.80)
\]

where the last equality follows from the special form of the determinant of a Vandermonde matrix (see e.g. Bellman [29]).

To compute the determinant of \( A_{j+1} \) we expand this matrix on its \( j + 1 \)-st column, which gives

\[
|A_{j+1}| = (-1)^{j+2}(s - \mu_{ij})|B|, \quad (2.81)
\]

where

\[
|B| = \begin{vmatrix}
  s & \ldots & s - j - 1 & s - j + 1 & \ldots & 1 \\
  z_1^s - 1 & \ldots & z_1^{j-1} & z_1^j - z_1^{j+1} & \ldots & z_1^{s-1} - z_1^{s-1} \\
  \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
  z_{s-1}^{s-1} - 1 & \ldots & z_{s-1}^{j-1} & z_{s-1}^j - z_{s-1}^{j+1} & \ldots & z_{s-1}^{s-1} - z_{s-1}^{s-1} \\
\end{vmatrix}. \quad (2.82)
\]
We then transpose the matrix $B$, and subtract column $k+1$ from column $k$ to obtain

$$|B| = |B^T| = egin{vmatrix} z_1 - 1 & z_2 - 1 & \cdots & z_{s-1} - 1 \\ z_1(z_1 - 1) & z_2(z_2 - 1) & \cdots & z_{s-1}(z_{s-1} - 1) \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{s-2}(z_1 - 1) & z_2^{s-2}(z_2 - 1) & \cdots & z_{s-1}^{s-2}(z_{s-1} - 1) \\ z_1^{s-1}(z_1 - 1) & z_2^{s-1}(z_2 - 1) & \cdots & z_{s-1}^{s-1}(z_{s-1} - 1) \end{vmatrix}. \quad (2.83)$$

Dividing each column $k+1$ by $z_k - 1$ for $k = 1, \ldots, s - 1$ then yields

$$|B| = \prod_{k=1}^{s-1} (z_k - 1) \begin{vmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_{s-1} \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{j-2} & z_2^{j-2} & \cdots & z_{s-1}^{j-2} \\ z_1^{j-1} & z_2^{j-1} & \cdots & z_{s-1}^{j-1} \\ z_1^{j} & z_2^{j} & \cdots & z_{s-1}^{j} \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{s-2} & z_2^{s-2} & \cdots & z_{s-1}^{s-2} \\ z_1^{s-1} & z_2^{s-1} & \cdots & z_{s-1}^{s-1} \end{vmatrix}. \quad (2.84)$$

Since the determinant is a linear operator we then obtain

$$|B| = \prod_{k=1}^{s-1} (z_k - 1) \begin{vmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_{s-1} \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{j-2} & z_2^{j-2} & \cdots & z_{s-1}^{j-2} \\ z_1^{j-1} & z_2^{j-1} & \cdots & z_{s-1}^{j-1} \\ z_1^{j} & z_2^{j} & \cdots & z_{s-1}^{j} \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{s-2} & z_2^{s-2} & \cdots & z_{s-1}^{s-2} \\ z_1^{s-1} & z_2^{s-1} & \cdots & z_{s-1}^{s-1} \end{vmatrix} + \prod_{k=1}^{s-1} (z_k - z_n)S_{s-j} + \prod_{1 \leq n < k \leq s-1} (z_k - z_n)S_{s-j-1}. \quad (2.85)$$

The matrices at the right-hand side of (2.85) are very similar to Vandermonde matrices, except that one row has been deleted. As for Vandermonde matrices, the determinants of such matrices have a nice form (see Pólya & Szegő [132], Exercise 10, p. 93, and Neagoe [123]):

$$|B| = \prod_{k=1}^{s-1} (z_k - 1) \left( \prod_{1 \leq n < k \leq s-1} (z_k - z_n)S_{s-j} + \prod_{1 \leq n < k \leq s-1} (z_k - z_n)S_{s-j-1} \right), \quad (2.86)$$
where the functions $S_{s-j}$ and $S_{s-j-1}$ are defined as in (2.13). Altogether, this gives that

$$x_j = \frac{|A_{j+1}|}{|A|} = (-1)^{j+2} (s - \mu_Y) \frac{S_{s-j} + S_{s-j-1}}{\prod_{k=1}^{s-1} (z_k - 1)},$$

which completes the proof. An alternative proof of Lemma 2.2.2 has been given in Zhao & Campbell [165].
Chapter 3

Fourier sampling

In Chapter 2 we have seen that the pgf of the stationary queue length in the discrete bulk service queue can be obtained in several ways, leading to expressions in terms of the roots of \( z^* = A(z) \) inside and outside the unit circle, or in terms of infinite series that involve convolutions of the distribution of \( A \). In this chapter we present a new technique to deal with the discrete bulk service queue that relies on Fourier sampling, with which we can derive the expressions in terms of infinite series from the expressions in terms of roots.

The roots of \( z^* = A(z) \) on or inside the unit circle can be considered to be located on a queue-specific curve, which we call the generalized Szegő curve. Under some mild conditions, and adopting a special parametrization of these Szegő curves, the roots occur as equidistant samples of a \( 2\pi \)-periodic function whose Fourier coefficients can be determined analytically. As such, we can for instance write the series that occur in the expressions for the moments of the stationary queue length as Fourier series with terms given in analytic form. Also, the stationary queue length distribution can be found in that way. This gives rise to root-free formulas for e.g. the mean, variance and the probability distribution of the stationary queue length that match the infinite series expressions obtained from random walk theory and the Wiener-Hopf technique, as presented in Secs. 2.3 and 2.4, respectively.

This chapter is organized as follows. In Sec. 3.1 we give a sketch of the approach that relies on the Lagrange inversion theorem and Fourier sampling. In Sec. 3.2 we discuss the queue-specific curve, in relation with the analytic and geometric conditions that are required in applying the Fourier sampling approach. With this approach, we derive analytical and root-free expressions for the mean and variance of the stationary queue length in Sec. 3.3. Similar expressions for the stationary distribution are derived in Sec. 3.4. Some conclusions are presented in Sec. 3.5. The chapter is based on Janssen & Van Leeuwaarden [P5].
3.1 A sketch of the approach

We now introduce the key elements of this chapter. In later sections we will discuss the various aspects in more detail.

Throughout this chapter, we make the following assumption:

Assumption 3.1.1  It should hold that $a_0 = P(A = 0) > 0$, $A'(1) < s$, and the pgf $A(z)$ is analytic in a disk $|z| < 1 + \epsilon$ with $\epsilon > 0$.

The assumption that $a_0 > 0$ is not restrictive, see p. 22. The fact that $A(z)$ is analytic in a disk $|z| < 1 + \epsilon$ with $\epsilon > 0$ excludes those $A(z)$ with radius of convergence 1, which were not excluded in Chapter 2. Examples of $A(z)$ with radius of convergence 1 were given in Sec. 2.6.

We now sketch our approach to obtain analytic expressions for series $\sum_k g(z_k)$ with $z_k$ the roots of $z^s = A(z)$ in $|z| \leq 1$ and some function $g$. The starting point of our approach is Lagrange’s inversion theorem:

**Theorem 3.1.2** (Lagrange inversion) For $f(z)$ and $g(z)$ analytic on and inside a contour $\mathcal{J}$ surrounding the origin, and for $w$ satisfying

$$|wf(z)| < |z|,$$

for every $z$ on $\mathcal{J}$, the equation $w = z/f(z)$ has a unique solution $z = z_0(w)$ inside $\mathcal{J}$ and

$$g(z_0(w)) = g(0) + \sum_{l=1}^{\infty} \frac{w^l}{l!} \left[ \left( \frac{d}{dz} \right)^{l-1} [f^l(z)g'(z)] \right]_{z=0}.$$  \hspace{1cm} (3.2)

**Proof** See Whittaker & Watson [162], § 7.32.  \hspace{1cm} □

The absolute convergence of (3.2) follows from the estimate

$$\left| \frac{w^l}{l!} \left[ \left( \frac{d}{dz} \right)^{l-1} [f^l(z)g'(z)] \right]_{z=0} \right| = \left| \frac{1}{2\pi i} \oint_{\mathcal{J}} \frac{w^lf^l(z)g'(z)}{z^{l+1}} dz \right| \leq \frac{1}{2\pi i} \text{length(}{\mathcal{J}}\text{)} M_1 M_2^l.  \hspace{1cm} (3.3)$$

Here, $M_1 = \max \{|g'(z)|: z \in \mathcal{J}\}$, $M_2 = \max \{|z^{-1}wf(z)|: z \in \mathcal{J}\}$ and it holds by (3.1) that $M_2 < 1$.

To simply write $g(z_0(w))$ as a power series, assumptions weaker than in Thm. 3.1.2 suffice. Consider for $w$ in a neighborhood of zero the equation

$$z A^{-1/s}(z) = w,$$  \hspace{1cm} (3.4)

where at the left-hand side of (3.4) we have taken the principal value of the root. Let $g$ be a function analytic in a neighborhood of $z = 0$. Then, by Lagrange’s inversion theorem with $f(z) = A^{1/s}(z)$, there is a neighborhood of $w = 0$ such
that the equation (3.4) has a unique solution \( z = z_0(w) \). Furthermore, the function \( g(z_0(w)) \) has the power series expansion

\[
g(z_0(w)) = g(0) + \sum_{l=1}^{\infty} c_l(g) w^l, \tag{3.5}
\]

where for \( l = 1, 2, ... \)

\[
c_l(g) = \frac{1}{l!} \left( \frac{d}{dz} \right)^{l-1} [A^{l/s}(z) g'(z)]_{z=0} = \frac{1}{l!} C_{z^l} [A^{l/s}(z) g'(z)]. \tag{3.6}
\]

We have used here the short-hand notation \( C_{z^l} [h(z)] \) for the coefficient of \( z^l \) in \( h(z) \).

In applying (3.5) to find an expression for the unique solution itself, we introduce

\[
g_0(z) = z; \quad c_l := c_l(g_0), \tag{3.7}
\]

and we let \( R \) be the radius of convergence of the series

\[
z_0(w) = \sum_{l=1}^{\infty} c_l w^l. \tag{3.8}
\]

We shall show in Sec. 3.2 that the mapping \( w, |w| < R \rightarrow z_0(w) \) is analytic and injective.

### 3.1.1 Fourier sampling

Now assume that \( R > 1 \). Then we can consider the equation (3.4) and its unique solution \( z_0(w) \) for \( w = e^{i\alpha}, \alpha \in [0, 2\pi] \). Accordingly, we let

\[
z(\alpha) := z_0(e^{i\alpha}), \quad \alpha \in [0, 2\pi]. \tag{3.9}
\]

The \( s \) roots \( z = z_k, \, k = 0, 1, ..., s - 1 \), of the equation \( z^s = A(z) \) with \( z_0 = 1, |z_k| \leq 1, k = 1, ..., s - 1 \), are distinct and are obtained as

\[
z_k = z(2\pi k/s) = z_0(e^{2\pi ik/s}), \quad k = 0, 1, ..., s - 1. \tag{3.10}
\]

Furthermore, with (3.9) we have a parametrization of a Jordan curve with 0 in its interior. Finally, if the function \( g \) is analytic in an open neighborhood of \( \{z_0(w) \, | \, |w| \leq 1\} \), then the \( 2\pi \)-periodic function \( \alpha \rightarrow g(z(\alpha)) \) has the Fourier series representation

\[
g(z(\alpha)) = g(0) + \sum_{l=1}^{\infty} c_l(g) e^{i\alpha l}, \quad \alpha \in [0, 2\pi], \tag{3.11}
\]

with \( c_l(g) \) given in (3.6).

The assumption \( R > 1 \) is, for instance, satisfied when \( A(z) \) is zero-free in \( |z| \leq 1 \). An example of this is the Poisson case,

\[
a_j = e^{-\lambda} \frac{\lambda^j}{j!}, \quad j = 0, 1, ...; \quad A(z) = e^{\lambda(z-1)}, \tag{3.12}
\]
with $0 \leq \mu_A = \lambda < s$. There are also non-trivial examples of distributions $A$ with pgf’s that do have zeros in the unit disk for which $R > 1$. See Example 3.2.6 where we consider the binomial distribution.

With $R > 1$ and $g$ analytic in an open neighborhood of $\{z_0(w) \mid |w| \leq 1\}$ it follows from the Fourier series representation (3.11) and elementary Fourier sampling theory that (for a more detailed treatment see p. 53)

$$
\sum_{k=0}^{s-1} g(z_k) = s \cdot g(0) + s \sum_{l=1}^{\infty} c_l(s) \cdot g(z) = s \cdot g(0) + \sum_{l=1}^{\infty} \frac{1}{l} \cdot C_{z_k}^{l-1} \cdot [A'(z) \cdot g'(z)].
$$

Note that (3.13) has terms that involve integral powers of $A$ only. In fact, when the function $g$ is analytic in a disk $|z| < 1 + \epsilon$ with $\epsilon > 0$ and $A$ satisfies Assumption 3.1.1, the numbers $C_{z_k}^{l-1} \cdot [A'(z) \cdot g'(z)]$ decay exponentially fast, irrespective of whether $R > 1$ or not. Hence in these cases the right-hand side of (3.13) makes sense regardless of whether $R > 1$ or not. It therefore seems a plausible conjecture that (3.13) holds for these more general $A$ and somewhat different functions $g$.

Some of the functions $g$ we are interested in fail to be analytic at $z = 1$, but become so after proper regularization. This is, for instance, the case with

$$
g(z) = \frac{1}{1-z} \cdot \frac{z}{(1-z)^2}
$$

which occur in the expression for the mean and variance of the stationary queue length, (2.21) and (2.22), respectively. In Sec. 3.3 we regularize the functions $g$ in (3.14) by subtracting

$$
\frac{B}{1-z \cdot A^{-1/s}(z)} + \frac{C}{(1-z \cdot A^{-1/s}(z))^2} + \frac{D}{1-z \cdot A^{-1/s}(z)}
$$

with properly chosen $B$ and $C, D$. Indeed, since $A(1) = 1$, proper choice of $B$ and $C, D$ cancels the poles of the functions $g$ at $z = 1$. Furthermore

$$
\left[z \cdot A^{-1/s}(z)\right]_{z=z_k} = e^{2\pi ik/s}, \quad k = 1, \ldots, s,
$$

and in Sec. 3.3 we present an identity for the series $\sum_{k=1}^{s-1} (1 - e^{2\pi ik/s})^{-m}$, $m = 1, 2$, which shows that regularization of the functions $g$ in (3.14) according to (3.15) may maintain the analytic nature of the expressions for $\sum g(z_k)$. For this it is also required that the $c_l(g)$, with regularized functions $g$, can still be expressed in analytic form. That this happens to be the case is also shown in Sec. 3.3. In Sec. 3.2 we give the details regarding the parametrization in (3.8) of the Jordan curve $\{z : |z| = |A^{1/s}(z)|, \ |z| \leq 1\}$ in case that $R > 1$. In Sec. 3.3 we give the details of our approach to find analytic expressions for $\sum_{k=1}^{s-1} (1 - z_k)^{-1}$, $\sum_{k=1}^{s-1} z_k (1 - z_k)^{-2}$, and we present the resulting expressions for $\mu_X$ and $\sigma_X^2$. In Sec. 3.4 we give explicit expressions for the probabilities $x_j$, $j = 0, 1, \ldots, s$, by using the Fourier sampling approach.
Generalized Szegö curves and Fourier sampling

In 1924, Szegö \cite{[150]} showed that the zeros of the normalized partial sums
\[ s_n(nz) = \sum_{k=0}^{n} \frac{(nz)^k}{k!}, \quad n = 0, 1, \ldots \] (3.17)
of \(e^z\) tend to what nowadays is called the Szegö curve
\[ S := \{ z \in \mathbb{C} : |z| = |e^z - 1|, |z| \leq 1 \}. \] (3.18)
This Szegö curve to this date continues to attract attention of researchers in approximation theory, see e.g. Pritsker & Varga \cite{[136]}, Walker \cite{[158]}, Wielonsky \cite{[163]}, and the references therein.

Curves of the Szegö type occur in the present context as follows. When \(A(z)\) satisfies Assumption 3.1.1, the roots \(z_k\) of \(z^s = A(z)\) in the unit disk all lie in the set
\[ S_{A,s} := \{ z \in \mathbb{C} : |z| = |A^{1/s}(z)|, |z| \leq 1 \}. \] (3.19)
In the Poisson case \(A(z) = \exp(\lambda(z - 1))\), with \(A'(1) = \lambda < s\), we obtain the set
\[ S_\theta := \{ z \in \mathbb{C} : |z| = |e^{\theta(z-1)}|, |z| \leq 1 \}, \] (3.20)
where \(\theta := \lambda/s\). The set \(S\) in (3.18) occurs as the limit case when \(\theta \uparrow 1\).

### 3.2.1 Analytical conditions for Fourier sampling

We now prove the claims on the mappings \(z_0(w)\) and \(z(\alpha)\) made in Sec. 3.1 under the assumption that the power series
\[ \sum_{l=1}^{\infty} c_l w^l, \quad c_l = \frac{1}{l} C_{z_l-1} [A^{1/s}(z)], \] (3.21)
has radius of convergence \(R > 1\). We have assumed that \(A(z)\) is analytic in a disk \(|z| < 1 + \epsilon\), and also that \(A'(1) < s\), and \(a_0 > 0\). Let
\[ G_\epsilon := \left\{ \sum_{l=1}^{\infty} c_l w^l : |w| < R \right\} \cap \{ z : |z| < 1 + \epsilon \}. \] (3.22)
Define
\[ z_0(w) := \sum_{l=1}^{\infty} c_l w^l, \quad |w| < R, \] (3.23)
and let
\[ H_\epsilon := z_0^{-1}(G_\epsilon) = \{ w \in \mathbb{C} : |w| < R, |z_0(w)| < 1 + \epsilon \}. \]
Lemma 3.2.1  With the above assumptions and definitions the following holds. The function $A$ is analytic and zero-free on $G_\epsilon$. When taking the principal $s^{-1}$-root of $A(z)$, $z \in G_\epsilon$, there holds that

$$z_0(w) A^{-1/s}(z_0(w)) = w, \quad w \in H_\epsilon,$$  

(3.24)

where $z_0(w)$ is the unique solution of the equation $z A^{-1/s}(z) = w$ with $w \in H_\epsilon$ and $z \in G_\epsilon$. This unique $z_0(w)$ is positive for $w \in (0,1]$ and satisfies $z_0(1) = 1$. For $\alpha \in [0,2\pi]$ we have that $z(\alpha)$ is the unique solution $z$ in $|z| \leq 1$ of

$$z A^{-1/s}(z) = e^{i\alpha}.$$  

(3.25)

The set $\{z(\alpha) | \alpha \in [0,2\pi]\}$ is a Jordan curve with $0$ in its interior. Finally, the roots $z_k$ of the equation $z^s = A(z)$, $k = 0,1,...,s-1$, occur as $z(2\pi k/s)$ and are distinct.

Proof Evidently, $A$ is analytic on $G_\epsilon \subset \{z | |z| < 1+\epsilon\}$. Since $z_0(w) A^{-1/s}(z_0(w)) = w$ holds in a neighborhood of $w = 0$, we have by analyticity that

$$z_0'(w) = w^s A(z_0(w)), \quad w \in H_\epsilon.$$  

(3.26)

Suppose that $w \in H_\epsilon$, $w \neq 0$, and that $A(z_0(w)) = 0$. Then it follows from (3.26) that $z_0(w) = 0$, whence that $A(0) = a_0 \neq 0$. Contradiction. So $A(z_0(w)) \neq 0$ for $w \in H_\epsilon$. We can therefore take the principal $s^{-1}$-root of $A(z)$ for $z \in G_\epsilon$ which is analytic on $G_\epsilon$. By analyticity we then have that $z_0(w) A^{-1/s}(z_0(w)) = w$ holds on all of $H_\epsilon$, and not just in a neighborhood of $w = 0$. That is, (3.24) holds. From (3.24) it readily follows that $z_0$ is injective on $H_\epsilon$. Also when $w \in H_\epsilon$ we have that $z_0(w)$ is the unique solution $z \in G_\epsilon$ of the equation $z A^{-1/s}(z) = w$.

The function $z \in [0,1+\delta] \to z A^{-1/s}(z)$ is strictly increasing for some $\delta > 0$. Indeed, when $z \in (0,1]$ we have that

$$(z A^{-1/s}(z))' = \frac{1}{s} A^{-\frac{1}{s}-1}(z) [s A(z) - z A'(z)]$$

$$= \frac{1}{s} A^{-\frac{1}{s}-1}(z) \sum_{j=0}^{\infty} (s-j) a_j z^j \geq \frac{1}{s} A^{-\frac{1}{s}-1}(z) z^s \sum_{j=0}^{\infty} (s-j) a_j > 0,$$

(3.27)

since $A'(1) < s$. Moreover $A(1) = 1$. It thus follows that $z_0(w)$ increases from 0 to 1 as $w$ increases from 0 to 1.

We consider now $w = e^{i\alpha}$ with $\alpha \in [0,2\pi]$, and let $z(\alpha)$ be the unique solution of (3.25). We shall show that $|z(\alpha)| \leq 1$. To that end we observe that there is a $\delta > 0$ such that $|A(z)| \neq |z|^s$ when $1 < |z| < 1 + \epsilon$ (this follows from the assumptions that $a_j \geq 0$, $A'(1) < s$). Since $z(0) = z_0(1) = 1$ and $z(\alpha)$ depends continuously on $\alpha \in [0,2\pi]$, we see that $|z(\alpha)| \leq 1$, $\alpha \in [0,2\pi]$. Furthermore, $z(0) = z(2\pi)$ and $z(\alpha) \neq z(\beta)$ when $0 \leq \alpha < \beta < 2\pi$, while the mapping $r \in [0,1] \to \{z_0(re^{i\alpha}) | \alpha \in [0,2\pi]\}$ is analytic on $[0,1]$. Finally, the roots $z_k$ of the equation $z^s = A(z)$, $k = 0,1,...,s-1$, occur as $z(2\pi k/s)$ and are distinct.
is analytic in
where
is indeed a Jordan curve with 0 in its interior.
Finally consider (3.25) with \( \alpha = 2\pi k/s, k = 0, 1, \ldots, s - 1 \). Evidently, the \( z(2\pi k/s) \)
are distinct and have modulus \( \leq 1 \), as follows from the above. Also, any \( z(2\pi k/s) \)
is a root of the equation \( z^s = A(z) \), see (3.25). Hence, the sets \( \{ z_k | k = 0, \ldots, s - 1 \} \)
and \( \{ z(2\pi k/s) | k = 0, \ldots, s - 1 \} \) coincide.

**Remark 3.2.2** (i) We have \( |A(z)|^{1/s}, =, < |z| \) as \( z, |z| \leq 1 \), is inside, on, outside the Jordan curve \( \{ z(\alpha) | \alpha \in [0, 2\pi] \} \).

(ii) As implicitly stated in the proof of Lemma 3.2.1, we have the following geometric condition which is equivalent with \( R > 1 \): There is a Jordan curve \( J \) with \( S_{A,s} \) in its interior such that \( A(z) \) is zero-free on and inside \( J \) while \( |A(z)| < |z|^s \) on \( J \). Also see Subsec. 3.2.2 for this matter.

In case \( A(z) \) is zero-free in \( |z| < 1 + \epsilon \), things can be simplified, as shown in the next lemma.

**Lemma 3.2.3** Assume that \( A \) satisfies Assumption 3.1.1 and that \( A(z) \) is zero-free in \( |z| < 1 + \epsilon \), where \( \epsilon > 0 \). Then \( C_{z,l-1}[A^{l/s}(z)] \) decays exponentially.

**Proof** By Cauchy’s theorem we have

\[
C_{z,l-1}[A^{l/s}(z)] = \frac{1}{2\pi i} \oint_{|z|=r} \frac{A^{l/s}(z)}{z} dz, \quad l = 1, 2, \ldots, \quad (3.28)
\]

for any \( r \in (0, 1 + \epsilon) \). Noting that there is a \( \delta > 0 \) such that

\[
\left| \frac{A(z)}{z^s} \right| < \frac{A(|z|)}{|z|^s} < 1, \quad |z| < 1 + \delta, \quad (3.29)
\]

we see that for any \( r \in (1, 1 + \delta) \)

\[
|C_{z,l-1}[A^{l/s}(z)]| \leq \frac{1}{2\pi} \text{length}(|z| = r) \left( \max_{|z|=r} \left| \frac{A(z)}{z^s} \right|^{1/s} \right)^l
= r \left( \frac{A(r)}{r^s} \right)^{l/s} = r \left( \frac{A(r)}{r^{s/l}} \right)^{l/s}, \quad (3.30)
\]

and this decays exponentially fast as \( l \to \infty \). □

The convergence rate of the series that appear in (2.39)-(2.42) can be deduced from the following result.

**Lemma 3.2.4** Assume that \( A \) satisfies Assumption 3.1.1 with \( \epsilon > \delta > 0 \) such that \( A \) is analytic in \( |z| < 1 + \epsilon \) and \( |A(z)| < |z|^s \) in \( 1 < |z| < 1 + \delta \) (no assumption on \( R \)). Let \( h \) be analytic in \( |z| < 1 + \epsilon \). Then for any \( r \in (1, 1 + \delta) \) we have

\[
|C_{z,j}[A^{l}(z) h(z)]| \leq \left( \frac{A(r)}{r^{s/l}} \right)^l M \frac{1}{r^{j/l}}, \quad l = 1, 2, \ldots, \quad j \geq ls, \quad (3.31)
\]

where \( M = \max \{|h(z)| : |z| = r\} \).
Proof This follows in a similar fashion as Lemma 3.2.3. By Cauchy's theorem and $A'(1) < s$ we have

$$C_{z_l}[A^l(z) h(z)] = \frac{1}{2\pi i} \oint_{|z|=r} \frac{A^l(z) h(z)}{z^{j+l+1}} \, dz,$$

for any $r \in (0, 1 + \epsilon)$. Hence, for any $r \in (1, 1 + \delta)$ we have

$$|C_{z_l}[A^l(z) h(z)]| \leq \frac{1}{2\pi} \text{length}(|z| = r) \max_{|z|=r} \left| \frac{A^l(z)}{z^{j+l+1}} \cdot \frac{h(z)}{z^{j+1-ls}} \right|$$

$$\leq r \left( \frac{A(r)}{r^{j+1-ls}} \right)^l \frac{M}{r^{j+1-ls}},$$

and this decays exponentially fast as $l \to \infty$. □

Example 3.2.5 Consider the Poisson case $A(z) = \exp(\lambda(z - 1))$ with $0 \leq \lambda < s$. We have plotted in Fig. 5.2 the set $S_{\theta}$ in (3.20) for a number of values of $\theta := \lambda/s$ (although not permitted, $\theta = 1$ is included). The dots on the curves indicate the roots $z_k$ for the case $s = 20$. We compute

$$c_l = \frac{1}{l} C_{z_l-1}[A^{l/s}(z)] = \frac{1}{l} C_{z_l-1}[e^{\theta l(z-1)}] = e^{-\theta l} \frac{(l\theta)^{l-1}}{l!}$$

for $l = 1, 2, \ldots$. Hence $S_{\theta}$ has the parametric representation (see (3.11))

$$z_{\theta}(\alpha) = \sum_{l=1}^{\infty} e^{-\theta l} \frac{(l\theta)^{l-1}}{l!} e^{il\alpha}, \quad \alpha \in [0, 2\pi].$$
3.2 Generalized Szegő curves and Fourier sampling

Figure 3.3: $S_{A, s=2n}$ for binomial case, $q = 2(\sqrt{2} - 1)$. The dots indicate $z_0, \ldots, z_{19}$ for $s = 20$.

Figure 3.4: $S_{A, s=2n}$ for binomial case, $q = .83$. The dots indicate $z_0, \ldots, z_{19}$ for $s = 20$.

Example 3.2.6 Consider the binomial case $A(z) = (p + qz)^n$ where $p, q \geq 0$, $p + q = 1$ and $A'(1) = nq < s$. We compute in this case

$$c_l = \frac{1}{l} C_{z_{l-1}}[A^{l/s}(z)] = \frac{1}{l} C_{z_{l-1}}[(p + qz)^{nl/s}]$$

$$= \frac{1}{l} p^{nl-1} q^{l-1} \left(\frac{nl}{l-1}\right), \quad l = 1, 2, ..., \tag{3.36}$$

where we have used the notation

$$\binom{\alpha}{k} = \frac{\alpha(\alpha - 1) \cdot \cdots \cdot (\alpha - k + 1)}{k!} = \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha - k + 1) \Gamma(k + 1)}. \tag{3.37}$$

Let $\beta := n/s$. When $\beta = 1$ we have

$$c_l = pq^{l-1}, \quad l = 1, 2, ..., \tag{3.38}$$

and there is exponential decay (when $\beta = 1$ we have $q < s/n = 1$). When $\beta > 1$, Stirling’s formula $\Gamma(n + 1) \approx n^{n+1/2}e^{-n}\sqrt{2\pi}$ (see e.g. De Bruijn [41], p. 70) yields

$$c_l \approx \frac{p}{q} \frac{1}{\beta - 1} \frac{1}{l\sqrt{2\pi}l} p^{(\beta-1)} q^{l} \left(\frac{\beta}{\beta - 1}\right)^{1/2} \left[\frac{\beta^\beta}{(\beta - 1)^{\beta - 1}}\right]^l, \tag{3.39}$$

so there is exponential decay when

$$\frac{\beta^\beta}{(\beta - 1)^{\beta - 1}} p^{\beta - 1} q < 1. \tag{3.40}$$
For fixed $p$, $q$, the quantity at the left-hand side of (3.40) is maximal as a function of $\beta$ at $\beta = 1/q$, with the value 1. Hence, since $\beta = n/s < 1/q$, we have exponential decay. Finally, when $0 < \beta < 1$, Stirling’s formula and the formula $\Gamma(x)\Gamma(1-x) = \pi/(\sin \pi x)$ yield

$$c_l \approx \frac{p}{q} \frac{1}{(1-\beta)^2} \frac{(-1)^l \sin \pi l \beta}{l(\frac{1}{2} \pi l)^{1/2}} p^{l(1-\beta)} q'^{(1-\beta)l}((1-\beta)^{1-\beta} \beta^l)^l,$$  

(3.41)

so there is exponential decay when

$$p^{\beta-1} q(1-\beta)^{1-\beta} \beta^l < 1.$$  

(3.42)

Note that the left-hand side of (3.42) increases from 0 to $\infty$ when $q$ increases from 0 to 1 ($p = 1 - q$). In the critical case, where we have = instead of < in (3.42), $c_l$ still decays as $l^{-3/2}$. This critical case also arises in the following way. With $\beta = n/s$ we consider the equation

$$|p + qz|^\beta = |z|$$  

(3.43)

for negative $z = -r [-1,0)$. When $0 < \beta < 1$ and $p/q < 1$ this equation has at least one and at most three roots $z \in [-1,0]$. The critical case now occurs when (3.43) has three roots of which two coincide.

In Figs. 3.2-3.4 we consider the case that $\beta = 1/2$ and $s = 20$. The critical case now occurs for $q_0 = 2(\sqrt{2} - 1) = 0.828427125$. We have plotted the set

$$S_{A,s=2n} = \{ z : |z| \leq 1, |p + qz|^{1/2} = |z| \}$$  

(3.44)

for $q = 0.82, 2(\sqrt{2} - 1), 0.83$. We observe that $S_{A,s}$ turns from a smooth Jordan curve containing 0 (Fig. 3.2) into two separate closed curves when $q$ passes $q_0$ (Fig. 3.4). For a more extensive treatment of the binomial case, we refer to Sec. 4.3.

**Lemma 3.2.7** Assume that $A$ satisfies Assumption 3.1.1 and that the radius of convergence, $R$, of the series in (3.21) $> 1$. Also assume that $g$ is analytic in an open neighborhood of $\{ z_0(w) : |w| \leq 1 \}$. Then $c_l(g) = l^{-1} C_{l-1}(A^{1/s}(z) g'(z))$ has exponential decay, and there is an $R_g > 1$ such that

$$g(z_0(w)) = g(0) + \sum_{l=1}^{\infty} c_l(g) w^l, \quad |w| < R_g,$$  

(3.45)

with absolute convergence at the right-hand side of (3.45). In particular, we have

$$g(z(\alpha)) = g(0) + \sum_{l=1}^{\infty} c_l(g) e^{il\alpha}, \quad \alpha \in [0,2\pi],$$  

(3.46)

with absolute convergence at the right-hand side of (3.46).
3.2 Generalized Szegö curves and Fourier sampling

Proof There is an $R_1$, $1 < R_1 < R$, such that $g$ is analytic in $\{z_0(w) : |w| < R_1\}$. With $1 < R_2 < R_1$ and $C_2 = \{z_0(R_2 e^{i\alpha}) | \alpha \in [0, 2\pi]\}$ a Jordan curve with 0 in its interior, Cauchy’s theorem yields for $l = 1, 2, \ldots$

$$C_{z_{l-1}}[A^{l/s}(z)g'(z)] = \frac{1}{2\pi i} \oint_{C_2} \frac{A^{l/s}(z)g'(z)}{z^l} \, dz. \quad (3.47)$$

On $C_2$ we have $|A(z)/z^s| = R_2^{-s}$, whence

$$|C_{z_{l-1}}[A^{l/s}(z)g'(z)]| \leq M R_2^{-l}, \quad l = 1, 2, \ldots, \quad (3.48)$$

where $M = \max \{|g'(z)| : z \in C_2\}$. This shows exponential decay of $c_l(g)$. From this (3.45) easily follows with $R_g = R_2$ since $g(z_0(w)) = g(0) + \sum_{l=1}^{\infty} c_l(g) w^l$ holds in a neighborhood of 0 by Lagrange’s theorem. Finally, (3.46) is a direct consequence of (3.45).

We now make some comments on equidistant sampling of functions $g(z(\alpha))$ under the conditions of Lemma 3.2.7, for which we use the following property

**Property 3.2.8** The following identity holds:

$$\sum_{k=0}^{s-1} e^{2\pi ik/s} = \begin{cases} 
  s, & l = 0 \text{(mod } s), \\
  0, & l \neq 0 \text{(mod } s). 
\end{cases} \quad (3.49)$$

Proof If $l = 0 \text{(mod } s)$, $e^{2\pi ik/s}$ equals 1 for each $k$, and thus $\sum_{k=0}^{s-1} e^{2\pi ik/s} = s$. For the case $l \neq 0 \text{(mod } s)$, take $l = ms + r$, with $m \in \mathbb{Z}$ and $r \in \{1, \ldots, s-1\}$. Then

$$\sum_{k=0}^{s-1} e^{2\pi ik/s} = \sum_{k=0}^{s-1} e^{2\pi ikm} e^{2\pi ir/s} = \sum_{k=0}^{s-1} (e^{2\pi ir/s} k) = \frac{1 - (e^{2\pi ir/s})^s}{1 - e^{2\pi ir/s}} = 0,$$

which completes the proof.

From (3.46), we immediately see that when the conditions of Lemma 3.2.7 are satisfied, we have

$$\sum_{k=0}^{s-1} g(z_k) = \sum_{k=0}^{s-1} g(\{z(2\pi k/s)\}) = s \cdot g(0) + \sum_{k=0}^{s-1} \sum_{l=1}^{\infty} c_l(g) e^{2\pi ik/l/s}$$

$$= s \cdot g(0) + \sum_{l=1}^{\infty} c_l(g) \sum_{k=0}^{s-1} e^{2\pi ik/l/s} = s \cdot g(0) + s \sum_{l=1}^{\infty} c_{ls}(g). \quad (3.50)$$

3.2.2 Geometrical conditions for Fourier sampling

As one sees from Thm. 3.1.2 (Lagrange’s inversion theorem) and the proof of Lemma 3.2.7, the condition $R > 1$ is closely related to geometric conditions on $S_{A,s}$.
These conditions are addressed in this section. For the notions used below from complex function theory we refer to Duren [62] and Silverman [143].

We impose the following condition on the generalized Szegő curve $S_{A,s}$:

**Condition 3.2.9** $S_{A,s}$ is a Jordan curve with 0 in its interior, and $A(z)$ is zero-free on and inside $S_{A,s}$.

Recall that $a_0 > 0$ so that $|A(z)| > |z|^s$ for $z$ in the interior of $S_{A,s}$. Condition 3.2.9 is geometric in nature, and can be visually checked using some standard software package. A useful geometric formulation equivalent with Condition 3.2.9 is as follows:

**Lemma 3.2.10** Condition 3.2.9 is satisfied if and only if there is a Jordan curve $J$ with $S_{A,s}$ in its interior such that $A(z)$ is zero-free on and inside $J$ while $|A(z)| < |z|^s$ on $J$.

The proof that Condition 3.2.9 implies the existence of a $J$ as in Lemma 3.2.10 uses continuity of $A$ on $S_{A,s}$ and some basic considerations of Jordan curve theory. The proof of the reverse implication can be based on the considerations in the proof of Lemma 3.2.3, but we omit the details. We now present an equivalent form of Condition 3.2.9 of more analytic nature (where $C_{z,l}[h(z)]$ again denotes the coefficient of $z^l$ in $h(z)$):

**Lemma 3.2.11** Condition 3.2.9 is satisfied if and only if $C_{z,l-1}[A^{l/s}(z)]$ decays exponentially in $l$.

**Proof** Assume that Condition 3.2.9 holds. Letting $J$ as in Lemma 3.2.10 we see that we can define an analytic root $A^{l/s}(z)$ for $z$ on and inside $J$ that is positive at $z = 0$. By Cauchy’s theorem we thus have

$$C_{z,l-1}[A^{l/s}(z)] = \frac{1}{2\pi i} \oint_{z \in J} \frac{A^{l/s}(z)}{z^l} \, dz, \quad l = 1, 2, \ldots.$$  \hfill (3.51)

Since $|A(z)| < |z|^s$ for $z \in J$, it follows that

$$|C_{z,l-1}[A^{l/s}(z)]| \leq \frac{1}{2\pi \text{length}(J)} \left( \max_{z \in J} \left| \frac{A(z)}{z^s} \right|^{1/s} \right)^l,$$  \hfill (3.52)

and this decays exponentially, as required.

Now assume that $C_{z,l-1}[A^{l/s}(z)]$ decays exponentially. We consider for $w$ in a neighborhood of 0 the equation

$$zA^{-1/s}(z) = w,$$  \hfill (3.53)

where we have taken in a neighborhood of $z = 0$ the root $A^{-1/s}(z)$ of $A(z)$ that is positive at $z = 0$ (recall $a_0 > 0$). By Lagrange’s inversion theorem (Thm. 3.1.2), the solution $z_0(w)$ of (3.53) has the power series representation

$$z_0(w) = \sum_{l=1}^{\infty} c_l w^l.$$  \hfill (3.54)
for \( w \) in a neighborhood of 0 in which

\[
\frac{1}{l} \left[ \left( \frac{d}{dz} \right)^{l-1} \left( \frac{z}{zA^{-1/s}(z)} \right) \right]_{z=0} = \frac{1}{l} C_{z}^{-1}[A^{1/s}(z)]. \tag{3.55}
\]

By assumption, we have that \( c_l \to 0 \) exponentially, whence the power series in \( (3.54) \) for \( z_0(w) \) has a radius of convergence \( R > 1 \). It follows then from basic considerations in analytic function theory that \( A^{-1/s}(z) \) extends analytically to the open set \( \{ \sum_{l=1}^{\infty} c_l w^l : |w| < R \} \) and that \( z_0(w) \) extends according to \( (3.54) \) on the set \( |w| < R \). The Szegö curve \( S_{A,s} \) in \( (3.19) \) can alternatively be described as

\[
S_{A,s} = \{ z_0(e^{i\alpha}) : \alpha \in [0, 2\pi] \}, \tag{3.56}
\]

and it can be shown that the parametrization

\[
\alpha \in [0, 2\pi] \to z_0(e^{i\alpha}) = \sum_{l=1}^{\infty} c_l e^{il\alpha} \in S_{A,s} \tag{3.57}
\]

has no double points while a homotopy between \( \{0\} \) and \( S_{A,s} \) is obtained according to

\[
r \in [0, 1] \to \{ z_0(re^{i\alpha}) : \alpha \in [0, 2\pi] \}. \tag{3.58}
\]

From the latter facts it follows that \( S_{A,s} \) is a Jordan curve with 0 in its interior, and this completes the sketch of the proof of the converse statement. \( \square \)

**Remark 3.2.12** The \( z_0(w) \) of \( (3.54) \) is a univalent function of a special type on an open set containing the closed unit disk \( |w| \leq 1 \). Hence, the results of the theory of univalent functions, as presented for instance in Duren [62], Chapters 2-3 and Silverman [143], Chapter 12, become available. This leads to all kinds of observations on the shape of \( S_{A,s} \). We shall not elaborate on this point, except for a casual note in Sec. 4.2.

Condition 3.2.9 and its equivalent forms as given by Lemmas 3.2.10 and 3.2.11 are equally useful in deciding whether a given \( A(z) \) satisfies it. We now present some special cases for which Condition 3.2.9 is immediately satisfied.

i. \( A(z) \) is zero-free in \( |z| \leq 1 \). An appropriate Jordan curve \( J \) is found as \( |z| = 1 + \delta \) with sufficiently small \( \delta > 0 \). Indeed, the assumptions on \( A \) imply that there is a \( \delta > 0 \) such that \( 0 < |A(z)| < |z|^s \) for \( 1 < |z| \leq 1 + \delta \).

ii. \( A(z) \) is zero-free in \( |z| < 1 \). There may occur now a finite number of zeros of \( A \) on \( |z| = 1 \), necessitating a modification of the Jordan curve \( J \) in (i). We indent this \( J \) around the zeros such that the zeros are outside the new \( J \) while \( |A(z)| < |z|^s \) for all \( z \) on the new \( J \). This technique may also work in cases where there are zeros of \( A \) strictly inside \( |z| = 1 \). A class of examples follows from Kakeya’s theorem (see Kakeya [95]) as follows:

- when \( a_0 > a_1 > \ldots, \) we have that \( A(z) \) is zero-free in \( |z| \leq 1, \)
- when \( a_0 \geq a_1 \geq \ldots \), we have that \( A(z) \) is zero-free in \(|z| < 1\).

iii. The \( c_l \) in (3.55) are all non-negative. It follows from Pringsheim’s theorem (see Titchmarsh [154]) and the fact that \( z_0(w) \) is well-defined for \( w \in [0, 1 + \delta] \) with some \( \delta > 0 \), that the radius of convergence of the power series in (3.54) exceeds 1. Thus Lemma 3.2.11 applies and it follows that Condition 3.2.9 is satisfied.

**Example 3.2.13** (A.J.E.M. Janssen, private communication) Consider the case 
\[ A(z) = \left( \frac{1}{4} \left( z^2 + \frac{1}{2} z + \frac{1}{4} \right) \right)^4. \]
Then, the generalized Szegő curves show a typical behavior when increasing \( s \). In Figs. 3.5-3.8 \( S_{A,s} \) has been displayed for \( s = 20, 22, 23 \) and 26, respectively. Observe that the curve turns from a Jordan curve into three separate components, where two small components occur like two islands separated from the mainland, each containing exactly four roots. The roots displayed in Figs. 3.5-3.8 have been found numerically using the computer package QROOT (see Chaudhry et al. [46]). Note that the roots in Figs. 3.5 and 3.6 could have been determined with the Fourier series representation.

### 3.3 Moments of the queue length

In this section we express the series
\[
\sum_{k=1}^{s-1} \frac{1}{1-z_k}, \quad \sum_{k=1}^{s-1} \frac{z_k}{(1-z_k)^2},
\]
(3.59)
in terms of aliasing series under Assumption 3.1.1 on \( A \) and the assumption that the series in (3.21) has radius of convergence \( R > 1 \).

To apply Lemma 3.2.7, we need to regularize the functions
\[
g_1(z) = \frac{1}{1-z}, \quad g_2(z) = \frac{z}{(1-z)^2} = \frac{1}{(1-z)^2} - \frac{1}{1-z},
\]
(3.60)
at \( z = 1 \). We consider \( g_1 \) and \( g_2 \) on and around the Szegő curve \( S_{A,s} = \{ z_0(w) \mid w = e^{i\alpha}, \alpha \in [0, 2\pi) \} \), and so it would be convenient when the regularizing functions assume a simple form on \( S_{A,s} \). We thus propose to subtract from \( g_1 \) and \( g_2 \) in (3.60) the functions
\[
h_1(z) = \frac{B}{1-z A^{-1/s}(z)}, \quad h_2(z) = \frac{C}{(1-z A^{-1/s}(z))^2} + \frac{D}{1-z A^{-1/s}(z)},
\]
(3.61)
for which we have
\[
h_1(z_0(w)) = \frac{B}{1-w}, \quad h_2(z_0(w)) = \frac{C}{(1-w)^2} + \frac{D}{1-w},
\]
(3.62)
on account of (3.24). In (3.60) we must choose \( B, C, D \) such that \( g_1 - h_1 \) and \( g_2 - h_2 \) are regular at \( z = 1 \); this can be done indeed since \( 1-z A^{-1/s}(z) \) has a
3.3 Moments of the queue length

Figure 3.5: $S_{A,s}$ for $A(z) = (\frac{1}{4}(z^2 + \frac{1}{2}z + \frac{1}{4}))^4$ and $s = 20.$

Figure 3.6: $S_{A,s}$ for $A(z) = (\frac{1}{4}(z^2 + \frac{1}{2}z + \frac{1}{4}))^4$ and $s = 22.$

Figure 3.7: $S_{A,s}$ for $A(z) = (\frac{1}{4}(z^2 + \frac{1}{2}z + \frac{1}{4}))^4$ and $s = 23.$

Figure 3.8: $S_{A,s}$ for $A(z) = (\frac{1}{4}(z^2 + \frac{1}{2}z + \frac{1}{4}))^4$ and $s = 26.$

first-order zero at $z = 1$ (as $A'(1) < s$). Noting that $z_k = z_0(\exp(2\pi ik/s))$, we have from (3.61) that $\sum_{k=1}^{s-1} h_i(z_k)$ are computationally manageable. In fact there holds that

$$\sum_{k=1}^{s-1} \frac{1}{1 - e^{2\pi ik/s}} = \frac{1}{2}(s - 1), \quad \sum_{k=1}^{s-1} \frac{1}{(1 - e^{2\pi ik/s})^2} = \frac{1}{12}(s - 1)(s - 5).$$

(3.63)
Fourier sampling

The decisive reason to choose \( h_1 \) and \( h_2 \) of the above type is the following result that shows that subtraction of \( h_i \) does not lead to unmanageable expressions in the aliasing series.

**Lemma 3.3.1** Let \( f \) be analytic in a neighborhood of 0. Then

\[
\frac{1}{l} C_{z^{l-1}} \left[ A^{l/s}(z) \frac{d}{dz} (f(z A^{-1/s}(z))) \right] = C_{w^l} [f(w)]. \tag{3.64}
\]

**Proof** By Lagrange's theorem (see p. 44), we have

\[
\begin{align*}
\frac{1}{l} C_{z^{l-1}} \left[ A^{l/s}(z) \frac{d}{dz} (f(z A^{-1/s}(z))) \right] &= \frac{1}{l!} \left( \frac{d}{dz} \right)^{l-1} \left[ A^{l/s}(z) \frac{d}{dz} (f(z A^{-1/s}(z))) \right]_{z=0} \\
&= C_{w^l} [f(z A^{-1/s}(z))] \text{ where } z \text{ satisfies } z A^{-1/s}(z) = w \\
&= C_{w^l} [f(w)], \tag{3.65}
\end{align*}
\]

as required. \( \square \)

We finally consider the issue of choosing \( B \) and \( C, D \) properly in (3.61). Thus we let \( g_i^R := g_i - h_i, \ i = 1, 2 \). A lengthy but otherwise elementary computation shows that for \( g_1^R \) we need to take

\[
B = 1 - s^{-1} A'(1), \tag{3.66}
\]

so that \( g_1^R \) is indeed regular at \( z = 1 \), with value

\[
g_1^R(1) = \left[ \frac{1}{1-z} - \frac{B}{1-z A^{-1/s}(z)} \right]_{z=1} = s^{-1} A'(1) - \frac{1}{2} \left[ s^{-1} (s^{-1} + 1) (A'(1))^2 - s^{-1} A''(1) \right] \\
1 - s^{-1} A'(1) \tag{3.67}
\]
at \( z = 1 \). For the regularization of \( g_2 \) we need to take

\[
C = (1 + a^{[1]})^2, \quad D = -1 - 3a^{[1]} - a^{[2]}, \tag{3.68}
\]

where

\[
a^{[i]} = \left[ \left( \frac{d}{dz} \right)^i A^{-1/s}(z) \right]_{z=1}, \quad i = 1, 2, 3, \tag{3.69}
\]

and so

\[
g_2^R(1) = \left[ \frac{z}{(1-z)^2} - \frac{C}{(1-z A^{-1/s}(z))^2} - \frac{D}{1-z A^{-1/s}(z)} \right]_{z=1} \\
= \frac{a^{[2]} + \frac{1}{2} a^{[3]}}{1 + a^{[1]}} + \frac{a^{[1]} + \frac{1}{2} a^{[2]}}{1 + a^{[1]}} \left( 1 - \frac{a^{[1]} + \frac{1}{2} a^{[2]}}{1 + a^{[1]}} \right). \tag{3.70}
\]
3.3 Moments of the queue length

We are then in a position to prove the following theorem:

**Theorem 3.3.2** Under Assumption 3.1.1 on A and Condition 3.2.9 we have that

\[
\sum_{k=1}^{s-1} \frac{1}{1-z_k} = \frac{1}{2} (s-1) + \frac{1}{2} \mu_A - \frac{\sigma_A^2}{2(s-\mu_A)}
\]

\[+ \sum_{l=1}^{\infty} \frac{1}{l} \sum_{j=ls}^{\infty} (j - ls) C_{z^l} [A^l(z)]. \tag{3.71}\]

**Proof** By Lemma 3.2.7 we have with \( g = g_1^R \) and (3.63) that

\[
\sum_{k=1}^{s-1} \frac{1}{1-z_k} = \sum_{k=1}^{s-1} \frac{B}{1 - e^{2\pi ik/s}} + \sum_{k=1}^{s-1} g_1^R(z_k)
\]

\[= \frac{1}{2} B(s-1) - g_1^R(1) + \sum_{k=0}^{s-1} g_1^R(z_k)
\]

\[= \frac{1}{2} B(s-1) - g_1^R(1) + s g_1^R(0) + s \sum_{l=1}^{\infty} c_{ls}(g_1^R). \tag{3.72}\]

Furthermore,

\[c_{ls}(g_1^R) = \frac{1}{ls} C_{z^{ls-1}} [A^l(z)(g_1^R)'(z)]
\]

\[= \frac{1}{ls} C_{z^{ls-1}} [A^l(z) \frac{1}{(1-z)^2}] - \frac{B}{ls} C_{z^{ls-1}} [A^l(z) \left( \frac{1}{1-s^{-1}A(z)} \right)']. \tag{3.73}\]

Using \((1-z)^{-2} = \sum_{j=0}^{\infty} (j+1) z^j\) and applying Lemma 3.3.1 we then get that

\[c_{ls}(g_1^R) = \frac{1}{ls} \sum_{j=0}^{ls-1} (ls-j) C_{z^j} [A^l(z)] - C_{w^{ls}} \left[ \frac{B}{1-w} \right]
\]

\[= \frac{1}{ls} \sum_{j=0}^{ls-1} (ls-j) C_{z^j} [A^l(z)] - B. \tag{3.74}\]

To bring the right-hand side of (3.74) in its final form, we observe that \( c_{ls}(g_1^R) \to 0 \) as \( l \to \infty \) and that

\[1 = A^l(1) = \sum_{j=0}^{\infty} C_{z^j} [A^l(z)], \tag{3.75}\]
\[
A'(1) = \frac{d}{dz} [A'(z)] (z = 1) = \sum_{j=0}^{\infty} j C_{zj}[A'(z)]. 
\]

(3.76)

This implies that \(B = 1 - s^{-1} A'(1)\), which agrees with (3.66), and that

\[
c_{ls}(g_1^R) = \frac{1}{ls} \sum_{j=ls}^{\infty} (j - ls) C_{zj}[A'(z)]. 
\]

(3.77)

Therefore we arrive at (noting that \(g_1^R(0) = 1 - B\))

\[
\sum_{k=1}^{s-1} \frac{1}{1 - z_k} = (1 - \frac{1}{2} B) s - \frac{1}{2} B - g_1^R(1) + \sum_{l=1}^{\infty} \frac{1}{l} \sum_{j=ls}^{\infty} C_{zj}[A'(z)]. 
\]

(3.78)

The proof of Thm. 3.3.2 is completed by a rather long but otherwise elementary computation, using the expressions in (3.66) and (3.67) for \(B\) and \(g_1^R(1)\) and the fact that \(A'(1) = \mu_A\) and \(A''(1) = \sigma^2_A + \mu^2_A - \mu_A\). \(\square\)

For the mean stationary queue length, Thm. 3.3.2 together with (2.21) now yields the expression in terms of infinite series (2.39). A similar result can be obtained for the variance of the stationary queue length:

**Theorem 3.3.3** Under Assumption 3.1.1 on \(A\) and Condition 3.2.9 we have that

\[
\sum_{k=1}^{s-1} \frac{z_k}{(1 - z_k)^2} = -\frac{1}{12} C(s-1)(s-5) + \frac{1}{2} D(s-1) - g_2^R(1) - s(C + D)
\]

\[-\sum_{l=1}^{\infty} \frac{1}{l} \sum_{j=ls}^{\infty} (j - ls)^2 C_{zj}[A'(z)], \]

(3.79)

where \(C\) and \(D\) are given by

\[
C = \left(1 - \frac{1}{s} \mu_A\right)^2, \quad C + D = \frac{1}{s} \sigma^2_A, \quad (3.80)
\]

and \(g_2^R(1)\) as in (3.70). Alternatively, the constant on the first line of (3.79) may be expressed as

\[
-\frac{1}{12} C(s-1)(s-5) + \frac{1}{2} D(s-1) - g_2^R(1) - s(C + D)
\]

\[= \frac{A'''(1) - s(s-1)(s-2)}{3(s - \mu_A)} + \frac{A''(1) - s(s-1)}{2(s - \mu_A)} + \left(\frac{A''(1) - s(s-1)}{2(s - \mu_A)}\right)^2, \]

(3.81)
3.3 Moments of the queue length

Proof The procedure for computation of \( \sum_{k=1}^{s-1} \frac{z_k}{(1-z_k)^2} \) is entirely the same as the one for \( \sum_{k=1}^{s-1} (1-z_k)^{-1} \), although quite a bit more elaborate. Accordingly, using both items in (3.63) and Lemma 3.2.7 with \( g = g_2^R \) we get as in (3.78) that (using \( g_2^R(0) = -C - D \))

\[
\sum_{k=1}^{s-1} \frac{z_k}{(1-z_k)^2} = - \frac{1}{12} C(s-1)(s-5) + \frac{1}{2} D(s-1) - g_2^R(1) - s(C + D)
\]

\[+ s \sum_{l=1}^{\infty} c_{sl}(g_2^R). \quad (3.82)
\]

Using that \( (z(1-z)^{-2})' = \sum_{j=0}^{\infty} j^2 z^{j-1} \) we find in a similar fashion as in (3.74) from Lemma 3.3.1 that

\[
c_{ls}(g_2^R) = \frac{1}{ls} C_{zls-1}
\]

\[\times \left[ A'(z) \left( \frac{z}{(1-z)^2} \right)' - C_w \left[ \frac{C}{(1-w)^2} + \frac{D}{1-w} \right] \right]
\]

\[= \frac{1}{ls} \sum_{j=0}^{ls-1} (ls-j)^2 C_{z}(A'(z)) - C(l+1) - D. \quad (3.83)
\]

To bring (3.83) in its final form, we observe that \( c_{ls}(g_2^R) \to 0 \) as \( l \to \infty \) by Lemma 3.2.7, and we use (3.75) and (3.76) together with

\[
\sum_{j=0}^{\infty} j^2 C_{z}(A'(z)) = l(l-1)(A'(1))^2 + l A''(1) + l A'(1). \quad (3.84)
\]

This yields (in agreement with (3.68))

\[
C + D = \frac{1}{s} \sigma_A^2, \quad C = \left( 1 - \frac{1}{s} \mu_A \right)^2, \quad (3.85)
\]

and

\[
c_{ls}(g_2^R) = - \frac{1}{ls} \sum_{j=ls}^{\infty} (j-ls)^2 C_{z}(A'(z)). \quad (3.86)
\]

We then find that

\[
\sum_{k=1}^{s-1} \frac{z_k}{(1-z_k)^2} = - \frac{1}{12} C(s-1)(s-5) + \frac{1}{2} D(s-1) - g_2^R(1) - s(C + D)
\]

\[\quad - \sum_{l=1}^{\infty} \frac{1}{l} \sum_{j=ls}^{\infty} (j-ls)^2 C_{z}(A'(z)). \quad (3.87)
\]

The proof of Thm. 3.3.3 is then completed by a long but otherwise elementary calculation in which the two members in (3.81) are shown to be equal. □

Thm. 3.3.3 together with (2.22) yields the infinite series expression (2.40) for the variance of the stationary queue length.
3.4 Stationary queue length distribution

In this section we derive explicit formulas for the \( x_j \), \( j = 0, \ldots, s \), and for

\[
\sum_{j=0}^{s-1} x_j, \quad \sum_{j=0}^{s} x_j.
\]  

(3.88)

As in Sec. 3.3, we derive these formulas using the Fourier-sampling approach sketched in Sec. 3.1. Again we work under Assumption 3.1.1 and the condition that the series in (3.21) has radius of convergence \( R > 1 \).

For the polynomial

\[
Q(z) := \sum_{j=0}^{s-1} (z^s - z^j) x_j =: \sum_{j=0}^{s} q_j z^j,
\]

(3.89)

we have shown in Subsec. 2.2.1 that the following identity holds:

\[
Q(z) = \gamma_1 (z - 1) \prod_{k=1}^{s-1} (z - z_k),
\]

(3.90)

where

\[
\gamma_1 = \frac{s - \mu_A}{\prod_{k=1}^{s-1} (1 - z_k)}.
\]

(3.91)

It follows from (3.89) that

\[
x_j = -q_j, \quad j = 0, \ldots, s - 1; \quad x_s = -q_s - a_{-1} \cdot q_0.
\]

(3.92)

Also, it holds that

\[
\gamma_1 = C_z [Q(z)] = \sum_{j=0}^{s-1} x_j.
\]

(3.93)

For computations later in this section we rewrite \( Q(z) \) as

\[
Q(z) = (-1)^s \gamma_1 \prod_{k=1}^{s-1} z_k \prod_{k=0}^{s-1} \left( 1 - \frac{z}{z_k} \right),
\]

(3.94)

and then (3.92) shows that it is enough to find explicit formulas for

\[
\gamma_1 = \frac{s - \mu_A}{\prod_{k=1}^{s-1} (1 - z_k)} \cdot \prod_{k=0}^{s-1} z_k, \quad C_z \left[ \prod_{k=0}^{s-1} \left( 1 - \frac{z}{z_k} \right) \right], \quad j = 1, \ldots, s - 1.
\]

(3.95)

The explicit formulas can be found, leading to the following results:
3.4 Stationary queue length distribution

Theorem 3.4.1 Under Assumption 3.1.1 on $A$ and Condition 3.2.9 we have that

\[ x_0 = a_0 \exp \left\{ -\sum_{l=1}^{\infty} \frac{1}{l} \sum_{j=l+1}^{\infty} C_{2l} \{ A^l(z) \} \right\}, \]  

(3.96)

\[ s^{-1} \sum_{j=0}^{s-1} x_j = \exp \left\{ -\sum_{l=1}^{\infty} \frac{1}{l} \sum_{j=l+1}^{\infty} C_{2l} \{ A^l(z) \} \right\}, \]  

(3.97)

\[ s \sum_{j=0}^{s-1} x_j = \exp \left\{ -\sum_{l=1}^{\infty} \frac{1}{l} \sum_{j=l+1}^{\infty} C_{2l} \{ A^l(z) \} \right\}. \]  

(3.98)

Proof We start by considering $\prod_{k=1}^{s-1} (1 - z_k)^{-1}$ and to that end we regularize $g_3(z) := \ln (1 - z)$ at $z = 1$ by setting

\[ g^R_3(z) = \ln (1 - z) - \ln (1 - z A^{-1/s}(z)). \]  

(3.99)

Then $g^R_3$ is analytic in an open neighborhood of \( \{ z : |z|^s \leq A(z) \} \), and

\[ g^R_3(1) = -\ln(1 - \mu_A/s), \quad g^R_3(0) = 0. \]  

(3.100)

Also, we have $z_k A^{-1/s}(z_k) = \exp(2\pi ik/s)$, $k = 0, 1, ..., s - 1$, and the identity

\[ s^{-1} \sum_{k=1}^{s-1} \ln(1 - e^{2\pi ik/s}) = \ln s. \]  

(3.101)

We thus obtain as before that

\[ s^{-1} \sum_{k=1}^{s-1} \ln(1 - z_k) = \sum_{k=1}^{s-1} \ln(1 - e^{2\pi ik/s}) + \sum_{k=1}^{s-1} g^R_3(z(2\pi k/s)) \]

\[ = \ln(s - \mu_A) + s \sum_{l=1}^{\infty} c_{ls}(g^R_3). \]  

(3.102)

Here we have, also as before, from Lemmas 3.2.7 and 3.3.1

\[ c_{ls}(g^R_3) = \frac{1}{ls} C_{ls+1} \{ A^l(z)(\ln(1 - z))^l \} - C_{ls} \{ \ln(1 - w) \} \]

\[ = \frac{1}{ls} \sum_{j=ls}^{\infty} C_{2j} \{ A^j(z) \}. \]  

(3.103)

Hence we get

\[ s^{-1} \sum_{k=1}^{s-1} \ln(1 - z_k) = \ln(s - \mu_A) + \sum_{l=1}^{\infty} \frac{1}{l} \sum_{j=ls}^{\infty} C_{2j} \{ A^j(z) \}, \]  

(3.104)
so that (3.97) follows.

We next compute \( \prod_{k=0}^{s-1} z_k \). To that end we note that

\[
z_k = e^{2\pi ik/s} A^{1/s}(z(2\pi k/s)), \quad (3.105)
\]

so that

\[
\prod_{k=0}^{s-1} z_k = (-1)^{s-1} \exp \left\{ \sum_{k=0}^{s-1} \ln \left[ A^{1/s}(z(2\pi k/s)) \right] \right\}. \quad (3.106)
\]

The function \( g_4(z) := \ln [A^{1/s}(z)] \) is analytic in a neighborhood of \( \{z : |z|^s \leq |A(z)|\} \), and we have

\[
g_4(1) = 0, \quad g_4(0) = \frac{1}{s} \ln a_0. \quad (3.107)
\]

Hence, it holds that

\[
\sum_{k=0}^{s-1} \ln \left[ A^{1/s}(z(2\pi k/s)) \right] = \ln a_0 + s \sum_{l=1}^{\infty} c_l(g_4). \quad (3.108)
\]

The \( c_l(g_4) \) follow from

\[
c_{ls}(g_4) = \frac{1}{(ls)!} \left[ \left( \frac{d}{dz} \right)^{ls-1} [A^{l}(z)(\ln [A^{1/s}(z)])'] \right]_{z=0}
\]

\[
= \frac{1}{s(ls)!} \left[ \left( \frac{d}{dz} \right)^{ls-1} [A^{-1}(z) A'(z)] \right]_{z=0} = \frac{1}{ls} C_{2ls}[A'(z)]. \quad (3.109)
\]

We thus have that

\[
\prod_{k=1}^{s-1} z_k = (-1)^{s-1} a_0 \exp \left\{ \sum_{l=1}^{\infty} \frac{1}{l} C_{2ls}[A^{l}(z)] \right\}. \quad (3.110)
\]

We then find from (3.92) and (3.94) that

\[
x_0 = -q_0 = (-1)^s \gamma_1 \prod_{k=1}^{s-1} z_k, \quad (3.111)
\]

which can be rewritten as (3.96). Moreover, we have that \( \sum_{j=0}^{s} x_j = x_0/a_0 \) and so we obtain (3.98). This completes the proof. \( \square \)

We conclude by computing the \( x_i, i = 1, \ldots, s - 1 \). Note that

\[
x_i = x_0 C_{z_i} \prod_{k=0}^{s-1} \left( 1 - \frac{z}{z_k} \right), \quad i = 1, \ldots, s - 1. \quad (3.112)
\]
We shall consider, using (3.105) and the Taylor expansion of \( \ln(1-x) \) around \( x = 0 \), the expression
\[
\sum_{k=0}^{s-1} \ln\left(1 - \frac{z}{2k}\right) = -\sum_{j=1}^{\infty} \frac{z^j}{j} \sum_{k=0}^{s-1} A^{-j/s}(z(2\pi k/s)) e^{-2\pi i j k/s}.
\] (3.113)

We can then prove the following result:

**Theorem 3.4.2** Under Assumption 3.1.1 on \( A \) and Condition 3.2.9 we have that
\[
x_i = \sum_{j=0}^{s} x_j C_v \left[ A(v) \exp\left\{\sum_{j=1}^{s-1} v^j \sum_{l=1}^{\infty} \frac{1}{l} C_{z^l+s}[A^l(z)]\right\}\right],
\] (3.114)

for \( i = 0, 1, \ldots, s - 1 \).

**Proof** The \( x_i \) in (3.112) are completely determined by the terms at the right-hand side of (3.113) with \( j = 1, \ldots, s - 1 \). Thus we consider for \( j = 1, \ldots, s - 1 \) the \( 2\pi \)-periodic functions
\[
A^{-j/s}(z(\alpha)) = A^{-j/s}(0) + \sum_{l=1}^{\infty} c_l[A^{-j/s}] e^{i l \alpha}.
\] (3.115)

The \( c_l[A^{-j/s}] \) are given here as
\[
c_l[A^{-j/s}] = \frac{-1}{l!} j \frac{1}{s} \left[ \left( \frac{d}{dz} \right)^{l-1} [A^{l/s}(z)(A^{-j/s})'(z)] \right]_{z=0}.
\] (3.116)

It is seen from (3.116) that
\[
c_j[A^{-j/s}] = -\frac{j}{s} C_z[\ln [A(z)]],
\] (3.117)
\[
c_l[A^{-j/s}] = -\frac{j}{l-j} C_{z^l}[A^{l-j/s}(z)], \quad l \neq j.
\] (3.118)

Since \( A(0) = a_0 \), we thus get that
\[
A^{-j/s}(z(\alpha)) e^{-ij\alpha} = a_0^{-j/s} e^{-ij\alpha} - \frac{j}{s} C_z[\ln A(z)] - j \sum_{l=-j+1, l \neq 0}^{\infty} \frac{1}{l} C_{z^{l+j}}[A^{l/s}(z)] e^{i l \alpha}.
\] (3.119)
Therefore, for \( j = 1, \ldots, s - 1 \), we obtain by sampling theory that
\[
\sum_{k=0}^{s-1} A^{-j/s}(z(2\pi k/s)) e^{-2\pi ijk/s} = -j C_z, \ln [A(z)] - \sum_{l=1}^{\infty} j C_{z^{l+j}}[A^l(z)].
\]
This gives, see (3.112) and (3.113), for \( i = 1, \ldots, s - 1 \) that
\[
x_i = x_0 C_{v^i} \left[ \exp \left\{ \sum_{j=1}^{s-1} v^j \left( C_z, \ln [A(z)] \right) + \sum_{l=1}^{\infty} \frac{1}{l} C_{z^{l+j}}[A^l(z)] \right\} \right]. \tag{3.120}
\]
Since we consider \( i = 1, \ldots, s - 1 \) in (3.120), the summation over \( j \) may be extended to all \( j = 1, 2, \ldots \). Noting that
\[
\sum_{j=1}^{\infty} v^j C_z, \ln [A(z)] = \ln A(v) - \ln a_0, \tag{3.121}
\]
and that \( x_0 = a_0 \sum_{j=0}^s x_j \), we arrive for \( i = 1, \ldots, s - 1 \) at (3.114).

### 3.5 Conclusions

For the mean, variance and probability distribution of the stationary queue length, Thms. 3.3.2, 3.3.3 and 3.4.2 show that the expressions in terms of infinite series can be obtained from the expressions in terms of roots. For proving these theorems we have presented a new technique that relies on Fourier sampling. The condition under which the technique can be applied, has been formulated both analytically and geometrically. In particular, the technique can be applied if the generalized Szegő curve is a Jordan curve with zero in its interior, which is the case for most probability distributions of \( A \). However, as shown in this chapter, there are distributions of \( A \) for which the generalized Szegő curve intersects itself or consists of more than one component. Hence, the infinite-series expressions derived in this chapter are less general than the infinite-series expressions in Chapter 2.

So what does this all bring us, except for confirmation? First of all, we argue that the technique outlined in this chapter is interesting, both from a methodological and conceptual viewpoint. Second, our detailed analysis of the roots and associated generalized Szegő curve provides expressions for rather general functions of the roots. After all, the expressions for the functions
\[
\sum_{k=1}^{s-1} \frac{1}{1 - z_k}, \sum_{k=1}^{s-1} \frac{z_k}{(1 - z_k)^2}, \prod_{k=1}^{s-1} (1 - z_k)^{-1}, \prod_{k=0}^{s-1} z_k, C_{z^i} \left[ \prod_{k=0}^{s-1} \left(1 - \frac{z}{z_k}\right) \right],
\]
that have led to the results in this chapter, were just a handful of examples that can be tackled. Perhaps the function \( z_k \) is both the simplest and nicest one, since this
gives an explicit expression for the roots. One might say that our detailed analysis was all about deriving this explicit characterization of the roots, and we therefore consider this to be the most important result derived in this chapter. This is also why we zoom into the issue of explicit root-finding in the next chapter.
Chapter 4

Back to the roots

In the previous two chapters we have seen that the pgf of the stationary queue length in the discrete bulk service queue can be expressed in terms of the roots of \( z^s = A(z) \). This technique of completing a transform solution with roots has become a classic one in queueing theory. It has been applied to numerous queueing models other than the discrete bulk service queue, and due to advanced numerical algorithms and increased computational power, root-finding has become a more or less straightforward numerical procedure (see Chaudhry et al. [46]).

For the discrete bulk service queue we have presented in Chapter 3 identities for the mean, variance and probability distribution in terms of roots on the one hand and in terms of infinite series on the other. Crucial was the fact that the roots of \( z^s = A(z) \) could be represented as sample values of a periodic function with analytically given Fourier coefficients, in which case we have an explicit solution for the roots. We consider this to be the most important result obtained in Chapter 3, and so we return to this topic in the present chapter. We investigate the explicit representation of the roots, both from the analytical and numerical viewpoint. We also compare the Fourier representation with an implicit representation that results from applying successive substitution to a fixed-point equation. This idea originates from the work of Harris et al. [85] on root-finding for the continuous-time \( G/E_k/1 \) queue, and was presented more formally by Adan & Zhao [23] who identified a class of continuous distributions for which the method works. We further investigate this iterative method in case that \( A \) is a discrete random variable. We present necessary conditions for the method to work and compare these to the conditions needed for the Fourier series representation of the roots.

The chapter is structured as follows. The Fourier series representation of the roots is discussed in Sec. 4.1. In Sec. 4.2 the iterative method is presented, followed by a comparison with the Fourier series representation. In Sec. 4.3 we provide a detailed treatment of the binomial case, which serves as an illustrative example of the theory presented in Sec. 4.2. In Sec. 4.4 we discuss a special case in which the roots have a probabilistic interpretation. Some numerical results are presented in Sec. 4.5. The chapter is based on Janssen & Van Leeuwaarden [P6].
4.1 Roots and Fourier series

We now turn to the representation of the $s$ roots of $z^s = A(z)$ in $|z| \leq 1$. Under Condition 3.2.9, these roots all lie inside the Jordan curve $J$ as discussed in Lemma 3.2.10, and are given by

$$z_k = w_k A^{1/s}(z)_{w_k} = w_k A_{1/s}(z), \quad k = 0, 1, \ldots, s - 1,$$

(4.1)

where $w_k = e^{2\pi ki/s}$. Hence, from (3.57) we have

$$z_k = \sum_{l=1}^{\infty} c_l w_k^l, \quad k = 0, 1, \ldots, s - 1,$$

(4.2)

where the $c_l$ are explicitly given as

$$c_l = \frac{1}{l} C_{z^l - 1}[A^{l/s}(z)].$$

(4.3)

Remark 4.1.1 When $A(z)$ is a polynomial of degree $m > s$, an expression similar to (4.2) can be derived for the $m - s$ roots of $z^s = A(z)$ outside the unit circle. Substituting $1/v$ into $z^s = A(z)$ and multiplying by $v^m$, we get

$$v^{m-s} = B(v),$$

(4.4)

where $B(v) = v^m A(1/v)$ is a polynomial of degree $m$. Note that $|A(z)| < |z|^s$ for $1 < |z| < 1 + \delta$ for some $\delta > 0$ implies that $|B(v)| < |v|^{m-s}$ for $(1 + \delta)^{-1} < |v| < 1$. Therefore, by Rouché’s theorem, there occur exactly $m - s$ roots $v_k, k = \{s, s + 1, \ldots, m - 1\}$ of (4.4) in $|v| \leq (1 + \delta)^{-1}$, obviously satisfying $v_k = 1/z_k$. When there exists a Jordan curve (within $|v| < 1$) such that $B(v)$ is zero-free on and inside this curve, while 0 lies inside this curve and $|B(v)| < |v|^{m-s}$ on this curve, we find that

$$v_k = \sum_{l=1}^{\infty} \frac{1}{l} C_{v^l - 1}[B^{l/(m-s)}(v)] e^{2\pi(k-s)i/(m-s)}, \quad k = s, s + 1, \ldots, m - 1.$$

(4.5)

In Example 4.5.5 the above scheme will be applied.

Examples 3.2.5 and 3.2.6 are specific cases for which the $c_l$ can be computed explicitly. In general, the $c_l$ can be determined using the following property:

Property 4.1.2 For $A(z) = \sum_{j=0}^{\infty} a_j z^j$ and $\alpha \in \mathbb{R}$, and $A^\alpha(z) = \sum_{j=0}^{\infty} b_j z^j$, the coefficients $b_j$ follow from the coefficients $a_j$ according to $b_0 = a_0^\alpha$ and

$$b_{j+1} = \alpha a_0 a_j - a_{j+1} + \frac{1}{(j+1)a_0} \sum_{n=0}^{j-1} [\alpha(n+1) - (j-n)]a_{n+1} b_{j-n}, \quad j = 0, 1, \ldots.$$

(4.6)
4.2 Roots through fixed-point iteration

**Proof** The proof consists of computing the $b_j$’s successively by equating coefficients in $A(z)(A^s)'(z) = \alpha A'(z)A^s(z)$. □

In Chaudhry et al. [46] it is shown that the condition that $A$ is infinitely-divisible, or the somewhat weaker condition that $A(z)$ has no zeros inside the unit circle, are sufficient for the roots of $z^s = A(z)$ on and within the unit circle to be distinct. However, examples exist of $A(z)$ having zeros inside the unit circle and at the same time having distinct roots. It is therefore that in both Chaudhry et al. [46] and Harris et al. [85] the need for finding a necessary condition for distinctness is expressed. In this respect, we have the following result:

**Lemma 4.1.3** When Condition 3.2.9 is satisfied, the $s$ roots of $z^s = A(z)$ on and within the unit circle are distinct.

**Proof** The roots lie inside $J$, and satisfy (4.1). Since $|A(z)|^{1/s} < |z|$ for all $z \in J$, it follows from Rouché’s theorem that for each $w_k$, the function $z - w_kA^{1/s}(z)$ has as many zeros inside $J$ as $z$. □

Although Condition 3.2.9 is not necessary for the roots to be distinct (as appears to be the case in Example 3.2.6 with $\beta = 1/2$ and $q = 0.83$), it covers a far larger class of distributions of $A$ than those for which $A(z)$ has no zeros within the unit circle.

4.2 Roots through fixed-point iteration

We now discuss a way to determine the roots by applying successive substitutions to a fixed-point equation. We present necessary conditions for the method to work and compare these to the conditions needed for the Fourier series representation of the roots introduced in the previous section.

When $A(z)$ is assumed to have no zeros for $|z| \leq 1$, we know that the $s$ roots of $z^s = A(z)$ in $|z| \leq 1$ satisfy

$$z = wG(z),$$

with $G(z) = A^{1/s}(z)$ and $w^s = 1$. For each feasible $w$, Equation (4.7) can be shown as in Lemma 4.1.3 to have one unique root in $|z| \leq 1$. One could try to find the roots by successive substitutions (see Adan & Zhao [23] and Harris et al. [85]) as

$$z_k^{(n+1)} = w_kG(z_k^{(n)}), \quad k = 0, 1, ..., s - 1,$$

(4.8)

with starting values $z_k^{(0)} = 0$ (one could choose different starting values).

**Lemma 4.2.1** When for $|z| \leq 1$, $A(z)$ is zero-free and $|G'(z)| < 1$, the fixed-point equations (4.8) converge to the desired roots.

**Proof** For $|z| \leq 1$, $|w| \leq 1$,

$$|wG(z)| \leq G(|z|) \leq G(1) = 1,$$

(4.9)
so \( wG(z) \) maps \(|z| \leq 1\) into itself. For \(|\tilde{z}|, |\hat{z}| \leq 1\) we have that
\[
|wG(\tilde{z}) - wG(\hat{z})| \leq |\tilde{z} - \hat{z}| \max_{0 \leq t \leq 1} |G'(\tilde{z} + t(\tilde{z} - \hat{z}))|.
\] (4.10)
Hence, from (4.10) and \(|G'(z)| < 1\) for all \(|z| \leq 1\), we conclude that \( wG(z) \) is a contraction on \(|z| \leq 1\).

For the Poisson distribution with \( \lambda < s \), it is readily seen that \( A(z) \neq 0 \) and \(|G'(z)| < 1\) for \(|z| \leq 1\), so that the iteration (4.8) works. We want to consider, however, also distributions for which \( A(z) \) has zeros within the unit circle (see e.g. Example 3.2.6). We restrict the attention here naturally to functions \( A(z) \) that allow a root \( G(z) = A^{1/s}(z) \) that is analytic around \( S_{A,s} \) and positive at 0. Hence we introduce the following condition:

**Condition 4.2.2** Condition 3.2.9 should be satisfied and for all points \( z \in S_{A,s} \) it should hold that \(|G'(z)| < 1\).

According to the maximum principle (see e.g. Whittaker & Watson [162]) we have that Condition 4.2.2 implies that \(|G'(z)| < 1\) holds for all points inside \( S_{A,s} \) as well. Condition 4.2.2 thus ensures that for \( \alpha \in [0, 2\pi] \) the point \( z_k \) is an attractor for the iteration (4.8).

Note that Condition 4.2.2 is what is minimally needed to ensure (4.8) to converge locally. However, under Condition 4.2.2 the iterates are by no means guaranteed to stay in \( S_{A,s} \) and its interior. This is already seen for the binomial case with \( \beta < 1 \), \( s \) even, and the iteration (4.8) for \( k = s/2 \), i.e.
\[
z_{s/2}^{(n+1)} = -1(p + qz_{s/2}^{(n)})^\beta.
\] (4.11)
For this iteration, the \( z_{s/2}^{(n)} \), \( n = 0, 1, \ldots \), are alternatingly inside and outside \( S_{A,s} \). The iteration, though, converges to the correct point when \( q \) is not too large. It is difficult, in general, to give guarantees for convergence; nevertheless, convergence seems to occur in most cases where Condition 4.2.2 holds.

While Condition 3.2.9 implies that \( S_{A,s} \) is a closed curve without double points, Condition 4.2.2 apparently does not hold for all such curves. We shall compare Condition 4.2.2 with the notions convexity and starshapedness from the geometric theory of univalent functions (see Remark 3.2.12).

**Definition 4.2.3** (i) A closed curve without double points is called starshaped with respect to a point in its interior if any ray from this point intersects the curve at exactly one point,

(ii) A closed curve without double points is called convex when it is starshaped with respect to any point in its interior.

We have the following result:
4.2 Roots through fixed-point iteration

Theorem 4.2.4

\( S_{A,s} \) is convex \( \Rightarrow \) Condition 4.2.2 holds \( \Rightarrow \) \( S_{A,s} \) is starshaped with respect to 0.

Proof We have \( z(\alpha) = z_0(e^{i\alpha}) \) where \( z_0(w) \) is the solution of

\[ z_0(w) = wG(z_0(w)), \quad G(z) = A^{1/s}(z), \tag{4.12} \]

see the proof of Lemma 3.2.3.

The notions of convexity and starshapedness vary slightly from one place to another in the literature. We use here the notions as can be found in Pólya & Szegö [131], p. 125, but we shall also use results from Duren [62] and Silverman [143] who use a somewhat different convention.

Assume that \( S_{A,s} \) is convex in the sense of [131]. According to [131], Exercise 108, we have

\[ \Re \left[ 1 + \frac{wz''_0(w)}{z'_0(w)} \right] > 0, \quad |w| = 1. \tag{4.13} \]

By continuity, the inequality (4.13) holds in an annulus \( 1 \leq |w| < 1 + \delta \) for some \( \delta > 0 \). Then, by the theory as given in [143], Chapter 12, Sec. 2 (in particular, the material on pp. 335-336), we have that \( z_0(w) \) is convex in \( |w| < 1 + \delta \) in the sense of [143], Chapter 12 and [62], Chapter 2. Now by [62], Corollary 1(a) on p. 251, we have that

\[ z_0(w) \text{ is convex in } |w| < 1 + \delta \Rightarrow \Re \left[ \frac{wz''_0(w)}{z'_0(w)} \right] > 1/2, \quad |w| < 1 + \delta. \tag{4.14} \]

From (4.12) we have upon computation

\[ \frac{wz'_0(w)}{z_0(w)} = \frac{1}{1 - wG'(z_0(w))}. \tag{4.15} \]

Now note that for \( v \in \mathbb{C}, v \neq 1 \) we have

\[ \Re \left[ \frac{1}{1 - v} \right] > \frac{1}{2} \Leftrightarrow |v| < 1. \tag{4.16} \]

Hence, when \( z_0(w) \) is convex, it holds that \( |wG'(z_0(w))| < 1 \) for \( |w| < 1 + \delta \) and thus Condition 4.2.2 is satisfied.

Next assume that Condition 4.2.2 holds. To prove that \( S_{A,s} \) is starshaped with respect to the origin, it is sufficient by [131], Exercise 109, to show that

\[ \phi'(\alpha) := \frac{d}{d\alpha} \arg z(\alpha) = \Re \left[ \frac{wz'_0(w)}{z_0(w)} \right] > 0, \tag{4.17} \]

for \( w = e^{i\alpha}, \alpha \in [0, 2\pi] \). Since \( |wG'(z_0(w))| < 1 \) by Condition 4.2.2, it is seen at once from (4.15) and (4.16) that \( \phi'(\alpha) > 1/2 > 0 \). This completes the proof. \( \square \)
4.3 Comparison of the two methods: Binomial case

The binomial case in Example 3.2.6 gives a nice demonstration of Thm. 4.2.4. From an inspection of \( S_{A,s} \) in Fig. 3.2 one sees that \( S_{A,s} \) is not starshaped with respect to 0, and one can thus immediately conclude that Condition 4.2.2 is not satisfied and hence the iteration (4.8) cannot be applied to determine the roots. We now take a closer look at the binomial case. We will provide a more detailed formulation of Condition 4.2.2 and compare this with Condition 3.2.9, which serves as an illustrative example of the theory presented in the previous section. Results were presented in Janssen & Van Leeuwaarden [P6]; the detailed treatment presented in this section is due to A.J.E.M. Janssen (private communication).

So we consider the binomial case, \( A(z) = (p + qz)^n \) where \( p, q \geq 0, p + q = 1 \) and \( \beta = n/s \). For Condition 3.2.9 we must check whether

\[
cl = 1 - l^{-1} q^{-1} \left( l - 1 \right), \quad l = 1, 2, ..., (4.18)
\]

has exponential decay, whereas for Condition 4.2.2 we should check whether

\[
\max_{z \in S_{A,s}} |G'(z)| = \max_{z \in S_{A,s}} |\beta q(p + qz)^{\beta - 1}| < 1 (4.19)
\]

is satisfied.

It was shown in Example 3.2.6 that \( c_l \) in (4.18) has exponential decay when \( \beta \geq 1 \). Also, it was shown that for \( 0 \leq \beta < 1 \) the \( c_l \) in (4.18) has exponential decay if and only if

\[
p^{\beta - 1} q(1 - \beta)^{1 - \beta} \beta^\beta < 1. (4.20)
\]

We shall assume that \( \beta < 1 \). We observe that the left-hand side of (4.20) increases from 0 to \( \infty \) when \( q \) increases from 0 to 1. For \( \beta \in (0, 1) \) we then define \( q_1(\beta) \in (0, 1) \) as the unique solution \( q \) of the equation

\[
p^{\beta - 1} q(1 - \beta)^{1 - \beta} \beta^\beta = 1. (4.21)
\]

It can be shown that \( q_1(\beta) \to 1/2 \) as \( \beta \downarrow 0 \), that \( q_1(\beta) \to 1 \) as \( \beta \uparrow 1 \), and that \( q_1(\beta) \) increases in \( \beta \in (0, 1) \). It then follows that, for the binomial case,

\[
\text{Condition 3.2.9 } \iff (4.20) \iff q < q_1(\beta). (4.22)
\]

Next we consider the contraction condition (Condition 4.2.2). We shall assume that \( \beta < 1 \). Indeed, when \( \beta \geq 1 \), we have that the maximum of \( |G'(z)| \) for \( z \in S_{A,s} \) occurs at \( z = 1 \) and equals \( \beta q < 1 \). Hence, when \( \beta \geq 1 \), the contraction condition is satisfied. Assuming that Condition 3.2.9 holds and \( \beta < 1 \), we know from Example 3.2.6 that there is exactly one \( r = r_0 \in (0, 1) \) such that

\[
|p - qr|^{\beta} = r, (4.23)
\]

and \(-r_0 \in S_{A,s}\) (see Fig. 4.1).
4.3 Comparison of the two methods: Binomial case

From (4.19) we see that (using the definition of $S_{A,s}$)

$$\max_{z \in S_{A,s}} |G'(z)| = \beta q \max_{z \in S_{A,s}} |z|^{1-1/\beta}. \quad (4.24)$$

We now claim that

$$\min_{z \in S_{A,s}} |z| = r_0. \quad (4.25)$$

Indeed, when $z \in S_{A,s}$ satisfies $|z| < r_0$ we have

$$|z| = |p + qz|^\beta > |p - qz|^\beta > |z|,$$

see Fig. 4.1. From this contradiction we see that (4.25) holds. Hence, since $\beta < 1$, we get that

$$\max_{z \in S_{A,s}} |G'(z)| = \beta q r_0^{1-1/\beta}. \quad (4.27)$$

When $\beta$ is fixed, one easily sees that $r_0$ decreases in $q \in (0, 1)$, so that

$$\beta q r_0^{1-1/\beta}$$

increases in $q \in (0, 1)$. Furthermore, $\beta q r_0^{1-1/\beta} \to 0$ as $q \downarrow 0$ and $\beta q r_0^{1-1/\beta} \to \infty$ as $q \uparrow 1$. Therefore, for any $\beta \in (0, 1)$ there is exactly one $q = q_2(\beta)$ such that

$$\beta q_2 r_0^{1-1/\beta} = 1. \quad (4.28)$$

Note that for the binomial case we thus have

$$\text{Condition 4.2.2} \iff \text{Condition 3.2.9 and } q < q_2(\beta). \quad (4.29)$$
We shall now show that \( q_2(\beta) < q_1(\beta) \) when \( 0 < \beta < 1 \). Indeed, we have \( r_0 < p/q \), and with \( q = q_2(\beta) \) we see that

\[
1 = \beta qr_0^{1-1/\beta} > \beta q \left( \frac{p}{q} \right)^{1-1/\beta} = \beta q^{1/\beta} p^{1-1/\beta}.
\] (4.30)

Hence \( p^{\beta-1} q q^\beta < 1 \), and this implies that

\[
p^{\beta-1} q (1 - \beta)^{1-\beta} q^\beta < 1.
\] (4.31)

Remembering that the left-hand side of (4.31) increases in \( q \) and the definition of \( q_1(\beta) \), we find that \( q_2(\beta) < q_1(\beta) \). It further follows that

\[
r_0 = (\beta q) \frac{q_1^\alpha}{q_1^\alpha}.
\] (4.32)

Substituting this into (4.28), we obtain that \( q_2(\beta) \) is the unique solution \( q \in (0, 1) \) of

\[
q = 1 - q^{1/\beta} \beta^{1/\beta} (1 + \beta).
\] (4.33)

Rewriting (4.21), we obtain that \( q_1(\beta) \) is the unique solution \( q \in (0, 1) \) of

\[
q = 1 - q^{1/\beta} \beta^{1/\beta} (1 - \beta).
\] (4.34)

We observe that the right-hand side of (4.33) is smaller than the right-hand side of (4.34), which confirms that \( q_2(\beta) < q_1(\beta) \) for \( \beta \in (0, 1) \).

We have plotted \( q_1(\beta) \) and \( q_2(\beta) \) for \( \beta \in [0, 1] \) in Fig. 4.2. It is clearly visible

![Figure 4.2: q₁(β) and q₂(β) for β ∈ (0, 1).](image)

that the set of values \( q \) for which the Fourier series representation (Condition 3.2.9, \( q < q_1(\beta) \)) holds is much larger than the set for which Condition 4.2.2 holds (\( q < q_2(\beta) \)). We remark that, although Condition 4.2.2 is not sufficient for the fixed-point iteration to work, we have numerical evidence that whenever \( q < q_2(\beta) \) the iteration
4.4 On a result for a right-continuous random walk

(4.8) does work. Finally note that the roots in Fig. 3.2 ($\beta = 1/2, q = 0.82$), which are computed using the Fourier series representation, cannot be obtained using the fixed-point iteration.

Although the fixed-point iteration is a very efficient method, the class of distributions $A$ for which it can be applied is clearly smaller than the class of distributions $A$ for which the Fourier series representation holds. That is, Condition 3.2.9 is much weaker than Condition 4.2.2.

4.4 On a result for a right-continuous random walk

We have seen that the derivation of the Fourier series expression for the roots involves the determination of a region of analyticity as defined by the interior of the Szegö curve, and the application of Lagrange’s inversion theorem. Due to this highly analytical character of the derivation, the coefficients

$$c_l = \frac{1}{l} C_{l-1} \left[ A^{1/s}(z) \right]$$  \hspace{1cm} (4.35)

have, in most cases, no probabilistic interpretation. An exception to this is when $G(z) = A^{1/s}(z)$ is again (like $A(z)$) a pgf. Then, obviously

$$c_l = \frac{1}{l} P(G^*l = l - 1),$$  \hspace{1cm} (4.36)

where

$$G^*l = G_1 + \ldots + G_l \; ; \; G_i \sim G \; \text{i.i.d.}$$  \hspace{1cm} (4.37)

For this case, an interesting result is the following:

**Lemma 4.4.1** For a discrete random variable $G$ that follows a distribution with pgf $G(z)$ and mean $G'(1) < 1$, it holds that

$$\sum_{l=1}^{\infty} P(G^*l = l - 1) = \frac{1}{1 - G'(1)}. $$  \hspace{1cm} (4.38)

**Proof** Follows from substituting $w = 1$ into (4.15). \hfill \Box

Lemma 4.4.1 leads to a variety of equalities. For instance, for the Poisson case with $G(z) = \exp(\lambda(z - 1))$ and $0 \leq \lambda < 1$, (4.38) yields

$$\sum_{l=1}^{\infty} e^{-\lambda l} \frac{(\lambda l)^{l-1}}{(l-1)!} = \frac{1}{1 - \lambda},$$  \hspace{1cm} (4.39)

for which an independent derivation seems hard to give. However, one does expect an equality of this type to have some probabilistic interpretation. Indeed, a direct proof of Lemma 4.4.1 can be given, both analytically and probabilistically. In that respect, we prove the following extended version of Lemma 4.4.1:
Lemma 4.4.2  For a discrete random variable $G$ that follows a distribution with probability generating function $G(z)$ and mean $G'(1) < 1$, we have that

\[
\sum_{l=1}^{\infty} P(G^*l = l - m) = \frac{1}{1 - G'(1)}, \quad m = 0, 1, \ldots
\]  

(4.40)

Proof  The analytic proof does not rely on Lagrange’s inversion theorem. Instead, we have that

\[
\sum_{l=1}^{\infty} P(G^*l = l - m) = \sum_{i=1}^{\infty} \frac{1}{2\pi i} \oint_{\mathcal{C}} \frac{G^i(z)}{z^{l-m+1}} dz, \quad m = 0, 1, \ldots
\]  

(4.41)

where $\mathcal{C}$ is an arbitrary contour around the origin within the analyticity region of $G(z)$. The radius of convergence of $G(z)$, denoted by $\delta$, satisfies $\delta > 1$. Take as the contour $\mathcal{C}$ the circle \{ $z \in \mathbb{C} : |z| = 1 + \epsilon$ \}, with $\epsilon$ such that $0 < \epsilon < \delta - 1$ and $|G(z)| < |z|$ for all $z$ with $|z| = 1 + \epsilon$. Note that the latter can be realized because $G'(1) < 1$ and taking $\epsilon$ sufficiently small. This gives

\[
\sum_{i=1}^{\infty} \frac{1}{2\pi i} \oint_{\mathcal{C}} \frac{G^i(z)}{z^{l-m+1}} dz = \frac{1}{2\pi i} \oint_{\mathcal{C}} \frac{z^{m-1} \sum_{i=1}^{\infty} \left( \frac{G(z)}{z} \right)^i dz}{z - G(z)} = \frac{1}{2\pi i} \oint_{\mathcal{C}} \frac{z^{m-1} G(z)}{z - G(z)} dz.
\]  

(4.42)

Since $|G(z)| < |z|$ for all $z$ with $|z| = 1 + \epsilon$, it follows from Rouché’s theorem that $z - G(z)$ has exactly one zero in $|z| \leq 1 + \epsilon$, which can be easily seen to be $z = 1$. Consequently, $z^{m-1} G(z)/(z - G(z))$ has one singularity in $|z| \leq 1 + \epsilon$, and we thus have by the Cauchy residue theorem that

\[
\frac{1}{2\pi i} \oint_{\mathcal{C}} \frac{z^{m-1} G(z)}{z - G(z)} dz = \lim_{z \to 1} (z - 1) \frac{z^{m-1} G(z)}{z - G(z)} = \frac{1}{1 - G'(1)},
\]

(4.43)

which concludes the proof.

Alternative proof of Lemma 4.4.2.  A probabilistic proof follows from random walk theory, and can be found in e.g. Spitzer [146], p. 33. For that, we introduce

\[
\bar{G}_i := 1 - G_i; \quad \bar{S}_l := \sum_{i=1}^{l} \bar{G}_i, \quad \bar{S}_0 = 0.
\]  

(4.44)

Then, $\bar{S}_l$ is a right-continuous random walk (cannot skip a point in going to the right) with positive mean $\mu := 1 - G'(1)$. For random walks of this type, we have that

\[
\sum_{l=1}^{\infty} P(G^*l = l - m) = \sum_{l=1}^{\infty} P(\bar{S}_l = m), \quad m = 0, 1, \ldots
\]  

(4.45)
represents the total number of expected visits of the random walk to point $m$. In Spitzer [146] it is proved, using $\bar{S}_l \to \infty$ with probability 1, that
\[ \sum_{l=1}^{\infty} P(\bar{S}_l = 0) = \sum_{l=1}^{\infty} P(\bar{S}_l = 1) = \sum_{l=1}^{\infty} P(\bar{S}_l = 2) = \ldots = \frac{1}{\mu}, \quad (4.46) \]
which concludes the proof. \hfill $\square$

### 4.5 Numerical results

We now present some numerical results. First, we discuss two examples for which the roots can be determined by fixed-point iteration, and compare the results with the roots obtained from the Fourier series representation for several truncation levels. Next, we present three examples of the discrete bulk service queue in which we compare the different expressions for stationary queue length characteristics as presented in Chapter 2. In particular, we investigate the sensitivity of the results with respect to the method used for root-finding.

For each root determined with the fixed-point iteration (4.8), we stop the iteration when
\[ |z_k^{(n+1)} - z_k^{(n)}| < 10^{-14}, \]
and we denote the resulting values by $\hat{z}_k$. Denote by $z_k(L)$ the estimated root value that results when we truncate the infinite series over $l$ in (4.2) at $l = L$. We would like to have some more insight in how fast $z_k(L)$ converges to $z_k$, where the $\hat{z}_k$ determined above are considered to be sufficiently accurate approximations of the $z_k$ to serve as references.

**Example 4.5.1** Consider the Poisson case, $A(z) = \exp(\lambda(z - 1))$, $\lambda = 8$, $s = 10$, Table 4.1 displays the roots $\hat{z}_k$, along with the distance between $\hat{z}_k$ and $z_k(L)$ for $L = 10, 20, 50$. As it appears, the Fourier series (4.2) converge quite rapidly. We further note that the series for $z_0(L)$ is most slowly convergent.

| $k$ | $\text{Re } \hat{z}_k$ | $\text{Im } \hat{z}_k$ | $|z_k(10) - \hat{z}_k|$ | $|z_k(20) - \hat{z}_k|$ | $|z_k(50) - \hat{z}_k|$ |
|-----|----------------|----------------|----------------|----------------|----------------|
| 0   | 1.000000      | 0.000000       | 0.110194       | 0.048179       | 0.009637       |
| 1   | 0.300438      | 0.486051       | 0.017461       | 0.005283       | 0.000694       |
| 2   | -0.017701     | 0.442657       | 0.009539       | 0.002817       | 0.000366       |
| 3   | -0.205881     | 0.320697       | 0.006988       | 0.002052       | 0.000266       |
| 4   | -0.308844     | 0.166704       | 0.005961       | 0.001747       | 0.000226       |
| 5   | -0.341824     | 0.000000       | 0.005673       | 0.001662       | 0.000215       |

Table 4.1: Poisson distribution, $\lambda = 8$, $s = 10$. The roots of $z^s = A(z)$ for $|z| \leq 1$ determined with (4.8) (denoted as $\hat{z}_k$), along with the distance between $\hat{z}_k$ and $z_k(L)$ for $L = 10, 20, 50$. 
Example 4.5.2 Consider the binomial case, \( A(z) = (p + qz)^n \) where \( p, q \geq 0 \), \( p + q = 1 \), for which we take \( n = 16, q = 0.5 \) and \( s = 10 \). Table 4.2 displays \( \hat{z}_k \) and \( |z_k(L) - \hat{z}_k|, L = 10, 20, 50 \). Again, the series \( z_0(L) \) is most slowly convergent.

Table 4.2: Binomial distribution, \( n = 16, q = 0.5, s = 10 \). The roots of \( z^{s} = A(z) \) for \( |z| \leq 1 \) determined with (4.8) (denoted as \( \hat{z}_k \)), along with the distance between \( \hat{z}_k \) and \( z_k(L) \) for \( L = 10, 20, 50 \).

| \( k \) | Re \( \hat{z}_k \) | Im \( \hat{z}_k \) | \( |z_k(10) - \hat{z}_k| \) | \( |z_k(20) - \hat{z}_k| \) | \( |z_k(50) - \hat{z}_k| \) |
|---|---|---|---|---|---|
| 0 | 1.000000 | 0.000000 | 0.118685 | 0.037943 | 0.003067 |
| 1 | 0.169044 | 0.439341 | 0.024368 | 0.006010 | 0.000364 |
| 2 | -0.066258 | 0.315413 | 0.013329 | 0.003216 | 0.000192 |
| 3 | -0.164522 | 0.199596 | 0.009766 | 0.002344 | 0.000140 |
| 4 | -0.208378 | 0.096590 | 0.008330 | 0.001996 | 0.000119 |
| 5 | -0.221147 | 0.000000 | 0.007928 | 0.001899 | 0.000113 |

In the above examples, the series for \( z_0(L) \) is most slowly convergent, which can be explained by the following result:

**Lemma 4.5.3** The truncation error \( |z_k(L) - z_k| \) is largest for \( k = 0 \) among all \( k = 0, 1, \ldots, s - 1 \), when the coefficients \( c_l \geq 0 \), \( l = 1, 2, \ldots \).

**Proof** Follows directly from (4.2). \( \square \)

For the Poisson case the \( c_l \) are non-negative, indeed. This is also the case for e.g. the geometric distribution \( a_j = (1 - p)p^j \) with \( 0 \leq p < 1 \), and for the binomial distribution in Sec. 4.3 with \( \beta \geq 1 \), but it fails to hold for the latter distribution with \( 0 < \beta < 1 \).

In general, if one applies (4.2) to a distribution \( A \) for which \( c_l \geq 0 \), then \( |z_0(L) - z_0| = |z_0(L) - 1| \) being small is a good test for convergence, since it reflects the maximum distance between the estimated and true values of the roots. We stress that there are many distributions \( A \) for which the iteration (4.8) fails to work, while (4.2) still holds, i.e. Condition 3.2.9 is satisfied (see e.g. Sec. 4.3). We simply chose the above examples so that we could obtain precise estimates of the real roots without invoking some other, less transparent, numerical method than (4.8).

Let us now investigate what the impact is of the root-finding on characteristics of the stationary queue length in the discrete bulk service queue.

**Example 4.5.4** Consider the Poisson case, \( A(z) = \exp(\lambda(z - 1)) \), \( \lambda < s \). For \( s = 10 \) and \( \lambda = 5, 8, 9 \), Table 4.3 displays the mean and variance of the stationary queue length in the discrete bulk service queue. For the results obtained from (2.21) and (2.22) we have determined the roots using (4.8) with stopping criterion \( |z_k^{(n+1)} - z_k^{(n)}| < 10^{-14} \). For the results obtained from the infinite series expressions (2.39) and (2.40) we have truncated the sum over \( l \) at \( l = 30 \) and the sum over \( i \) at \( i = 300 \). We observe that the higher the load, the higher we should choose the
Table 4.3: Mean and variance of $X$ for the Poisson case with $s = 10$, $\lambda = 5, 8, 9$.

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\mu_X$</th>
<th>$\sigma_X^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda = 5$</td>
<td>5.0237</td>
<td>5.0519</td>
</tr>
<tr>
<td>$\lambda = 8$</td>
<td>8.8786</td>
<td>11.6109</td>
</tr>
<tr>
<td>$\lambda = 9$</td>
<td>12.1012</td>
<td>29.9067</td>
</tr>
</tbody>
</table>

truncation level. For $\lambda = 5$ and $\lambda = 8$ the truncation levels chosen are sufficient, while for $\lambda = 9$ they should be taken somewhat higher.

Example 4.5.5 We take the example considered in Chaudhry & Kim [47], in which $A(z) = Y(z)^6$ where

$$Y(z) = 0.1+0.15z+0.2z^2+0.2z^3+0.15z^4+0.1z^5+0.05z^6+0.01z^7+0.01z^8+0.03z^{10}. \tag{4.48}$$

In [47] the stationary queue length distribution is determined from (2.30), for which the zeros outside the unit circle are determined numerically using the computer package QROOT. The iteration (4.8) does not work for this example. We calculate the stationary queue length distribution from (2.12), (2.30) and (2.41). For (2.12) and (2.30) we calculate the roots of $z^* = A(z)$ inside and outside the unit circle using (4.2) and (4.5), respectively. For (4.2), (4.5) and (2.41) we truncate the sum over $l$ at $l = 50$. The results are displayed in Table 4.4.

Table 4.4: Stationary queue length distribution for $A(z) = Y(z)^6$, with $Y(z)$ given in (4.48), $s = 30$.

<table>
<thead>
<tr>
<th>$j$</th>
<th>$x_j$ from [47]</th>
<th>$x_j$ from (2.12)</th>
<th>$x_j$ from (2.30)</th>
<th>$x_j$ from (2.41)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000000098</td>
<td>0.000000098</td>
<td>0.000000098</td>
<td>0.000000098</td>
</tr>
<tr>
<td>1</td>
<td>0.000000885</td>
<td>0.000000885</td>
<td>0.000000885</td>
<td>0.000000885</td>
</tr>
<tr>
<td>2</td>
<td>0.000004501</td>
<td>0.000004501</td>
<td>0.000004501</td>
<td>0.000004501</td>
</tr>
<tr>
<td>3</td>
<td>0.000016868</td>
<td>0.000016868</td>
<td>0.000016868</td>
<td>0.000016868</td>
</tr>
<tr>
<td>4</td>
<td>0.000049733</td>
<td>0.000049715</td>
<td>0.000049733</td>
<td>0.000049733</td>
</tr>
<tr>
<td>5</td>
<td>0.00125555</td>
<td>0.00125503</td>
<td>0.00125555</td>
<td>0.00125555</td>
</tr>
<tr>
<td>6</td>
<td>0.00277138</td>
<td>0.00277010</td>
<td>0.00277138</td>
<td>0.00277138</td>
</tr>
<tr>
<td>7</td>
<td>0.00546268</td>
<td>0.00545988</td>
<td>0.00546269</td>
<td>0.00546268</td>
</tr>
<tr>
<td>8</td>
<td>0.00976060</td>
<td>0.00975504</td>
<td>0.00976060</td>
<td>0.00976060</td>
</tr>
<tr>
<td>9</td>
<td>0.01598541</td>
<td>0.01597540</td>
<td>0.01598541</td>
<td>0.01598541</td>
</tr>
<tr>
<td>10</td>
<td>0.02420260</td>
<td>0.02418598</td>
<td>0.02420260</td>
<td>0.02420260</td>
</tr>
<tr>
<td>20</td>
<td>0.06498585</td>
<td>0.06487376</td>
<td>0.06498585</td>
<td>0.06498585</td>
</tr>
<tr>
<td>30</td>
<td>0.00728773</td>
<td>0.00661255</td>
<td>0.00728773</td>
<td>0.00728773</td>
</tr>
<tr>
<td>40</td>
<td>0.00015022</td>
<td>0.00049559</td>
<td>0.00015022</td>
<td>0.00015022</td>
</tr>
<tr>
<td>50</td>
<td>0.00000080</td>
<td>0.00072575</td>
<td>0.00000080</td>
<td>0.00000080</td>
</tr>
</tbody>
</table>
We see that both (2.30) and (2.41) lead to similar results as obtained in [47].

Determining the probabilities from (2.12) gives problems when moving into the tail of the distribution. Although these problems might be resolved by truncating the sum over \( l \) in (4.2) at a higher level, (2.30) and (2.41) seem more stable. The truncation level of \( l = 50 \) is by far sufficient, although it is no problem to increase it from a numerical point of view.

**Example 4.5.6** Consider the binomial case, \( A(z) = (p + qz)^n \) where \( p, q \geq 0, p + q = 1 \), for which we take \( n = 16, q = 0.5, s = 10 \). Table 4.5 displays some of the \( x_j \), calculated by \( x_j(L) \) for \( L = 10, 20, 30 \). Additionally, the \( x_j \) have been determined from (2.30) where the roots of \( z^* = A(z) \) outside the unit circle follow from (4.5) (with the sum over \( l \) truncated at \( l = 50 \)). Note that for \( x_{50} \) and \( x_{100} \) we need some higher level of \( L \) to determine these small probabilities up to a reasonable accuracy. Increasing \( L \) would give no numerical difficulties, so that the accuracy is just a matter of choice. This makes this approach well-suited for calculating tail probabilities.

<table>
<thead>
<tr>
<th>( j )</th>
<th>( x_j(10) )</th>
<th>( x_j(20) )</th>
<th>( x_j(30) )</th>
<th>(2.30)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.13132067(\times 10^{-4})</td>
<td>0.13131227(\times 10^{-4})</td>
<td>0.13131225(\times 10^{-4})</td>
<td>0.13131228(\times 10^{-4})</td>
</tr>
<tr>
<td>10</td>
<td>0.12967227(\times 10^{-0})</td>
<td>0.12967413(\times 10^{-0})</td>
<td>0.12967413(\times 10^{-0})</td>
<td>0.12967413(\times 10^{-0})</td>
</tr>
<tr>
<td>20</td>
<td>0.10032061(\times 10^{-4})</td>
<td>0.10120244(\times 10^{-4})</td>
<td>0.10120578(\times 10^{-4})</td>
<td>0.10120543(\times 10^{-4})</td>
</tr>
<tr>
<td>30</td>
<td>0.25745503(\times 10^{-9})</td>
<td>0.29484924(\times 10^{-9})</td>
<td>0.29484924(\times 10^{-9})</td>
<td>0.29484924(\times 10^{-9})</td>
</tr>
<tr>
<td>50</td>
<td>0.07585901(\times 10^{-18})</td>
<td>0.11512297(\times 10^{-18})</td>
<td>0.11512297(\times 10^{-18})</td>
<td>0.11512297(\times 10^{-18})</td>
</tr>
<tr>
<td>70</td>
<td>0.01219941(\times 10^{-27})</td>
<td>0.09314437(\times 10^{-41})</td>
<td>0.09314437(\times 10^{-41})</td>
<td>0.09314437(\times 10^{-41})</td>
</tr>
<tr>
<td>100</td>
<td>0.00126202(\times 10^{-41})</td>
<td>0.00126202(\times 10^{-41})</td>
<td>0.00126202(\times 10^{-41})</td>
<td>0.00126202(\times 10^{-41})</td>
</tr>
</tbody>
</table>
Chapter 5

Relaxation time

In the previous chapters we have seen that characteristics of the stationary queue length in the discrete bulk service queue can be expressed in terms of either the roots of $z^* = A(z)$ or an infinite series that involves convolutions of the distribution of $A$. In Chapter 4 we have thoroughly investigated both the explicit description and the numerical determination of the roots.

In this chapter we investigate the infinite series that involve convolutions of the distribution of $A$. For practical purposes, the infinite series should be truncated. We therefore seek for some means to characterize the speed with which these series converge. In queueing theory, such a characterization is related to the notion of relaxation time, a generic term for the time required for transient system characteristics to tend to their stationary values. We derive asymptotic expressions for the relaxation time in a purely analytical way, mostly relying on the saddle point method. We present a simple and useful upper bound which may serve as a stopping criterion for the level at which one should truncate the infinite series in case the load on the system is not very high. A more detailed upper bound is then developed that continues to be sharp for high loads.

Researchers that have dealt with the relaxation time in queueing systems have presented their theory along the lines of the waiting time (see Cohen [51], p. 600, Blanc & Van Doorn [35], Heathcote [87] and Heathcote & Winer [88]). It is therefore that we choose to present the results in this chapter along the lines of the waiting time in the $D/G/1$ queue. We have seen that the discrete $D/G/1$ queue is essentially equivalent with the discrete bulk service queue, and so the results presented in this chapter lead to similar results for the discrete bulk service queue. The results also hold for the discrete approximation to the $G/G/1$ queue as discussed in Remark 2.3.5.

To make this chapter self-contained, we briefly repeat some of the arguments and results presented in Chapter 2. An overview is given at the end of the next section. The chapter is based on Janssen & Van Leeuwaarden [P9].
5.1 Introduction and motivation

The discrete $D/G/1$ queue refers to a single-server queue at which customers arrive with discrete and deterministic interarrival times. We assume that customers are served on a first-come-first-served basis and that their service requirements are i.i.d. according to a discrete random variable $A$. The waiting time of the $n$th customer, denoted by $W_n$, then satisfies (see e.g. Servi [142])

$$W_{n+1} = (W_n + A_n - s)^+, \quad n = 0, 1, \ldots. \quad (5.1)$$

Here, $x^+ = \max\{0, x\}$, $W_0 = 0$, $A_n$ denotes the service time of customer $n$ and the integer $s$ denotes the fixed interarrival time between two consecutive customers. When $E(A) < s$, the stationary waiting time, denoted by $W$, exists.

The stationary waiting time in the discrete $D/G/1$ queue as defined by (5.1) is then fully specified by its pgf

$$W(z) = \frac{s - E(A)}{z^s - A(z)} (z - 1) \prod_{k=1}^{s-1} \left( \frac{z}{1-z_k} \right), \quad (5.2)$$

where $z_0 = 1, z_1, \ldots, z_{s-1}$ are the $s$ roots of $z^s = A(z)$ in $|z| \leq 1$. We have stressed earlier (p. 27) that $X(z) = W(z)A(z)$ where $X(z)$ is the pgf of the stationary queue length distribution in the discrete bulk service queue.

We now show that Spitzer’s identity results for the discrete $D/G/1$ queue in manageable expressions for both transient and stationary waiting-time characteristics.

Using (5.1), the distribution of $W_{n+1}$ follows from the convolution of the distribution of $W_n$ and that of $A_n - s$, corrected for the maximum operator. The idea of iterating (5.1) to obtain transient waiting-time characteristics can be made more rigorous using random walk theory. When we assume that the first customer (referred to by subscript 0) arrives at an empty queue ($W_0 = 0$), the joint probability generating function of all $W_n$ is given by Spitzer’s identity (see Thm. 2.3.1), which says that for $0 \leq t < 1$, $|z| \leq 1$,

$$\sum_{n=0}^{\infty} t^n E[z^n W_n] = \exp \left\{ \sum_{l=1}^{\infty} t^{l-1} E[z^l S_l^+] \right\}, \quad (5.3)$$

with $S_l = \sum_{i=1}^{l} (A_i - s)$, $A_i$ i.i.d. as $A$. It follows from manipulating (5.3) that the mean of $W_n$ is given by

$$E[W_n] = \sum_{l=1}^{n} l^{-1} E[S_l^+]$$
$$= \sum_{l=1}^{n} l^{-1} \sum_{i=ls+1}^{\infty} (i - ls) P(A^* = i), \quad (5.4)$$

where $A^* = \sum_{i=1}^{l} A_i$ with $A_i$ i.i.d. as $A$. 

From (5.3) the stationary waiting-time distribution can be obtained as well. When we write (5.3) as

\[(1 - t) \sum_{n=0}^{\infty} t^n E_z W_n = \exp \left\{ \sum_{l=1}^{\infty} l^{-1} (E_z S_l^+ - 1) \right\}, \tag{5.5}\]

it follows from Abel’s theorem (see Spitzer [146], p. 207, Cohen [51], p. 650), that \(W(z)\) is given by

\[W(z) = \lim_{t \to 1} (1 - t) \sum_{n=0}^{\infty} t^n E_z W_n = \exp \left\{ \sum_{l=1}^{\infty} l^{-1} (E_z S_l^+ - 1) \right\} = \exp \left\{ - \sum_{l=1}^{\infty} l^{-1} P(S_l > 0) \right\} \exp \left\{ \sum_{l=1}^{\infty} l^{-1} E(z S_l) \mathbb{1}\{S_l > 0\} \right\}, \tag{5.6}\]

where \(\mathbb{1}\{x\}\) equals 1 if \(x\) true and 0 otherwise. Again using the short-hand notation \(C_z[h(z)]\) for the coefficient of \(z^j\) in \(h(z)\), and \(w_j\) for \(P(W = j)\), it is readily seen that the stationary waiting-time distribution is given by

\[w_j w_0 = C_z \left\{ \exp \left\{ \sum_{l=1}^{\infty} l^{-1} \sum_{i=ls+1}^{\infty} P(A^{*l} = i) z^{-i} \right\} \right\}, \quad j = 0, 1, \ldots, \tag{5.7}\]

where

\[w_0 = \exp \left\{ - \sum_{l=1}^{\infty} l^{-1} \sum_{i=ls+1}^{\infty} P(A^{*l} = i) \right\}. \tag{5.8}\]

Expressions (5.4), (5.7) and (5.8) provide explicit representations of waiting-time characteristics solely in terms of infinite series of convolutions of \(A\). Calculating these characteristics is a matter of brute force and the applicability strongly depends on the ability of computing the discrete convolutions involved. An easy way would be to determine the distribution of \(A^{*l}\) from the distribution of \(A^{*(l-1)}\). As suggested in Ackroyd [21], it is better, though, to apply a fast Fourier transform algorithm. In that way, given the pgf \(A(z)\), the probability distribution of the \(l\)-fold convolution can be obtained directly from its pgf \(A^l(z)\). In Ackroyd [21] it is shown that the gain in computational speed is considerable. For a description of the fast Fourier transform approach to invert a pgf we refer to Abate & Whitt [17].

### 5.1.1 Relaxation time

Irrespective of the method used to compute the convolutions, the issue of truncating the infinite series should be addressed. It is therefore that we seek for some means to characterize the speed at which these series converge. The issue of truncating turns out to be intimately related with a notion in the queueing literature known as relaxation time. The relaxation time is a generic term for time required for transient system characteristics to tend to their steady-state values. When the
Relaxation time

Relaxation time would be defined in terms of the mean waiting time, it could be expressed as the speed at which the difference (with $\mathbb{E}W = \mathbb{E}W_\infty$)

$$\mathbb{E}W - \mathbb{E}W_{L-1} = \sum_{l=L}^{\infty} l^{-1} S^+_l$$

$$= \sum_{l=L}^{\infty} l^{-1} \sum_{i=ls+1}^{\infty} (i - ls) \mathbb{P}(A^*l = i)$$

(5.9)

tends to zero for increasing values of $L$.

Another common way to define the relaxation time is in terms of the probability that a customer has zero waiting time (see e.g. Blanc & Van Doorn [35]), since determining $w_0$ is often the bottleneck. For the discrete $D/G/1$ queue this can be seen as follows. Denote by $w_j(L)$ the estimated value of $w_j$ that results from truncating the series over $l$ at $l = L - 1$ in (5.7) and (5.8), respectively. The relative error made in estimating $w_0$ then equals

$$\frac{w_0(L) - w_0}{w_0} = \exp \left\{ \sum_{l=L}^{\infty} l^{-1} \sum_{i=ls+1}^{\infty} \mathbb{P}(A^*l = i) \right\} - 1$$

$$\approx \sum_{l=L}^{\infty} l^{-1} \sum_{i=ls+1}^{\infty} \mathbb{P}(A^*l = i),$$

(5.10)

where the far right-hand side of (5.10) sums all truncation errors $\sum_{l=L}^{\infty} l^{-1} \mathbb{P}(A^*l = i)$ that appear in (5.7) when estimating $w_j$ by $w_j(L)$. Hence, when the left-hand side of (5.10) is small enough, the accuracy of the estimated values of all $w_j$ seems guaranteed.

A third way to define the relaxation time is in terms of the variance of the waiting time, whose stationary value $\sigma^2_W$ follows from (2.35) by $\sigma^2_W = W''(1) + W'(1) - W'(1)^2$ yielding

$$\sigma^2_W = \sum_{l=1}^{\infty} l^{-1} \sum_{i=ls+1}^{\infty} (i - ls)^2 \mathbb{P}(A^*l = i).$$

(5.11)

When we denote by $\sigma^2_W(L)$ the series (5.11) over $l$ truncated at $l = L - 1$, the relaxation time can be expressed as the speed at which the difference

$$\sigma^2_W - \sigma^2_W(L) = \sum_{l=L}^{\infty} l^{-1} \sum_{i=ls+1}^{\infty} (i - ls)^2 \mathbb{P}(A^*l = i)$$

(5.12)

tends to zero.

In order to extract information from the above measures on the relaxation time, we need insight in the behavior as $L \to \infty$ of the tail series

$$R_m(L) := \sum_{l=L}^{\infty} S_m(l), \quad m = 0, 1, 2,$$

(5.13)
where

\[ S_m(l) = t^{-1} \sum_{i=ls+1}^{\infty} (i-ls) \mathbb{P}(A^*l = i). \]  

(5.14)

Formally, the relaxation time \( T(R_m; \epsilon) \) for \( R_m \) at level \( \epsilon > 0 \) can be defined as

\[ T(R_m; \epsilon) = \min \{ L \mid R_m(L) < \epsilon \}, \]  

(5.15)

although this definition is not very practical since it requires computation of all terms in the series defining \( R_m(L) \). In this chapter we present easily computable asymptotic approximations of \( R_m(L) \). In particular, we prove, see Thm. 5.2.3, that

\[ R_m(L) \approx f(x, m, L)x^L, \quad x \in [0, 1), \]  

(5.16)

where \( \approx \) in (5.16) indicates a relative error \( O(x/(L(1-x))) \). We present a detailed description of the function \( f(x, m, L) \), along with a sharpening of (5.16) for values of \( x \) close to 1. Then, one can replace the \( R_m(L) \) in the definition of \( T(R_m; \epsilon) \) in (5.15) by (5.16) (or its sharpened version) to obtain information on the relaxation time.

For the continuous \( G/G/1 \) queue, the relaxation time in terms of the virtual waiting time has been studied extensively by Cohen [51], p. 600, based on analytic continuation of a Laplace transform and the saddle point method. An overview and continuation of this work is given in Blanc & Van Doorn [35]. In terms of moments of the actual waiting time, expressions for the relaxation time using a change of measure or large-deviations technique are obtained in Heathcote [87] and Heathcote & Winer [88] (also see Asmussen [25], p. 355).

The main contribution in this chapter is that we derive relaxation time asymptotics for the discrete \( D/G/1 \) queue in a concise and purely analytical way. We start from a simple asymptotic approximation of the \( \mathbb{P}(A^*l = i) \) that appear in (5.14) obtained from using the saddle point method. From this classical result, using the specific structure of the discrete \( D/G/1 \) queue, we derive asymptotic expressions for \( S_m(l) \) and \( R_m(L) \), where the latter will allow us to calculate a good approximation of \( T(R_m; \epsilon) \) in (5.15). As a first result, we present an asymptotic expression for \( R_m(L) \) based on this asymptotic approximation. This expression admits a simple and useful upper bound. A sharpening of this upper bound, which involves the complementary error function, is then developed and this covers both the cases of low and high loads. For an overview of previous work on queueing models that relies on the saddle point method we refer to Abate et al. [16] and Asmussen [25] and the references therein. Also, in Abate & Whitt [18] the saddle point method has been applied to investigate truncation of infinite series representations of Laplace transforms.

In Sec. 5.2 we present the main results, which are proved in Secs. 5.3, 5.4 and 5.5. Examples are provided in Sec. 5.6.
5.2 Results

Denote $P(A = n)$ by $a_n$. Let $z_\infty$ be the radius of convergence of the series $\sum_{n=0}^{\infty} a_n z^n$, and let

$$L_A = \lim_{z \rightarrow z_\infty} \frac{z A'(z)}{A(z)}, \quad (5.17)$$

In Sec. 5.7 we show that the limit in (5.17) always exists as a finite or infinite number, and that $A'(1) < L_A$ unless $A$ is a monomial (i.e. of the form $z^n$ for some $n \geq 0$). In Sec. 5.3 we obtain an asymptotic approximation of

$$P(A^i = i) = \frac{1}{2\pi i} \oint_C \frac{A'(z)}{z^{i+1}} \, dz, \quad (5.18)$$

where $i = \sqrt{-1}$ and $C$ is any contour around 0 within the analyticity region of $A(z)$. Using the saddle point method, see De Bruijn [41], we find the following theorem:

**Theorem 5.2.1** Assume that $|A(e^{i\theta})|$ is strictly maximal at $\theta = 0$ as a function of $\theta \in [-\pi, \pi]$. Let $i, l \geq 0$ be integers such that $A'(1) \leq i/l < L_A$, and denote $h(z) = i \ln A(z) - i \ln z$. Then there is a unique solution $z = z_0 \in [1, z_\infty)$ of the equation $h'(z) = 0$, we have $h''(z_0) > 0$ and

$$P(A^i = i) \approx \frac{1}{z_0 \sqrt{2\pi h''(z_0)}} \frac{A'(z_0)}{z_0^i}, \quad (5.19)$$

where $\approx$ means to indicate a relative error $O(1/l)$ uniformly in $i$ and $l$ when there is a $\delta > 0$ such that $i/l \in [s, (1 + \delta)s]$.

Thm. 5.2.1 is a standard classical result. It concerns the probability distribution of the sum of i.i.d. random variables far from its mean, which is classically treated via saddle point methods. Related material can be found in e.g. Asmussen [25], Jensen [92] and Lauwerier [110]. Details of the derivation of Thm. 5.2.1 are given in Sec. 5.3.

The assumption $A'(1) \leq i/l < L_A$ in Thm. 5.2.1 ensures the existence of a saddle point on the positive real axis for the integral in (5.18). The fact that we have to consider integers $i, l \geq 0$ such that $A'(1) \leq i/l < L_A$ is not very restrictive in the present context. To see this, first note that the series defining $S_m(l)$ in (5.14) involve $i \geq ls$ while we have made the assumption $A'(1) < s$. Secondly, we have in many cases that $L_A = \infty$, see the examples in Sec. 5.6. Finally, the cases where the load $\rho = A'(1)/s$ is not far away from the maximum sustainable value 1 are the more interesting ones. We shall see in Sec. 5.4 that, for an accurate approximation of $S_m(l)$ in (5.14), it is sufficient to consider $i$ for which $i/l$ is not much larger than $s$. Hence, even in the case of finite $L_A$, the more interesting cases allow one to restrict to $s$ and $i, l$ satisfying $A'(1) < s \leq i/l < L_A$.

The assumption that $|A(e^{i\theta})|, \theta \in [-\pi, \pi]$, is strictly maximal at $\theta = 0$ allows us to restrict the attention to the immediate vicinity of the saddle point on the
positive real axis when the contour $C$ in (5.18) is taken to be the circle around zero passing through the saddle point. This condition is not restrictive either. Due to the non-negativity of the $a_n$ and the fact that $a_0 > 0$, the condition is contravened only for $A(z)$ of the form $B(z^p) = \sum_{i=0}^{\infty} b_i z^p$, where

$$p = \min\{|n_1 - n_2| : n_1, n_2 = 0, 1, \ldots, n_1 \neq n_2, a_{n_1} \neq 0 \neq a_{n_2}\} > 1. \quad (5.20)$$

This $B$ is a pgf, just like $A$, and it does satisfy the condition that $|B(e^{i\theta})|, \theta \in [-\pi, \pi]$, is strictly maximal at $\theta = 0$. If $s$ is a multiple of $p$, it suffices to consider $B$ and $s/p$ instead of $A$ and $s$. Much of the analysis given in this chapter applies when $A(z) = B(z^p)$ where $s$ is not a multiple of $p$, but the administration required for the series over $i$ in (5.7), (5.8), (5.9) and (5.11) becomes somewhat complicated due to the fact that $\mathbb{P}(A^s = i) \neq 0$ only when $i$ is a multiple of $p$; we shall exclude such $A$’s.

In all cases, irrespective of whether the conditions in Thm. 5.2.1 on $A$ and $i/l$ are satisfied or not, we have the following bound. For $i, l = 0, 1, \ldots$ we have that

$$\mathbb{P}(A^s = i) \leq \inf_{1 \leq s < z_\infty} \frac{A^s(z)}{z^s}. \quad (5.21)$$

In case that the conditions in Thm. 5.2.1 on $A$ and $i/l$ are satisfied, the number at the right-hand side of (5.21) equals $A'(z_0)/z_0'$ with $z_0$ as in Thm. 5.2.1. We note that in that case the right-hand sides of (5.19) and (5.21) basically differ by the factor $1/\sqrt{2\pi z_0^2 h''(z_0)}$. Normally, this factor is quite innocent, the key features of the bounds and approximations being determined by the crucial quantity $A'(z_0)/z_0'$.

For simplicity we shall assume now that $L_A = \infty$, and we denote

$$\hat{S}_m(l) = \sum_{i=1}^{\infty} \frac{1}{l} (i - ls)^m \frac{A'(z_0)}{z_0 \sqrt{2\pi h''(z_0)}} \frac{A^s(z_0)}{z_0^s}. \quad (5.22)$$

**Theorem 5.2.2** We have

$$S_m(l) \approx \hat{S}_m(l) \approx \frac{l^{-3/2}}{\sqrt{2\pi z_0 \phi(z)^s}} \left(\frac{A(\hat{z})}{\hat{z}^s}\right) \sum_{i=1}^{\infty} \frac{A^s(z_0)}{z_0^s} \hat{z}^{-i}, \quad (5.23)$$

where $\phi(z) = zA'(z)/A(z)$ and $\hat{z}$ is the unique $z \geq 1$ such that $\phi(z) = s$. Both $\approx$-signs mean to indicate a relative error $O(1/l)$.

This result is proved in Sec. 5.4. In fact, in many cases, the second $\approx$ in (5.23) holds as an upper bound on $\hat{S}_m(l)$. Moreover, we briefly consider the issue of how to modify Thm. 5.2.2 for the case that $L_A$ is finite.

The $\phi$ of Thm. 5.2.2 is considered in some detail in Sec. 5.7 and is related to $z_0$ of Thm. 5.2.1 as follows. When $A'(1) \leq t < L_A$ and $z_0(t)$ denotes the unique root $z \in [1, z_\infty)$ of $\phi(z) = t$, then $z_0 = z_0(i/l)$ for integer $i, l \geq 0$ such that $A'(1) \leq i/l < L_A$. 
The series $K_m(v) = \sum_{i=1}^{\infty} i^m v^i$, $m = 0, 1, 2, \ldots$, have been studied in some detail, see Lawden [111] and Stalley [148]. We only need the first few $K_m$’s. We have

$$K_0(v) = \frac{v}{1 - v}, \quad K_1(v) = \frac{v}{(1 - v)^2}, \quad K_2(v) = \frac{v^2}{(1 - v)^3} + \frac{v}{(1 - v)^2}.$$ \hspace{1cm} (5.24)

**Theorem 5.2.3** When $A'(1)/s$ stays away from 1 and $L \to \infty$, there holds for the $R_m(L)$ in (5.13) that

$$R_m(L) \approx \hat{R}_m(L) := \sum_{l=L}^{\infty} S_m(l) \approx \frac{K_m(\hat{z}^{-1}) x^L}{\sqrt{2\pi \hat{z}'(\hat{z}) L^{3/2} (1 - x)}},$$ \hspace{1cm} (5.25)

where $x = A(\hat{z})/\hat{z}^s$. The first $\approx$ in (5.25) indicates a relative error $O(1/l)$, while the second $\approx$ in (5.25) indicates a relative error $O(x/(L(1 - x)))$.

For the case that $A'(1)$ can be close to $s$ (so that both $\hat{z}$ and $x$ can be close to 1), we have the more powerful result

$$\hat{R}_m(L) \approx \frac{K_m(\hat{z}^{-1})}{\sqrt{2\pi \hat{z}'(\hat{z})}} \left[ 2x^{L-1} \sqrt{L^{3/2}} (1 - x) e^{\beta^2 \text{erfc}(\beta)} \right],$$ \hspace{1cm} (5.26)

with a relative error of the order $1/(L(1 + \beta^2/2))$, and where $\beta = \sqrt{(1 - x)L/x}$, and

$$\text{erfc}(\beta) = \frac{2}{\sqrt{\pi}} \int_{\beta}^{\infty} e^{-t^2} dt, \quad \beta \geq 0,$$ \hspace{1cm} (5.27)

denotes the complementary error function.

In Sec. 5.5 we present the proof of this result, and we pay more attention to the $\approx$ in (5.25) and (5.26). The sharpening in (5.26) requires a detailed study of the function $\sum_{l=L}^{\infty} l^{-3/2} x^l$ in which $L \to \infty$ and $x$ is allowed to vary through all values of $[0, 1]$.

**5.3 Details for Theorem 5.2.1**

In this section we use the saddle point method to prove Thm. 5.2.1, and we discuss the conditions on $A$ and $i, l$ that appear in the formulation of Thm. 5.2.1.

$A(z)$ is assumed to be an analytic function in a disk $|z| < z_\infty$, where the radius of convergence $z_\infty$ of $\sum_{n=0}^{\infty} a_n z^n$ exceeds 1. Hence

$$P(A^{i'l} = i) = \frac{1}{2\pi i} \oint_{C_r} \frac{A^i(z)}{z^{i+l}} dz,$$ \hspace{1cm} (5.28)

where $\iota = \sqrt{-1}$ and $C_r$ is any contour around 0 with radius $r \in [1, z_\infty)$. On such a circle we have by non-negativity of all $a_n$ that $|A^i(z)/z^l|$ is maximal at $z = r$. Hence we get at once that for all $i, l = 0, 1, \ldots$

$$P(A^{i'l} = i) \leq \frac{1}{2\pi} \cdot 2\pi r \cdot \frac{A^l(r)}{r^l} = \frac{A^l(r)}{r^l}.$$ \hspace{1cm} (5.29)
for any $r \in [1, z_\infty)$.

We consider now $i, l$ such that $i/l \geq A'(1)$. Under this assumption we have that $\frac{d}{dz} [A'(z)/z^i] \leq 0$ at $z = 1$. The information on $P(A^i = i)$ can now be made more precise when the infimum over $r \in [1, z_\infty)$ of the numbers at the right-hand side of (5.29) is actually assumed as a minimum at a point $r = z_0 \in [1, z_\infty)$. In that case the point $z_0$ is a saddle point of

$$h(z) := l \ln A(z) - i \ln z,$$

(5.30)

and it is tempting to apply the saddle point method, see De Bruijn [41], Chapter 5, to obtain an approximation of $P(A^i = i)$ of the form

$$P(A^i = i) \approx \frac{e^{h(z_0)}}{z_0 \sqrt{2\pi h''(z_0)}},$$

(5.31)

In order that this saddle point approach is valid, we must make some assumptions. First of all, we need to ask ourselves whether the saddle point $z_0$ really exists. Thus, with $\phi(z) = zA'(z)/A(z)$ we want

$$h'(z) = \frac{1}{z} \left( \phi(z) - \frac{i}{l} \right)$$

(5.32)

to have one zero $z = z_0(i/l) \in [1, z_\infty)$ at which we have $h''(z) > 0$. In Sec. 5.7 it will be shown that (unless $A$ is monomial) $\phi(z)$ is strictly increasing in $z \in [1, z_\infty)$. Hence $h'(z)$ has exactly one zero in $[1, z_\infty)$, provided that

$$\frac{i}{l} \in [\phi(1), \lim_{z \to z_\infty} \phi(z)] = [A'(1), L_A]$$

(5.33)

with $L_A$ as in (5.17). Furthermore, since

$$h''(z) = \frac{1}{z} \phi'(z) - \frac{1}{z^2} \left( \phi(z) - \frac{i}{l} \right),$$

(5.34)

we have that

$$h''(z_0(i/l)) = \frac{1}{z_0(i/l)} \phi'(z_0(i/l)) > 0.$$  

(5.35)

The second issue in validating the saddle point approach as embodied by (5.31) is the fact that we should be allowed to restrict the attention to only a small portion of the integration contour $C_{z_0}$ in (5.28) close to the saddle point $z_0$. To that end we make the assumption that $|A(e^{j\theta})|$, $\theta \in [-\pi, \pi]$, is strictly maximal at $\theta = 0$. Due to the non-negativity of $a_n$, this assumption is not really restrictive. Indeed, assuming that for some $\theta \neq 0$, $\theta \in [-\pi, \pi]$

$$|A(e^{j\theta})| = \sum_{n=0}^{\infty} a_n e^{jn\theta} = \sum_{n=0}^{\infty} a_n = A(1),$$

(5.36)
we see that there is a \( \gamma \in [-\pi, \pi] \) such that \( e^{in\theta} = e^{i\gamma} \) for all \( n = 0, 1, \ldots \) with \( a_n \neq 0 \).

It is easily seen that this strict maximality of \( |A(e^{i\theta})|, \theta \in [-\pi, \pi], \) at \( \theta = 0 \) is equivalent with strict maximality of \( |A(re^{i\theta})|, \theta \in [-\pi, \pi], \) at \( \theta = 0 \) for any \( r \in [1, z_{\infty}] \). As a consequence we can replace the integral along \( C_{z_0} \) by an integral along small circle segments \( \{z_0e^{i\theta}||\theta| \leq \delta \} \) at the expense of exponentially small errors of order

\[
\max_{\delta \leq |\theta| \leq \pi} \left| \frac{A(z_0)e^{i\theta}}{A(z_0)} \right|.
\] (5.37)

The term exponentially small used here can be somewhat deceptive as the following example shows. Choose

\[
A(z) = \frac{\cosh \lambda z + \epsilon z}{\cosh \lambda + \epsilon},
\] (5.38)

where \( \lambda > 0 \) is large and \( \epsilon > 0 \) is small. The ratio \( A(-z_0)/A(z_0) \) is extremely close to 1, and it is only for very large \( l \) that one can ignore the contribution to the integral of \( z \)'s near \(-z_0\).

The further details of applying the saddle point method for the present case follow to a large extent the discussion in De Bruijn [41], p. 92 on the range of the saddle point. Here, it is important to note that considerations in Sec. 5.4 show that we can restrict the attention to integers \( i, l \geq 0 \) such that \( i/l \) is not much larger than \( s \). Thus we can write

\[
h(z) = lg_t(z); \quad g_t(z) = \ln A(z) - t \ln z,
\] (5.39)

where \( t \in [s, (1 + \delta)s] \) with \( \delta > 0 \) not large and certainly such that \( (1 + \delta)s < L_A \). This implies that the \( z_0 \) are in a compact interval in \([1, z_{\infty}]\), whence the \( g''_t(z_0) \) are uniformly bounded away from 0 while higher derivatives, such as \( g^{(4)}_t(z_0) \) and \( g^{(6)}_t(z_0) \), are uniformly bounded away from infinity. Following the discussion in [41], p. 92, we then replace the integral along \( C_{z_0} \) by an integral along the line segment between \( z_0 - u^{-\gamma} \) and \( z_0 + u^{-\gamma} \), where \( \gamma \) is a real number between 1/3 and 1/2, at the expense of an exponentially small error like \( \exp(-\frac{1}{2}u^{1-2\gamma}g''_t(z_0)) \). On this line segment the remainder of \( lg_t(z) \), after splitting off the constant and quadratic term

\[
-1g_t(z_0) + \frac{1}{3}g''_t(z_0)(z - z_0)^2,
\]

\[
l(\frac{1}{3}g''_t(z_0)(z - z_0)^3 + \frac{1}{8}g^{(3)}_t(z_0)(z - z_0)^4 + \ldots)
\] (5.40)

tends to zero as \( l \to \infty \). Hence, at the expense of smaller errors, we can linearize \( \exp(h(z)) \) on the line segment \( z_0 + u, [u] \leq l^{-\gamma} \), as

\[
e^{h(z)} = e^{h(z_0)-\frac{1}{2}u^{2\gamma}h''(z_0)}(1 - \frac{1}{6}u^2g''_t(z_0)u^3 + \frac{1}{24}u^6g^{(3)}_t(z_0)u^4).
\] (5.41)

Now note that the term involving \( g''_t(z_0) \) cancels upon integration over \( u \in [-l^{-\gamma}, l^{-\gamma}] \) since \( u^3 \) is odd. The integral of the term involving \( g^{(3)}_t(z_0) \) can be estimated as

\[
e^{h(z_0)}\frac{1}{24}g^{(3)}_t(z_0)\int_{-\infty}^{\infty} e^{-\frac{1}{2}u^{2\gamma}h''(z_0)u^4}du = \Gamma(5/2)\frac{1}{24}g^{(3)}_t(z_0)e^{h(z_0)}\left(\frac{2}{h''(z_0)}\right)^{5/2}.
\] (5.42)
5.4 Details for Theorem 5.2.2

It follows that the relative error due to this latter term has the order

\[ l \left( \frac{2}{h''(z_0)} \right)^2 = \frac{4}{l(g''_1(z_0))^2} = O(1/l) \]  

(5.43)

uniformly in \( t \in [s, (1 + \delta)s] \). Similarly, the lowest-order deleted quadratic term \(-\frac{1}{3} l^2(g'''_1(z_0))^2u^6\) at the right-hand side of (5.41) produces a relative error \( O(1/l) \) as well, and higher-order terms produce smaller errors, etc. In all this, the additional factor \( 1/z \) that appears in the integral in (5.28) according to

\[ A(l(z) \equiv \frac{z^{l+1}}{z} e^{h(z)} \]  

(5.44)

has been considered as a constant \( 1/z_0 \). As in the above, this can be shown to be allowed, at the expense of a relative error of order \( O(1/l) \). We conclude that when we restrict \( i/l \) to a range in \([s, (1 + \delta)s] \subset [A'(1), L_A]\), the relative error for the approximation in (5.31) is \( O(1/l) \) uniformly in \( i \).

We conclude this section by a consideration of \( A \) for which \( L_A \) is finite (in many cases \( L_A = \infty \) so that the assumption \( i/l < L_A \) in Thm. 5.2.1 presents no restriction). First assume that \( z_\infty = \infty \), so that \( A(z) \) is an entire function. From \( L_A < \infty \) and the fact that \( a_n \geq 0 \) it then follows that \( A(z) \) is a polynomial of degree \( L_A \). Hence in this case \( P(A^1 = i) = 0 \) when \( i > L_A \). Next consider the case that \( z_\infty < \infty \) and \( L_A < \infty \). It is easy to see that then

\[ A(z_\infty) := \lim_{z \uparrow z_\infty} A(z), \quad A'(z_\infty) := \lim_{z \uparrow z_\infty} A'(z) \]  

(5.45)

exist as finite numbers. When now \( i/l > L_A \), a precise approximation is feasible only when additional information about the nature of the singular point \( z_\infty \) is available. However, the bound in (5.29) remains valid, and this is normally enough for our purposes where we may restrict to integers \( i/l \) such that \( i/l \) is not much larger than \( s \) while \( s < L_A \).

5.4 Details for Theorem 5.2.2

In this section we present the proof of Thm. 5.2.2 and detail some of its claims. We exclude the case that \( A \) is a polynomial (only for the sake of a smoother presentation with \( z_\infty \) below). Hence, when \( L_A < \infty \), we assume that \( z_\infty < \infty \) so that \( A(z_\infty), A'(z_\infty) \) are given by (5.45) as finite numbers, while

\[ L_A = \frac{z_\infty A'(z_\infty)}{A(z_\infty)} = \phi(z_\infty). \]  

(5.46)

Here, with \( L_A \) finite or not,

\[ \phi(z) = zA'(z)/A(z), \quad |z| < z_\infty, \]  

(5.47)
as in (5.32). We show in Sec. 5.7 that \( \phi \) is strictly increasing in \( z \in [0, z_\infty) \), unless \( A \) is a monomial.

We let for \( t > A'(1) \)

\[
z_0(t) = \begin{cases} 
  \text{unique } z \geq 1 \text{ such that } \phi(z) = t, & A'(1) \leq t < L_A, \\
  z_\infty, & t \geq L_A.
\end{cases}
\]  
(5.48)

Thus \( z_0(t) \) is strictly increasing in \( t \in [A'(1), L_A) \) and constant \( z_\infty \) for \( t \geq L_A \).

In terms of \( \phi \) and \( z_0 \) we can express the saddle point approximation of \( P(A^*l = i) \) in Thm. 5.2.1 as

\[
P(A^*l = i) \approx \frac{1}{\sqrt{2\pi l z_0(i/l) \phi'(z_0(i/l))}} \left( \frac{A(z_0(i/l))}{z_0(i/l)^{1/l}} \right)^l
\]  
(5.49)

when \( A'(1) \leq i/l < L_A \). Also, the bound (5.29) can be expressed in terms of \( z_0 \) as

\[
P(A^*l = i) \leq \left( \frac{A(z_0(i/l))}{z_0(i/l)^{1/l}} \right)^l, \quad i/l \geq A'(1).
\]  
(5.50)

For the analysis that follows we introduce the function

\[
G(t) := \ln \left\{ \frac{A(z_0(t))}{z_0(t)^{1/l}} \right\}, \quad t \geq A'(1).
\]  
(5.51)

Note that \( G(t) = g_t(z_0(t)) \), see (5.39). The function \( G \) is considered in some detail in Sec. 5.7. It is shown that \( G \) is a non-positive, strictly decreasing, concave function of \( t \geq A'(1) \) for which the \( t \)-axis is a tangent of the graph \( (t, G(t)) \), \( t \geq A'(1) \) at the point \( (t = A'(1), G(A'(1)) = 0) \). Moreover, it is shown that

\[
G'(t) = -\ln z_0(t), \quad t \geq A'(1).
\]  
(5.52)

In particular, \( G \) is strictly concave on \( [A'(1), L_A) \) with

\[
G''(t) = -\frac{1}{z_0(t) \phi'(z_0(t))}, \quad t \in [A'(1), L_A),
\]  
(5.53)

and \( G \) is linear on \( [L_A, \infty) \) with \( G'(t) = -\ln z_\infty \) (when \( L_A < \infty \)). Also see Fig. 5.1.

We restrict for the moment to \( L_A = \infty \), and we consider

\[
\hat{S}_m(l) = \frac{1}{l\sqrt{2\pi l}} \sum_{i=ls+1}^{\infty} \frac{(i-ls)^m}{\sqrt{z_0(i/l) \phi'(z_0(i/l))}} \left( \frac{A(z_0(i/l))}{z_0(i/l)^{1/l}} \right)^l
\]
\[
= \frac{1}{l\sqrt{2\pi l}} \sum_{i=ls+1}^{\infty} (i-ls)^m \sqrt{-G''(i/l)} e^{G(i/l)}
\]  
(5.54)

as an approximation of \( S_m(l) \), see (5.14). In the first line of (5.54) we have inserted the saddle point approximation (5.49) of \( P(A^*l = i) \) into the series (5.14) at the
When we now use (5.51)-(5.53), we get (5.23) in Thm. 5.2.2. The crucial step in getting the approximation (5.55) is the linearization of the function $G(t)$ around $t = s$. In Fig. 5.1 we display this linearization for the case that $A(z) = e^{\lambda (z-1)}$ with $\lambda = 9$ and $s = 9, 15, 20$. We note that in many cases the approximation (5.55) holds as an upper bound on $\hat{S}_m(l)$. This is certainly so when $z\phi'(z)$ is an increasing function of $z$ (as often happens, see the examples in Sec. 5.6). In that case, replacing $-G''(i/l)$ by $-G''(s)$ and $G$ by its linearization in (5.54) comes with a $\leq$-sign. The condition $(z\phi'(z))' = \phi'(z) + z\phi''(z) \geq 0$ is not very restrictive; it excludes functions $A$ that grow slower than $z^{a+b\ln z}$ with some $a > 0, b > 0$.

We next make a brief error assessment for the approximation in (5.55). We note
that by Taylor’s formula

\[l[G(i/l) - G(s) - (i/l - s)G'(s)] = \frac{1}{2l}(i - ls)^2G''(\zeta) < 0, \quad (5.56)\]

where \(\zeta\) is a number \(\in [s, i/l]\). Also, \(G''(i/l) - G''(s) = O(i/l - s)\). Then, due to exponential decay, one can show that relative errors of order \(1/l\) occur.

Finally, in the case of finite \(L_A\) and \(s < L_A\), the above argument to approximate and bound \(S_m(l)\) remains basically the same (due to the bound in (5.50) and concavity of \(G\)) at the expense of exponentially small relative errors.

### 5.5 Proof of Theorem 5.2.3

We shall now approximate and bound the quantity

\[\hat{R}_m(L) = \frac{K_m(\hat{z}^{-1})}{\sqrt{2\pi \hat{\varphi}'(\hat{z})}} \sum_{l=L}^{\infty} l^{-3/2}\left(\frac{A(\hat{z})}{\hat{z}^s}\right)^l, \quad L \to \infty, \quad (5.57)\]

as it occurs as an upper bound of the approximation \(\sum_{l=L}^{\infty} \hat{S}_m(l)\) on \(R_m(L) = \sum_{l=L}^{\infty} S_m(l)\). It suffices to study the quantity

\[F_L(x) := \sum_{l=L}^{\infty} l^{-3/2}x^l, \quad (5.58)\]

for large \(L\) and \(x \in [0, 1]\). It is interesting to note that \(F_L(x) = x^L \Phi(z = x, s = 3/2, v = L)\), where \(\Phi(z, s, v)\) is Lerch’s transcendent as occurs in [64], §1.11 on pp. 27-31. Of the many formulas and representations developed in [64], §1.11 for \(\Phi\), the one in §1.11(3) is particularly convenient for getting a simple and accurate approximation of \(F_L(x)\) when \(L\) gets large and \(x \in [0, 1]\). When \(x\) is away from 1, one simply has

\[F_L(x) = \frac{x^L}{L^{3/2}} \sum_{l=0}^{\infty} \frac{1}{(1 + l/L)^{3/2}} x^l, \approx \frac{x^L}{L^{3/2}} \sum_{l=0}^{\infty} x^l \approx \frac{x^L}{L^{3/2}(1 - x)}, \quad (5.59)\]

with a relative error that is of the order \(-3x/(2L(1-x))\). The right-hand side of (5.59) is in fact an upper bound for \(F_L(x)\). This gives the result in (5.25).

When \(x\) may get close to 1 while \(L \to \infty\), we have to proceed more carefully: While \(F_L(x)\) is evidently bounded for \(x \in [0, 1]\), the last member of (5.59) tends to infinity as \(x\) tends to 1. From

\[y^{-3/2} = \frac{2}{\sqrt{\pi}} \int_0^{\infty} u^{1/2}e^{-yu}du, \quad (5.60)\]
5.5 Proof of Theorem 5.2.3

we obtain [64], §1.11(3),

\[ F_L(x) = \frac{2x^L}{\sqrt{\pi}} \int_0^\infty \frac{u^{1/2}e^{-Lu}}{1 - xe^{-u}} du. \]  

(5.61)

With \( L \to \infty \), we may restrict attention in the integral in (5.61) to small \( u \geq 0 \), and we expand

\[ \frac{1}{1 - xe^{-u}} = \frac{1}{1 - x + xu - x(e^{-u} - 1 + u)} \]

\[ = \frac{1}{1 - x + xu} \left( 1 + \frac{x(e^{-u} - 1 + u)}{1 - x + xu} + \left( \frac{x(e^{-u} - 1 + u)}{1 - x + xu} \right)^2 + \ldots \right). \]

(5.62)

Since

\[ 0 \leq \frac{x(e^{-u} - 1 + u)}{1 - x + xu} \leq \frac{\frac{1}{2}xu^2}{1 - x + xu} \leq \frac{1}{2}u, \]

(5.63)

we see that the leading term of \( F_L(x) \) is given as

\[ \frac{2x^L}{\sqrt{\pi}} \int_0^\infty \frac{u^{1/2}e^{-Lu}}{1 - x + xu} du, \]

(5.64)

while the error is accurately estimated at

\[ \frac{2x^L}{\sqrt{\pi}} \int_0^\infty \frac{\frac{1}{2}xu^{5/2}e^{-Lu}}{(1 - x + xu)^2} du. \]

(5.65)

The analysis given above can be made more precise as follows. We restrict the integration range in (5.61) to \( u \in [0, L^{-1/2}] \) at the expense of exponentially small errors \( \mathcal{O}(\exp(-L^{-1/2})) \). On the relevant integration range we have from (5.62) and (5.63)

\[ \frac{1}{1 - xe^{-u}} = \frac{1}{1 - x + xu} + \frac{\frac{1}{2}xu^2}{(1 - x + xu)^2} \left( 1 + \mathcal{O}(L^{-1/2}) \right), \]

(5.66)

where the \( \mathcal{O} \) holds uniformly in \( x \in [0, 1], u \in [0, L^{-1/2}] \). Then the integration range for the two functions of \( u \) at the right-hand side of (5.66) is restored to \( [0, \infty) \), again at the expense of exponentially small errors \( \mathcal{O}(\exp(-L^{-1/2})) \).

We shall now express the integrals in (5.64), (5.65) in terms of the complementary error function (see (5.27) and Abramowitz & Stegun [20], p. 297). We have that

\[ \int_0^\infty \frac{u^{1/2}e^{-Lu}}{1 - x + xu} du = \frac{1}{xL^{1/2}} \int_0^\infty \frac{t^{1/2}e^{-t}}{\beta^2 + t} dt, \]

(5.67)

where we have set

\[ \beta = \left( \frac{1 - x}{x} \right)^{1/2}. \]

(5.68)
Then, by [20], p. 302
\[
\int_0^\infty \frac{t^{1/2}e^{-t}}{\beta^2 + t} \, dt = \int_0^\infty t^{-1/2}e^{-t}dt - \beta^2 \int_0^\infty \frac{e^{-t}}{\sqrt{t} \,(\beta^2 + t)} \, dt
= \sqrt{\pi}(1 - \sqrt{\pi} \beta e^{\beta^2} \text{erfc}(\beta)). \tag{5.69}
\]
Therefore, with \(\beta\) given in (5.68), we have
\[
\int_0^\infty \frac{u^{1/2}e^{-Lu}}{1 - x + xu} \, du = \frac{\sqrt{\pi}}{xL^{1/2}}(1 - \sqrt{\pi} \beta e^{\beta^2} \text{erfc}(\beta)) \tag{5.70}
\]
This yields the right-hand side of (5.26) as a first order approximation of \(\hat{R}_m(L)\).
Similarly, we have by partial integration
\[
\int_0^\infty \frac{1}{2} xu^{5/2}e^{-Lu} \, du = \frac{1}{2} \int_0^\infty \frac{1}{1 - x + xu} \left( \frac{5}{2}u^{3/2}/2 - u^{5/2}/2 \right) e^{-Lu} \, du
= -\frac{1}{2} \left( \frac{5}{2} \frac{d}{dL} \right) \left( \frac{d}{dL} \right)^2 \int_0^\infty \frac{u^{1/2}e^{-Lu}}{1 - x + xu} \, du. \tag{5.71}
\]
Then, using (5.70) and the definition of \(\beta\) in (5.68), we see that the integral in (5.65) can be expressed in terms of elementary functions and the erfc, although the resulting expression is rather unwieldy, i.e.
\[
\int_0^\infty \frac{1}{2} xu^{5/2}e^{-Lu} \, du = \frac{\sqrt{\pi}}{8xL^{3/2}} \left\{ (5 + 2\beta^2)(1 - 2\beta^2)(1 - \sqrt{\pi} \beta e^{\beta^2} \text{erfc}(\beta)) - 3 \right\}. \tag{5.72}
\]
The function \(\exp(\beta^2)\text{erfc}(\beta)\) is known as Mills’ ratio, see Abramowitz & Stegun [20], 7.1.13 on p. 298. Using the asymptotic series, [20], p. 298,
\[
\sqrt{\pi}\beta e^{\beta^2} \text{erfc}(\beta) \sim 1 - \frac{1}{2\beta^2} + \frac{3}{4\beta^4} - \frac{15}{8\beta^6} + \frac{105}{16\beta^8} - \ldots, \quad \beta \to \infty, \tag{5.73}
\]
we get from (5.70)
\[
\int_0^\infty \frac{u^{1/2}e^{-Lu}}{1 - x + xu} \, du \sim \frac{\sqrt{\pi}}{x} \left( \frac{1}{2\beta^2} - \frac{3}{8(1 - x)^2L^{5/2}} + \frac{15}{16(1 - x)^2L^{7/2}} - \ldots \right). \tag{5.74}
\]
as \((1 - x)L/x \to \infty\). Furthermore, from (5.71) and (5.74) (repeated differentiation of the asymptotic series is allowed)
\[
\int_0^\infty \frac{1}{2} xu^{5/2}e^{-Lu} \, du \sim \frac{\sqrt{\pi}}{x} \left( \frac{15}{16(1 - x)^2L^{7/2}} - \frac{105}{16(1 - x)^3L^{9/2}} + \ldots \right), \tag{5.75}
\]
as \((1 - x)L/x \to \infty\). Note that the first term at the right-hand side of (5.74) agrees with the approximation given in (5.59). Also note that the leading term in (5.75) is a factor \(\frac{8}{15}\frac{1-x}{x}L^2\) smaller than the leading term in the right-hand side of (5.74).
5.6 Examples

The asymptotics in (5.74), (5.75) are valid when \((1-x)L/x \to \infty\). We complement these results by presenting lower and upper bounds for the integrals in (5.64), (5.65) that show that the second integral is roughly a factor \(L(1 + \frac{1}{2} \beta^2)\) smaller than the first integral for all values \(\beta = \sqrt{(1-x)L/x} \geq 0\). As to the first integral we have

\[
\int_0^\infty \frac{u^{1/2}e^{-Lu}}{1-x+ux} \, du = \frac{1}{xL^{1/2}} \int_0^\infty \frac{t^{1/2}e^{-t}}{\beta^2 + t} \, dt = \frac{\sqrt{\pi}}{2xL^{1/2}} \int_0^\infty e^{-\beta^2v} \frac{v}{(1+v)^{3/2}} \, dv. \tag{5.76}
\]

Here we have inserted

\[
\Gamma(\alpha) \left( \frac{1}{\beta^2 + 1} \right)^{\alpha} = \int_0^\infty v^{\alpha - 1}e^{-v(t+\beta^2)} \, dv \tag{5.77}
\]

with \(\alpha = 1\) into the second integral in (5.76), interchanged the order of integration and used \(\sqrt{\pi}/2 = \Gamma(3/2) = \int_0^\infty t^{1/2}e^{-t} \, dt\). Then by the inequality

\[
e^{-\left(\beta^2 + 3/2\right)v} \leq \frac{e^{-\beta^2v}}{(1+v)^{3/2}} \leq \left( \frac{1}{1+v} \right)^{3/2 + \beta^2}, \tag{5.78}
\]

we immediately get

\[
\frac{\sqrt{\pi}}{2xL^{1/2}} \frac{1}{3/2 + \beta^2} \leq \int_0^\infty \frac{u^{1/2}e^{-Lu}}{1-x+ux} \, du \leq \frac{\sqrt{\pi}}{2xL^{1/2}} \frac{1}{1/2 + \beta^2}. \tag{5.79}
\]

In an entirely similar way, using (5.77) with \(\alpha = 2\), we get

\[
\int_0^\infty \frac{1}{2} xu^{5/2}e^{-Lu} \, du = \frac{15\sqrt{\pi}}{16xL^{3/2}} \int_0^\infty \frac{ve^{-\beta^2v}}{(1+v)^{7/2}} \, dv, \tag{5.80}
\]

from which it follows that

\[
\frac{15\sqrt{\pi}}{16xL^{3/2}} \frac{1}{(7/2 + \beta^2)^2} \leq \int_0^\infty \frac{1}{2} xu^{5/2}e^{-Lu} \, du \leq \frac{15\sqrt{\pi}}{16xL^{3/2}} \frac{1}{(3/2 + \beta^2)(5/2 + \beta^2)}. \tag{5.81}
\]

We may, finally, note that continued fraction expansions for the integrals in (5.64) and (5.65) can be obtained from Wall [159], pp. 352-355; also see Temme [152], Sec. 11.2 where asymptotics of integrals of type (5.64) and (5.65) are considered in connection with the incomplete Gamma function.

5.6 Examples

In this section we consider several examples for which we determine characteristics of the relaxation time. For each example, the load on the system is defined as \(\rho = A'(1)/s\) and assumed to be less than one.
Example 5.6.1 Poisson case. \( A(z) = e^{\lambda(z-1)} \), \( A'(1) = \lambda \), \( \phi(z) = \lambda z \), \( z_\infty = \infty \), \( L_A = \infty \), \( z \phi'(z) = \lambda z \) increasing and

\[
z_0 = \frac{i}{\lambda}, \quad \hat{z} = z_0(s) = \frac{s}{\lambda}, \quad x = \frac{A(z_0(s))}{(z_0(s))^s} = \left( \frac{\lambda}{s} e^{1-\lambda/s} \right)^s.
\] (5.82)

From Thm. 5.2.1 we thus have

\[
\mathbb{P}(A^*l = i) \approx \frac{1}{i/(\lambda)} \cdot \frac{1}{\sqrt{2\pi \frac{1}{i}(\lambda)^2}} \cdot \exp\left(\frac{\lambda}{\lambda} - 1\right) \left(\frac{\lambda}{i} \right)^i = \frac{1}{\sqrt{2\pi}} \left( \frac{\lambda}{i} e^{1-\lambda/s} \right)^i. \quad (5.83)
\]

Observe that \( te^{-t} \in [0,1] \) when \( t \in [0,1) \). In the present case we have, explicitly,

\[
\mathbb{P}(A^*l = i) = \frac{e^{-1\lambda}}{i!} (\lambda)^i \approx e^{-1\lambda}(i^i e^{-i} \sqrt{2\pi})^{-1}(\lambda)^i = \frac{1}{\sqrt{2\pi}} \left( \frac{\lambda}{i} e^{1-\lambda/s} \right)^i, \quad (5.84)
\]

where Stirling’s formula \( i! \approx i^{i+1/2} e^{-i} \sqrt{2\pi} \) has been used. It is thus seen that the approximation as obtained per Thm. 5.2.1 amounts to replacing \( i! \) in the exact expression (5.84) by its Stirling approximation. Accordingly, the approximation given by (5.83) has relative error \( O(1/i) \) independent of \( \lambda \).

Example 5.6.2 Geometric case. \( A(z) = (1-p)/(1-pz) \), \( A'(1) = p/(1-p) \), \( \phi(z) = p z/(1-pz) \), \( z_\infty = 1/p \), \( L_A = \infty \), \( z \phi'(z) = p z/(1-pz)^2 \) increasing for \( z \leq 1 \) and decreasing for \( z \geq 1 \), and

\[
z_0 = \frac{i}{p(i+l)}, \quad \hat{z} = z_0(s) = \frac{1}{p} \frac{s}{s+1}, \quad x = \frac{A(z_0(s))}{(z_0(s))^s} = (1-p)p^s \frac{(s+1)^s+1}{s^s}. \quad (5.85)
\]

From Thm. 5.2.1 it follows that

\[
\mathbb{P}(A^*l = i) \approx \frac{1}{\sqrt{2\pi}} (1-p)^i p^i (i+l)^{i+l-1/2} i^{-i-1/2} l^{-i+1/2}. \quad (5.86)
\]

From the explicit representation

\[
\mathbb{P}(A^*l = i) = (1-p)^i p^i (i+l)! \frac{i}{i+l}, \quad (5.87)
\]

we obtain by Stirling’s formula exactly (5.86). Accordingly, as in Example 5.6.1, the approximation given by (5.86) has relative error \( O(1/i) \) independent of \( p \).

Example 5.6.3 Binomial case. \( A(z) = (p+qz)^n \), \( p + q = 1 \), \( A'(1) = nq \), \( \phi(z) = nqz/(p+qz) \), \( z_\infty = \infty \), \( L_A = n \), \( z \phi'(z) = npqz/(p+qz)^2 \) increasing for \( z \leq q/p \) and decreasing for \( z \geq q/p \), and

\[
z_0 = \frac{1}{q n^l - i}, \quad \hat{z} = z_0(s) = \frac{1}{q} \frac{ps}{n-s}, \quad x = \frac{A(z_0(s))}{(z_0(s))^s} = q^n p^{n-s} n^s (n-s)^{(n-s)}. \quad (5.88)
\]
5.6 Examples

From Thm. 5.2.1 it follows that
\[ P(A^*l = i) \approx \frac{1}{\sqrt{2\pi}} \frac{(nl)^{nl+i+1/2}}{(nl-i)!} p^{nl-i} q^i. \] (5.89)

From the explicit representation
\[ P(A^*l = i) = \frac{(nl)!}{(nl-i)!} p^{nl-i} q^i, \] (5.90)
we obtain by Stirling’s formula exactly (5.89). For the remainder of this section we set \( n = 4s \).

Thm. 5.2.3 gives asymptotic expressions for \( \hat{R}_m(L) \) from which we can extract information on the relaxation time. Expression (5.25) yields an upper bound on \( \hat{R}_m(L) \), which is expected to be sharp for loads well below one. Expression (5.26) sharpens (5.25) and should be useful when \( \rho \) tends to 1. The complementary error function \( \text{erfc}(\beta) \) needed to calculate (5.26) is a standard function available in most software packages.

For the Poisson case, Figs. 5.2 and 5.3 depict characteristics of these asymptotic approximations for \( m = 1 \) (corresponding to the mean waiting time), for \( s = 10 \) and \( \rho = 0.8, 0.9 \), respectively. The true value of \( R_1(L) \) results from \( E(W) - EW_L \), where we approximate \( EW \) using extremely high truncation levels (something we want to avoid). For \( \rho = 0.8 \), \( EW - EW_L \) decreases rapidly for increasing values of \( L \), indicating that the transient behavior of the waiting time converges rapidly to its steady-state. (5.26) improves upon (5.25), although the improvement is marginal relative to the true value \( EW - EW_L \).

For \( \rho = 0.9 \), \( EW - EW_L \) again decreases rapidly, although we need a much higher value of \( L \) in order to achieve the same accuracy as for \( \rho = 0.8 \). Again, (5.26) improves upon (5.25), where now the absolute improvement is much larger.

For each of the three examples, we calculate
\[ T(\hat{R}_m; \epsilon) = \min\{L \mid \hat{R}_m(L) < \epsilon\}, \] (5.91)
where we replace \( \hat{R}_m(L) \) in (5.91) by either (5.25) or (5.26). Remember that (5.25) is an upper bound on \( \hat{R}_m(L) \), where (5.26) is, although asymptotically sharp, an asymptotic approximation of \( \hat{R}_m(L) \). We set \( \epsilon \) equal to 0.001. When we want, for example, to determine the mean stationary waiting time, we could approximate \( EW \) using (5.4) with \( n = T(\hat{R}_1; 0.001) \), knowing that \( EW - EW_n \) is of order 0.001. Of course, (5.4) still contains an infinite series over \( i \), but truncating this series at \( ls + M \) for some large value \( M \) gives a truly negligible error, for reasons addressed in Sec. 5.3.

Results are displayed in Tables 5.1-5.3. We first make some general observations. For low values of \( \rho \), a small value of \( T(\hat{R}_m; \epsilon) \) is sufficient. For high values of \( \rho \), though, the \( T(\hat{R}_m; \epsilon) \) required increases enormously. Using (5.26) instead of (5.25) leads to moderate reductions in \( T(\hat{R}_m; \epsilon) \), mostly for high values of \( \rho \). Further, \( T(\hat{R}_0; \epsilon) \leq T(\hat{R}_1; \epsilon) \leq T(\hat{R}_2; \epsilon) \), which is obvious from the \( K_m \) functions given in (5.24).
Figure 5.2: Exact values and approximations for $E_W - EW_L$ for the Poisson case with $s = 10$, $\lambda = 8$.

Figure 5.3: Exact values and approximations for $E_W - EW_L$ for the Poisson case with $s = 10$, $\lambda = 9$.

5.6.1 On the impact of the distribution of $A$

The geometric distribution results in much higher values of $T(\hat{R}_m; \epsilon)$ than does the binomial and Poisson distribution. The reason for this is that the geometric distribution is more volatile. To be more precise, the crucial quantity (as it appears in (5.51))

$$A(z_0(t)) = \exp \left( \min_{z \geq 1} \ln A(z) - t \ln z \right),$$

(5.92)

is far larger for the geometric distribution. To give a comparison with a relatively light-tailed distribution, we introduce a fourth distribution of $A$.

Example 5.6.4 Light-tailed case.

$$P(A = n) = \frac{\theta^{2n}}{(2n)! \cosh \theta}, \quad n = 0, 1, \ldots,$$

(5.93)

and

$$A(z) = \frac{\cosh \sqrt{z \tanh \theta}}{\cosh \theta}, \quad A'(1) = \frac{1}{2} \theta \tanh \theta, \quad \phi(z) = \frac{1}{2} \theta \sqrt{z \tanh(\theta \sqrt{z})},$$

(5.94)

$z_\infty = \infty$, $L_A = \infty$ and $z \phi'(z)$ increasing. Also, let $z^{(1)}_0(t) \equiv z_0(t)$ denote the solution of $\theta \sqrt{z \tanh(\theta \sqrt{z})} = 2t$.

We denote by $z^{(2)}_0(t)$ and $z^{(3)}_0(t)$ the $z_0(t)$ for the Poisson and geometric case, respectively, i.e. $z^{(2)}_0(t) = t/\lambda$ and $z^{(3)}_0(t) = t/(p(1+t))$. Fig. 5.4 displays $z^{(i)}_0(t)$, $i = 1, 2, 3$, for a common value $A'(1)$ of unity (i.e. $\theta = 2.065$, $\lambda = 1$, $p = 1/2$) and $t = 3$. The three heavy line segments above $z^{(i)}_0(t)$ indicate the difference between $\ln A(z)$ and $t \ln z$ at the minimizing $z = z_0(t)$, see (5.92). It is thus seen that the magnitude
5.6 Examples

Table 5.1: $T(\hat{R}_m; \epsilon)$ for $m = 0$, $\epsilon = 0.001$, using either (5.25) or (5.26), for the binomial, Poisson and geometric case.

<table>
<thead>
<tr>
<th>$s = 10$</th>
<th>binomial</th>
<th>Poisson</th>
<th>geometric</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>(5.25)</td>
<td>(5.26)</td>
<td>(5.25)</td>
</tr>
<tr>
<td>0.5</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>0.6</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0.7</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>0.8</td>
<td>14</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>0.9</td>
<td>54</td>
<td>51</td>
<td>70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$s = 30$</th>
<th>binomial</th>
<th>Poisson</th>
<th>geometric</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.6</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>0.7</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>0.8</td>
<td>5</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>0.9</td>
<td>19</td>
<td>18</td>
<td>24</td>
</tr>
</tbody>
</table>

Table 5.2: $T(\hat{R}_m; \epsilon)$ for $m = 1$, $\epsilon = 0.001$, using either (5.25) or (5.26), for the binomial, Poisson and geometric case.

<table>
<thead>
<tr>
<th>$s = 10$</th>
<th>binomial</th>
<th>Poisson</th>
<th>geometric</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>(5.25)</td>
<td>(5.26)</td>
<td>(5.25)</td>
</tr>
<tr>
<td>0.5</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>0.6</td>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>0.7</td>
<td>7</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>0.8</td>
<td>16</td>
<td>16</td>
<td>21</td>
</tr>
<tr>
<td>0.9</td>
<td>75</td>
<td>72</td>
<td>101</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$s = 30$</th>
<th>binomial</th>
<th>Poisson</th>
<th>geometric</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.6</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>0.7</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>0.8</td>
<td>6</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>0.9</td>
<td>25</td>
<td>25</td>
<td>34</td>
</tr>
</tbody>
</table>

of the quantity in (5.92) strongly depends on the type of distribution. This effect becomes even more manifest when we increase $s$ and keep $\rho$ fixed, as we discuss in the next subsection.

5.6.2 On the effect of increasing $s$ at fixed load $\rho$

We now compare the values of $T(\hat{R}_m; \epsilon)$ in Tables 5.1-5.3 for $s = 10$ and $s = 30$. Observe that by increasing $s$ from 10 to 30, the values $T(\hat{R}_m; \epsilon)$ decrease for the binomial and Poisson distribution (and the geometric distribution for $m = 0$), while
Table 5.3: $T(\hat{R}_m; \epsilon)$ for $m = 2$, $\epsilon = 0.001$, using either (5.25) or (5.26), for the binomial, Poisson and geometric case.

<table>
<thead>
<tr>
<th>$s$ = 10</th>
<th>binomial</th>
<th>Poisson</th>
<th>geometric</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>(5.25)</td>
<td>(5.26)</td>
<td>(5.25)</td>
</tr>
<tr>
<td>0.5</td>
<td>3</td>
<td>3</td>
<td>29</td>
</tr>
<tr>
<td>0.6</td>
<td>4</td>
<td>5</td>
<td>60</td>
</tr>
<tr>
<td>0.7</td>
<td>8</td>
<td>10</td>
<td>139</td>
</tr>
<tr>
<td>0.8</td>
<td>20</td>
<td>27</td>
<td>406</td>
</tr>
<tr>
<td>0.9</td>
<td>98</td>
<td>135</td>
<td>2156</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$s$ = 30</th>
<th>binomial</th>
<th>Poisson</th>
<th>geometric</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>(5.25)</td>
<td>(5.26)</td>
<td>(5.25)</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td>0.6</td>
<td>2</td>
<td>2</td>
<td>67</td>
</tr>
<tr>
<td>0.7</td>
<td>3</td>
<td>4</td>
<td>154</td>
</tr>
<tr>
<td>0.8</td>
<td>7</td>
<td>10</td>
<td>448</td>
</tr>
<tr>
<td>0.9</td>
<td>33</td>
<td>46</td>
<td>2347</td>
</tr>
</tbody>
</table>

As a consequence, we see that it does not help to increase $s$ to substantially decrease the value of the crucial quantity (5.92). In the Poisson case, we find, for $\alpha \geq 1$ and a fixed value of $\rho = A'(1)/s$, that

$$
\frac{A(z_0(\alpha s))}{(z_0(\alpha s))^{\alpha s}} = \frac{1}{1 + \rho s} \left( 1 + \frac{1}{\rho s} \right)^{-\alpha s} \left( 1 + \frac{1}{\alpha s} \right)^\alpha (1 + \alpha s) \\
\to \frac{\alpha}{\rho} e^{1-\alpha/\rho}, \quad s \to \infty. 
$$

(5.95)

As a consequence, we see that it does not help to increase $s$ to substantially decrease the value of the crucial quantity (5.92). In the Poisson case, we find, for $\alpha \geq 1$ and a fixed value of $\rho = A'(1)/s$,

$$
\frac{A(z_0(\alpha s))}{(z_0(\alpha s))^{\alpha s}} = \left( \frac{\rho}{\alpha} e^{1-\rho/\alpha} \right)^\alpha, 
$$

(5.96)

and this decays exponentially in $s$ since $te^{1-t} < 1$ for $t \in (0,1)$. Thus in the Poisson case it does pay to increase $s$. This observation continues to be valid for distributions with lighter tails, such as the binomial distribution or the distribution in Example 5.6.4.
5.7 Results on the functions $\phi$ and $G$

5.7.1 The function $\phi$

We consider the function $\phi(z) = zA'(z)/A(z)$, and we show that $\phi$ strictly increases on $[0, z_\infty)$ unless $A$ is a monomial.

Let $z > 0$ and set

$$t := \phi(z) = zA'(z)/A(z).$$  \hfill(5.97)

Using $A'(z)/A(z) = t/z$ we compute

$$\phi'(z) = \frac{A'(z)}{A(z)} + z \frac{A''(z)}{A(z)} - z \left( \frac{A'(z)}{A(z)} \right)^2 = \frac{z^2 A''(z) - t(t - 1)A(z)}{zA(z)}. \hfill(5.98)$$
With \( A(z) = \sum_{j=0}^{\infty} a_j z^j \) we can write
\[
z^2 A''(z) - t(t-1)A(z) = \sum_{j=0}^{\infty} (j(j-1) - t(t-1)) a_j z^j
\]
\[
= \sum_{j=0}^{\infty} (j-t)(j+t-1)a_j z^j. \tag{5.99}
\]
Subtracting
\[
0 = (2t-1)(zA'(z) - tA(z)) = (2t-1) \sum_{j=0}^{\infty} (j-t)a_j z^j \tag{5.100}
\]
from either side of (5.99) we obtain
\[
z^2 A''(z) - t(t-1)A(z) = \sum_{j=0}^{\infty} (j-t)^2 a_j z^j. \tag{5.101}
\]
Hence
\[
\phi'(z) = \frac{1}{zA(z)} \sum_{j=0}^{\infty} (j-t)^2 a_j z^j \geq 0. \tag{5.102}
\]
There is equality in (5.102) only when \( t \) is a non-negative integer and \( A(z) = z^t \).
As a consequence we have that (excluding the monomial case)
\[
A'(1) = \phi(1) < \lim_{z \to \infty} \phi(z) = L_A. \tag{5.103}
\]
We observe furthermore from (5.102) that \( \phi'(1) = \sigma_A^2 \) since \( t = A'(1) = EA \) when \( z = 1 \).
Interestingly, one computes in a similar fashion as above that
\[
(z\phi'(z))' = \frac{1}{zA(z)} \sum_{j=0}^{\infty} (j-t)^3 a_j z^j, \tag{5.104}
\]
with \( t \) as in (5.97). This is of some relevance to the approximation made going from (5.54) to (5.55): one cannot assert monotonicity of \( z\phi'(z) \) in general.

### 5.7.2 The function \( G \)

We consider the function
\[
G(t) = \ln[A(z_0(t))/(z_0(t))]', \quad t \geq A'(1), \tag{5.105}
\]
where \( z_0(t) \) is given by (5.48) for which we assume that \( A \) is not a polynomial. In terms of \( g_t \) in (5.39) we have \( G(t) = g_t(z_0(t)) \).
5.7 Results on the functions $\phi$ and $G$

We compute for $t \in [A'(1), L_A)$ from

$$\phi(z_0(t)) = z_0(t) \frac{A'(z_0(t))}{A(z_0(t))} = t, \quad \phi'(z_0(t))z_0'(t) = 1 \quad (5.106)$$

that

$$G'(t) = \frac{A'(z_0(t))}{A(z_0(t))} z_0'(t) - t \frac{z_0'(t)}{z_0(t)} - \ln z_0(t) = -\ln z_0(t). \quad (5.107)$$

Note that the identity $G'(t) = -\ln z_0(t)$ continues to hold for $t \geq L_A$ by the definition of $G$, $z_0$ on $[L_A, \infty)$, and it holds that $G'(t) = -\ln z_\infty$ for $t \geq L_A$.

When $t \in [A'(1), L_A)$ it furthermore follows from (5.106) and (5.107) that

$$G''(t) = -\frac{z_0'(t)}{z_0(t)} = \frac{-1}{z_0(t)\phi'(z_0(t))} < 0. \quad (5.108)$$

Observe that $\phi(1) = A'(1)$, whence $z_0(A'(1)) = 1$, and that $\phi'(1) = \sigma_A^2$. It then follows that

$$G(A'(1)) = 1, \quad G'(A'(1)) = 0, \quad G''(A'(1)) = -1/\sigma_A^2. \quad (5.109)$$

From (5.108) and (5.109) all further claims made about $G$ in this chapter follow.
In Chapter 2 we have seen that the mean and variance of the stationary queue length in the discrete bulk service queue can be expressed in terms of the roots of $z^* = A(z)$ or in terms of an infinite series that involves convolutions of the probability distribution of $A$. From a practical viewpoint, both alternatives have their own limitation: The roots should be calculated numerically while the infinite series should be truncated. It is therefore that we paid specific attention to the root-finding in Chapter 4 and truncating the infinite series in Chapter 5. Still, the numerical work that goes with each of the two alternatives is considerable.

In this chapter we will derive bounds on the mean and variance of the stationary queue length. The bounds derived have added value as compared to the exact expression in that they give more intuitive insight in the behavior of the performance characteristics and can be used for back-of-the-envelope computations. Furthermore, when one only knows the first two or three moments of $A$, the exact approaches cannot be applied while the presented bounds retain their value. Additionally, we identify the distributions of $A$ for which the bounds are attained, which gives additional insight into the behavior of the estimated values. We pay considerable attention to the case in which the arrivals follow a Poisson distribution. For this case, additional properties of the series are proved leading to even sharper bounds. The Poisson case serves as a pilot study for a broader range of distributions. The chapter is based on Denteneer et al. [P4].
6.1 Preliminaries

The mean and variance of the stationary queue length in the discrete bulk service queue are given by (see (2.21) and (2.22))

\[
\mu_X = \frac{\sigma_A^2}{2(s - \mu_A)} + \frac{1}{2} \mu_A - \frac{1}{2} (s - 1) + \sum_{k=1}^{s-1} \frac{1}{1 - z_k}, \quad (6.1)
\]

\[
\sigma_X^2 = \sigma_A^2 + \frac{A''(1) - s(s - 1)(s - 2)}{3(s - \mu_A)} + \frac{A''(1) - s(s - 1)}{2(s - \mu_A)}
+ \left( \frac{A''(1) - s(s - 1)}{2(s - \mu_A)} \right)^2 - \sum_{k=1}^{s-1} \frac{z_k}{(1 - z_k)^2}, \quad (6.2)
\]

The series

\[
\sum_{k=1}^{s-1} \frac{1}{1 - z_k}, \quad \sum_{k=1}^{s-1} \frac{z_k}{(1 - z_k)^2}, \quad (6.3)
\]

will be called the \( \mu \)-series and \( \sigma^2 \)-series, respectively. Evidently, both series are real since the zeros \( z_k \) are either real or come in conjugate pairs. For \( s = 1 \) both series equal zero and one obtains the exact expressions for \( \mu_X \) and \( \sigma_X^2 \) from (6.1) and (6.2), respectively. It is therefore that we will henceforth consider \( s \geq 2 \).

We have pointed out before that the discrete bulk service queue is closely related to the discrete \( D/G/1 \) queue (see p. 27). The latter model falls into the more general class of the \( G/G/1 \) queue, and ever since the publication of Kingman [98], a vast literature on bounding waiting-time characteristics for the \( G/G/1 \) queue has been developed. Daley et al. [58] give a comprehensive treatment of most of this research. Simple bounds for the mean and variance of the waiting time can be constructed by observing that the deterministic interarrival times \( s \) belong to the class of increasing failure rate (see e.g. Daley et al. [58], p. 200). For the mean stationary queue length, we then obtain a lower and an upper bound (see Kleinrock [102], (2.51), and Kingman [98], respectively) which, translated to the current setting, read

\[
\frac{\sigma_A^2}{2(s - \mu_A)} + \frac{\mu_A}{2} \leq \mu_X \leq \frac{\sigma_A^2}{2(s - \mu_A)} + \mu_A, \quad (6.4)
\]

i.e.

\[
\frac{s - 1}{2} \leq \sum_{k=1}^{s-1} \frac{1}{1 - z_k} \leq \frac{s - 1}{2} + \frac{\mu_A}{2}. \quad (6.5)
\]

The right-hand side of (6.4) is known as Kingman’s upper bound. We will show that it is relatively easy to further sharpen the Kingman upper bound, although the gain is marginal.

Largely paralleling the approach used for the mean, bounds for the variance of the stationary queue length were derived as well. The lower bound in Daley et al. [58]
and the upper bound derived by Fainberg [66] yield for the discrete $D/G/1$ queue
\begin{equation}
-s^2/4 + 1/12 \leq \sum_{k=1}^{s-1} \frac{z_k}{(1-z_k)^2} \leq -\frac{1}{12}(s-\mu_A)^2 + \frac{1}{12},
\end{equation}
where, for reasons of brevity, we give here the bounds on the $\sigma^2$-series. Together with (6.2), they yield bounds on $\sigma_X^2$. We strengthen these bounds on $\sigma_X^2$ and further derive some new bounds on $\sigma_X^2$ that will be shown to be very sharp.

In this chapter we derive relatively simple bounds for the $\mu$-series and the $\sigma^2$-series. Here simple means bounds requiring knowledge on the arrival distribution by at most the first three moments. We do so by representing the moment series (6.3) in terms of random variables related to the idle time of the system. One of the features of this study is that we extend the bounding techniques to a discrete setting. In doing so, we obtain simple bounds on the mean and variance of the stationary queue length that are sharper than comparably simple bounds (6.4) and (6.6). Finally, we show that the bounds can be further strengthened by combining them with specific properties of the moment series (6.3). This is done for the Poisson distribution, which serves as a sort of pilot study for other distributions. It is worth mentioning that for the $E_k/G/1$ queue, Daley [57] also proves properties of roots on a particular curve in order to derive bounds.

In Sec. 6.2 we give a detailed account of the main results, a comparison with the bounds (6.4) and (6.6), along with an overview of the chapter.

### 6.2 Overview and results

For the discrete $D/G/1$ queue, the stationary distribution of the length of the idle periods, denoted by $I$, is completely determined by the probabilities $x_0, \ldots, x_s$, where $x_j$ denotes the probability that a customer has a sojourn time of length $j$. That is, once a customer has a sojourn that is less than $s$, the slots remaining until the arrival of the next customer remain idle, i.e.
\begin{equation}
\mathbb{P}(I = j) = \frac{x_{s-j}}{\sum_{i=0}^{s} x_i}, \quad j = 0, 1, \ldots, s.
\end{equation}
Note that the idle period can be zero. For a convenient presentation of our results we now define two auxiliary random variables $Y$ and $Z$ that are closely related to $I$ and take values in \{0, 1, \ldots, s\} according to
\begin{equation}
\mathbb{P}(Y = j) = \frac{x_j}{\sum_{i=0}^{s} x_i}, \quad \mathbb{P}(Z = j) = \frac{(s-j)x_j}{s-\mu_A}, \quad j = 0, 1, \ldots, s,
\end{equation}
and $\mathbb{P}(Y = j) = \mathbb{P}(Z = j) = 0$, $j = s+1, s+2, \ldots$. Observe that $\mathbb{P}(Z = j)$, $j = 0, 1, 2, \ldots$ defines a probability distribution since (see (2.10))
\begin{equation}
s - \mu_A = \sum_{j=0}^{s-1} x_j(s-j).
\end{equation}
Also note that $Y$ represents both $s-I$ and $X$ conditional on $X \leq s$. Further note that the $k$th moment of $Z$ can be expressed in terms of the first $k+1$ moments of $I$. For example, $\mu_Z = s - \mathbb{E}(I^2)/\mu_I$. The random variables $Y$ and $Z$ are studied in detail in Sec. 6.3. In particular, it holds that

$$
\mu_Y \leq \mu_A; \quad 0 \leq \mu_Z \leq s - 1,
$$

(6.10)

with equality in the first inequality if and only if $A$ is concentrated on $\{0, 1, \ldots, s\}$. In Sec. 6.3 we also present representations for the $\mu$-series and $\sigma^2$-series in terms of $Y$ and $Z$. From these representations, one can obtain various inequalities, as well as insight into the matter when equality occurs in these.

We show the following bounds on the $\mu$-series in Sec. 6.4.

**Theorem 6.2.1** (i) We have

$$
\sum_{k=1}^{s-1} \frac{1}{1 - z_k} \geq \frac{1}{2} (s - 1) + \frac{1}{2} \mu_A - \frac{\sigma_A^2}{2(s - \mu_A)},
$$

and there is equality if and only if $A$ is concentrated on $\{0, 1, \ldots, s\}$.

(ii) Define $f : [0, s) \to [0, \infty)$ by

$$
f(\mu) = \frac{1}{2} (s - 1) + \frac{1}{2} \mu - \frac{(\mu) - (\mu)^2}{2(s - \mu)},
$$

(6.12)

where we have defined $(\mu) = \mu - \lfloor \mu \rfloor$ and $\lfloor \mu \rfloor$ = largest integer $\leq \mu$. Then we have

$$
\sum_{k=1}^{s-1} \frac{1}{1 - z_k} \leq f(\mu_A),
$$

and there is equality if and only if $A$ is concentrated on $\{j, j+1\}$ with $j = 0, 1, \ldots, s-2$ or $A$ is concentrated on $\{s-1, s, s+1, \ldots\}$.

In Sec. 6.4 we present somewhat sharper forms of Thm. 6.2.1 that explicitly involve $\mu_Y$ and $\sigma_Y^2$. The result in Thm. 6.2.1(i) presents a sharpening of the first inequality in (6.5) in case that $\sigma_A^2 \leq \mu_A(s - \mu_A)$. The inequality in Thm. 6.2.1(ii) exploits the discrete nature of the involved random variables. In Fig. 6.1, we have plotted the graphs of both $f(\mu)$ and $\mu - \frac{1}{2} (s - 1) + \frac{1}{4} \min\{\mu, s-1\}$ for $s = 5$. As one sees, the graph of $f$ hangs down from the second graph as a sort of guirlande with nodes at all integers $\mu = 0, 1, \ldots, s-1$.

In Sec. 6.5 we show the following result:

**Theorem 6.2.2** We have

$$
\frac{-s^2}{3(4 - \mu_A/s)} + \frac{1}{12} \leq \sum_{k=1}^{s-1} \frac{z_k}{(1 - z_k)^2} \leq \frac{1}{12} (s - \mu_A)^2 + \frac{1}{12},
$$

(6.14)
The upper bound in (6.14) is the same as in (6.6), but the lower bound in (6.14) is far sharper than the one in (6.6).

In Sec. 6.5 we present a more precise and sharper result in which the $\sigma^2$-series is bounded in terms of $\mu_Y$ and $\sigma^2_Y$, and from which one can infer the cases of equality in (6.14). This requires a result, communicated to us by E. Verbitskiy, on the extreme values of the third central moment of a random variable taking all real values between 0 and $s$, whose mean and variance are prescribed. The bounds in Thm. 6.2.2 disregard the discrete nature of the involved random variable, and, indeed, there is again a guirlande phenomenon that is detailed in Sec. 6.5. The bounds in (6.14) can be sharpened somewhat:

$$\frac{1}{3} (s - 1) + \frac{1}{2} \mu \leq f(\mu) \leq 0,$$

and this improves the bounds in (6.14) when $\mu_A \uparrow s$.

We further prove that

**Theorem 6.2.3** (i) We have

$$\sum_{k=1}^{s-1} \frac{z_k}{(1-z_k)^2} \leq \frac{A''(1) - s(s-1)(s-2)}{3(s - \mu_A)} + \frac{A''(1) - s(s-1)}{2(s - \mu_A)} + \left( \frac{A''(1) - s(s-1)}{2(s - \mu_A)} \right)^2,$$

and there is equality if and only if $A$ is concentrated on $\{0, 1, \ldots, s\}$. 

**Figure 6.1**: Universal bounds for the $\mu$-series, $s = 5$. 

- $\frac{1}{2} (s - 1) + \frac{1}{2} \mu$
- $\frac{1}{2} (s - 1)$
(ii) Defining \( h : [0, s) \to [0, \infty) \) by

\[
h(\mu) = \begin{cases} 
0, & 0 \leq \mu \leq 2, \\
\mu(\mu - 1)(\mu - 2), & \mu > 2,
\end{cases}
\] (6.17)

it holds that

\[
\sum_{k=1}^{s-1} \frac{z_k}{(1 - z_k)^2} \geq \frac{h(\mu_A) - h(s)}{3(s - \mu_A)} + \frac{A''(1) - s(s - 1)}{2(s - \mu_A)} + \left( \frac{A''(1) - s(s - 1)}{2(s - \mu_A)} \right)^2. \tag{6.18}
\]

Here \( \sigma_A^2 \) and \( \mu_A \) must be constrained according to

\[
\sigma_A^2 \leq (s - \mu_A)(\mu_A + 2s - 4). \tag{6.19}
\]

There is equality in (6.18) if and only if \( A \) is concentrated on \( \{0, 1, 2\} \) or on \( \{j\} \) with \( j = 2, \ldots, s - 1 \).

The proof of this result uses the representation (6.2) together with \( \sigma_X^2 \geq \sigma_A^2 \) for Thm. 6.2.3(i), and representation (6.27) in conjunction with Jensen’s inequality and \( \mu_Y \leq \mu_A \) for Thm. 6.2.3(ii).

In Sec. 6.6 we study in considerable detail the case that \( A \) is distributed according to the Poisson distribution. Among other things, it is shown that both the \( \mu \)-series and \( \sigma^2 \)-series increase in \( \mu_A \in [0, s) \) in the Poisson case, which can be exploited to derive the following theorems:

**Theorem 6.2.4** For \( A \) distributed according to the Poisson distribution, i.e. \( A(z) = e^{\lambda(z-1)} \), that satisfies \( \lambda < s \), the corresponding \( \mu \)-series can be bounded as

\[
\sum_{k=1}^{s-1} \frac{1}{1 - z_k} \geq \frac{1}{2} (s - 1) + m_1(\lambda), \tag{6.20}
\]

\[
\sum_{k=1}^{s-1} \frac{1}{1 - z_k} \leq \frac{1}{2} (s - 1) + \frac{1}{2} \lambda - \langle \lambda \rangle + \frac{1}{2} \langle \lambda \rangle^2. \tag{6.21}
\]

where \( m_1(\lambda) = \max\{ \frac{\tau}{2} - \frac{\tau^2}{2(s - \tau)} | 0 \leq \tau \leq \lambda \} \).

**Theorem 6.2.5** For \( A \) distributed according to the Poisson distribution, i.e. \( A(z) = e^{\lambda(z-1)} \), that satisfies \( \lambda < s \), and when Condition (6.19) holds, the corresponding \( \sigma^2 \)-series can be bounded as

\[
\sum_{k=0}^{s-1} \frac{z_k}{(1 - z_k)^2} \geq m_2(\lambda), \tag{6.22}
\]

\[
\sum_{k=0}^{s-1} \frac{z_k}{(1 - z_k)^2} \leq -\frac{1}{12} (s - \lambda)^2 - \frac{1}{2} \lambda + \frac{s(s + 2\lambda)}{12(s - \lambda)^2}, \tag{6.23}
\]

where \( m_2(\lambda) = \max\{ -\frac{1}{12} (s - \tau)^2 - \frac{1}{2} \tau + \frac{s(s + 2\tau)}{12(s - \tau)^2} - \frac{\tau}{s - \tau} (\tau - \frac{2}{3}) | 0 \leq \tau \leq \lambda \} \).
The functions $m_1(\lambda)$ and $m_2(\lambda)$ are strictly increasing for $\lambda \in [0, s - \sqrt{s}]$ and $\lambda \in [0, \lambda_2(s)]$, respectively, where $\lambda_2(s)$ is a point close to $s - (6(s^2 - \frac{1}{2}s))^{1/3}$ (this follows from elementary but somewhat lengthy considerations on the function $m_2(\lambda)$).

In Sec. 6.7 we present examples of distributions $A$ to illustrate the bounds on the $\mu$-series and $\sigma^2$-series. For the Poisson case, we use the bounds in Thms. 6.2.4 and 6.2.5. For other distributions, we employ for the $\mu$-series the bounds in Thm. 6.2.1 together with $\frac{1}{2}(s - 1)$ as an overall lower bound. For the $\sigma^2$-series we employ the bounds in Thm. 6.2.3, where the lower bound (6.38) is only used when condition (6.19) is satisfied. If not, we use the overall lower bound $-\frac{1}{9}(s - \frac{1}{2})^2$, and the overall upper bound 0.

The bounds on the $\mu$-series and $\sigma^2$-series provide more insight in the behavior of the model. However, we are primarily interested in bounds on $\mu_X$ and $\sigma^2_X$. In Sec. 6.8 we present the bounds on $\mu_X$ and $\sigma^2_X$ for the same distributions as in Sec. 6.7. These bounds will be shown to be sharp, both for the low and high load situations.

6.3 Representations of the $\mu$-series and $\sigma^2$-series

In this section we take a closer look at the random variables $Y$ and $Z$ as defined by (6.8), and we show that they give rise to the representations

\[ \sum_{k=1}^{s-1} \frac{1}{1 - z_k} = \frac{1}{2}(s - 1) + \frac{1}{2}\mu_Y - \frac{\sigma^2_Y}{2(s - \mu_Y)} \]  \hspace{1cm} (6.24)

\[ s(s - 1) - \frac{Y''(1)}{2(s - \mu_Y)} = \frac{s^2 - E(Y^2)}{2(s - \mu_Y)} - \frac{1}{2} \]  \hspace{1cm} (6.25)

\[ = \frac{1}{2}(s - 1) + \frac{1}{2}\mu_Z. \]  \hspace{1cm} (6.26)

for the $\mu$-series, and

\[ \sum_{k=1}^{s-1} \frac{z_k}{(1 - z_k)^2} = \frac{Y''(1) - s(s - 1)(s - 2)}{3(s - \mu_Y)} + \frac{Y''(1) - s(s - 1)}{2(s - \mu_Y)} + \frac{(Y''(1) - s(s - 1))^2}{2(s - \mu_Y)} \]  \hspace{1cm} (6.27)

\[ = \frac{1}{4} \left( \frac{s^2 - E(Y^2)}{s - \mu_Y} \right)^2 - \frac{1}{3} \frac{s^3 - E(Y^3)}{s - \mu_Y} + \frac{1}{12} \]  \hspace{1cm} (6.28)

\[ = -\frac{1}{12}(s - \mu_Z)^2 - \frac{1}{3}\sigma^2_Z + \frac{1}{12}, \]  \hspace{1cm} (6.29)

for the $\sigma^2$-series.

We note that $Y(z)$ has degree $s$ and that the roots of $Y(z) = z^s$ are precisely $z_0 = 1, z_1, \ldots, z_{s-1}$. The latter statement follows from the fact that the numerator
\( A(z) \sum_{j=0}^{s} x_j (z^s - z^j) \) at the right-hand side of (2.8) has to cancel the \( s \) zeros of the denominator \( z^s - A(z) \) within the closed unit disk \( |z| \leq 1 \) (when \( A(0) = 0 \) some trivial modifications are required). As a consequence, the random variables \( Y \) and \( A \) give rise to the same \( \mu \)-series and \( \sigma^2 \)-series while \( \mathbb{P}(Y > s) = 0 \). It follows from (2.10) that
\[
s - \mu_A = (s - \mu_Y) \mathbb{P}(X \leq s),
\]
and thus \( \mu_Y \leq \mu_A \) with equality if and only if \( \mathbb{P}(X > s) = 0 \). From the process definition we see furthermore that
\[
A = X = Y \iff \mathbb{P}(A > s) = 0.
\]

We now derive the representations (6.24)-(6.26) and (6.27)-(6.29). The representations (6.24), (6.27) follow from the observation that \( A \) and \( Y \) yield the same \( \mu \)-series and \( \sigma^2 \)-series, and the fact that \( \mathbb{P}(Y > s) = 0 \), so that (6.24) and (6.27) result from consideration of the process definition and application of (6.1), (6.2) with \( Y \) instead of \( A \). The derivation of (6.25) and (6.28) follows from straightforward rewriting.

Finally, we show the representations (6.26), (6.29). The former follows from
\[
\frac{s^2 - \mathbb{E}(Y^2)}{s - \mu_Y} = \frac{1}{s - \mu_Y} \sum_{j=0}^{s} (s^2 - j^2) \mathbb{P}(Y = j)
= \frac{1}{(s - \mu_Y) \mathbb{P}(X \leq s)} \sum_{j=0}^{s} (s + j)(s - j)x_j
= \frac{s - \mu_A}{(s - \mu_Y) \mathbb{P}(X \leq s)} \mathbb{E}(s + Z) = s + \mu_Z,
\]
where we have used the definitions of \( Y \) and \( Z \) together with (6.30). Similarly, we have
\[
\frac{s^3 - \mathbb{E}(Y^3)}{s - \mu_Y} = \mathbb{E}(s^2 + sZ + Z^2) = s^2 + s\mu_Z + \mathbb{E}(Z^2),
\]
and (6.29) follows after some manipulation.

We shall now be concerned with the question how certain concentration properties of \( Y \) (and \( Z \)) are reflected by corresponding properties of \( A \). The result given below is vital in Secs. 6.4, 6.5 for settling cases of equality in our theorems.

**Definition 6.3.1** Let \( B \) be a random variable with values in \( \{0, 1, \ldots\} \) and let \( S \) be a subset of \( \{0, 1, \ldots\} \). We say that \( B \) is concentrated on \( S \) when \( \mathbb{P}(B \notin S) = 0 \).

According to this definition \( Y \) is concentrated on \( \{0, 1, \ldots, s\} \) while \( Z \) is concentrated on \( \{0, 1, \ldots, s-1\} \). Moreover, we have the following result.

**Lemma 6.3.2** (i) Let \( j = 0, 1, \ldots, s-1 \). Then \( Y \) concentrated on \( \{j\} \) \( \iff \) \( A \) concentrated on \( \{j\} \).
6.4 Bounds for the $\mu$-series

(ii) Let $j = 0, 1, \ldots, s - 2$. Then $Y$ concentrated on $\{j, j + 1\} \iff A$ concentrated on $\{j, j + 1\}$.

(iii) $Y$ concentrated on $\{s - 1, s\} \iff A$ concentrated on $\{s - 1, s, s + 1, \ldots\}$.

(iv) $Y$ concentrated on $\{0, s\} \iff Z$ concentrated on $\{0\} \iff A$ concentrated on $\{0, s, 2s, \ldots\}$.

We omit the proof of Lemma 6.3.2, but it follows from carefully analyzing the process definition.

6.4 Bounds for the $\mu$-series

In this section we prove (the claims associated with) Thm. 6.2.1. From the process definition in (2.3) we see that

\[
\mu_X \geq \mu_A.
\]

So from (6.1) it follows that

\[
\sum_{k=1}^{s-1} \frac{1}{1 - z_k} \geq \frac{1}{2} (s - 1) + \frac{1}{2} \mu_A - \frac{\sigma_A^2}{2(s - \mu_A)},
\]

with equality if and only if $A$ is concentrated on $\{0, \ldots, s\}$. We further see from representation (6.26) that

\[
\sum_{k=1}^{s-1} \frac{1}{1 - z_k} \geq \frac{1}{2} (s - 1),
\]

and there is equality if and only if $A$ is concentrated on $\{0, s, 2s, \ldots\}$ (see Lemma 6.3.2(iv)). Next we consider the representation (6.24) in which the $\mu$-series is expressed in terms of the mean and variance of $Y$. Observe that for any random variable $B$ concentrated on $\{0, \ldots, s\}$ with mean $\mu$ the smallest value of $\sigma_B^2$ is given by $\langle \mu \rangle - \langle \mu \rangle^2$ (as defined in Thm. 6.2.1), and is assumed when

\[
P(B = \lfloor \mu \rfloor) = 1 - \langle \mu \rangle, \quad P(B = \lfloor \mu \rfloor + 1) = \langle \mu \rangle.
\]

The function $f$ as defined by (6.12) is strictly increasing in $\mu \in [0, s - 1]$, and constant, $s - 1$, for $\mu \in [s - 1, s)$. We thus have

\[
\sum_{k=1}^{s-1} \frac{1}{1 - z_k} \leq f(\mu_Y) \leq f(\mu_A) \leq \frac{1}{2} (s - 1) + \frac{1}{2} \min \{\mu_A, s - 1\}.
\]

In the first inequality there is equality if and only if $\mu_Y = 0, 1, \ldots, s - 1$ and $Y$ is concentrated on $\{\mu_Y\}$, or $\mu_Y$ is non-integer and $Y$ is concentrated on $\{\lfloor \mu_Y \rfloor, \lfloor \mu_Y \rfloor + 1\}$. In the second inequality there is equality if and only if $\mu_Y < s - 1$ and $\mu_A = \mu_Y$, or $s - 1 \leq \mu_Y < s$. In the third inequality there is equality if and only if $\mu_A = 0, 1, \ldots, s - 2$ or $\mu_A \geq s - 1$. The inequalities (6.34) and (6.35) together with the second inequality in (6.37) prove Thm. 6.2.1. Also, the case of equality in the third inequality in (6.37) is settled now: It holds if and only if $A$ is concentrated on $\{j\}$ with $j = 0, 1, \ldots, s - 2$ or $A$ is concentrated on $\{s - 1, s, s + 1, \ldots\}$.
6.5 Bounds for the \( \sigma^2 \)-series

In this section we prove Thms. 6.2.2-6.2.3. We first derive bounds for the \( \sigma^2 \)-series that depend on the mean and the variance of \( Y \), from which we derive bounds that depend on \( \mu_A \). We consider the representation (6.28) in which the \( \sigma^2 \)-series is expressed in terms of \( \mu_Y \), \( \sigma^2_Y \) and \( E(Y^3) \). We are interested in the smallest and largest value of (6.28) under the condition that \( \mu_Y \) and \( \sigma^2_Y \) take prescribed values. For convenience we assume \( Y \) takes, not necessarily integer, values between 0 and \( s \), and that \( 0 < \mu_Y < s \). Under these assumptions, we have

\[
0 < \theta := \frac{\mu_Y}{s} < 1, \quad 0 \leq \omega := \frac{\sigma^2_Y}{\mu_Y(s - \mu_Y)} \leq 1,
\]

(6.38)

and equality in the last inequality occurs if and only if \( Y \) is concentrated on \( \{0, s\} \). We start by presenting a lemma.

**Lemma 6.5.1** Let \( D \) be a random variable with values in \([-c, d]\), where \( c \geq 0 \), \( d \geq 0 \), and assume that \( \mu_D = 0 \), \( \sigma^2_D = \sigma^2 \) is fixed. Then the minimum and maximum value of \( E(D^3) \) are given by

\[
\frac{\sigma^4}{c} - c\sigma^2, \quad d\sigma^2 - \frac{\sigma^4}{d},
\]

(6.39)

respectively. The minimum and maximum value occur when \( D \) is concentrated on \( \{-c, \sigma^2/c\} \) and \( \{-\sigma^2/d, d\} \), respectively.

The proof of this result follows from Thm. 2.4 in Krein & Nudelman [106], as was kindly communicated to us by E. Verbitskiy.

We next present three results from which Thm. 6.2.2 follows. In Thms. 6.5.2-6.5.4 the random variable \( Y \) is allowed to take non-integer values in \([0, s]\) and \( \theta, \omega \) are as in (6.38).

**Theorem 6.5.2** We have

\[
\sum_{k=1}^{s-1} \frac{z_k}{(1 - z_k)^2} \geq -\frac{1}{12} s^2(1 - \theta + \omega \theta)^2 + \frac{1}{12} - \frac{1}{3} s^2(1 - \omega)\theta \omega,
\]

(6.40)

\[
\sum_{k=1}^{s-1} \frac{z_k}{(1 - z_k)^2} \leq -\frac{1}{12} s^2(1 - \theta + \omega \theta)^2 + \frac{1}{12}.
\]

(6.41)

The lower bound is assumed if and only if \( Y \) is concentrated on

\[
\{0, \mu_Y + \frac{\sigma^2_Y}{\mu_Y}\} = \{0, s\omega + s(1 - \omega)\theta\},
\]

(6.42)

and the upper bound is assumed if and only if \( Y \) is concentrated on

\[
\{\mu_Y - \frac{\sigma^2_Y}{s - \mu_Y}, s\} = \{s(1 - \omega)\theta, s\}.
\]

(6.43)
6.5 Bounds for the $\sigma^2$-series

Theorem 6.5.3 We have

$$-\frac{s^2}{3(4-\theta)} + \frac{1}{12} \leq \sum_{k=1}^{s-1} \frac{z_k}{(1-z_k)^2} \leq -\frac{1}{12} (1-\theta)^2 + \frac{1}{12}. \quad (6.44)$$

The lower bound is assumed if and only if $Y$ is concentrated on the set in (6.42) with $\omega = (3-\theta)/(4-\theta)$, and the upper bound is assumed if and only if $Y$ is concentrated on the set in (6.43) with $\omega = 0$.

Theorem 6.5.4 We have

$$-\frac{1}{9} s^2 + \frac{1}{12} \leq \sum_{k=1}^{s-1} \frac{z_k}{(1-z_k)^2} \leq \frac{1}{12}. \quad (6.45)$$

The lower bound is assumed if and only if $Y$ is concentrated on the set in (6.42) with $\omega = (3-\theta)/(4-\theta) \rightarrow \frac{2}{3}$ and $\theta \uparrow 1$, and the upper bound is assumed if and only if $Y$ is concentrated on the set in (6.43) with $\omega = 0$ and $\theta \uparrow 1$.

Proofs It is convenient to combine the proofs of the above results. We rewrite representation (6.28) using

$$E(Y^2) = \sigma_Y^2 + \mu_Y^2, \quad E(Y^3) = m^3_Y + 3\mu_Y\sigma_Y^2 + \mu_Y^3, \quad (6.46)$$

where $m^3_Y = E((Y-\mu_Y)^3)$. This yields

$$\sum_{k=1}^{s-1} \frac{z_k}{(1-z_k)^2} = -\frac{1}{12} (s - \mu_Y)^2 - \frac{1}{2} \sigma_Y^2 + \left(\frac{\sigma_Y^2}{2(s - \mu_Y)}\right)^2 + \frac{m^3_Y}{3(s - \mu_Y)} + \frac{1}{12}. \quad (6.47)$$

We then use Lemma 6.5.1 with $D = Y - \mu_Y$, $c = \mu_Y$, $d = s - \mu_Y$ and some rewriting, to see that

$$\sum_{k=1}^{s-1} \frac{z_k}{(1-z_k)^2} \geq -\frac{1}{12} \left(s - \mu_Y + \frac{\sigma_Y^2}{s - \mu_Y}\right)^2 + \frac{1}{12}$$

$$- \frac{s\sigma_Y^2}{3(s - \mu_Y)} \left(1 - \frac{\sigma_Y^2}{\mu_Y(s - \mu_Y)}\right), \quad (6.48)$$

$$\sum_{k=1}^{s-1} \frac{z_k}{(1-z_k)^2} \leq -\frac{1}{12} \left(s - \mu_Y + \frac{\sigma_Y^2}{s - \mu_Y}\right)^2 + \frac{1}{12}, \quad (6.49)$$

with equality if and only if $Y$ is concentrated on $\{0, \mu_Y + \sigma_Y^2/\mu_Y\}$ and on $\{\mu_Y - \sigma_Y^2/(s - \mu_Y), s\}$, respectively. The inequalities in (6.48) and (6.49) can be written succinctly, in terms of $\theta$, $\omega$ as (6.40) and (6.41), respectively, and this shows Thm. 6.5.2.

For fixed $\theta \in (0, 1)$, the minimum of (6.40) as a function of $\omega \in [0, 1]$ equals $-s^2/(4(3-\theta)) + 1/12$ and occurs uniquely at $\omega = (3-\theta)/(4-\theta)$. The maximum
of (6.41) equals \(-s^2(1 - \theta)^2/12 + 1/12\) and occurs uniquely at \(\omega = 0\). This shows Thm. 6.5.3.

Finally, the minimum of the first member of (6.44) as a function of \(\theta \in (0, 1)\) equals \(-\frac{1}{s}\left(s^2 + 1/12\right)\) and occurs uniquely when \(\theta \uparrow 1\) and thus \(\omega = (3 - \theta)/(4 - \theta) \to 2/3\), while the maximum of the third member of (6.44) equals 1/12 and occurs uniquely when \(\theta \uparrow 1\) and thus \(\omega = 0\). This then also shows Thm. 6.5.4. □

The bounds in Thm. 6.2.2 are in terms of \(\mu_A\). They can be obtained straightforwardly from Thm. 6.5.3 by noting that \(\mu_Y \leq \mu_A\) and the fact that the first member in (6.44) is decreasing in \(\theta\) while the third member in (6.44) is increasing in \(\theta\). A corresponding result for the inequalities in (6.40) and (6.41) is unlikely to hold since the relation between \(\sigma_Y^2\) and \(\sigma_A^2\) seems much more awkward. Note once more that \(Y = A\) when \(A\) is concentrated on \(\{0, 1, \ldots, s\}\), and then Thms. 6.5.2-6.5.4 hold with \(Y\) replaced by \(A\).

In Thms. 6.5.2-6.5.4 the discrete nature of the random variables has been disregarded. Accordingly, the two bounds in (6.40) and (6.41) are achieved by some integer-valued \(Y\) if and only if
\[
\mu_Y + \frac{\sigma_Y^2}{\mu_Y} = s\omega + s(1 - \omega)\theta \in \mathbb{Z}, \quad (6.50)
\]
\[
\mu_Y - \frac{\sigma_Y^2}{s - \mu_Y} = s(1 - \omega)\theta \in \mathbb{Z}, \quad (6.51)
\]
respectively. In general, when these integrality conditions are not met, slight improvement of the bounds in Thm. 6.5.2 can be achieved by invoking an appropriate discrete version of Lemma 6.5.1 in Formula (6.47). This then gives rise to two guirlanded \((\mu, \sigma)\)- or \((\theta, \omega)\)-surfaces, with contact curves described by (6.50) and (6.51), just as we had a guirlanded graph in Thm. 6.2.1 for the upper bound for the \(\mu\)-series (since the lower bound is constant and achievable by \(Y\) concentrated on \(\{0, s\}\), no guirlande phenomenon occurs for the lower bound of the \(\mu\)-series).

A slight improvement of the upper bound in (6.44) can be obtained by observing that \(\sigma_Y^2 \geq \langle \mu_Y \rangle - \langle \mu_Y \rangle^2\) when \(Y\) is integer-valued. Thus we find, see (6.49), in a similar fashion as in Sec. 6.4 for the \(\mu\)-series
\[
\sum_{k=1}^{s-1} \frac{z_k}{(1 - z_k)^2} \leq -\frac{1}{12} \left(s - \mu_Y + \langle \mu_Y \rangle - \langle \mu_Y \rangle^2\right)^2 + \frac{1}{12}
\]
\[
= -\frac{1}{12} (2s - 1 - 2f(\mu_Y))^2 + \frac{1}{12}
\]
\[
\leq -\frac{1}{12} (2s - 1 - 2f(\mu_A))^2 + \frac{1}{12} =: g(\mu_A) \leq 0, \quad (6.52)
\]
with \(f\) as in Thm. 6.2.1.

We may also observe the bounds
\[
-\frac{1}{9} (s - \frac{1}{2})^2 \leq \sum_{k=1}^{s-1} \frac{z_k}{(1 - z_k)^2} \leq 0, \quad (6.53)
\]
and their simple proofs from the representation (6.29) in terms of \( W \). Indeed, consider an arbitrary random variable \( C \) concentrated on \( \{0, 1, \ldots, s-1\} \) with mean \( \mu \) and variance \( \sigma^2 \). When \( \mu \) is fixed, the minimum value of
\[
-\frac{1}{12} (s - \mu)^2 - \frac{1}{3} \sigma^2 + \frac{1}{12}
\]
occurs when \( C \) is concentrated on \( \{0, s-1\} \) and equals
\[
-\frac{1}{9} (s - \frac{1}{2})^2 + \frac{1}{4} (\mu - \frac{1}{3} (s-2))^2 \geq -\frac{1}{9} (s - \frac{1}{2})^2.
\]
Similarly, the maximum value of (6.54) occurs when \( C \) is concentrated on \( \{\mu\} \) or on \( \{\lfloor \mu \rfloor, \lfloor \mu \rfloor + 1\} \) (\( \mu \) non-integer) and equals
\[
-\frac{1}{12} (s - \mu)^2 - \frac{1}{3} (\langle \mu \rangle - (\langle \mu \rangle)^2) + \frac{1}{12} \leq 0,
\]
with equality if and only if \( \mu = s - 1 \).

In Fig. 6.2 we have plotted the bounds in (6.44), (6.52) and (6.53) for \( s = 5 \) and \( 0 \leq \mu_A < s \). Observe that the graph of \( g \) hangs down from \( -\frac{1}{12} (s - \mu)^2 + \frac{1}{12} \) as a guirlande with nodes at all integers \( \mu = 0, 1, \ldots, s - 1 \).

We conclude this section by proving Thm. 6.2.3. Theorem 6.2.3(i) follows at once from (6.2) and the fact that \( \sigma^2_X \geq \sigma^2_3 \), with equality if and only if \( A \) is concentrated on \( \{0, 1, \ldots, s\} \). As to Thm. 6.2.3(ii), we start from the representation (6.27) in which we write
\[
Y'''(1) = \mathbb{E}(Y(Y-1)(Y-2)) = \mathbb{E}(h(Y)),
\]
with $h$ given in (6.17). In (6.57) the last identity follows from the fact that $Y$ is integer-valued. The function $h$ is convex on $[0, \infty)$ and strictly convex on $[2, \infty)$, whence by Jensen’s inequality we have that

$$\mathbb{E}(h(Y)) \geq h(\mathbb{E}(Y)) = h(\mu_Y),$$

(6.58)

with equality if and only if $Y$ is concentrated on $\{0, 1, 2\}$ or $Y$ is concentrated on $\{j\}$ with $j = 2, 3, \ldots, s - 1$. Next we observe from convexity of $h$ that the function

$$\frac{Y''(1) - s(s-1)(s-2)}{3(s-\mu_Y)} \geq \frac{h(\mu_Y) - h(s)}{3(s-\mu_Y)} \geq \frac{h(\mu_A) - h(s)}{3(s-\mu_A)},$$

(6.59)

with equality if and only $\mu_A = \mu_Y$. We next turn to the quantity

$$\frac{(s(s-1) - Y''(1))}{2(s-\mu_Y)} - \frac{s(s-1) - Y''(1)}{2(s-\mu_Y)},$$

(6.60)

which occurs at the right-hand side of (6.27). We note from (6.25) that

$$\frac{s(s-1) - Y''(1)}{2(s-\mu_Y)} \geq \frac{1}{2}(s-1).$$

(6.61)

Furthermore, we have from (6.25) and Thm. 6.2.1(i) that

$$\frac{s(s-1) - Y''(1)}{2(s-\mu_Y)} \geq \frac{1}{2}(s-1) + \frac{1}{2}\mu_A - \frac{\sigma_A^2}{2(s-\mu_A)} = \frac{s(s-1) - A''(1)}{2(s-\mu_A)}.$$ (6.62)

Denoting the far left-hand side of (6.62) by $x_Y$ and the far right-hand side of (6.62) by $x_A$, we have $x_Y \geq \frac{1}{2}(s-1)$ and $x_A \geq \frac{1}{2}(s-1)$, whence

$$(x_Y^2 - x_Y) - (x_A^2 - x_A) = (x_Y - x_A)(x_Y + x_A - 1) \geq 0,$$

(6.63)

whenever $x_A \geq -\frac{1}{2}(s-1) + 1$. This latter condition can be worked out to yield constraint (6.19). Hence, under this constraint, (6.16) follows. The cases with equality easily follow from what has been said in connection with occurrence of equality in (6.58) and (6.59).

### 6.6 Detailed results for the Poisson distribution

In case one has, or wants to use, more knowledge on the distribution of $A$, sharper bounds can be derived. For example, the Kingman upper bound in case of the discrete $D/G/1$ queue (6.4) can be sharpened by using the quantity $\mathbb{P}(A < s)$ to give (see Daley et al. [58], (3.11))

$$\sum_{k=1}^{r-1} \frac{1}{1 - z_k} \leq \frac{1}{2}(s-1) + \frac{1}{2}\mu_A - \frac{1}{2}(\mathbb{P}(A < s)^{-1} - 1)(s-\mu_A).$$

(6.64)
In this section we show for the case that $A$ is distributed according to a Poisson distribution, i.e.

$$a_j = e^{-\lambda} \frac{\lambda^j}{j!}, \quad j = 0, 1, \ldots; \quad A(z) = e^{\lambda(z-1)},$$

(6.65)

that the $\mu$-series and $\sigma^2$-series are monotone functions of $\mu_A$, which facilitates a sharpening of the lower bounds for both series. For that, we consider the curve on which the roots of $A(z) = z^s$ lie, and prove some properties for all points on this curve.

We have

$$\mu_A = \sigma_A^2 = \lambda; \quad A^{(k)}(1) = \lambda^k,$$

(6.66)

with $A^{(k)}(1)$ the $k$-th derivative of $A(z)$ evaluated at $z = 1$. The roots $z_0, z_1, \ldots, z_{s-1}$ now occur on, what we have called, the generalized Szegő curve

$$S_\theta = \{ z \in \mathbb{C} : |z| \leq 1, |z| = |e^{\theta(z-1)}| \}, \quad \theta := \lambda/s.$$

(6.67)

In Fig. 6.3 some examples of $S_\theta$ are plotted.

We now introduce two useful parameterizations of $S_\theta$. First, we represent a point $z$ on $S_\theta$ as

$$z = r_\theta(\varphi)e^{i\varphi}, \quad 0 \leq \varphi \leq 2\pi,$$

(6.68)

where $0 \leq r_\theta(\varphi) \leq 1$. In (6.67) and (6.68) we allow $\theta = 1$, i.e. $\lambda = s$. It holds that

$$r_\theta(\varphi) = \exp\{\theta(r_\theta(\varphi) \cos \varphi - 1)\}, \quad 0 \leq \varphi \leq 2\pi,$$

(6.69)

and so it follows that

$$\frac{d}{d\theta} (r_\theta(\varphi)) = \frac{r_\theta(\varphi)(r_\theta(\varphi) \cos \varphi - 1)}{1 - \theta r_\theta(\varphi) \cos \varphi} \leq 0,$$

(6.70)
which yields the result that for $0 \leq \theta \leq 1$

$$r_1(\varphi) \leq r_\theta(\varphi), \quad 0 \leq \varphi \leq 2\pi. \quad (6.71)$$

This proves that the interior of $S_1$ is a root-free region for any $\theta \leq 1$.

A second parametrization of $S_\theta$ is obtained by solving for $\alpha \in [0, 2\pi]$ the equation

$$ze^{\theta(1-z)} = e^{i\alpha}. \quad (6.72)$$

Denoting the solution of (6.72) by $z_\theta(\alpha)$, we have the following Fourier series representation, see (3.57), where this is done for more general $A$ as well,

$$z_\theta(\alpha) = \sum_{l=1}^{\infty} e^{-i\varphi} \frac{(\theta l)l^{-1}}{l!} e^{il\alpha}, \quad \alpha \in [0, 2\pi]. \quad (6.73)$$

This allows convenient computation of all $z_k$’s, since

$$z_k = z_{k,\theta} = z_\theta(2\pi k/s), \quad k = 0, 1, \ldots, s-1. \quad (6.74)$$

Using the parameterizations of $S_\theta$, we derive the following results.

**Lemma 6.6.1** For any $z$ on the generalized Szegő curve $S_\theta$, it holds that

$$\text{Re} \left[ \frac{z}{(1-z)(1-\theta z)} \right] \leq 0, \quad (6.75)$$

with equality if and only if $z \to 1$.

**Proof** With $z = re^{i\varphi}$, we get

$$\text{Re} \left[ \frac{z}{(1-z)(1-\theta z)} \right] = \frac{r}{|1-z|^2|1-\theta z|^2} \text{Re}[e^{i\varphi}(1-re^{-i\varphi})(1-\theta re^{-i\varphi})]$$

$$= \frac{r}{|1-z|^2|1-\theta z|^2} (\cos \varphi - (1 + \theta)r + \theta r^2 \cos \varphi),$$

and it suffices to show that, omitting the subindex $\theta$ in $r_\theta$ for notational convenience,

$$g(\varphi) := (1 + \theta r^2(\varphi)) \cos \varphi - (1 + \theta)r(\varphi) \leq 0, \quad (6.76)$$

with equality if and only if $\varphi = 0$. Here it is evidently sufficient to consider the case that $\cos \varphi > 0$, $\varphi \geq 0$, i.e. $\varphi \in [0, \frac{1}{2}\pi]$. There is indeed equality in (6.76) when $\varphi = 0$ since $r(0) = 1$. It follows from (6.69) that

$$\frac{d}{d\varphi} r(\varphi) = \frac{-\theta r^2(\varphi) \sin \varphi}{1 - \theta r(\varphi) \cos \varphi}, \quad (6.77)$$
and hence
\[
g'(\varphi) = -(1 + \theta r^2(\varphi))\sin \varphi + (2\theta r(\varphi) \cos \varphi - 1 - \theta)r'(\varphi)
\]
\[
= -(1 + \theta r^2(\varphi))\sin \varphi - \frac{(2\theta r(\varphi) \cos \varphi - 1 - \theta)\theta^2 r^2(\varphi) \sin \varphi}{1 - \theta r(\varphi) \cos \varphi}
\]
\[
= \frac{-\sin \varphi}{1 - \theta r(\varphi) \cos \varphi}[(1 + \theta r^2(\varphi))(1 - \theta r(\varphi) \cos \varphi)
\]
\[
+ (2\theta r(\varphi) \cos \varphi - 1 - \theta)\theta^2 r^2(\varphi)]
\]
\[
= \frac{-\sin \varphi}{1 - \theta r(\varphi) \cos \varphi}[1 - \theta r(\varphi) \cos \varphi - \theta^2 r^2(\varphi)(1 - r(\varphi) \cos \varphi)] \quad (6.78)
\]

Now, as \(\cos \varphi > 0\) and \(\theta \leq 1\),
\[
1 - \theta r(\varphi) \cos \varphi - \theta^2 r^2(\varphi)(1 - r(\varphi) \cos \varphi)
\]
\[
\geq 1 - r(\varphi) \cos \varphi - \theta^2 r^2(\varphi)(1 - r(\varphi) \cos \varphi)
\]
\[
= (1 - r(\varphi) \cos \varphi)(1 - \theta^2 r^2(\varphi)) \geq 0 \quad (6.79)
\]
with equality in the last inequality if and only if \(\varphi = 0\). Thus \(g'(\varphi) < 0\) for \(\varphi > 0\), and it follows that (6.76) is smaller than or equal to zero, with equality if and only if \(\varphi = 0\). This completes the proof. \(\square\)

Lemma 6.6.2 The \(\mu\)-series in case of \(A(z) = e^{\theta z(z-1)}\) is increasing in \(\theta \in [0,1]\).

Proof From
\[
z_\theta(\alpha) = e^{iz}e^{\theta(z_\theta(\alpha)-1)}, \quad \frac{dz_\theta(\alpha)}{d\theta} = \frac{z_\theta(\alpha)(z_\theta(\alpha) - 1)}{1 - \theta z_\theta(\alpha)}, \quad (6.80)
\]
we obtain
\[
\frac{d}{d\theta}(1 - z_\theta(\alpha))^{-1} = \frac{1}{(1 - z_\theta(\alpha))^2} \frac{dz_\theta(\alpha)}{d\theta} = \frac{-z_\theta(\alpha)}{(1 - z_\theta(\alpha))(1 - \theta z_\theta(\alpha))}. \quad (6.81)
\]
Applying Lemma 6.6.1 then shows that the real part of (6.81) is non-negative for each point on \(S_\theta\), and thus for all roots \(z_1, \ldots, z_{s-1}\). \(\square\)

Lemma 6.6.3 The \(\sigma^2\)-series in case of \(A(z) = e^{\theta z(z-1)}\) is increasing in \(\theta \in [0,1]\).

Proof It is readily seen that
\[
\frac{d}{d\theta} \left( \frac{z_\theta(\alpha)}{(1 - z_\theta(\alpha))^2} \right) = \frac{-z_\theta(\alpha)}{(1 - z_\theta(\alpha))(1 - \theta z_\theta(\alpha))} \frac{1 + z_\theta(\alpha)}{1 - z_\theta(\alpha)}, \quad (6.82)
\]
and thus
\[
\text{Re} \left[ \frac{d}{d\theta} \left( \frac{z_\theta(\alpha)}{(1 - z_\theta(\alpha))^2} \right) \right] = \text{Re} \left[ \frac{-z_\theta(\alpha)}{(1 - z_\theta(\alpha))(1 - \theta z_\theta(\alpha))} \right] \text{Re} \left[ \frac{1 + z_\theta(\alpha)}{1 - z_\theta(\alpha)} \right]
\]
\[
- \text{Im} \left[ \frac{-z_\theta(\alpha)}{(1 - z_\theta(\alpha))(1 - \theta z_\theta(\alpha))} \right] \text{Im} \left[ \frac{1 + z_\theta(\alpha)}{1 - z_\theta(\alpha)} \right]. \quad (6.83)
\]
First note that with $z = re^{i\varphi}$
\[
\text{Im}\left[\frac{z}{(1-z)(1-\theta z)}\right] = \frac{r}{|1-z|^2|1-\theta z|^2} \cdot \text{Im}[e^{i\varphi}(1-re^{-i\varphi})(1-\theta re^{-i\varphi})]
\]
\[
= \frac{r(1-\theta r^2)}{|1-z|^2|1-\theta z|^2} \cdot \sin \varphi.
\] (6.84)

Furthermore, we have
\[
\frac{1+z}{1-z} = \frac{1}{|1-z|^2}(1-r^2+2ir \sin \varphi),
\] (6.85)
whence
\[
\text{Re}\left[\frac{1+z}{1-z}\right] = \frac{1-r^2}{|1-z|^2}, \quad \text{Im}\left[\frac{1+z}{1-z}\right] = \frac{2r}{|1-z|^2} \cdot \sin \varphi.
\] (6.86)

This, together with Lemma 6.6.1, shows that both members at the right-hand side of (6.83) are $\geq 0$, and thus the real part of (6.82) is non-negative for each point on $\mathbb{S}_\theta$, including all roots $z_1, \ldots, z_{s-1}$.

Combining the monotonicity of the $\mu$-series and $\sigma^2$-series, as proven in Lemmas 6.6.2 and 6.6.3, and the bounds in Thms. 6.2.1 and 6.2.3 yield the proofs of Thms. 6.2.4 and 6.2.5.

Figs. 6.4 and 6.6 display the $\mu$-series and the bounds in Thm. 6.2.4 for $s = 20$ and $s = 100$, respectively, with $\frac{1}{3}(s-1)$ as an overall lower bound. The more general lower bound arising from Thm. 6.2.1 is also plotted. Figs. 6.5 and 6.7 display the $\sigma^2$-series and the bounds in Thm. 6.2.5 for $s = 20$ and $s = 100$, respectively, where $-\frac{1}{5}(s-\frac{1}{2})^2$ holds as an overall lower bound and as the lower bound when condition (6.19), i.e. $\lambda \leq 19.64$ for $s = 20$ and $\lambda \leq 99.66$ for $s = 100$, is not met. The more general lower bound arising from Thm. 6.2.3 is also plotted. In Figs. 6.4-6.7 it is nicely demonstrated that the lower bound is sharpened substantially when monotonicity can be proven. We conjecture that monotonicity of the $\mu$-series and $\sigma^2$-series also hold for distributions of $A$ other than Poisson, e.g. the binomial and geometric distribution.

### 6.7 Numerical examples of series bounds

In this section we first present some more examples of distributions of $A$ to illustrate the behavior of the $\mu$-series and $\sigma^2$-series and the sharpness of the bounds in Thms. 6.2.1 and 6.2.3.

The $\mu$-series and $\sigma^2$-series can be computed numerically by finding the roots $z_1, \ldots, z_{s-1}$, which is feasible in the cases below. We display the $\mu$-series and $\sigma^2$-series, with corresponding lower and upper bounds, for a number of parametrically given $A$ in which $\mu_A$ covers the whole range of permitted values below $s = 5$. For these cases we also give $\mu_A$, $\sigma_A^2$, $A''(1)$ and $A'''(1)$, as required in the various bounds.
6.7 Numerical examples of series bounds

For the $\mu$-series we employ the bounds in Thm. 6.2.1 together with $\frac{1}{2}(s-1)$ as an overall lower bound. For the $\sigma^2$-series we employ the bounds in Thm. 6.2.3, where the lower bound (6.18) is only used when condition (6.19) is satisfied. If not, we use the overall lower bound $-\frac{1}{9}(s-\frac{1}{2})^2$, and the overall upper bound 0.

**Example 6.7.1** Take $a_n = 1 - a$, $a_{n+1} = a$ where $a \in [0,1]$ and $n = 0, 1, \ldots$, so that

$$A(z) = (1-a)z^n + az^{n+1}. \quad (6.87)$$

We have

$$\mu_A = n + a, \quad \sigma^2_A = a - a^2, \quad (6.88)$$

and for $k = 2, 3, \ldots$

$$A^{(k)}(1) = n(n-1) \cdots (n-k+2)(n+1 - (1-a)k). \quad (6.89)$$
Fig. 6.8 and Fig. 6.9 display the $\mu$-series and $\sigma^2$-series for $s = 10$, $\mu_A \in [0, s)$, i.e.
$0 \leq n \leq s - 1$, $a \in [0, 1)$. Note that the $\mu$-series and its lower and upper bound equal
the guirlande upper bound. The graph of the $\sigma^2$-series is given by the right-hand
side of (6.52), and coincides with the upper bound. In this case, we have
\[
\mathbb{P}(Z = n) = \frac{(s - n)(1 - a)}{s - n - a}, \quad \mathbb{P}(Z = n + 1) = \frac{(s - n - 1)a}{s - n - a},
\tag{6.90}
\]
and so there is no need for numerical determination of the roots. Instead, since
$X = A = I$, we could use representation (6.24) and (6.29).

**Example 6.7.2** Take $a_0 = 1 - \mu/s$, $a_s = \mu/s$ with $\mu \in [0, s)$, so that
\[
A(z) = (1 - \frac{\mu}{s}) + \frac{\mu}{s}z^s.
\tag{6.91}
\]
We have
\[
\mu_A = \mu, \quad \sigma^2_A = \mu(s - \mu),
\tag{6.92}
\]
and for $k = 2, 3, \ldots$
\[
A^{(k)}(1) = \mu(s - 1)(s - 2) \cdots (s - k + 1).
\tag{6.93}
\]
Note that $z_k = \exp(2\pi ik/s)$, and thus
\[
\sum_{k=1}^{s-1} \frac{1}{1 - z_k} = \frac{1}{2}(s - 1), \quad \sum_{k=1}^{s-1} \frac{z_k}{(1 - z_k)^2} = -\frac{1}{12}(s^2 - 1),
\tag{6.94}
\]
which can also be found using (6.24) and (6.29), and the fact that $\mathbb{P}(Z = 0) = 1$.

**Example 6.7.3** Take $a_0 = 1 - \mu/(s - 1)$, $a_{s-1} = \mu/(s - 1)$ with $\mu \in [0, s - 1]$, so that
\[
A(z) = (1 - \frac{\mu}{s - 1}) + \frac{\mu}{s - 1}z^{s-1}.
\tag{6.95}
\]
We have
\[
\mu_A = \mu, \quad \sigma^2_A = \mu(s - 1 - \mu),
\tag{6.96}
\]
and for $k = 2, 3, \ldots$
\[
A^{(k)}(1) = \mu(s - 2)(s - 3) \cdots (s - k).
\tag{6.97}
\]
We also compute
\[
\mathbb{P}(Z = 0) = \frac{s(s - 1) - \mu s}{(s - 1)(s - \mu)}, \quad \mathbb{P}(Z = s - 1) = \frac{\mu}{(s - 1)(s - \mu)},
\tag{6.98}
\]
so that
\[
\mu_Z = \frac{\mu}{s - \mu}, \quad \sigma^2_Z = \frac{\mu s}{s - \mu} \left(1 - \frac{1}{s - \mu}\right) = \mu_Z(s - 1 - \mu_Z).
\tag{6.99}
\]
6.7 Numerical examples of series bounds

Therefore,
\[
\sum_{k=1}^{s-1} \frac{1}{1-z_k} = \frac{1}{2}(s-1) + \frac{1}{2}\mu_Z, \quad (6.100)
\]
\[
\sum_{k=1}^{s-1} \frac{z_k}{(1-z_k)^2} = -\frac{1}{12}s^2 - \frac{1}{2}\mu_Z\left(\frac{1}{3}s - \frac{2}{3}\right) + \frac{1}{4}\mu_Z^2 + \frac{1}{12}, \quad (6.101)
\]
and these quantities are displayed in Figs. 6.10 and 6.11 for \(s = 10, \mu_A \in [0, s-1]\).

The least value, \(-\frac{1}{9}(s-1)^2\), of the \(\sigma^2\)-series occurs for \(\mu_Z = \frac{1}{3}(s-2)\), i.e. for \(\mu = s(s-2)/(s+1) = 2\frac{1}{2}\). The \(\mu\)-series and \(\sigma^2\)-series coincide with their lower and upper bounds, respectively.

**Example 6.7.4** Let \(A\) be uniformly distributed on \(\{0, 1, \ldots, n-1\}\) so that
\[
A(z) = \frac{1}{n} (1 + z + \ldots + z^{n-1}) = \frac{1}{n} \frac{z^n - 1}{z - 1}. \quad (6.102)
\]
We have
\[
\mu_A = \frac{1}{2}(n-1), \quad \sigma_A^2 = \frac{1}{12}(n^2 - 1), \quad (6.103)
\]
and for \(k = 2, 3, \ldots\)
\[
A^{(k)}(1) = \frac{1}{k+1} (n-1)(n-2) \cdots (n-k). \quad (6.104)
\]
Figs. 6.12 and 6.13 display the \(\mu\)-series and \(\sigma^2\)-series for \(s = 10, \mu_A \in [0, s-\frac{1}{2}]\), i.e. \(1 \leq n \leq 2s\). As an aside, we mention that the values of the \(\mu\)-series and \(\sigma^2\)-series at \(n = s, s+1\) are identical, viz. \(\frac{2}{3}(s-1)\) and \(-\frac{1}{18}(s-1)(s+2)\), respectively. Condition (6.19) is satisfied for \(\mu_A \leq 8.83\).

**Example 6.7.5** Take a symmetric binomially distributed \(A\), i.e.,
\[
a_j = \frac{1}{2^{n-1}} \binom{n-1}{j}, \quad j = 0, 1, \ldots, n-1; \quad a_j = 0, \quad j = n, n+1, \ldots, \quad (6.105)
\]
so that
\[
A(z) = \left(\frac{1+z}{2}\right)^{n-1}. \quad (6.106)
\]
We now have
\[
\mu_A = \frac{1}{2}(n-1), \quad \sigma_A^2 = \frac{1}{4}(n-1), \quad (6.107)
\]
and for \(k = 2, 3, \ldots\)
\[
A^{(k)}(1) = \left(\frac{1}{2}\right)^k (n-1)(n-2) \cdots (n-k). \quad (6.108)
\]
Figure 6.8: $\mu$-series, Ex. 6.7.1, $s = 10$.

Figure 6.9: $\sigma^2$-series, Ex. 6.7.1, $s = 10$.

Figure 6.10: $\mu$-series, Ex. 6.7.3, $s = 10$.

Figure 6.11: $\sigma^2$-series, Ex. 6.7.3, $s = 10$.

Figure 6.12: $\mu$-series, Ex. 6.7.4, $s = 10$.

Figure 6.13: $\sigma^2$-series, Ex. 6.7.4, $s = 10$. 
6.7 Numerical examples of series bounds

Figure 6.14: $\mu$-series, Ex. 6.7.5, $s = 10$.

Figure 6.15: $\sigma^2$-series, Ex. 6.7.5, $s = 10$.

Figure 6.16: $\mu$-series, Ex. 6.7.6, $s = 10$.

Figure 6.17: $\sigma^2$-series, Ex. 6.7.6, $s = 10$.

Figs. 6.14 and 6.15 display the $\mu$-series and $\sigma^2$-series for $s = 10$, $\mu_A \in [0, s - \frac{1}{2}]$, i.e. for $1 \leq n \leq 2s$, and we observe a qualitatively similar behavior for the two series as in the Poisson case, see Sec. 6.6. Condition (6.19) is satisfied for $\mu_A \leq 9.81$.

Example 6.7.6 Take $a_0 = 1/2$, $a_{n-1} = 1/2$ where $n \in [1, 2s]$, so that

$$A(z) = \frac{1}{2} + \frac{1}{2} z^{n-1}.$$  \hfill (6.109)

We have

$$\mu_A = \frac{1}{2} (n - 1), \quad \sigma^2_A = \frac{1}{4} (n - 1)^2,$$  \hfill (6.110)

and for $k = 2, 3, \ldots$

$$A^{(k)}(1) = \frac{1}{2} (n - 1)(n - 2) \cdots (n - k).$$  \hfill (6.111)
Figs. 6.16 and 6.17 display the $\mu$-series and $\sigma^2$-series for $s = 10$, $\mu_A \in [0, s - \frac{1}{2}]$, i.e. for $1 \leq n \leq 2s$. Note that the $\mu$-series starts decreasing as a function of $n - 1$ around $n - 1 = s(2 - \sqrt{2})$, which is well before $n - 1 = s$. Condition (6.19) is satisfied for $\mu_A \leq 7.57$.

### 6.8 Numerical examples of moment bounds

We now present bounds on $\mu_X$ and $\sigma^2_X$ for the Poisson case and the examples given in Sec. 6.7, to see how well these bounds perform as approximations. We denote the lower and upper bounds on $\mu_X$ by $\hat{\mu}_X$ and $\check{\mu}_X$, respectively, and similarly for $\sigma^2_X$. These bounds are simply the addition of the bounds on the $\mu$-series and $\sigma^2$-series as described at the beginning of Sec. 6.7. For comparison we also display the known bounds given by Expressions (6.4) and (6.6). The actual values $\mu_X$ and $\sigma^2_X$ are computed numerically by finding the roots $z_1, \ldots, z_{s-1}$, which is feasible in the cases below.

<table>
<thead>
<tr>
<th>$s$ = 5</th>
<th>Poisson</th>
<th>Example 6.7.4</th>
<th>$s$ = 10</th>
<th>Poisson</th>
<th>Example 6.7.5</th>
<th>Example 6.7.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta$</td>
<td>$\text{lb}(6.4)$ $\mu_X$</td>
<td>$\hat{\mu}_X$</td>
<td>$\check{\mu}_X$</td>
<td>$\text{ub}(6.4)$</td>
<td>$\text{lb}(6.4)$ $\mu_X$</td>
<td>$\hat{\mu}_X$</td>
</tr>
<tr>
<td>0.2</td>
<td>0.63</td>
<td>0.97</td>
<td>0.97</td>
<td>1.06</td>
<td>1.13</td>
<td>0.58</td>
</tr>
<tr>
<td>0.4</td>
<td>1.33</td>
<td>1.98</td>
<td>2.27</td>
<td>2.33</td>
<td>1.33</td>
<td>2.00</td>
</tr>
<tr>
<td>0.6</td>
<td>2.25</td>
<td>3.01</td>
<td>3.17</td>
<td>3.68</td>
<td>3.75</td>
<td>2.50</td>
</tr>
<tr>
<td>0.8</td>
<td>4.00</td>
<td>4.74</td>
<td>5.14</td>
<td>5.91</td>
<td>6.00</td>
<td>5.33</td>
</tr>
<tr>
<td>0.9</td>
<td>6.75</td>
<td>7.51</td>
<td>7.98</td>
<td>8.75</td>
<td>9.00</td>
<td>10.50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$s$ = 5</th>
<th>Poisson</th>
<th>Example 6.7.4</th>
<th>$s$ = 10</th>
<th>Poisson</th>
<th>Example 6.7.5</th>
<th>Example 6.7.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta$</td>
<td>$\text{lb}(6.4)$ $\mu_X$</td>
<td>$\hat{\mu}_X$</td>
<td>$\check{\mu}_X$</td>
<td>$\text{ub}(6.4)$</td>
<td>$\text{lb}(6.4)$ $\mu_X$</td>
<td>$\hat{\mu}_X$</td>
</tr>
<tr>
<td>0.2</td>
<td>1.13</td>
<td>1.96</td>
<td>1.96</td>
<td>2.07</td>
<td>2.13</td>
<td>1.13</td>
</tr>
<tr>
<td>0.4</td>
<td>2.33</td>
<td>3.96</td>
<td>3.97</td>
<td>4.28</td>
<td>4.33</td>
<td>2.56</td>
</tr>
<tr>
<td>0.6</td>
<td>3.75</td>
<td>5.98</td>
<td>6.06</td>
<td>6.69</td>
<td>6.75</td>
<td>4.75</td>
</tr>
<tr>
<td>0.8</td>
<td>6.00</td>
<td>8.07</td>
<td>8.85</td>
<td>9.93</td>
<td>10.00</td>
<td>10.00</td>
</tr>
<tr>
<td>0.9</td>
<td>9.00</td>
<td>9.41</td>
<td>12.08</td>
<td>13.41</td>
<td>15.00</td>
<td>15.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$s$ = 10</th>
<th>Poisson</th>
<th>Example 6.7.4</th>
<th>$s$ = 10</th>
<th>Poisson</th>
<th>Example 6.7.5</th>
<th>Example 6.7.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta$</td>
<td>$\text{lb}(6.4)$ $\mu_X$</td>
<td>$\hat{\mu}_X$</td>
<td>$\check{\mu}_X$</td>
<td>$\text{ub}(6.4)$</td>
<td>$\text{lb}(6.4)$ $\mu_X$</td>
<td>$\hat{\mu}_X$</td>
</tr>
<tr>
<td>0.2</td>
<td>1.06</td>
<td>2.00</td>
<td>2.00</td>
<td>2.06</td>
<td>2.06</td>
<td>1.25</td>
</tr>
<tr>
<td>0.4</td>
<td>2.17</td>
<td>4.00</td>
<td>4.00</td>
<td>4.17</td>
<td>4.17</td>
<td>2.33</td>
</tr>
<tr>
<td>0.6</td>
<td>3.38</td>
<td>6.00</td>
<td>6.00</td>
<td>6.38</td>
<td>6.38</td>
<td>3.75</td>
</tr>
<tr>
<td>0.8</td>
<td>5.00</td>
<td>8.00</td>
<td>8.24</td>
<td>9.00</td>
<td>9.00</td>
<td>10.00</td>
</tr>
<tr>
<td>0.9</td>
<td>6.75</td>
<td>9.00</td>
<td>10.22</td>
<td>11.25</td>
<td>13.25</td>
<td>15.00</td>
</tr>
</tbody>
</table>

Table 6.1 displays the bounds on $\mu_X$ for various load values $\theta := \mu_A/s$ and $s = 5, 10$. For $s = 5$, the bounds are sharp, irrespective of the load values $\theta$. For
low load values the lower bound on the $\mu$-series is extremely sharp, as we have seen in Sec. 6.7, leading to sharp bounds on $\mu_X$ as well. For higher load values, the bounds on the $\mu$-series tend to be less sharp. However, comparing with the $\mu$-series, the other terms in (6.1) containing the first two moments of $A$ get more dominant for an increasing load $\theta$, which gradually diminishes the influence of the $\mu$-series. This makes that the bounds on $\mu_X$ are sharp, even asymptotically for $\theta \uparrow 1$. We further note that while the series in Sec. 6.7 have similar order of magnitudes for all four examples, the mean and variance of $X$ show large differences, in particular for high load values. Again, this is due to the dominance of the term $\sigma^2_A/(2(s - \mu_A))$. We observe that in the above examples the bounds derived in this chapter slightly improve the known bounds.

Although for $s = 10$ the values are somewhat larger than those for $s = 5$, the bounds remain sharp. In particular, the lower bound $\hat{\mu}_X$ substantially improves the known lower bound.

Table 6.2 displays the bounds on $\sigma_X$ for various load values and $s = 5$ and $s = 10$. As for the bounds on $\mu_X$, the bounds are sharp, both in the lower and higher load regime. The improvement of the upper bound $\tilde{\sigma}_X^2$ in comparison with the known bound is considerable.
Moment bounds
Part II

Shared capacity models
Chapter 7

Periodic scheduling

For modeling cable networks organized via a request-grant procedure, we propose two models referred to as fixed boundary model and flexible boundary model (both models have been introduced in Subsec. 1.2.1). For these models, time is divided into slots, and each slot can either be a request slot, in which new packets can arrive, or a data slot, in which packets can be transmitted and hence depart from the queue. The decision whether a slot is a request or a data slot is periodically made for a fixed number of slots (referred to as frame). For both models the queue length process embedded at the beginning of frames can be described as a type of discrete bulk service queue.

The fixed boundary model divides each frame into a fixed number of request and data slots. The flexible boundary model also uses this division, but additionally, the unused data slots (due to lack of data packets) are turned into request slots. For both models, we first consider the queue length at frame boundaries. The pgf of the stationary queue length in either model can be determined from a rather standard application of the generating function technique (demonstrated in Sec. 2.2 for the discrete bulk service queue). Due to the periodic scheduling, however, it is far less straightforward to analyze the stationary delay. By adopting a technique developed in Bruneel & Kim [43] and Kang & Steyaert [96], we succeed in deriving the pgf of the stationary delay. The chapter is based on work in Van Leeuwaarden [P1] and Van Leeuwaarden et al. [P3].
7.1 Introduction

In Chapter 1 we elaborated on how the fixed and flexible boundary models arise in the context of data transmission procedures in cable access networks regulated by a request-grant mechanism in which actual data transmission is preceded by a reservation procedure. Both the reservation procedure and the actual data transmission take place on the same upstream channel, and hence each slot can either be used for the reservation procedure or for the data transmission.

The fixed and flexible boundary models both serve as models for the data queue, defined as those data packets for which transmission has already been requested, but that are still waiting to be transmitted. Clearly, if a slot is used for reservation (request slot), new packets can enter this queue, and if a slot is used for data transmission (data slot), a packet can leave this queue.

Because the transmission delay requires that scheduling decisions are taken in advance, one is naturally led to consider frame-based scheduling. The nature of each slot in the frame is periodically determined and broadcast to all the stations. In this chapter we assume that the timing is such that each station is aware of the layout of a frame before it actually starts. The situation where not all stations are aware of the layout of a frame is covered in Chapter 8.

The chapter is structured as follows. In Sec. 7.2, we describe and motivate the models in more detail. In Sec. 7.3, we derive the pgf of the stationary queue length. A closed-form expression for the pgf of the packet delay for both models is derived in Sec. 7.4. A numerical comparison of the two models is given in Sec. 7.5, followed by some conclusions in Sec. 7.6.

7.2 Models

The fixed and flexible boundary models were introduced in Subsec. 1.2.1. We will repeat their exact definitions and refer to Subsec. 1.2.1 for a discussion of the model assumptions in view of the cable network application.

Time is assumed to be slotted, with a given slot duration. In case of the fixed boundary model the schedule of each frame is fixed. That is, a frame defined as \( f \) consecutive slots consists of \( c \) request slots followed by \( s := f - c \) data slots. Let the random variable \( Y_{ti} \) denote the number of arriving packets during the \( i \)th request slot of frame \( t \), and assume that the sequence \( Y_{ti} \) is i.i.d. for all \( t \) and \( i \). We further assume that packets that arrive during frame \( t \) cannot depart from the queue until the beginning of frame \( t + 1 \). This leads to the recursion

\[
X_{t+1} = (X_t - s)^+ + \sum_{i=1}^{c} Y_{ti}, \tag{7.1}
\]

where \( X_t \) denotes the queue length at the beginning of frame \( t \) and \( x^+ := \max(0, x) \).

The fixed boundary mechanism seems inefficient, in the sense that if the queue length is smaller than \( s \), it leaves slots unused which could alternatively be scheduled...
7.3 Stationary queue length

as request slots. This motivates the flexible boundary model in which these unused slots are designated as request slots, yielding the recursion

$$\tilde{X}_{t+1} = (\tilde{X}_t - \tilde{s})^+ + \sum_{i=1}^{c+(s-\tilde{X}_t)^+} Y_{ti},$$  \hspace{1cm} (7.2)$$

where, for notational purposes, we add a tilde to the random variables related to the flexible boundary model. We refer to the $c$ request slots that are scheduled at the beginning of every frame as forced request slots.

7.2.1 Scheduling parameter $c$

The number of forced request slots $c$ in (7.1) and (7.2) can be interpreted as the amount of bandwidth guaranteed for the request-grant procedure: In each frame there are at least $c$ request slots. For the flexible boundary model there are two, unfortunately conflicting, heuristics that guide a judicious choice of $c$. On the one hand, setting $c$ small implements a greedy schedule which empties the data queue as quickly as possible, which suggests that this is the appropriate schedule to minimize the data queue size. On the other hand, setting $c$ large smoothens the arrival process, and intuition suggests that this also helps to reduce the data queue size. In choosing the right value of $c$, one should strike the proper balance between these two considerations. One of the goals of this chapter is to investigate the impact of $c$ through a mathematical analysis of the models. Numerical results are presented in Sec. 7.5.

7.3 Stationary queue length

In this section we derive the pgf of the stationary queue length for both the fixed and flexible boundary model. For each model, we first present the results for the queue length at frame boundaries, from which the results for the queue length throughout a frame follow.

7.3.1 Stationary queue length in fixed boundary model

Let us denote by $Y$ a random variable that has the same distribution as the number of arriving packets during one request slot (i.e. $Y_{ti}$ are i.i.d. copies of a discrete random variable $Y$ for all $t$ and $i$). Let $Y(z)$ be the pgf of $Y$ and denote the mean and variance of $Y$ by $\mu_Y$ and $\sigma_Y^2$, respectively. Clearly, to have stability, it is required that the expected number of arriving packets in a frame is less than the maximum number of packets that can be transmitted in a frame, i.e.

$$c\mu_Y < s.$$  \hspace{1cm} (7.3)$$

We have denoted the queue length at the beginning of frame $t$ by $X_t$. Then $\{X_t, t \in \mathbb{Z}^+\}$ constitutes a discrete-time Markov chain, with transitions governed by (7.1).
As is easily verified, the following conditional expectation holds
\[
E(z^{X_{t+1}} | X_t = k) = \begin{cases} 
  Y(z)^c, & k < s, \\
  z^{k-s}Y(z)^c, & k \geq s.
\end{cases} \tag{7.4}
\]

For reasons of brevity, we introduce the random variable \( A \) denoting the \( c \)-fold convolution of the distribution of \( Y \), that is, the pgf of \( A \) is given by \( A(z) = Y(z)^c \). We denote the mean and variance of \( A \) by \( \mu_A \) and \( \sigma_A^2 \).

Let \( X \) be a random variable distributed as the stationary distribution of the queue length, with
\[
x_k = P(X = k) = \lim_{t \to \infty} P(X_t = k), \quad k = 0, 1, 2, \ldots
\]
From (7.4) it follows that the pgf of \( X \) is given by (see Expression (2.8) for the pgf of the stationary queue length in the discrete bulk service queue)
\[
X(z) = \frac{A(z) \sum_{k=0}^{s-1} x_k (z^s - z^k)}{z^s - A(z)}. \tag{7.5}
\]
In this expression there are still \( s \) unknowns \( x_0, \ldots, x_{s-1} \), which can be found using the classical approach discussed in Sec. 2.2 of this monograph. In Thm. 2.2.1 it has been proven that \( z^s = A(z) \) has \( s \) roots on or within the unit circle. Since a pgf is analytic and well-defined in \( |z| \leq 1 \), the numerator of \( X(z) \) should vanish at each of the roots. This gives \( s \) equations. One of the roots equals 1, and leads to a trivial equation. However, the normalization condition \( X(1) = 1 \) provides an additional equation. Using l'Hôpital's rule, this condition is found to be
\[
s - \mu_A = \sum_{k=0}^{s-1} x_k (s - k), \tag{7.6}
\]
which equates two expressions for the mean number of unused data slots per frame. In case some of the roots have a multiplicity higher than one, still a set of linear equations can be constructed that yields the unique solution \( x_0, x_1, \ldots, x_{s-1} \), see Remark 2.2.3 on p. 24.

Explicit expressions for the moments of the queue length can be obtained by taking derivatives of \( X(z) \). For example, evaluating the first derivative of \( X(z) \) at \( z = 1 \) yields
\[
EX = \frac{\sigma_A^2}{2(s - \mu_A)} + \frac{s + \mu_A}{2} - \sum_{k=0}^{s-1} \frac{x_k (s - k)^2}{2(s - \mu_A)}. \tag{7.7}
\]
So far we looked at the queue length at the beginning of a frame. We can also model the behavior of the queue length throughout a frame. Denote by \( X_{[n]} \), \( n = 1, 2, \ldots, f \), the steady-state queue length at the end of the \( n \)th slot of a frame. The first \( c \) slots of a frame are request slots. This implies that the pgf of \( X_{[n]} \) is given by
\[
X_{[n]}(z) = X(z)Y(z)^n, \quad n = 1, \ldots, c. \tag{7.8}
\]
7.3 Stationary queue length

The remaining \( s \) slots are data slots, yielding \((n = 1, 2, \ldots, s)\)

\[
\mathbb{E}(z^{X_{c+1}} | X = k) = \begin{cases} A(z), & k < n, \\ A(z)z^{k-n}, & k \geq n. \end{cases} \tag{7.9}
\]

Summing over all possible values of \( X \) then gives

\[
X_{c+1}(z) = A(z) \left( \sum_{k=0}^{n-1} x_k + \frac{1}{z^n} (X(z) - \sum_{k=0}^{n-1} x_k z^k) \right), \quad n = 1, \ldots, s. \tag{7.10}
\]

The expectation of the stationary queue length throughout a frame then follows from evaluating the first derivative of (7.8) and (7.10) at \( z = 1 \). That is

\[
\mathbb{E}X[n] = \begin{cases} \mathbb{E}X + n\mu_Y, & n = 1, \ldots, c, \\ \mathbb{E}X + \mu_A - n + c + \sum_{k=0}^{n-c-1} x_k(n-c-k), & n = c+1, \ldots, f. \end{cases} \tag{7.11}
\]

Observe that \( \mathbb{E}X[f] \) equals \( \mathbb{E}X \) due to the normalization condition (7.6).

7.3.2 Stationary queue length in flexible boundary model

For the flexible boundary mechanism, unused data slots are turned into request slots. So, within a frame, the \( c \) forced request slots are scheduled first, then the data slots (if any), and finally the additional request slots (if any). We choose this type of scheduling to simplify the analysis of the packet delay later on. The stability condition (7.3) still applies and is equivalent to requiring \( c \) to be smaller than \( f/(\mu_Y + 1) \).

With \( \tilde{X}_t \) representing the queue length at the beginning of frame \( t \), \( \{\tilde{X}_t, t \in \mathbb{Z}^+\} \) constitutes a discrete-time Markov chain, with transitions governed by (7.2). Note that the following conditional expectation holds

\[
\mathbb{E}(z^{\tilde{X}_{t+1}} | \tilde{X}_t = k) = \begin{cases} Y(z)^{f-k}, & k < s, \\ z^{k-s}A(z), & k \geq s. \end{cases} \tag{7.12}
\]

Because in the flexible boundary model all slots are used, the mean number of request slots per frame, denoted by \( c^\star \), is fixed and independent of \( c \), i.e.

\[
c^\star = \frac{f}{\mu_Y + 1}, \tag{7.13}
\]

as each request slot requires \( 1 + \mu_Y \) slots in total: The request slot itself and \( \mu_Y \) slots for transmitting the packets.

Let \( \tilde{X} \) denote a random variable distributed as the stationary queue length distribution, with

\[
\tilde{x}_k = \mathbb{P}(\tilde{X} = k) = \lim_{t \to \infty} \mathbb{P}(\tilde{X}_t = k), \quad k = 0, 1, 2, \ldots.
\]
From (7.12), it follows that the pgf of $\tilde{X}$ is given by

$$\tilde{X}(z) = \frac{A(z) \sum_{k=0}^{s-1} \tilde{x}_k (z^s Y(z)^{s-k} - z^k)}{z^s - A(z)}. \quad (7.14)$$

As in Subsec. 7.3.1, the $s$ roots of $z^s = A(z)$ on or within the unit circle can be used to determine $\tilde{x}_0, \ldots, \tilde{x}_{s-1}$. Using l'Hôpital’s rule, the normalization condition $\tilde{X}(1) = 1$ reads

$$s - \mu_A = \sum_{k=0}^{s-1} \tilde{x}_k (s - k)(\mu_Y + 1), \quad (7.15)$$

which equates two expressions for the mean number of slots per frame that are used for arrivals and departures of packets that arrived in other than the $c$ forced request slots.

The mean queue length in case of the flexible boundary model is given by

$$\mathbb{E}\tilde{X} = \frac{\sigma^2_A}{2(s - \mu_A)} + \frac{\sigma^2_Y}{2(\mu_Y + 1)} + \frac{s + \mu_A}{2} - (1 - \mu_Y) \sum_{k=0}^{s-1} \tilde{x}_k (s - k)^2 \frac{(1 + \mu_Y)}{2(s - \mu_A)}. \quad (7.16)$$

Using the same notation as for the fixed boundary model, the behavior of the queue length throughout a frame follows from

$$\tilde{X}[n](z) = \tilde{X}(z) Y(z)^n, \quad n = 1, \ldots, c, \quad (7.17)$$

and (for $n = 1, 2, \ldots, s$)

$$\mathbb{E}(z^{\tilde{X}[c+n]} | \tilde{X} = k) = \begin{cases} Y(z)^{c+n-k}, & k < n, \\ A(z) z^{k-n}, & k \geq n, \end{cases} \quad (7.18)$$

and consequently (for $n = 1, 2, \ldots, s$),

$$\tilde{X}[c+n](z) = A(z) \left( \sum_{k=0}^{n-1} \tilde{x}_k Y(z)^{n-k} + \frac{1}{z^n} (\tilde{X}(z) - \sum_{k=0}^{n-1} \tilde{x}_k z^k) \right), \quad n = 1, \ldots, s. \quad (7.19)$$

Hence,

$$\mathbb{E}\tilde{X}[n] = \begin{cases} \mathbb{E}\tilde{X} + n \mu_Y, & n = 1, \ldots, c, \\ (1 + \mu_Y) \sum_{k=0}^{n-c-1} \tilde{x}_k (n - c - k) + c \mu_Y + \mathbb{E}\tilde{X} - n + c, & n = c + 1, \ldots, f. \end{cases} \quad (7.20)$$

Observe that $\mathbb{E}\tilde{X}[f]$ equals $\mathbb{E}\tilde{X}$ due to the normalization condition (7.15).
Example 7.3.1 Consider a frame length of 18 slots, and $Y$ distributed according to a Poisson and geometric distribution

\[ \mathbb{P}(Y = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad \mathbb{P}(Y = k) = (1-p)p^k, \quad k = 0, 1, \ldots, \]

respectively, both with mean 1 ($\lambda = 1$, $p = 1/2$). The mean and variance of $\tilde{X}$ that correspond to these distributions are shown in Figs. 7.1 and 7.2 for increasing $c$. In terms of the mean queue length, having forced request slots at the beginning of the frame is disadvantageous. However, the variance of the queue length is reduced by increasing $c$ (except for high values of $c$).

Remark 7.3.2 In Jacquet et al. [91] a scheduling strategy called implicit framing is studied. For this strategy, no frame structure is used and priority is given to data slots. Periods of consecutive request slots are implicitly closed by the first data packet to be transmitted. When all data packets have been transmitted, a new period of consecutive request slots, in which reservation takes place, is restarted. Note that such an implicit framing strategy yields in fact the flexible boundary model with $f = 1$ and $c = 0$. Jacquet et al. [91] demonstrate that within the framework of the flexible boundary model, implicit framing minimizes the average delay. In the case of implicit framing, the pgf of $\tilde{X}$ reduces to

\[ \tilde{X}(z) = \frac{\tilde{x}_0(zY(z) - 1)}{z - 1} = \frac{1 - zY(z)}{(1 - z)(\mu_Y + 1)}, \quad (7.21) \]

where $\tilde{x}_0$ equals $1/((\mu_Y + 1)$ according to the normalization condition (7.15). Note that $\tilde{X}$ can be interpreted as the residual lifetime of the random variable $Y + 1$. To see this, divide the time axis in cycles of one request slot plus the number of transmission slots $Y$ granted during that request slot. The residual lifetime is an arbitrary point in a cycle, and since in every slot during this cycle exactly one packet
is transmitted, the residual lifetime equals the queue length. We stress, though, that implicit framing is less useful to model the data queue in cable networks, because the transmission delay prevents that a request made by a station in slot \( s \) is granted in slot \( s + 1 \).

### 7.4 Packet delay

In deriving the packet delay distribution, the periodic scheduling causes some difficulties, as shown next. We first present a basic result that holds for both the fixed and flexible boundary model, after which we complete the analysis for both models separately.

Assume that the packets are transmitted in order of arrival. Tag an arbitrary packet, and let the random variable \( T \) denote the slot within the frame in which this packet arrives, \( T \in \{1, 2, \ldots, f\} \). Assume that the packet arrives during slot \( m \), i.e. \( T = m \). Introduce \( U[m] \) as the number of packets present at the end of the frame that contribute to the tagged packet’s delay. Then \( U[m] \) consists of the queue length at the end of the frame that was already present at the end of the previous frame, the packets that arrive in the same frame in request slots before \( T \), plus the packets that arrive within the same request slot but before the tagged packet. We then express \( U[m] \) in terms of two integer random variables \( F[m] \) and \( R[m] \):

\[
U[m] = SF[m] + R[m], \quad F[m] \geq 0, \quad 0 \leq R[m] \leq s - 1,
\]

(7.22)

where \( F[m] \) denotes the number of complete frames included in the tagged packet’s delay, and \( R[m] \) the number of packets that will be transmitted during the same frame as the tagged packet, but before it. Introduce \( D[m] \) as the random variable representing the delay of a packet that arrives during arrival slot \( m \), defined as

\[
D[m] = f - m + F[m] + c + R[m] + 1.
\]

(7.23)

That is, \( f - m \) slots until the beginning of the next frame, \( F[m] \) frames, \( c \) forced request slots, \( R[m] \) slots within the frame of transmission, and the actual transmission slot of the tagged packet. The pgf of \( D[m] \) then reads

\[
D[m](z) = \sum_{i=0}^{\infty} \mathbb{P}(D[m] = i)z^i
\]

\[
= z^{f-m+c+1} \sum_{j=0}^{\infty} \sum_{k=0}^{s-1} \mathbb{P}(F[m] = j, R[m] = k)z^{fj+k}
\]

\[
= z^{f-m+c+1} \sum_{j=0}^{\infty} \sum_{k=0}^{s-1} \mathbb{P}(U[m] = sj + k)z^{fj+k}.
\]

(7.24)

From (7.24) it follows that

\[
D[m](z^s) = z^{s(f-m+c+1)} \sum_{k=0}^{s-1} z^{sk} n_k(z),
\]

(7.25)
7.4 Packet delay

where the functions \( \vartheta_{mk}(z) \) are defined as

\[
\vartheta_{mk}(z) = \sum_{j=0}^{\infty} z^{sfj} P(U_{[m]} = sj + k).
\] (7.26)

The problem now is that (7.26) cannot be formulated directly in terms of the pgf of \( U_{[m]} \). To achieve this, we use a basic approach that can be found in e.g. Bruneel & Kim [43] or Kang & Steyaert [96].

7.4.1 Basic approach

Substituting \( l = sj + k \) in (7.26) yields

\[
\vartheta_{mk}(z) = \sum_{l=0}^{\infty} z^{(l-k)f} P(U_{[m]} = l) \sum_{j=-\infty}^{\infty} \delta(l-sj-k),
\] (7.27)

with \( \delta(n) \) the Kronecker delta function, which equals 1 for \( n = 0 \) and 0 for all other \( n \). Now invoke the following property

**Property 7.4.1** For any two integers \( k \) and \( s \),

\[
\frac{1}{s} \sum_{t=0}^{s-1} a^{tk} = \sum_{j=-\infty}^{\infty} \delta(k - js),
\]

where \( a = \exp(2\pi i/s) \), \( i = \sqrt{-1} \).

**Proof** Follows from rewriting Property 3.2.8 on p. 53. \( \square \)

Using Property 7.4.1 we obtain

\[
\vartheta_{mk}(z) = \sum_{l=0}^{\infty} z^{(l-k)f} P(U_{[m]} = l) \frac{1}{s} \sum_{t=0}^{s-1} a^{tk} \sum_{j=-\infty}^{\infty} \delta(j) z^{sfj} a^{tf} U_{[m]}(a^t z^f).
\] (7.28)

Substituting (7.28) into (7.25) yields

\[
D_{[m]}(z^s) = z^{(s-fm+c+1)} \sum_{k=0}^{s-1} \frac{z^{sk} z^{-kf} s-1}{s} \sum_{t=0}^{s-1} a^{-tk} U_{[m]}(a^t z^f)
\]

\[
= z^{(s-fm+c+1)} \sum_{t=0}^{s-1} U_{[m]}(a^t z^f) \sum_{k=0}^{s-1} (z^{-c} a^{-t})^k
\]

\[
= z^{(s-fm+c+1)} \sum_{t=0}^{s-1} U_{[m]}(a^t z^f) \frac{1 - (z^{-c} a^{-t})^s}{1 - z^{-c} a^{-t}}.
\] (7.29)
Expression (7.29) gives an explicit formula for the pgf of the packet delay once the pgf of $U_{[m]}$ is known. This leaves us to specify the latter, for which we give separate derivations for the fixed and flexible boundary models.

### 7.4.2 Packet delay in fixed boundary model

Let $D$ denote the packet delay for an arbitrary packet. Let $Z_0$ denote the number of packets at the end of a frame that were already present the frame before, and $Z_1$ the number of packets within the tagged packet’s arrival slot arriving before it. The pgf’s of $Z_0$ and $Z_1$ are given by

$$Z_0(z) = \frac{1}{z^s}(X(z) + \sum_{k=0}^{s-1} x_k(z^s - z^k)),$$

(7.30)

and

$$Z_1(z) = \frac{1 - Y(z)}{(1 - z)\mu_Y}.$$

(7.31)

The pgf of $U_{[m]}$ is then simply given by

$$U_{[m]}(z) = Z_0(z)Y(z)^{m-1}Z_1(z), \quad m = 1, \ldots, c. \quad (7.32)$$

Since $P(T = m) = 1/c$ for $m = 1, \ldots, c$, we have that

$$D(z^s) = \frac{1}{c} \sum_{m=1}^{c} D_{[m]}(z^s), \quad (7.33)$$

which, combined with (7.29) and (7.32), yields the following result:

**Theorem 7.4.2** The pgf of the stationary packet delay in the fixed boundary model (7.1) is given by

$$D(z^s) = \frac{1}{sc} \sum_{t=0}^{s-1} \left\{ z^{s(t+1)}Z_0(a^t z^c)Z_1(a^t z^f) \frac{z^{sc} - A(a^t z^f)}{z^s - Y(a^t z^f)} \right\},$$

(7.34)

where $a = \exp(2\pi i/s)$, $i = \sqrt{-1}$, and $Z_0(z)$ and $Z_1(z)$ as given in (7.30) and (7.31).

The mean packet delay follows from

$$\mathbb{E}D = \left[ \frac{1}{s} \frac{d}{dz} D(z^s) \right]_{z=1}, \quad (7.35)$$

which gives after tedious but straightforward calculations

$$\mathbb{E}D = f + \frac{f \sigma_A^2}{2\mu_A(s - \mu_A)} + \frac{1 + \mu_Y}{\mu_Y} \left( \frac{s}{2} - \sum_{k=0}^{s-1} \frac{x_k(s - k)^2}{2(s - \mu_A)} \right).$$

(7.36)
Remark 7.4.3 Alternatively, the mean delay can be derived using Little’s law. The queue length at the beginning of an arbitrary slot is given by

\[ \frac{1}{f} \sum_{n=1}^{f} \mathbb{E}X_{[n]}, \]  

(7.37)

where \( \mathbb{E}X_{[n]} \) is given by (7.11). The average arrival rate of packets per slot equals \( c\mu_Y/f \). Dividing (7.37) by this rate then yields (7.36).

7.4.3 Packet delay in flexible boundary model

For the flexible boundary model, the derivation of \( \hat{U}_{[m]}(z) \) is somewhat more involved, since all slots within a frame are potential request slots. We first consider the case that \( c \geq 1 \), while \( c = 0 \) is covered at the end of this section. Distinguish two events: (a) the tagged packet arrives in one of the forced request slots, and (b) the tagged packet arrives in one of the additional request slots. Event (a) provides no extra information about the queue length at the beginning of a frame, since the \( c \) forced request slots are scheduled every frame. Thus

\[ \hat{U}_{[m]}(z) = \hat{Z}_0(z)Y(z)^{m-1}Z_1(z), \quad m = 1, \ldots, c, \]  

(7.38)

where

\[ \hat{Z}_0(z) = \frac{1}{z^s}(\hat{X}(z) + \sum_{k=0}^{s-1} \hat{x}_k(z^s - z^k)). \]  

(7.39)

Event (b) does provide extra information about the queue length at the beginning of the frame. We know that \( \hat{Z}_0 \) equals zero, otherwise there would be no extra request slots. Further, consider the case that the tagged packet arrives in slot \( c+1 \). This implies that \( \hat{X} \) equals zero. Hence, \( \hat{U}_{[c+1]} = A + Z_1 \). Now consider the packet arriving in slot \( c+2 \). This implies that \( \hat{X} \) equals either zero or one. In the first case it holds that \( \hat{U}_{[c+2]} = A + Y + Z_1 \), and in the latter case \( \hat{U}_{[c+2]} = A + Z_1 \). Similar reasoning leads to the following expression

\[ \hat{U}_{[m]}(z) = A(z)Z_1(z) \sum_{k=0}^{m-1} \hat{x}_kY(z)^{m-1-k} \sum_{k=0}^{m-c-1} \hat{x}_k, \quad m = c+1, \ldots, f. \]  

(7.40)

Finally, the distribution of \( T \) can be determined as follows. Remember that the extra request slots are scheduled at the end of a frame. If a packet arrives in slot \( m \in \{c+1, \ldots, f\} \) of a frame, this particular frame has at least \( f - m + c + 1 \) request slots, and thus at most a queue length of \( m - c - 1 \) packets at the beginning of the frame. This gives (recall that \( c^* \) is the mean number of request slots per frame)

\[ P(T = m) = \begin{cases} \frac{1}{c}, & m = 1, \ldots, c, \\ \frac{1}{c^*} \sum_{k=0}^{m-c-1} \hat{x}_k, & m = c+1, \ldots, f. \end{cases} \]  

(7.41)
Combining (7.29), (7.38) and (7.40), and conditioning on the request slot distribution given by (7.41) yields an explicit expression for the pgf of the packet delay. We have

$$
\tilde{D}(z^s) = \sum_{m=1}^{f} P(T = m) \tilde{D}_{[m]}(z^s),
$$

$$
= \sum_{m=1}^{c} \frac{1}{c^*} \tilde{D}_{[m]}(z^s) + \sum_{m=c+1}^{f} \frac{1}{c^*} \sum_{k=0}^{m-c-1} \tilde{x}_k \tilde{D}_{[m]}(z^s),
$$

(7.42)

where \( \tilde{D}_{[m]}(z^s) \) is defined as \( D_{[m]}(z^s) \) in (7.29), except with \( U_{[m]}(a^t z^f) \) replaced by \( \tilde{U}_{[m]}(a^t z^f) \). This gives the following result:

**Theorem 7.4.4** The pgf of the stationary packet delay in the flexible boundary model (7.2) is given by

$$
\tilde{D}(z^s) = \frac{1}{sc^*} \sum_{i=0}^{s-1} \frac{1 - (a^t z^f)^{-s}}{1 - (a^t z^f)^{-1}} \{ z^{s(f+1)} \tilde{Z}_0(a^t z^f) Z_1(a^t z^f) \frac{s^c - A(a^t z^f)}{z^s - Y(a^t z^f)} + z^{s(c+1)} A(a^t z^f) Z_1(a^t z^f) \sum_{k=0}^{s-1} \tilde{x}_k [z^{s(s-k)} - Y(a^t z^f)^s - k] \}.
$$

(7.43)

where \( a = \exp(2\pi i/s), \, i = \sqrt{-1}, \) and \( \tilde{Z}_0(z) \) and \( Z_1(z) \) are given in (7.39) and (7.31).

From (7.43), it follows that

$$
\mathbb{E} \tilde{D} = \frac{\mu_Y + 1}{\mu_Y} \left\{ \mathbb{E} \tilde{X} + \frac{(s + 1)\mu_A - s^2}{2f} + \sum_{k=0}^{s-1} \tilde{x}_k [z^{s(s-k)} - Y(a^t z^f)^s - k] \right\}.
$$

(7.44)

As in case of the fixed boundary model (see Remark 7.4.3), an alternative derivation of \( \mathbb{E} \tilde{D} \) follows from applying Little’s law.

In case \( c = 0 \), the basic approach as described in Sec. 7.4.1 is not needed. It is then straightforward to derive that

$$
\tilde{U}_{[m]}(z) = Z_1(z) \frac{\sum_{k=0}^{m-1} \tilde{x}_k Y(z)^{m-1-k}}{\sum_{k=0}^{m-1} \tilde{x}_k}; \quad P(\tilde{T} = m) = \frac{1}{c^*} \sum_{k=0}^{m-1} \tilde{x}_k,
$$

and that the pgf of the packet delay is given by

$$
\tilde{D}_0(z) = \frac{Z_1(z)}{f c^*} \sum_{m=1}^{f} z^{f-m+1} \sum_{k=0}^{m-1} \tilde{x}_k Y(z)^{m-1-k}.
$$

(7.45)

**Remark 7.4.5** We have derived the pgf’s of the packet delay for the fixed and flexible boundary models in (7.34) and (7.43), respectively. To find the underlying
7.5 Numerical results

packet delay distribution we use a numerical technique of Abate & Whitt [17]. A distribution \( \{p_k\} \) can be recovered from its pgf \( P(z) \) via the inversion

\[
p_k = \frac{1}{2\pi i} \oint_{C_r} \frac{P(z)}{z^{k+1}} dz,
\]

(7.46)

where \( C_r \) is a circle about the origin of radius \( r, \ 0 < r < 1 \). Abate & Whitt [17] approximate (7.46) using the trapezoidal rule with a step size of \( \pi/k \) as

\[
\hat{p}_k = \frac{1}{2kr^k} \sum_{j=1}^{2k} (-1)^j \text{Re}(P(re^{ij\pi/k})),
\]

(7.47)

and derive for \( 0 < r < 1, \ k \geq 1 \) the following error bound

\[
|p_k - \hat{p}_k| \leq \frac{r^{2k}}{1 - r^{2k}}.
\]

(7.48)

For practical purposes one can think of the error bound as \( r^{2k} \), because \( r^{2k}/(1 - r^{2k}) \approx r^{2k} \) for \( r^{2k} \) small. To have accuracy up to the \( \gamma \)th decimal, we let \( r = 10^{-\gamma/2k} \). In the numerical examples below, we set \( \gamma \) equal to 7.

Example 7.4.6 The distribution of the packet delay has a typical form. For \( f = 9 \), \( Y \) geometrically distributed with mean 1, Fig. 7.3 displays the packet delay distribution for \( c = 0, 2, 4 \), where we have used the method described in Remark 7.4.5. First note that the minimum delay corresponds to a packet that arrives in the last slot of a frame and is immediately transmitted in the slot \( c+1 \) of the next frame. The oscillating effect is due to the frame structure, and becomes stronger for higher values of \( c \).

7.5 Numerical results

In this section we first present a numerical comparison between the fixed and flexible boundary models. Next, for the flexible boundary model, we investigate the impact of different values of \( c \) on various queue length and delay characteristics.

7.5.1 Fixed versus flexible boundary model

We assume that the load, defined as the mean number of packets arriving per frame, is the same for the fixed and flexible boundary models, being \( c\mu_Y \) and \( c^*\mu_Y = f\mu_Y/(1 + \mu_Y) \), respectively. Thus, for a fair comparison, we choose the appropriate values of \( \mu_Y \) for which the load is the same for both models. For convenience, we further assume that \( Y \) is Poisson distributed.

Figs. 7.4 and 7.5 display the mean and variance of the packet delay for \( f = 9, \ c = 2 \) and various load values. For a load of 2, \( \mu_Y = 1 \) for the fixed boundary
model and $\mu_Y = 2/7$ ($c^* = 7$, $c^* \mu_Y = 2$) for the flexible boundary model. In terms of the mean packet delay, the flexible boundary model clearly outperforms its fixed counterpart.

For the flexible boundary model, a low load yields relatively many unused data slots that are used as additional request slots. The variation in request slots per frame inherent to such type of scheduling then causes a higher packet delay variance than the fixed boundary model.

### 7.5.2 Influence of $c$ in flexible boundary model

We now investigate the impact of different values of $c$ for the flexible boundary model on various queue length and delay characteristics. Table 7.1 contains queue length characteristics for $f = 9$, $c = 0, 2, 4$, and $Y$ is Poisson or geometrically distributed with mean 1. Table 7.2 contains delay characteristics for the same settings.

<table>
<thead>
<tr>
<th></th>
<th>$EX$</th>
<th>$\text{Var}X$</th>
<th>$P(X &gt; 10)$</th>
<th>$P(X &gt; 20)$</th>
<th>$P(X &gt; 50)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c = 0$</td>
<td>4.75</td>
<td>11.75</td>
<td>.0639</td>
<td>.0003</td>
<td>.0000</td>
</tr>
<tr>
<td>$c = 2$</td>
<td>4.95</td>
<td>7.97</td>
<td>.0408</td>
<td>.0001</td>
<td>.0000</td>
</tr>
<tr>
<td>$c = 4$</td>
<td>6.75</td>
<td>10.93</td>
<td>.1245</td>
<td>.0019</td>
<td>.0002</td>
</tr>
<tr>
<td>geometric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c = 0$</td>
<td>5.00</td>
<td>16.67</td>
<td>.1042</td>
<td>.0026</td>
<td>.0001</td>
</tr>
<tr>
<td>$c = 2$</td>
<td>5.40</td>
<td>14.07</td>
<td>.0995</td>
<td>.0020</td>
<td>.0000</td>
</tr>
<tr>
<td>$c = 4$</td>
<td>9.00</td>
<td>34.63</td>
<td>.3197</td>
<td>.0471</td>
<td>.0064</td>
</tr>
</tbody>
</table>

As we have seen in Example 7.3.1, increasing $c$ is disadvantageous in terms of the
7.5 Numerical results

![Graph 1](image1.png)  ![Graph 2](image2.png)

**Figure 7.4:** Mean packet delay, fixed vs. flexible boundary model, \( f = 9, c = 2, Y \) Poisson distributed.

**Figure 7.5:** Packet delay variance, fixed vs. flexible boundary model, \( f = 9, c = 2, Y \) Poisson distributed.

**Table 7.2:** Characteristics of the packet delay for \( f = 9 \) and \( \mu_Y = 1 \).

<table>
<thead>
<tr>
<th></th>
<th>( E \tilde{D} )</th>
<th>( Var \tilde{D} )</th>
<th>( P(\tilde{D} &gt; 10) )</th>
<th>( P(\tilde{D} &gt; 20) )</th>
<th>( P(\tilde{D} &gt; 30) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( c = 0 )</td>
<td>6.92</td>
<td>7.60</td>
<td>.0926</td>
<td>.0020</td>
<td>.0000</td>
</tr>
<tr>
<td>( c = 2 )</td>
<td>8.57</td>
<td>8.82</td>
<td>.5437</td>
<td>.0039</td>
<td>.0001</td>
</tr>
<tr>
<td>( c = 4 )</td>
<td>13.66</td>
<td>32.03</td>
<td>.9550</td>
<td>.2800</td>
<td>.0327</td>
</tr>
<tr>
<td>geometric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( c = 0 )</td>
<td>7.63</td>
<td>11.84</td>
<td>.1767</td>
<td>.0028</td>
<td>.0003</td>
</tr>
<tr>
<td>( c = 2 )</td>
<td>11.46</td>
<td>17.10</td>
<td>.6075</td>
<td>.0353</td>
<td>.0014</td>
</tr>
<tr>
<td>( c = 4 )</td>
<td>21.40</td>
<td>96.86</td>
<td>.9568</td>
<td>.4855</td>
<td>.1812</td>
</tr>
</tbody>
</table>

mean queue length, while the variance of the queue length is often reduced due to the stabilizing effect on the arrival process. The same can be seen from the results in Table 7.1. Increasing \( c \) reduces the flexibility of the system, which is the reason that both the mean and variance of the queue length increase when \( c \) gets large, \( c = 4 \) in this example.

The mean and variance provide only partial information on the underlying distribution function. We therefore consider some tail probabilities. Note that in Table 7.1, the probability that \( \tilde{X} \) gets larger than 10 is the smallest for \( c = 2 \), for both the Poisson and geometric distribution. Depending on the relevant performance characteristic, one can determine the optimal value of \( c \).

For the delay characteristics in Table 7.2 we do not see a stabilizing effect. This is mainly due to our definition of delay that includes the delay of the request slot until the beginning of the next frame. In this way small values of \( c \) are favored, because of the many additional request slots that bring along a smaller delay until the beginning of the next frame.
7.6 Conclusions

We have presented time-slotted queueing models that describe the frame-based scheduling of request and data slots. The fixed boundary model uses a fixed number of request slots per frame, whereas the flexible boundary model designates in addition the unused (due to lack of data packets) data slots as request slots. For both models we presented the pgf of the stationary queue length and the packet delay. For deriving the pgf of the packet delay we used a technique specifically designed to deal with the periodic scheduling. In Van Leeuwaarden [P14] this technique has been applied to derive the delay distribution of vehicles in the fixed-cycle traffic light (FCTL) queue.

For the flexible boundary model we use an optimistic scenario, in the sense that the packets that arrive in the additional request slots can all be transmitted at the beginning of the next frame. If this is not the case, the delayed flexible model, which is the topic of Chapter 8, might be more appropriate.

The impact of the forced request slots on the performance characteristics has been briefly touched upon in Subsec. 7.5.2. This issue will be pursued in much greater depth in Chapter 8.
In Chapter 7 we presented the fixed and flexible boundary models for cable networks organized via a request-grant mechanism. Both models are discrete-time (slotted) queueing models and incorporate periodic scheduling and forced request slots (see Subsec. 1.1.4). In Subsec. 1.2.2 we argued that in case of a large transmission delay, we need a more advanced model introduced as the delayed flexible boundary model. This model extends the flexible boundary model with the feature that there is a fixed minimum delay (possibly larger than the frame length) between the instant that a request for sending data packets gets granted and the instant that a packet can be actually transmitted.

The delayed flexible boundary model defines a higher-dimensional Markov chain for which deriving the stationary queue length distribution is harder than for the fixed and flexible boundary models. We therefore combine a method of Kingman [99], bounding techniques, and a heuristic argument to derive approximations for the mean queue length. These approximations suggest several interesting properties of the mean queue length as a function of the transmission delay and the number of forced request slots. The properties are verified and substantiated by simulation.

Based on these properties, we develop an adaptive scheduling strategy which uses detailed information in order to deal with the transmission delay, and designates for every frame the number of request slots. It is shown that the adaptive scheduling strategy, in comparison with the delayed flexible boundary model, leads to significant reductions in both the mean and the variance of the stationary queue length. This chapter is for a large part based on Denteneer & Van Leeuwaarden [P13].
8.1 Model description and overview

We first show how we arrive at the delayed flexible boundary model starting from the classical discrete bulk service queue that has been the topic of Chapters 2-6. The discrete bulk service queue is defined by the recursion

\[ X_{t+1} = (X_t - s)^+ + A_t. \]  

(8.1)

Here, \( x^+ := \max\{0, x\} \), time is divided into frames, \( X_t \) denotes the queue length at the beginning of frame \( t \), \( A_t \) denotes the number of newly arriving packets during frame \( t \), and \( s \) denotes the maximum number of packets that can be transmitted in one frame. Packets that arrive at the queue in frame \( t \) can be transmitted at the earliest in frame \( t + 1 \).

For usage in cable networks we propose two modifications to the basic bulk service queue. Firstly, following the flexible boundary model in Chapter 7, we couple the arrivals \( A_t \) to the queue length so that the arrival intensity is high if the queue size is small and the arrival intensity is low if the queue size is large. Secondly, we include a delay parameter in the model so that there is a fixed minimum delay between the instant of arrival and the instant that a packet can be transmitted. These two modifications lead to the delayed flexible boundary model defined by the recursion

\[ X_{t+1} = (X_t - s)^+ + c + (s - X_{t-d})^+ \sum_{i=1}^{c + (s - X_{t-d})^+} Y_{t-d,i}. \]  

(8.2)

Here, \( X_t \) denotes the size of the data queue at the beginning of frame \( t \), \( Y_{t,i} \) denotes the random variable distributed as the number of arriving packets during the \( i \)th request slot of frame \( t \), \( d \) represents the transmission delay such that a request made in frame \( t \) can be scheduled at the earliest in frame \( t + d + 1 \), \( c \) denotes the number of forced request slots per frame. For the remaining \( s = f - c \) slots in a frame (frame length is \( f \) slots), we assume that data transmission takes precedence over reservation. Then, the delayed flexible boundary model (8.2) serves as a model for the data queue. The quantity \( c + (s - X_{t-d})^+ \) can be interpreted as the number of slots used for handling request messages in frame \( t-d \). The actual data transmissions for these requests cannot be scheduled before frame \( t + 1 \) so that the data packets associated with these requests join the data queue at the beginning of slot \( t + 1 \). The sum in (8.2) thus represents the total number of new data packets that can be transmitted. As for the fixed and flexible boundary models, the \( Y_{t,i} \) are assumed to be i.i.d. copies of some integer-valued random variable \( Y \). The delayed flexible boundary model incorporates many of the key characteristics of cable networks regulated by a request-grant mechanism discussed in Subsec. 1.1.4.

8.1.1 Scheduling parameter \( c \) and transmission delay

For the fixed and flexible boundary models, we have argued in Subsec. 7.2.1 that there is no obvious choice of \( c \). For the delayed flexible boundary model, this choice
8.2 Mean queue length

gets even more difficult. Sala et al. [141] investigated the strategy that gives priority to the data queue \( c = 0 \) by simulating a cable access network with transmission delay, in which data transfer was organized by a reservation mechanism. They observe that this type of scheduling results in a very bursty arrival process, and a cyclic queue behavior. They compared this priority strategy with \( c = 0 \) to strategies that reduce the cycle length by forcing capacity to handle requests \( c > 0 \). These strategies, which guarantee some of the capacity to the request queue, lead to a smoother process and to shorter delays.

The relevance of the choice of \( c \) has also been observed in other simulations of cable access networks, see e.g. Hekstra-Nowacka et al. [89] and Golmie et al. [80]. One of the goals of this chapter is to better understand the appropriate choice of \( c \) in relation with the transmission delay \( d \) through a mathematical analysis of the delayed flexible boundary model.

8.1.2 Approach

The recursion (8.2) with \( d = 0 \) has been thoroughly studied in Chapter 7. In this case, the recursion defines a one-dimensional Markov chain, and the pgf of the stationary distribution of the queue length can be obtained. However, for \( d > 0 \), the recursion defines a \((d + 1)\)-dimensional Markov chain, and the approach from Chapter 7 does not carry over. We therefore give an analysis of (8.2) partly based on heuristic arguments. In particular, we derive approximating bounds for the mean stationary queue length. We exploit a method by Kingman [99] to express the mean queue length in terms of moments of the arrival distribution, a term related to the idle time, and a correlation term. We then use a technique from Chapter 6 to bound the term related to the idle time. Finally, we invoke a heuristic derivation to approximate the correlation term. The bounds and the approximation together yield approximations for the mean stationary queue length.

These approximations suggest some properties of the mean queue length. Bearing these properties in mind, we develop an adaptive scheduling strategy that is specifically designed to deal with the transmission delay.

In Sec. 8.2 we derive the exact expression for the mean queue length containing the idle time and correlation terms. We also give the bounds for the idle-time term. The approximation for the correlation term is presented in Sec. 8.3. The properties suggested by the approximation are stated in Sec. 8.4, and the adaptive scheduling algorithm is presented in Sec. 8.5. The properties are verified by simulation in Sec. 8.6. Also in Sec. 8.6, a simulation-based comparison is made between the performance of the delayed bulk service queue and adaptive scheduling.

8.2 Mean queue length

Throughout this chapter we denote the mean and variance of the random variable \( Y \) by \( \mu_Y \) and \( \sigma_Y^2 \). The following result holds:
Lemma 8.2.1. Denote by $Z_t$ the $(d + 1)$-dimensional Markov chain
\begin{equation}
Z_t = (X_{t-d}, X_{t-d+1}, \ldots, X_t),
\end{equation}
with $X_t$ as in (8.2). If $\{Z_t\}$ is irreducible, and $c\mu Y < f - c$, then there exists a unique stationary distribution for $\{Z_t\}$.

**Proof** See Denteneer & Van Leeuwaarden [P13] and Denteneer [60], p. 143. □

To get some insight in the mean queue length for general $d \geq 0$, we apply a method used by Kingman [99]. This method is based on the manipulation of
\begin{equation}
M_t = (s - X_t)^+, \quad P_t = (X_t - s)^+.
\end{equation}
For these variables, the following obvious relations hold:
\begin{equation}
X_t - s = P_t - M_t, \quad (X_t - s)^2 = P_t^2 + M_t^2.
\end{equation}

We will use $X^d$ to denote a random variable distributed according to the stationary distribution of the queue length process as defined by (8.2), and $M^d$ to denote a random variable that follows the same distribution as $(s - X^d)^+$. We then have that

**Theorem 8.2.2** The mean queue length in the delayed flexible boundary model with delay parameter $d$ is given by
\begin{equation}
\mathbb{E}(X^d) = \frac{c\sigma^2 Y}{2(s - c\mu Y)} + \frac{\sigma^2 Y^2}{2(1 + \mu Y)} + \frac{s + c\mu Y}{2} + \mathbb{E}((M^d)^2) \frac{\mu Y^2 - 1}{2(s - c\mu Y)}
+ \mathbb{E}(R^d) \frac{\mu Y}{s - c\mu Y},
\end{equation}
where
\begin{equation}
\mathbb{E}(R^d) = \lim_{t \to \infty} \mathbb{E}(P_t M_{t-d}).
\end{equation}

**Proof** See Sec. 8.7. The proof can also be found in Denteneer & Van Leeuwaarden [P13] and Denteneer [60], p. 144. □

Expression (8.6) for $\mathbb{E}(X^d)$ contains two unknown terms: A term $\mathbb{E}((M^d)^2)$ related to the idle time and a correlation term $\mathbb{E}(R^d)$. Now
\begin{equation}
\mathbb{E}((M^d)^2) = \sum_{j=0}^{s-1} \mathbb{P}(X^d = j)(s - j)^2
\end{equation}
can be satisfactorily bounded in the following way. Since
\begin{equation}
\left( \sum_{j=0}^{s-1} \mathbb{P}(X^d = j)(s - j) \right)^2 \leq \sum_{j=0}^{s-1} \mathbb{P}(X^d = j)(s - j)^2 \leq s \sum_{j=0}^{s-1} \mathbb{P}(X^d = j)(s - j),
\end{equation}
and
\[
\sum_{j=0}^{s-1} P(X^d = j)(s - j) = \frac{s - c\mu_Y}{1 + \mu_Y},
\]
we have
\[
\left(\frac{s - c\mu_Y}{1 + \mu_Y}\right)^2 \leq \mathbb{E}((M^d)^2) \leq s \frac{s - c\mu_Y}{1 + \mu_Y}.
\]
This is one of the techniques used in Chapter 6. Sharper bounds can be obtained along the lines of the theory presented in that chapter.

In case \( d = 0 \), we obviously have that \( R^d = 0 \). Then, combining (8.6) with (8.11) yields bounds for the mean queue length. The bounds are sharp for the heavy-traffic case in which \( c\mu_Y \to s \). For \( d \geq 1 \), no simple bounds on the correlation term can be derived. Instead, we derive a fluid approximation for \( \mathbb{E}(R^d) \) in the next section.

**Remark 8.2.3** For \( d = 0 \), (8.6) reduces to the mean queue length in the flexible boundary model given by (7.16). For the flexible boundary model, we were able to derive the pgf of the stationary queue length (7.14), and from differentiating this pgf we obtained (7.16). For the delayed flexible boundary model we are not able to derive the pgf of the stationary queue length. In this case, the technique applied in the proof of Thm. 8.2.2, leading to the mean stationary queue length is extremely valuable. This technique has also been applied for the fixed-cycle traffic light queue by Miller [119] and in the analysis of cell-based switches and routers in Leonardi et al. [112]. Also, the technique applied in the proof of Thm. 8.2.2 can be used to derive the mean stationary queue length in the fixed boundary model (7.7).

### 8.3 An approximation for the correlation term

In this section we use a heuristic argument to construct an approximation for \( \mathbb{E}(R^d) \), which, together with the bounds (8.11), yields approximations for \( \mathbb{E}(X^d) \). The argument is based on the inspection of the sample paths of various realizations of the process defined by (8.2).

One such sample path is shown in Fig. 8.1, where \( d = 100 \), \( c = 0 \), and \( Y \) geometrically distributed with \( \mu_Y = 1.25 \). The figure shows the queue length evolution after a long initial warm-up period. The sample path has settled on a cyclic pattern. Each cycle can be subdivided into three distinct parts. First, there is an interval, of length \( d + 1 \), in which the queue length equals 0. In the second interval, also of length \( d + 1 \), the queue length increases. Finally, in the third interval (the length of this interval is specified below), the data queue size is drained until it hits zero. Thereafter, a new cycle starts.

We conjecture that this is the typical behavior of the sample paths in case \( \mu_Y > 1 \) and \( d > 0 \), irrespective of the actual distribution of \( Y \). This conjecture suggests that we can construct a deterministic approximation of the sample path. Our heuristic approximation of \( \mathbb{E}(R^d) \) is then obtained by evaluating \( \mathbb{E}(R^d) \) for this deterministic approximation. More formally, we define the deterministic process \( x_t \) via (8.2) with
Periodic scheduling with transmission delay

Figure 8.1: Sample path of the process defined by (8.2), for \( f = 18, d = 100, c = 0 \), and \( Y \) geometrically distributed with \( \mu_Y = 1.25 \). The sample path comes from a simulation that started with an empty queue, and the results are displayed for frame 38,000 until frame 40,000.

Figure 8.2: Deterministic approximation \( x_t \) of the sample path of a realization of (8.2) for \( \mu_Y > 1 \).

\( Y_{t,i} \) replaced by its expected value (see Fig. 8.2):

\[
x_{t+1} = (x_t - f + c)^+ + \sum_{i=1}^{c+(f-c-x_{t-d})^+} \mu_Y
\]  

(8.12)

Given initial values \( x_1 = \ldots = x_{d+1} = c\mu_Y \), it is easy to see that (8.12) yields for \( j = 1, \ldots, d+1 \)

\[
x_{d+1+j} = j(f - c\mu_Y)\mu_Y - (j-1)(f-c),
\]

because in this period those packets join the data queue that were generated in the \( f - c\mu_Y \) request slots \( d+1 \) frames earlier, while packets are transmitted from the queue at maximum rate \( f - c \) packets per frame. At the end of this period, the queue has built up to the level \( (d+1)(f - c\mu_Y)\mu_Y - d(f-c) \), after which the queue
8.4 Properties

is drained at rate \((f - c) - c\mu_Y\). This yields
\[ x_{2(d+1)+j} = (d + 1)(f - c\mu_Y)\mu_Y - (d + j)(f - c) + j\mu_Y, \]
for \(j = 1, \ldots, L^*.\) Here \(L^*\) is the smallest value \(l\) for which \(x_{2(d+1)+l}\) hits \(c\mu_Y\). Consequently, \(L^*\) can be calculated from \(x_{2(d+1)+L^*} = c\mu_Y\), i.e.
\[ L^* = \frac{(d + 1)(f - c \mu_Y)\mu_Y - d(f - c)}{f - c - c\mu_Y}. \] (8.13)
After instant \(2(d + 1) + L^* - 1\) the sequence repeats itself. Hence the cycle length equals \(L = 2(d + 1) + L^* - 1 = (d + 1)(\mu_Y + 1)\). We therefore approximate \(E(R^d)\) as follows
\[ E(R^d) \approx \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} (x_t - s)^-(x_{t+d} - s)^+ \]
\[ \approx \frac{1}{L} \sum_{t=1}^{L} (x_t - f + c)^-(x_{t+d} - f + c)^+, \] (8.14)
where \(x^- := \min\{0, x\}\). Now for \(\mu_Y > 1\), we can approximate the second sum in (8.14) by the terms \(t = 2, \ldots, d + 1\), so that
\[ E(R^d) \approx \frac{1}{L} \sum_{j=1}^{d} (c\mu_Y - f + c)^-(j(f - c\mu_Y)\mu_Y - (j - 1)(f - c) - f + c)^+ \]
\[ = \frac{1}{L} \frac{d(d + 1)}{2} (f - c - c\mu_Y)((f - c\mu_Y)\mu_Y - f + c)^+ \]
\[ = \frac{1}{2\mu_Y + 1} (f - c - c\mu_Y)((f - c\mu_Y)\mu_Y - f + c)^+. \] (8.15)
Substituting (8.15) into (8.6) yields the following approximation for \(E(X^d)\):
\[ E(X^d) \approx \frac{\sigma_Y^2}{2(s - c\mu_Y)} + \frac{\sigma_Y^2}{2(1 + \mu_Y)} + \frac{s + c\mu_Y}{2} + E((M^d)^2) \frac{\mu_Y^2 - 1}{2(s - c\mu_Y)} \]
\[ + \frac{1}{2\mu_Y + 1} ((f - c\mu_Y)\mu_Y - f + c)^+. \] (8.16)
The bounds in (8.11) for \(E((M^d)^2)\) can again be used to obtain explicit expressions. In Denteneer [60] it is shown that the approximation (8.16) of the mean queue length is in general sharp, but breaks down in heavy-traffic conditions for \(c > 0\). The latter is because the deterministic approximation of the sample path is less suitable for \(c > 0\) and heavy-traffic conditions.

8.4 Properties

The approximation (8.16) suggests various interesting properties for \(E(X^d)\). Firstly, we consider \(E(X^d)\) as a function of \(d\). The approximation then suggests that \(d\) has
no impact on the mean queue length in case $\mu_Y \leq 1$. However, in case $\mu_Y > 1$, $\mathbb{E}(X^d)$ increases linearly with $d$. It follows in particular that the correlation term $\mathbb{E}(R^d)$ is the dominating term in the expression for $\mathbb{E}(X^d)$ and that $\mathbb{E}(X^d)$ grows without bounds for $d$ tending to infinity.

Secondly, we consider $\mathbb{E}(X^d)$ as function of $c$ (see Subsec. 8.1.1). In order to set $c$ such that the mean queue length is minimized, there are two considerations: The smaller $c$, the quicker the data queue is emptied, while the larger $c$, the more the arrival process is smoothened. The approximation (8.16) can be used to strike the proper balance between these two considerations (see also the discussion in Subsec. 8.1.1).

Thirdly, the approximation (8.16) suggests that the mean queue length is not necessarily monotonic in the traffic intensity for $d > 0$ and $c > 0$. To see this, observe that the approximation (8.15) of the correlation term $\mathbb{E}(R^d)$ is not monotonic in $\mu_Y$.

This non-monotonicity can be explained informally as follows. Observe that the input to the data queue consists of two sources: $(X_t - s)^+$ and a sum which increases in $(s - X_t - d)^+$. As the traffic intensity approaches the stability bound, the cyclic behavior of the sample paths vanishes which decorrelates the two input sources to the data queue. Hence, increasing the traffic intensity causes the input to be less bursty, and this results in a smaller mean queue length. Another way to see this is by observing that the bursts following periods in which the system is, relatively, empty are caused by an inflow of magnitude $(f - c - c\mu_Y)\mu_Y + c\mu_Y = (f - c\mu_Y)\mu_Y$. Now this latter expression is non-monotonic in $\mu_Y$.

Finally, we consider the heavy-traffic limit. The correlation term will vanish in the heavy-traffic limit where $\mu_Y$ approaches $(f - c)/c$. Thus, Thm. 8.2.2 implies that there exists a heavy-traffic limit in case $c > 0$. In fact, for $c > 0$, the heavy-traffic limit for the delayed flexible boundary model equals the heavy traffic limit for the ordinary bulk service queue and will be dominated by the term $c\sigma_Y^2/(2(f - c - c\mu_Y))$.

In case $c = 0$ there is no stability bound, and the expected inflow following empty periods equals $f\mu_Y$ which always increases in $\mu_Y$.

### 8.5 Adaptive scheduling

We have seen in Secs. 8.3 and 8.4 that the transmission delay results in a cyclic behavior and a strongly correlated arrival process. This might have severe consequences for the mean queue length (see Expression (8.6)), since the correlation term $\mathbb{E}(R^d)$ becomes dominant in high-load situations. We aim at smoothing the arrival process and reducing the correlation of the arrival process. We will do this by introducing a scheduling strategy that does not only allow to vary the number of request slots per frame (as for the delayed flexible boundary model), but also allows for the
number of request slots in a frame to depend on the queue length at the beginning
of the frame and the number of request slots scheduled in the previous \(d\) frames. We have referred to this strategy as adaptive scheduling.

Denote by \(c_t\) the number of request slots scheduled in frame \(t\). The evolution
equation of the queue length at frame boundaries then becomes

\[
X_{t+1} = (X_t - (f - c_t))^+ + \sum_{i=1}^{c_{t-d}} Y_{t-d,i}. \tag{8.17}
\]

Let us now recall where the cyclic behavior observed in Sec. 8.3 comes from. The
arrival process is coupled to the queue length such that more packets arrive when
the queue is small, and less packets arrive when the queue is long. This type of
control is expected to lead to a smoother distribution of the number of arriving
packets over time. However, the transmission delay upsets the balance. The impact
of a corrective decision, like more arrivals if the system is less busy, is only seen \(d\)
frames later. If the system is busier \(d\) frames later, the extra arrivals might have just
the opposite effect. This phenomenon of control decisions that have the opposite
effect as one had in mind is precisely what is captured by the correlation term:

\[
\mathbb{E}(R^d) = \lim_{t \to \infty} \mathbb{E}(P_t M_{t-d})
\]

\[
= \lim_{t \to \infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \mathbb{P}(M_{t-d} = j) \mathbb{P}(P_t = k|M_{t-d} = j)jk
\]

\[
= \lim_{t \to \infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \mathbb{P}(X_t = s - j) \mathbb{P}(X_t = s + k|X_{t-d} = s - j)(s-j)(s+k).
\]

So, \(\mathbb{E}(R^d)\) might be viewed as a measure for the performance of a scheduling strategy: A high value of \(\mathbb{E}(R^d)\) indicates that the scheduling strategy balances the input poorly, and ideally \(\mathbb{E}(R^d)\) equals zero. Obviously, it holds that the larger the transmission delay \(d\), the less unlikely it is that the relatively simple scheduling adopted by the flexible boundary model balances the input well.

Our primary goal is to reduce the mean queue length by choosing an adaptive
scheduling strategy that balances the input properly despite a substantial delay. Denote by \(c^*\) the mean number of request slots per frame, given by

\[
c^* = \frac{f}{1 + \mu_Y}. \tag{8.18}
\]

The mean number of arriving data packets per frame, denoted by \(\Lambda\), is then given by

\[
\Lambda = c^* \mu_Y = \frac{f \mu_Y}{1 + \mu_Y}. \tag{8.19}
\]

In balancing the input, one would want \(\Lambda\) packets to arrive to the queue per frame. This is not feasible, since we are dealing with a stochastic process, but it might serve
as a guiding principle. Say we are at the beginning of frame $t$. What do we know about the number of arriving packets in the next $d$ frames? We have scheduled $c_{t-d} + c_{t-d+1} + \cdots + c_{t-1}$ request slots in the previous $d$ frames and we are still free to choose $c_t$, which makes that the number of arriving packets in the next $d$ frames is given by

$$
\sum_{k=1}^{d} \sum_{i=1}^{c_{t-k}} Y_{t-k,i} + \sum_{i=1}^{c_t} Y_{t,i}.
$$

(8.20)

Ideally, there will be $f - c_t$ packets at the beginning of each frame, so that in each frame all waiting packets can be transmitted. In that case, we would have

$$
X_t = f - c_t; \quad \sum_{k=1}^{d} \sum_{i=1}^{c_{t-k}} Y_{t-k,i} + \sum_{i=1}^{c_t} Y_{t,i} = (d + 1)\Lambda.
$$

(8.21)

In reality, this will not be the case, but we will take these values as a benchmark. So, we aim at choosing $c_t$ such that

$$
X_t - (f - c_t) + \sum_{k=1}^{d} \sum_{i=1}^{c_{t-k}} Y_{t-k,i} + \sum_{i=1}^{c_t} Y_{t,i} \approx (d + 1)\Lambda.
$$

(8.22)

This benchmark provides a useful scheduling strategy when we replace the $Y_{t,i}$ in (8.22) by their expectation $\mu_Y$. Some rewriting then gives the value $\bar{c}_t$ as a target level for $c_t$, i.e.

$$
\bar{c}_t = \frac{1}{1 + \mu_Y} \left[ f + (d + 1)\Lambda - X_t - \mu_Y \sum_{k=1}^{d} c_{t-k} \right]
$$

$$
= c^* + \frac{1}{1 + \mu_Y} \left[ (d + 1)\Lambda - X_t - \mu_Y \sum_{k=1}^{d} c_{t-k} \right].
$$

(8.23)

To make sure that $c_t$ is integer-valued, and that all unused data slots are turned into request slots, we then choose $c_t$ according to

$$
c_t = \max\{0, \lfloor \bar{c}_t \rfloor, f - X_t \}.
$$

(8.24)

### 8.6 Numerical evaluation

In order to assess the merit of various scheduling strategies, we have carried out a number of simulations. We consider two different distributions for $Y$: The geometric and the Poisson distribution. The frame length has been set to $f = 18$, and we varied the traffic intensity $\Lambda$, the transmission delay $d$, and the number of forced request slots $c$. In all simulation results presented below, the performance measures have been evaluated on an interval of 1,000,000 frames, after an initial warm-up period of 200,000 frames.
Let us first look at the properties of the delayed flexible boundary model that were stated in Sec. 8.4. We do that by an example. We consider Poisson arrivals and a transmission delay of $d = 3$. Figs. 8.3-8.6 display results for regular scheduling (as in the delayed flexible boundary model) with $c = 0, 1, 2, 3, 4$ and for adaptive scheduling, where $\Lambda$ ranges from 9 to its maximum sustainable value.

Fig. 8.3 displays the mean stationary queue length. The curves obtained from regular scheduling all have an asymptote at $\Lambda = f - c$. For most values of $\Lambda$, $c = 0$ results in the largest mean queue length, while the adaptive scheduling results in the smallest mean queue length. For regular scheduling the non-monotonic behavior is clearly visible for $c = 1$. This behavior, though remarkable, is not uncommon in systems that involve control and feedback delay. Situations in which these characteristics lead to unwanted oscillations and increased delay occur if the traffic dynamics can be expressed via a difference equation or differential equation that involves a
delayed response, see e.g. Johari & Tan [93], Gyori & Ladas [83] and Fendick & Rodrigues [71].

Fig. 8.4 displays the variance of the stationary queue length. What strikes is that the figure is quite similar to Fig. 8.3, including the good performance of adaptive scheduling and the non-monotonic behavior.

Fig. 8.5 displays the correlation term. As \( \Lambda \) approaches its maximum sustainable value, the correlation term decreases for regular scheduling with \( c \geq 1 \). The adaptive scheduling succeeds in keeping the correlation term small, except for high values of \( \Lambda \). As mentioned before, a small correlation term indicates that the scheduling strategy balances the input well. We investigate this further in Subsec. 8.6.2.

Fig. 8.6 displays the idle-time term. Clearly, the term is bounded and decreases for increasing values of \( c \). It is confirmed that the idle-time term vanishes when \( \Lambda \) approaches its maximum sustainable value.

### 8.6.1 Delayed flexible boundary model

The parameter \( c \) has been interpreted as a scheduling parameter. Moreover, in Sec. 8.4, it was observed that the minimization of the mean queue length with respect to \( c \) yields a non-trivial function of the traffic intensity. We now investigate this phenomenon and illustrate the effect of changing \( c \). We have plotted in Fig. 8.7 the mean queue length as a function of traffic intensity for geometric arrivals, various values of \( c \) and \( d = 3 \).

**Figure 8.7**: \( \mathbb{E}X^d \) for geometric arrivals and \( d = 3 \).

Fig. 8.7 shows that we can choose \( c \) so as to improve system performance. For low traffic intensities, set \( c = 4 \) to minimize the mean queue length. For medium traffic intensities, i.e. from \( \Lambda \approx 12.8 \) to \( \Lambda \approx 14.5 \), set \( c = 3 \). From \( \Lambda \approx 14.5 \) to \( \Lambda \approx 15.8 \), set \( c = 2 \). For high loads, not shown in the figure, one should set \( c = 0 \).
8.6.2 Adaptive scheduling and delay

We now further investigate the performance of adaptive scheduling in relation with the transmission delay and the arrival process. In Figs. 8.8-8.13 we have plotted $\mathbb{E}X^d$ and $\mathbb{E}R^d$ for static and adaptive scheduling, for Poisson arrivals and $d = 1, 10, 25$.

First note that in all cases the adaptive scheduling performs well, in the sense that it minimizes the mean queue length for almost all values of $\Lambda$. An exception is for $\Lambda$ ranging from 16.7 to 17 for $d = 1$. In this case, regular scheduling with $c = 1$ gives a smaller mean queue length.

For increasing values of $d$, the relative performance of the adaptive scheduling becomes better. The reason for this can be seen from the figures that display the correlation term. The higher the transmission delay becomes, the more (relatively) the correlation term is lowered by adaptive scheduling. For $d = 25$, the correlation term for adaptive scheduling is almost negligible. This can be explained as follows. The adaptive scheduling determines the appropriate number of request slots by estimating the number of packets that will arrive in the future. Denote the total number of request slots scheduled in the $d$ previous frames by $m$. The estimated number of future arrivals is then $m\mu_Y$. Hence, the larger $d$, the larger $m$, and the more precise the estimation of the number of future arrivals will be. A similar argument can be used for describing the influence of the arrival distribution. The more volatile the distribution, the less accurate the estimation of the future arrivals will be. In Figs. 8.14 and 8.15 we have plotted the results for geometric arrivals and $d = 1$. Compare these results with those for Poisson arrivals in Figs. 8.8 and 8.9. Indeed, since the geometric distribution is more volatile than the Poisson distribution, the adaptive scheduling results for the geometric distribution in higher values of $\mathbb{E}R^d$. As a consequence, the reduction in mean queue length due to adaptive scheduling is much less than in case of Poisson arrivals.
Figure 8.10: $E X^d$ for Poisson arrivals and $d = 10$.

Figure 8.11: $E R^d$ for Poisson arrivals and $d = 10$.

Figure 8.12: $E X^d$ for Poisson arrivals and $d = 25$.

Figure 8.13: $E R^d$ for Poisson arrivals and $d = 25$.

Figure 8.14: $E X^d$ for geometric arrivals and $d = 1$.

Figure 8.15: $E R^d$ for geometric arrivals and $d = 1$.
8.7 Proof of Theorem 8.2.2

Define

\[ S_{t-d} = \sum_{i=1}^{c+M_{t-d}} Y_{t-d,i}. \]  

(8.25)

It holds that \( X_t - s = P_t - M_t = X_{t+1} - S_{t-d} - M_t \). Take expectations, limits for \( t \to \infty \), and rearrange to obtain

\[ \mathbb{E}(M^d) = (s - c \mu_Y)/(1 + \mu_Y). \]  

(8.26)

Next, we use that \( P_t = X_{t+1} - S_{t-d} \) together with (8.5) to obtain

\[
(X_t - s)^2 = P_t^2 + M_t^2
= (X_{t+1} - S_{t-d})^2 + M_t^2
= X_{t+1}^2 - 2X_{t+1}S_{t-d} + S_{t-d}^2 + M_t^2
= X_{t+1}^2 - 2(P_t + S_{t-d})S_{t-d} + S_{t-d}^2 + M_t^2
= X_{t+1}^2 - 2P_tS_{t-d} - S_{t-d}^2 + M_t^2. 
\]  

(8.27)

It follows in particular that

\[ 2sX_t = s^2 + X_t^2 - X_{t+1}^2 + 2P_tS_{t-d} + S_{t-d}^2 - M_t^2. \]  

(8.28)

Furthermore, it holds that

\[ \mathbb{E}(S_{t-d}^2) = (c + \mathbb{E}(M_{t-d})) \sigma_Y^2 + (c^2 + 2c\mathbb{E}(M_{t-d}) + \mathbb{E}(M_{t-d}^2)) \mu_Y^2, \]  

(8.29)

and

\[
\mathbb{E}(P_tS_{t-d}) = c\mathbb{E}(P_t)\mu_Y + \mathbb{E}(P_tM_{t-d})\mu_Y
= c\mathbb{E}(X_t - s + M_t)\mu_Y + \mathbb{E}(P_tM_{t-d})\mu_Y. 
\]  

(8.30)

Take expectations in (8.28), substitute (8.29) and (8.30), take limits for \( t \to \infty \), and rearrange to obtain

\[
\mathbb{E}(X^d)^2(s - c \mu_Y) = (s - c \mu_Y)^2 + 2c\mathbb{E}(M^d)(\mu_Y + \mu_Y^2) + (c + \mathbb{E}(M^d)) \sigma_Y^2
+ \mathbb{E}((M^d)^2)\mu_Y^2 - 1 + 2\mathbb{E}(R^d)\mu_Y, \]  

(8.31)

with \( \mathbb{E}(R^d) \) given by (8.7). Finally, substituting (8.26) into (8.31) yields (8.6). \( \square \)
Chapter 9

Tandem queue with coupled processors

In this chapter we investigate the two-stage tandem queue with coupled processors, which has been suggested as a model for cable access networks regulated by a request-grant mechanism in Sec. 1.3. It is assumed that jobs arrive at the first station according to a Poisson process and require service at both stations before leaving the system. The amounts of work that a job requires at each of the stations are independent, exponentially distributed random variables. When both stations are nonempty, the total service capacity is shared between the stations according to fixed proportions. When one of the stations becomes empty, the total service capacity is given to the nonempty station.

We study the two-dimensional Markov process that represents the numbers of jobs at the two stations. The problem of finding the generating function of the stationary distribution can be reduced to two different Riemann-Hilbert boundary value problems. Although both problems yield a complete analytical solution, they have different features from the numerical viewpoint. We discuss the similarities and differences between the two problems, and relate them to the computational aspects of obtaining performance measures. The chapter is based on Van Leeuwaarden & Resing [P7].
9.1 Introduction

One application of the tandem queue with coupled processors is a cable access network regulated by a request-grant mechanism. Another application of the model would be an assembly line for which two operations on each job must be performed using a limited service capacity. By coupling the service at each of the operations, and thus using the service capacity of an operation for which no jobs are waiting for the other operation, imbalance in the assembly line can be reduced and the throughput can be increased (see e.g. Andradottir et al. [24]).

Resing & Örmeci [139] have shown for the two-dimensional Markov process representing the numbers of jobs at the two stations, that the problem of finding the bivariate generating function of the stationary distribution can be reduced to a Riemann-Hilbert boundary value problem. In [139] the issue of how to obtain performance measures has not been discussed. In general, obtaining performance measures from the formal solution of a Riemann-Hilbert boundary value problem is not straightforward. In this chapter we discuss the numerical issues that arise when computing performance measures. In particular, we consider the fraction of time a station is empty and the mean stationary queue length at a station. The reduction of the problem of finding the generating function to a boundary value problem usually follows from considering a specific zero-set of the kernel of the functional equation. This can be done in more than one way. We discuss, next to the zero-set considered in [139], one other zero-set that leads to a second Riemann-Hilbert boundary value problem. From the analytical viewpoint, the second formulation has little added value, since solving either one of the two problems gives a full solution to the model. However, in determining performance measures numerically, the two problems have different features.

We describe the model and the key functional equation for the model in Sec. 9.2. In Sec. 9.3 we analyze the kernel of the functional equation. The results are used in Secs. 9.4 and 9.5 to reduce the problem of solving the functional equation to two different Riemann-Hilbert boundary value problems. Specific attention is paid to determining the conformal map that is required for the solution of the second Riemann-Hilbert boundary value problem. In Sec. 9.6 we derive some performance measures for the model. In Sec. 9.7 we discuss issues that arise when numerically determining the performance measures from the formal solutions of the Riemann-Hilbert boundary value problems. Among other things, we show that we can determine the performance measures for the whole set of allowed parameter values. Finally, we give some numerical results in Sec. 9.8.

9.2 Model description and functional equation

Consider a two-stage tandem queue, where jobs arrive at queue 1 according to a Poisson process with rate $\lambda$, each job demanding service at both queues before leaving the system. Each job requires an exponential amount of work with parameter
\( \nu_j \) at station \( j, j = 1, 2 \). The total service capacity of the two stations together is fixed. Without loss of generality we assume that this total service capacity equals one unit of work per time unit. Whenever both stations are nonempty, a proportion \( p \) of the capacity is allocated to station 1, and the remaining part \( 1 - p \) is allocated to station 2. Thus, when there is at least one job at each station, the departure rate of jobs at station 1 is \( \nu_1 p \) and the departure rate of jobs at station 2 is \( \nu_2 (1 - p) \).

Here we assume that \( 0 < p < 1 \), so that there is a real capacity sharing between the two stations. For the cases \( p = 0 \) and \( p = 1 \), the system can be seen as a tandem queue with a single server moving between the two queues and giving priority to one of the queues. The solutions for these cases are given in Resing & O’rmeć [139].

When one of the stations becomes empty, the total service capacity is allocated to the nonempty station. Hence, the departure rate at that station, station \( j \) say, is temporarily increased to \( \nu_j \). With \( X_j(t) \) the number of jobs at station \( j \) at time \( t \), the two-dimensional process \( \{(X_1(t), X_2(t)), t \geq 0\} \) is a Markov process. The condition under which this Markov process has a unique stationary distribution is given by

\[
\frac{\lambda}{\nu_1} + \frac{\lambda}{\nu_2} < 1. \tag{9.1}
\]

This can be explained by the fact that, independent of \( p \), the two stations together always work at capacity 1 (if there is work in the system), and that \( \lambda/\nu_1 + \lambda/\nu_2 \) equals the amount of work brought into the system per time unit. We henceforth assume that the ergodicity condition is satisfied.

Let us denote by \( \pi(n, k) \) the stationary probability of having \( n \) customers at station 1 and \( k \) customers at station 2, i.e. \( \pi(n, k) = \lim_{t \to \infty} \mathbb{P}(X_1(t) = n, X_2(t) = k) \). The following set of balance equations can then be derived:

\[
\begin{align*}
\lambda \pi(0, 0) &= \nu_2 \pi(0, 1), \\
(\lambda + \nu_2) \pi(0, 1) &= \nu_1 \pi(1, 0) + \nu_2 \pi(0, 2), \\
(\lambda + \nu_2) \pi(0, k) &= p \nu_1 \pi(1, k - 1) + \nu_2 \pi(0, k + 1), \quad k \geq 2,
\end{align*}
\]

and for \( n \geq 1 \)

\[
\begin{align*}
(\lambda + \nu_1) \pi(n, 0) &= \lambda \pi(n - 1, 0) + (1 - p) \nu_2 \pi(n, 1), \\
\lambda + \nu_1 + (1 - p) \nu_2 \pi(n, 1) &= \lambda \pi(n - 1, 1) + \nu_1 \pi(n + 1, 0) + (1 - p) \nu_2 \pi(n, 2), \\
\lambda + \nu_1 + (1 - p) \nu_2 \pi(n, k) &= \lambda \pi(n - 1, k) + p \nu_1 \pi(n + 1, k - 1) \\
& \quad + (1 - p) \nu_2 \pi(n, k + 1), \quad k \geq 2.
\end{align*}
\]

We define the joint probability generating function

\[
P(x, y) := \sum_{n \geq 0} \sum_{k \geq 0} \pi(n, k) x^n y^k, \quad |x| \leq 1, \quad |y| \leq 1,
\]

which is, for every fixed \( y \), regular for \( |x| < 1 \) and continuous for \( |x| \leq 1 \). A similar statement holds for \( x \) and \( y \) interchanged. From the balance equations it follows
that \( P(x, y) \) satisfies the functional equation

\[
    h_1(x, y)P(x, y) = h_2(x, y)P(x, 0) + h_3(x, y)P(0, y) + h_4(x, y)P(0, 0),
\]  

(9.2)

where

\[
\begin{align*}
    h_1(x, y) &= \left( \lambda + p \nu_1 + (1 - p) \nu_2 \right) xy - \lambda x^2 y - p \nu_1 y^2 - (1 - p) \nu_2 x, \\
    h_2(x, y) &= (1 - p) \left[ \nu_1 y(y - x) + \nu_2 x(y - 1) \right], \\
    h_3(x, y) &= p \left[ \nu_2 x(1 - y) + \nu_1 y(x - y) \right], \\
    h_4(x, y) &= p \nu_2 x(y - 1) + (1 - p) \nu_1 y(x - y).
\end{align*}
\]

The constant \( P(0, 0) \) can be determined by substituting \( x = \gamma(y) := \nu_1 y^2 / (\nu_1 y - \nu_2 y + \nu_2) \) into (9.2). For this choice of \( x \), both \( h_2(x, y) \) and \( h_3(x, y) \) equal zero, and hence (9.2) reduces to

\[
    P(\gamma(y), y) = \frac{h_4(\gamma(y), y)}{h_1(\gamma(y), y)} P(0, 0).
\]

(9.3)

Letting \( y \uparrow 1 \) in (9.3), we obtain \( P(0, 0) = 1 - \lambda / \nu_1 - \lambda / \nu_2 \). This result can again be explained by the fact that, independent of \( p \), the two stations together always work at capacity 1 (if there is work in the system), and that \( \lambda / \nu_1 + \lambda / \nu_2 \) equals the amount of work brought into the system per time unit.

### 9.3 Analysis of the kernel

In the analysis of the functional equation (9.2) a crucial role is played by the kernel \( h_1(x, y) \). Due to the regularity properties of \( P(x, y) \), for each pair \((x, y)\) on or within the unit circle for which \( h_1(x, y) \) equals zero, the right-hand side of (9.2) must vanish. This provides us with a relation between the unknown functions \( P(0, y) \) and \( P(x, 0) \). From the observation that \( h_1(x, y) \) is a polynomial in either \( x \) or \( y \), we can construct two Riemann-Hilbert boundary value problems, one for the function \( P(x, 0) \) and one for the function \( P(0, y) \).

Blanc [33] has investigated the transient behavior of the ordinary tandem queue without coupled processors, for which the kernel \( h_1(x, y) \) is of the exact same form. Since Blanc has studied \( h_1(x, y) \) as a polynomial in \( y \), most of the results presented in this section stem from his work. Using these results, the problem of finding the stationary queue length distribution can be reduced to a Riemann-Hilbert boundary value problem for \( P(0, y) \), as presented in Sec. 9.4. In Sec. 9.5 we derive a Riemann-Hilbert boundary value problem for \( P(x, 0) \).

We introduce

\[
    r_1 = \frac{\lambda}{\nu_1}, \quad r_2 = \frac{\lambda}{(1 - p)\nu_2},
\]

as the loads on each of the stations if they would work in isolation (no coupling). For notational convenience, we also introduce

\[
\hat{r} = 1 + \frac{1}{r_1} + \frac{1}{r_2},
\]
such that
\[ h_1(x, y) = \lambda \left[ \hat{r}xy - x^2y - \frac{1}{r_1}y^2 - \frac{1}{r_2}x \right]. \quad (9.4) \]
Observe that \( h_1(x, y) \) is, for each \( x \), a polynomial of degree 2 in \( y \). We thus have that for every value of \( x \) there are two possible values of \( y \), say \( y_1(x) \) and \( y_2(x) \), such that \( h_1(x, y_1(x)) = h_1(x, y_2(x)) = 0 \). These can be described by the two-valued function
\[ y(x) = \frac{r_1}{2} \left[ s_1(x) \pm \sqrt{D_1(x)} \right], \quad (9.5) \]
where
\[ s_1(x) = (\hat{r} - x)x, \quad D_1(x) = s_1(x)^2 - \frac{4x}{r_1r_2}. \]
We then obtain the following result:

**Lemma 9.3.1** The algebraic function \( y(x) \), defined by \( h_1(x, y(x)) = 0 \), has four real branch points \( 0 = x_1 < x_2 \leq 1 < x_3 < x_4 \).

**Proof** The branch points of \( y(x) \) are zeros of the discriminant \( D_1(x) \). Clearly, \( D_1(0) = 0 \), \( \lim_{x \to 0} D_1(x) < 0 \), \( D_1(1) \geq 0 \), \( D_1(\hat{r}) < 0 \) and \( \lim_{x \to \infty} D_1(x) = \infty \). Furthermore, if \( D_1(1) = 0 \) (i.e. \( r_1 = r_2 < 1 \)) then \( D_1'(1) > 0 \). \( \square \)

For later use, we present the following lemma which shows that the mapping \( y(x) \) for \( x \in [0, x_2] \) gives rise to a smooth and closed contour \( L \) (see Fig. 9.1).

**Lemma 9.3.2** For each \( x \in [0, x_2] \), \( y(x) \) lies on the closed contour \( L \), which is symmetric with respect to the real line, and defined by
\[ |y|^2 = \frac{r_1}{2r_2} (\hat{r} - \sqrt{\hat{r}^2 - 8\Re(y)/r_1}). \quad (9.6) \]
It further holds that
\[ |y|^2 \leq \frac{r_1}{r_2} x_2. \]  \hspace{1cm} (9.7)

**Proof** For \( x \in [0, x_2] \), \( D_1(x) \) is negative, so \( y_1(x) \) and \( y_2(x) \) are complex conjugates. It also follows that
\[ \text{Re}(y) = \frac{r_1}{2} (\hat{r} - x). \]  \hspace{1cm} (9.8)

Furthermore, from \( h_1(x, y(x)) = 0 \) we have \( |y|^2 = r_1 x / r_2 \leq r_1 x_2 / r_2 \). Since (9.8) is a quadratic equation in \( x \), substituting one of the two solutions into \( |y|^2 = r_1 x / r_2 \) yields (9.6). Of course, we choose the solution of \( x \) for which \( y(0) = 0 \) and \( y(x_2) = \sqrt{r_1 x_2 / r_2} \) lie on the contour. \( \square \)

We will henceforth denote the interior of \( L \) by \( L^+ \), and set
\[ \alpha := y(x_2) = \sqrt{r_1 x_2 / r_2}, \]  \hspace{1cm} (9.9)
representing the point on \( L \) with the largest modulus. With respect to \( \alpha \), the following assertions hold.

**Lemma 9.3.3** If \( r_1 = r_2 \), then \( \alpha = 1 \). If \( r_1 < r_2 \), then \( \alpha < 1 \). If \( r_1 > r_2 \), then \( \alpha > 1 \).

**Proof** For \( r_1 = r_2 \), we have that \( D_1(1) = 0 \), so \( x_2 = \alpha = 1 \). For \( r_1 < r_2 \), knowing \( x_2 < 1 \), it follows that \( \alpha < 1 \). For \( r_1 > r_2 \), knowing \( x_2 < 1 \), we have that \( D_1(r_2 / r_1) < 0 \) since \( r_2 + r_2 (1 - r_2) / r_1 < 1 \), and thus \( r_2 / r_1 < x_2 \) and \( \alpha > 1 \). \( \square \)

We note that \( \alpha = 1 \) (respectively \( \alpha < 1 \), \( \alpha > 1 \)) implies \( 1 \in L \) (respectively \( 1 \notin L \cup L^+, \ 1 \in L^+ \)), which plays a crucial role in the numerical work to be presented in Sec. 9.7.

Paralleling the approach above, the kernel \( h_1(x, y) \) is, for each \( y \), a polynomial of degree 2 in \( x \). Thus for each \( y \) there are two possible values of \( x \), say \( x_1(y) \) and \( x_2(y) \), such that \( h_1(x_1(y), y) = h_1(x_2(y), y) = 0 \). These can be described by the two-valued function
\[ x(y) = \frac{1}{2y} \left[ s_2(y) \pm \sqrt{D_2(y)} \right], \]  \hspace{1cm} (9.10)
where
\[ s_2(y) = \hat{r} y - \frac{1}{r_2}, \quad D_2(y) = s_2(y)^2 - \frac{4y^3}{r_1}. \]

The following then holds:

**Lemma 9.3.4** The algebraic function \( x(y) \) defined by \( h_1(x(y), y) = 0 \) has three real branch points \( 0 < y_1 < y_2 \leq y_3 \).

**Proof** The branch points of \( x(y) \) are zeros of the discriminant \( D_2(y) \). Clearly, \( D_2(0) = 1 / r_2^2 > 0 \), \( D_2(1) = (1 - 1 / r_1)^2 \geq 0 \) and \( \lim_{y \to -\infty} D_2(y) = -\infty \). For \( \hat{y} = 1 / (r_2 \hat{r}) \in (0, 1) \), it holds that \( D_2(\hat{y}) = -4\hat{y}^3 / r_1 < 0 \). Also, if \( D_2(1) = 0 \)
9.3 Analysis of the kernel

Figure 9.2: The mapping \( x = x(y) : [y_1, y_2] \to R \).

(which implies \( r_1 = 1 \) and, due to the ergodicity condition, \( r_2 < 1 \) then \( D'_2(1) = 4(1/r_2 - 1) > 0 \). □

We now study the mapping \( x(y) \) for \( y \in [y_1, y_2] \) in some more detail. This mapping can be shown to give rise to a smooth and closed contour \( R \), as specified in the next lemma and illustrated in Fig. 9.2.

**Lemma 9.3.5** For each \( y \in [y_1, y_2] \), \( x(y) \) lies on the closed and smooth contour \( R \), which is symmetric with respect to the real line, and defined by:

\[
|x|^2 = \frac{1}{r_1 r_2(\hat{r} - 2\text{Re}(x))},
\]

\[
|x|^2 \leq \frac{y_2}{r_1}.
\]

**Proof** Similar to the proof of Lemma 9.3.2. □

We set

\[ \beta := x(y_2) = \sqrt{y_2/r_1}, \]

the point on \( R \) with the largest modulus, for which it holds that

**Lemma 9.3.6** (i) When either \( r_1 = 1 \) or \( r_2 = 1 \) we have that \( \beta = 1 \). (ii) When both \( r_1 < 1 \) and \( r_2 < 1 \) we have that \( \beta > 1 \). (iii) When either \( r_1 > 1 \) or \( r_2 > 1 \) we have that \( \beta < 1 \).

**Proof** (i) If \( r_1 = 1 \), then \( y_2 = 1 \) and thus \( \beta = 1 \). If \( r_2 = 1 \), then \( y_2 = r_1 \) and thus \( \beta = 1 \).

(ii) For \( \beta > 1 \) we should prove that \( r_1 < y_2 \). Consider the function \( f(r_1) := D_2(r_1) = -4r_1^2 + (1 + r_1 + r_1/r_2 - 1/r_2)^2 \). The solutions to \( f(r_1) = 0 \) are given by
$r_1 = 1$ and $r_1 = \hat{r}_1 = (1 - r_2)/(1 + 3r_2)$. For $r_1 = \hat{r}_1$ it holds that $r_1 = y_1 < y_2$. Assume that there exists a value $r_1 \in (0, 1)$ for which it holds that $r_1 > y_2$. Then, since $y_2$ is a continuous function of $r_1$, there should be a value in $(0, 1)$ other than $\hat{r}_1$ for which $r_1 = y_2$ and hence $f(r_1) = 0$. This is not the case, and thus $r_1 < y_2$ for all values $r_1 \in (0, 1)$.

(iii) If $r_1 > 1$, then obviously $r_1 > y_2$ and thus $\beta < 1$. Now assume $r_2 > 1$. Then, for $\beta < 1$ we should prove that $r_1 > y_2$. Note that $f(r_1)$ is positive for all values $r_1 \in (0, 1)$. This implies that $r_1 < y_1$ or $r_1 > y_2$ (see the proof of Lemma 9.3.4). Furthermore, $y_1 < \hat{y} < 1/2$ for $r_2 > 1$ and $r_1 \in (0, 1)$, when $\hat{y}$ as defined in the proof of Lemma 9.3.4. Hence, for $r_1 \geq 1/2$ it clearly holds that $r_1 > y_1$. Assume that there exists a value $r_1 \in (0, 1)$ for which it holds that $r_1 < y_1$. Then, since $y_1$ is a continuous function of $r_1$, there should be a value in $(0, 1)$ for which $r_1 = y_1$ and hence $f(r_1) = 0$. This is not the case, and thus $r_1 > y_2$ for all values $r_1 \in (0, 1)$. □

We again note that $\beta = 1$ (respectively $\beta < 1$, $\beta > 1$) implies $1 \in R$ (respectively $1 \not\in R \cup R^+$, $1 \in R^+$), which plays a crucial role in the numerical work to be presented in Sec. 9.7.

### 9.4 Boundary value problem I

In the previous section we considered the kernel as a polynomial in either $y$ or $x$, which may lead to the curves $L$ and $R$, respectively. In this section we describe how the curve $L$ leads to a Riemann-Hilbert boundary value problem for the function $P(0, y)$.

**Lemma 9.4.1** The function $P(0, y)$ is regular in the domain $L^+$ and satisfies for $y \in L$ the condition

$$\text{Im}[P(0, y)] = \text{Im} \left[ -P(0,0) \frac{h_4(r_2 |y|^2/r_1, y)}{h_3(r_2 |y|^2/r_1, y)} \right].$$

(9.14)

**Proof** For zero-pairs $(x, y)$ of the kernel $h_1(x, y)$ for which $P(x, y)$ is finite, we have

$$h_2(x, y)P(x, 0) + h_3(x, y)P(0, y) + h_4(x, y)P(0, 0) = 0,$$

(9.15)

from which it follows that, for those zero-pairs,

$$P(0, y) = \frac{1}{p} P(x, 0) - \frac{h_4(x, y)}{h_3(x, y)} P(0, 0).$$

(9.16)

Thus, (9.14) follows from the fact that $P(x, 0)$ is real for $x \in [0, x_2]$ and $|y|^2 = r_1 x/r_2$ for $y \in L$. If $\alpha \leq 1$, $L$ lies entirely within the unit circle. Hence, $P(0, y)$ is regular in $L^+$. If $\alpha > 1$, $P(0, y(x))$ can be continued analytically over the interval $[0, x_2]$ via (9.15), because $P(x, 0)$ is regular on this interval. Hence, the analytic continuation
of $P(0, y)$ is finite at $y = y(x_2)$. Because $P(0, y)$ has a power series expansion at $y = 0$ with positive coefficients, this implies that $P(0, y)$ is regular for $|y| < y(x_2)$ and hence in $L^+$. □

Lemma 9.4.1 shows that the determination of $P(0, y)$ reduces to the determination of the solution of the following Riemann-Hilbert boundary value problem on the contour $L$: Determine a function $P(0, y)$ such that

1. $P(0, y)$ is regular for $y \in L^+$ and continuous for $y \in L \cup L^+$.
2. $\text{Re} \{iP(0, y)\} = c(y)$, for $y \in L$,

where

$$c(y) = \text{Im} \left[ P(0, 0) \frac{h_4(r_2|y|^2/r_1, y)}{h_3(r_2|y|^2/r_1, y)} \right].$$

The standard way to solve this type of boundary value problem is to transform the boundary condition (9.14) to a condition on the unit circle (see e.g. Muskhelishvili [121], p. 108). Denote the unit circle by $C$ and its interior by $C^+$. We introduce the conformal mapping

$$z = f(y) : L^+ \to C^+, \quad (9.17)$$

and its inverse

$$y = f_0(z) : C^+ \to L^+. \quad (9.18)$$

Using these mappings, we can reduce the Riemann-Hilbert problem on $L$ to the following problem: Determine a function $G(z)$ such that

1. $G(z)$ is regular for $z \in C^+$ and continuous for $z \in C \cup C^+$.
2. $\text{Re} \{iG(z)\} = \tilde{c}(z)$, for $z \in C$, where $\tilde{c}(z) = c(f_0(z))$,

which is known as the Dirichlet problem on a circle. Its solution is given by (see Muskhelishvili [121], p. 108)

$$G(z) = -\frac{1}{2\pi} \oint_C \tilde{c}(w) \frac{w + z}{w - z} \frac{dw}{w} + K_1, \quad z \in C \cup C^+, \quad (9.19)$$

where $K_1$ is some constant. In this way, $P(0, y) = G(f(y))$ has been formally determined as

$$P(0, y) = -\frac{1}{2\pi} \oint_C c(f_0(w)) \frac{w + f(y)}{w - f(y)} \frac{dw}{w} + K_1, \quad y \in L \cup L^+. \quad (9.20)$$

We can rewrite the contour integral (9.20) as a real integral on $[0, x_2]$. That is, for $y \in L^+ \cup L$, we have that

$$P(0, y) = \frac{1}{2\pi} \int_0^{x_2} \left[ c(y_1(x)) \frac{f(y_1(x)) + f(y)}{f(y_1(x)) - f(y)} - \int_0^{x_2} c(y_2(x)) \frac{f(y_2(x)) + f(y)}{f(y_2(x)) - f(y)} \right] dx + K_1. \quad (9.21)$$
Remark 9.4.2 For this specific problem, an explicit expression for the conformal mapping $f(y)$ can be found (see Blanc [33]). It is given by

$$f(y) = \frac{yk(\eta) - \eta k(y)}{yk(\eta) + \eta k(y)},$$

(9.22)

where

$$k(y) = (\alpha - y) \sqrt{r_1 - r_2^2 \alpha^2 y},$$

and $\eta$ is some unspecified constant in the interval $(0, \alpha)$. For our computations we set $\eta = \alpha/2$. With the explicit expression for $f(y)$ we have all ingredients for calculating the integral (9.21), as will be further discussed in Sec. 9.7.

9.5 Boundary value problem II

In this section we will show how the second zero-set discussed in Sec. 9.3 that leads to the curve $R$ gives rise to a Riemann-Hilbert problem for the function $P(x, 0)$. The approach is similar to the one followed in Sec. 9.4.

Lemma 9.5.1 The function $P(x, 0)$ is regular in the domain $R^+$ and satisfies for $x \in R$ the condition

$$\text{Im}[P(x, 0)] = \text{Im} \left[ - P(0, 0) \frac{h_4(x, r_1|x|^2)}{h_2(x, r_1|x|^2)} \right].$$

(9.23)

Proof Similar to the proof of Lemma 9.4.1. □

Lemma 9.5.1 shows that the determination of $P(x, 0)$ reduces to the determination of the solution of the following Riemann-Hilbert boundary value problem on the contour $R$: Determine a function $P(x, 0)$ such that

1. $P(x, 0)$ is regular for $x \in R^+$ and continuous for $x \in R \cup R^+$.
2. $\text{Re} \ [iP(x, 0)] = d(x)$, for $x \in R$,

where

$$d(x) = \text{Im} \left[ P(0, 0) \frac{h_4(x, r_1|x|^2)}{h_2(x, r_1|x|^2)} \right].$$

(9.24)

Note that this problem is inherently different from the Riemann-Hilbert problem for $P(0, y)$ discussed in the previous section, in the sense that there is no symmetry in $x$ and $y$. Moreover, the contours on which the problems have been defined have different features as well (see Lemmas 9.3.2 and 9.3.5).

We introduce the conformal mapping

$$z = g(x) : R^+ \to C^+,$$

(9.25)
9.5 Boundary value problem II

and its inverse

\[ x = g_0(z) : C^+ \rightarrow R^+, \]  \hfill (9.26)

which again allows us to reduce the Riemann-Hilbert problem to a Dirichlet problem on the unit circle: Determine a function \( H(z) \) such that

1. \( H(z) \) is regular for \( z \in C^+ \) and continuous for \( z \in C \cup C^+ \).
2. Re \([iH(z)] = \tilde{d}(z)\), for \( z \in C \), where \( \tilde{d}(z) = d(g_0(z)) \).

This implies that the solution of \( P(x, 0) \) is given by

\[ P(x, 0) = H(g(x)) = -\frac{1}{2\pi} \oint_C d(g_0(w)) \frac{w + g(x)}{w - g(x)} \frac{dw}{w} + K_2, \quad x \in R \cup R^+, \] \hfill (9.27)

where \( K_2 \) is some constant.

For the particular case that \( r_1 = 1 \), our contour \( R \) coincides with a contour in Blanc [32], in which a paired service model is studied using boundary value theory. In this case, an explicit expressions for the conformal mapping \( g(x) \) is given by (see [32], p. 882):

\[ g(x) = 1 - \frac{2\delta(1 - x)^2(1 - xr_2)}{x(1 - \delta)^2(1 - \delta r_2)} \left( 1 + \frac{x - \delta}{\delta(1 - x)} \sqrt{\frac{1 - x\delta^2 r_2}{1 - xr_2}} \right), \] \hfill (9.28)

where \( \delta = \frac{(1 - \sqrt{1 + 8r_2^2})}{4r_2} \). Unfortunately, we have not been able to derive an exact expression for \( g(x) \) in the case that \( r_1 \neq 1 \). When an explicit expression for \( g(x) \) is not available, the standard approach is to determine the inverse mapping \( g_0(z) \) using a well-known method from the theory of conformal mappings. This is sufficient to calculate (9.27), since we show in Sec. 9.7 that we do not need the mapping \( g(x) \) to evaluate \( P(x, 0) \) in \( x \).

For this approach, we need a representation of \( R \) in terms of polar coordinates, i.e.

\[ R = \{ x : x = \rho(\phi) \exp(i\phi), \quad 0 \leq \phi \leq 2\pi \}, \] \hfill (9.29)

which can be obtained in the following way. Since \( 0 \in R^+ \), we have by (9.11) that for each point \( x \) on \( R \) the relation between its absolute value and its real part is given by \( |x|^2 = m(\text{Re}(x)) \), where

\[ m(\delta) := \frac{1}{r_1 r_2 (\hat{r} - 2\delta)}. \] \hfill (9.30)

So, given the angle \( \phi \) belonging to some point on \( R \), the real part of this point, to be denoted by \( \delta(\phi) \), is the solution of

\[ \delta - \cos \phi \sqrt{m(\delta)} = 0, \quad 0 \leq \phi \leq 2\pi. \] \hfill (9.31)

The question arises when the solution to (9.31) is unique. This is the case when \( R \) is a Jordan curve for which it holds that every ray from the point 0 intersects the curve
Figure 9.3: Finding a boundary correspondence point through the mapping \( x = g_0(z) : C \to R \).

\( R \) exactly once. In fact, this is the notion of starshapedness (see Pólya & Szegő [131], p. 125, Exercise 109). In all cases we have considered, \( R \) is a smooth and egg-shaped contour, and thus a starshaped Jordan curve. We see that \( \rho(\phi) = \delta(\phi)/\cos \phi \), and so the parametrization in (9.29) is fully specified.

For a contour that can be described in polar coordinates, the mapping from \( C^+ \) to the interior of this contour is formally given by (cf. Cohen & Boxma [54], Sec. I.4.4, Gaier [77], Sec. 2.1):

\[
g_0(z) = z \exp \left[ \frac{1}{2\pi} \int_0^{2\pi} \log \left\{ \rho(\theta(\omega)) \right\} \frac{e^{i\omega} + z}{e^{i\omega} - z} \, d\omega \right], \quad |z| < 1, \tag{9.32}
\]

with the angular deformation \( \theta(\cdot) \) uniquely determined as the solution of Theodorsen’s integral equation

\[
\theta(\phi) = \phi - \int_0^{2\pi} \log \left\{ \rho(\theta(\omega)) \right\} \cot \left\{ \frac{1}{2} (\omega - \phi) \right\} d\omega, \quad 0 \leq \phi \leq 2\pi. \tag{9.33}
\]

Here, \( \theta(\phi) \) is a strictly increasing and continuous function of \( \phi \), and \( \theta(\phi) = 2\pi - \theta(2\pi - \phi) \). According to the corresponding-boundaries theorem (see Evgrafov [65]), \( g_0(z) \) is continuous in \( C \cup C^+ \). Equation (9.33) is nonlinear and cannot be solved in closed form, though a unique solution can be proven to exist.

We use (9.33) to determine boundary correspondence points. That is, for a point on the unit circle given by its angle \( \phi \), we solve (9.33) numerically to obtain the corresponding point on \( R \), given by its angle \( \theta(\phi) \), see Fig. 9.3. The numerical issues of this procedure are discussed in Sec. 9.7.

9.6 Performance measures

In this section we present exact expressions for two performance measures: the fraction of time a station is empty and the mean stationary queue length at a station. Furthermore, we show for both performance measures that a relation exists between its value at station 1 and station 2. As a consequence, an expression for a
9.6 Performance measures

A performance measure for one of the stations yields the performance measure for the other station as well.

The fractions of time stations 1 and 2 are empty are given by \( P(0, 1) \) and \( P(1, 0) \), respectively. Determining either \( P(0, 1) \) or \( P(1, 0) \) is sufficient to obtain both, since they are related in the following way. Setting \( x = y \) in (9.2) and taking the limit \( x \uparrow 1 \) gives

\[
P(1, 0) = 1 - \frac{\lambda}{\nu_2} - \frac{p}{1 - p} \left[ 1 - \frac{\lambda}{\nu_1} - P(0, 1) \right]. \tag{9.34}
\]

Equation (9.34) alternatively follows from one of the equations

\[
\lambda = \nu_1 (P(1, 0) - P(0, 0)) + \nu_2 (1 - P(1, 0) - P(0, 1) + P(0, 0)),
\]

\[
\lambda = \nu_2 (P(0, 1) - P(0, 0)) + (1 - p) \nu_2 (1 - P(1, 0) - P(0, 1) + P(0, 0)).
\]

These equations stem from the following reasoning: \( P(1, 0) - P(0, 0) \) is the fraction of time station 1 is nonempty while station 2 is empty, and \( 1 - P(1, 0) - P(0, 1) + P(0, 0) \) is the fraction of time both stations are nonempty. Thus, the first equation states that, for station 1, the arrival rate equals the departure rate. Similarly, the second equation corresponds to the equality of arrival-departure rates for station 2. Note that the equations are dependent and therefore do not yield an explicit solution for \( P(1, 0) \) and \( P(0, 1) \).

We will now derive expressions for the mean queue length at both stations, to be denoted by \( \mathbb{E}X_1 \) and \( \mathbb{E}X_2 \). First, we show how these mean queue lengths are related. Differentiating both sides of (9.3) w.r.t. \( y \), and letting \( y \uparrow 1 \), yields

\[
\mathbb{E}X_1 \left[ \frac{1}{\nu_1} + \frac{1}{\nu_2} \right] + \mathbb{E}X_2 \frac{1}{\nu_2} = \frac{\lambda(\nu_1^2 + \nu_1 \nu_2 + \nu_2^2)}{\nu_1 \nu_2 (\nu_1 \nu_2 - \lambda (\nu_1 + \nu_2))}. \tag{9.35}
\]

Again, an interpretation can be given. The left-hand side of (9.35) counts the mean amount of work in the system by multiplying the mean number of jobs at each station by the mean service time they still require before leaving the system. The right-hand side of (9.35) corresponds to the mean amount of work in an \( M/G/1 \) queue (see e.g. Cohen [51]), with Poisson arrivals with rate \( \lambda \) and service times distributed as the sum of two independent and exponentially distributed random variables with mean \( 1/\nu_1 \) and \( 1/\nu_2 \), respectively. Both sides of (9.35) are equal due to the work conservation property of the system. By (9.35) it suffices to calculate either \( \mathbb{E}X_1 \) or \( \mathbb{E}X_2 \) to obtain them both. We will show how \( \mathbb{E}X_2 \) and \( \mathbb{E}X_1 \) follow from the solution of the Riemann-Hilbert boundary value problems discussed in Sec. 9.4 and 9.5, respectively.

When setting \( x = 1 \) in (9.2), the factor \( (y - 1) \) cancels from all terms leaving

\[
P(1, y) = \frac{\nu_1 y + \nu_2}{p \nu_1 y - (1 - p) \nu_2} \left( (1 - p) P(1, 0) + p P(0, y) + \frac{(1 - p) \nu_1 y - p \nu_2}{p \nu_1 y - (1 - p) \nu_2} P(0, 0) \right). \tag{9.36}
\]
Taking derivatives w.r.t. $y$ at both sides of (9.36) yields
\[
\frac{d}{dy} P(1, y) = \frac{\nu_1 \nu_2}{(pv_1 y - (1 - p) \nu_2)^2} ((1 - p) P(1, 0) - p P(0, y))
+ \frac{(\nu_1 y + \nu_2) p}{pv_1 y - (1 - p) \nu_2} \frac{d}{dy} P(0, y) + \frac{(2 p - 1) \nu_1 \nu_2 P(0, 0)}{(pv_1 y - (1 - p) \nu_2)^2}.
\]  \hspace{1cm} (9.37)

Plugging (9.34) into (9.37) and setting $y = 1$ then gives for $p \nu_1 \neq (1 - p) \nu_2$:
\[
\mathbb{E} X_2 = \left[ \frac{d}{dy} P(1, y) \right]_{y=1} = -\frac{\lambda}{pv_1 - (1 - p) \nu_2} + \frac{(\nu_1 + \nu_2) p}{pv_1 - (1 - p) \nu_2} \left[ \frac{d}{dy} P(0, y) \right]_{y=1}.
\]  \hspace{1cm} (9.38)

Thus, to determine $\mathbb{E} X_2$, we only need to compute $\left[ \frac{d}{dy} P(0, y) \right]_{y=1}$. Note that from (9.20) we have that, for $y \in L \cup L^+$,
\[
\frac{d}{dy} P(0, y) = -\frac{1}{\pi} \oint_C f_o(w) \frac{f'(y)}{(w - f(y))^2} dw.
\]  \hspace{1cm} (9.39)

Similarly, when setting $y = 1$ in (9.2), the factor $(x - 1)$ cancels from all terms, which gives
\[
P(x, 1) = \frac{(1 - p) \nu_1}{pv_1 - \lambda x} \left[ \frac{p}{1 - p} P(0, 1) - P(x, 0) + P(0, 0) \right].
\]  \hspace{1cm} (9.40)

Taking derivatives w.r.t. $x$ at both sides of (9.40) yields
\[
\frac{d}{dx} P(x, 1) = \frac{\lambda \nu_1 (1 - p)}{(pv_1 - \lambda x)^2} \left[ \frac{p}{1 - p} P(0, 1) - P(x, 0) + P(0, 0) \right]
- \frac{(1 - p) \nu_1}{pv_1 - \lambda x} \frac{d}{dx} P(x, 0).
\]  \hspace{1cm} (9.41)

Plugging (9.34) into (9.41) and letting $x = 1$ then gives for $\lambda \neq p \nu_1$:
\[
\mathbb{E} X_1 = \left[ \frac{d}{dx} P(x, 1) \right]_{x=1} = \frac{\lambda}{pv_1 - \lambda} - \frac{(1 - p) \nu_1}{pv_1 - \lambda} \left[ \frac{d}{dx} P(x, 0) \right]_{x=1}.
\]  \hspace{1cm} (9.42)

Note that from (9.27) we have that, for $x \in R \cup R^+$,
\[
\frac{d}{dx} P(x, 0) = -\frac{1}{\pi} \oint_C d(g_o(w)) \frac{g'(x)}{(w - g(x))^2} dw.
\]  \hspace{1cm} (9.43)

**Remark 9.6.1** When $p \nu_1 = (1 - p) \nu_2$, setting $y = 1$ in (9.36) gives (after applying l'Hôpital)
\[
\left[ \frac{d}{dy} P(0, y) \right]_{y=1} = \frac{\lambda}{(\nu_1 + \nu_2) p} = \frac{\lambda}{\nu_2}.
\]

We then have an exact expression for $\left[ \frac{d}{dy} P(0, y) \right]_{y=1}$, but we cannot use (9.38) to determine $\mathbb{E} X_2$. We can, though, use (9.42) to find $\mathbb{E} X_1$, and $\mathbb{E} X_2$ through (9.35),...
since $\lambda$ is always smaller than $p\nu_1$ (when $p\nu_1 = (1 - p)\nu_2$) due to the ergodicity condition (9.1). Likewise, for $\lambda = p\nu_1$, setting $x = 1$ in (9.40) gives (after applying l'Hôpital)

$$\left[ \frac{d}{dx} P(x, 0) \right]_{x=1} = \frac{\lambda}{(1 - p)\nu_1} = \frac{\lambda}{\nu_1 - \bar{\lambda}},$$

and we cannot use (9.42) to determine $E_X_1$. We can use (9.38) to find $E_X_2$, since $(1 - p)\nu_2$ is always smaller than $p\nu_1$ (when $\lambda = p\nu_1$) due to the ergodicity condition. We can thus conclude that we can calculate either one of the integrals (9.39) or (9.43) for all allowed parameter values.

9.7 Computational issues

We will now discuss some issues that arise in computing the performance measures from the formal solutions of the Riemann-Hilbert boundary value problems. In Subsec. 9.7.1 we discuss how the location of $\alpha$ and $\beta$ is related to the set of parameter values for which we can actually determine the performance measures. In Subsec. 9.7.2 we discuss a way to determine the performance measures for all allowed parameter values. In Subsec. 9.7.3 we discuss how the integrals involved in computing the performance measures can be determined numerically. Finally, we present some conclusions in Subsec. 9.7.4.

9.7.1 Remarks on $\alpha$ and $\beta$

For calculating the performance measures described in Sec. 9.6, we have to evaluate $P(0, y)$ and $\frac{d}{dy} P(0, y)$ in $y = 1$ or $P(x, 0)$ and $\frac{d}{dx} P(x, 0)$ in $x = 1$. We first discuss the first option. The integration constant $K_1$ can be determined by calculating $P(0, 0)$ from the integral (9.20), and using that $P(0, 0) = 1 - \lambda/\nu_1 - \lambda/\nu_2$. The integrals (9.20), (9.39), however, follow from the solution of a Dirichlet problem that is only defined on or within the unit circle. So, in order to evaluate the integrals, $f(1)$ should lie on or within the unit circle, which is the same as requiring 1 to lie on or within the contour $L$.

The above problem is very common in queueing applications for which the boundary value technique is applied (see e.g. Boxma & Groenendijk [39], Cohen & Boxma [54], p. 360, De Klein [100], p. 89, Feng et al. [72], Mikou [117] and Mikou et al. [118]). In the present context, a key role is played by $\alpha$. In Lemma 9.3.3 we saw that, when $p\nu_1 \leq (1 - p)\nu_2$ (i.e. $r_1 \geq r_2$), it follows that $\alpha \geq 1$ and thus $1 \in L \cup L^+$. Hence, for these parameter values the integrals (9.20), (9.39) can be calculated. To obtain results for parameter values for which it holds that $p\nu_1 > (1 - p)\nu_2$, we might consider analytic continuations for the functions (9.20), (9.39), see e.g. Nauta [122]. However, this would most probably result in numerical difficulties. Alternatively, we can use Taylor series expansion of the corresponding functions around some point in $L^+$, as suggested by Cohen & Boxma [54], p. 360. For a Taylor series expansion
of order \( n \) around \( \hat{y} \in L^+ \), we then have that

\[
P(0,1) \approx \sum_{k=0}^{n} \frac{(1-\hat{y})^k}{k!} \left[ \frac{d^k}{dy^k} P(0,y) \right]_{y=\hat{y}}.
\]

(9.44)

The exact same problem applies for boundary value problem II. In that case, we have to evaluate the integrals (9.27), (9.43) in \( x = 1 \), which is only allowed when \( 1 \in R \cup R^+ \). In Lemma 9.3.6 we have seen that this is the case when \( pv_1 \) and \((1-p)v_2\) are both larger than \( \lambda \) (i.e. \( r_1 < 1 \) and \( r_2 < 1 \)).

To summarize, we show in Fig. 9.4 how the values of \( \alpha \) and \( \beta \) are related to the parameter values \( \lambda, \nu_1, \nu_2 \) (for \( p = 1/2 \)). So, starting from boundary value problem I, we can determine the performance measures for parameter values that fall within areas I and II. By considering boundary value problem II, we can enlarge this set by area III. For area IV we can apply the Taylor series expansion. As will be shown in the next section, the use of Taylor series expansion can be circumvented by considering a third zero-set of the kernel \( h_1(x, y) \).

### 9.7.2 A third zero-set of the kernel

We now discuss an approach to determine \( P(0,1) \) and \( \left[ \frac{d}{dy} P(0,y) \right]_{y=1} \) directly from (9.20), (9.39) despite the fact that \( \alpha < 1 \). The approach has been suggested by De Klein [100], p. 89, and makes use of a zero-set of \( h_1(x, y) \) other than the ones we have considered so far. By establishing a relation between \( P(x,0) \) and \( P(0,y) \) for zero-pairs \((x,y)\) of this set, we are able to calculate the performance measures for all allowed parameter values.

The new zero-set is defined by

\[
\{(x, y^*(x)) \mid h_1(x, y^*(x)) = 0, |x| = 1\},
\]

(9.45)

where \( y^*(x) \) is the zero of the kernel with the smallest modulus. From the function
$y(x)$, as given in (9.5), it is easily seen that
\[
  y^*(1) = \min \left\{ \frac{r_1}{r_2}, 1 \right\},
\]  
(9.46)
for which we have the following result:

**Lemma 9.7.1** For $r_1 = r_2$ it holds that $y^*(1) = 1 = \alpha$. For $r_1 \neq r_2$ it holds that $y^*(1) < \alpha$.

**Proof** The first assertion follows immediately from Lemma 9.3.3. For the second assertion note that if $r_1 > r_2$ it holds that $y^*(1) = 1 < \alpha$. If $r_1 < r_2$ it holds that
\[
  D_1(r_1/r_2) = \left[ 1 + r_1 + \frac{r_1^2 - r_2^2}{r_2} \right]^2 - \left( \frac{1}{r_2} \right)^2.
\]
Since $r_1 < r_2$ implies that $r_1 < 1$, we have that $1 + r_1 + r_1/r_2 - r_1^2/r_2 \in (-2, 2)$, and $D_1(r_1/r_2) < 0$. So, $r_1/r_2 < x_2$, and thus $y^*(1) = r_1/r_2 < \sqrt{r_1x_2/r_2} = \alpha$. □

We exploit the result in Lemma 9.7.1 in the following way. Introducing the shorthand notation $h_k(x) := h_k(x, y^*(x))$, we obtain from (9.2) that
\[
  h_2(x)P(x, 0) + h_3(x)P(0, y^*(x)) + h_4(x)P(0, 0) = 0, \quad |x| = 1.
\]  
(9.47)
Setting $x = 1$ in (9.47) yields
\[
  P(1, 0) = -\frac{1}{h_2(1)} \left[ h_3(1)P(0, y^*(1)) + h_4(1)P(0, 0) \right].
\]  
(9.48)
Since for $r_1 \neq r_2$ it holds that $y^*(1) < \alpha$, the value of $P(0, y^*(1))$ can be computed directly from (9.20). Hence, for $r_1 < r_2$ we cannot obtain $P(0, 1)$ directly from (9.20), but we can obtain $P(1, 0)$ using (9.48), and find $P(0, 1)$ through (9.34).

By using a similar approach we can determine $[\frac{\partial}{\partial y} P(0, y)]_{y=1}$ through (9.39), despite the fact that $r_1 < r_2$. We do need some extra results concerning the zero-set (9.45) though. Observe that $y^*(1)$ is of multiplicity 1 unless $r_1 = r_2$ (for which $y^*(1)$ is of multiplicity two). We further have

**Lemma 9.7.2** The zero $y^*(x)$ is of multiplicity 1 and contained in the disk $|y| < 1$ for every $|x| = 1, x \neq 1$.

**Proof** For $|x| = 1$ it holds that $h_4(x, y) = \lambda x(f(x, y) + g(x, y))$ where
\[
  f(x, y) := (1 + \frac{1}{r_1} + \frac{1}{r_2} - x)y, \quad g(x, y) := -\left( \frac{1}{r_1} \bar{x}y^2 + \frac{1}{r_2} \right),
\]
and $\bar{x}$ the complex conjugate of $x$. We have for $|x| = 1, x \neq 1$,
\[
  |f(x, y)| = |1 + \frac{1}{r_1} + \frac{1}{r_2} - x||y| > \left( \frac{1}{r_1} + \frac{1}{r_2} \right)|y|,
\]
\[
  |g(x, y)| \leq \frac{1}{r_1} |\bar{x}||y|^2 + \frac{1}{r_2} = \frac{1}{r_1} |y|^2 + \frac{1}{r_2}.
\]
Then, for all points $y$ on $|y| = 1$ we have that
\[ |g(x, y)| \leq \frac{1}{r_1} + \frac{1}{r_2} < |f(x, y)|, \quad |y| = 1, \quad |x| = 1, \quad x \neq 1, \]
which implies by Rouché’s theorem (see e.g. Titchmarsh [154]) that $f(x, y) + g(x, y)$ (and thus $h_1(x, y)$) has as many zeros (counted according to their multiplicity) inside $|y| = 1$ as $f(x, y)$. Since $f(x, y)$ has only one zero of multiplicity 1 at $y = 0$, we find that for every $x$ with $|x| = 1$, $x \neq 1$, $h_1(x, y) = 0$ has one solution inside $|y| = 1$, i.e. $y^*(x)$.

From Lemma 9.7.2 it follows for $r_1 \neq r_2$ and $|x| = 1$ that
\[ \left[ \frac{d}{dy} h_1(x, y) \right]_{y=y^*(x)} \neq 0, \quad (9.49) \]
because otherwise $y^*(x)$ would be of multiplicity 2. From the implicit function theorem we then have that $y^*(x)$ is differentiable for $r_1 \neq r_2$ and $|x| = 1$. Differentiating $h_1(x, y^*(x)) = 0$ at both sides gives
\[ \left[ \frac{d}{dx} h_1(x, y) \right]_{y=y^*(x)} + \frac{d}{dx} y^*(x) \left[ \frac{d}{dy} h_1(x, y) \right]_{y=y^*(x)} = 0, \quad (9.50) \]
and thus
\[ \frac{d}{dx} y^*(x) = -\frac{\left[ \frac{d}{dx} h_1(x, y) \right]_{y=y^*(x)}}{\left[ \frac{d}{dy} h_1(x, y) \right]_{y=y^*(x)}}. \quad (9.51) \]
Consequently, differentiating (9.47) w.r.t. $x$ and setting $x = 1$ gives
\[ \left[ \frac{d}{dx} P(x, 0) \right]_{x=1} = -\frac{1}{h_2(1)} \left( h'_2(1) P(1, 0) + h'_3(1) P(0, y^*(1)) + h'_4(1) P(0, 0) \right) \]
\[ + h_3(1) \left[ \frac{d}{dx} y^*(x) \right]_{x=1} \left[ \frac{d}{dy} P(0, y) \right]_{y=y^*(1)}. \quad (9.52) \]
Again, since for $r_1 \neq r_2$ it holds that $y^*(1) < \alpha$, the value of $\frac{d}{dy} P(0, y)$ in $y = y^*(1)$ can be computed directly from (9.39), and through (9.52), (9.42), (9.35) and (9.38) we obtain $\frac{d}{dx} P(0, y)$ in $y = 1$. The approach outlined in this section can also be applied to determine $P(1, 0)$ and $\frac{d}{dx} P(x, 0)$ in $x = 1$ in case $\beta < 1$.

### 9.7.3 Evaluating the integrals

We will now describe how the involved integrals can be determined numerically. For boundary value problem I, we have rewritten the integral (9.20) as (9.21). The integral (9.39) can be rewritten in a similar way. We will evaluate the integrals (9.21) and (9.39) using the trapezium rule, for which we split the interval $[0, 2\pi]$ into $K$ parts of equal length $2\pi/K$. The fact that the whole integrand including the
mapping \( f(y) \) is known explicitly allows for a fine subdivision. For the numerical results to be presented in the next section we have set \( K \) to 250, which guarantees a high level of accuracy.

For boundary value problem II, we need to calculate the integrals (9.27) and (9.43). We will now outline how these integrals can be computed, along with the numerical determination of the mapping \( g_0(z) \). For a more detailed exposition we refer to Chapter IV.1 of Cohen & Boxma [54].

**Step 1: Rewriting the integrals (9.27) and (9.43)**
Substitution of \( w = e^{i\phi} \) into (9.27) yields

\[
P(x, 0) = -\frac{i}{2\pi} \int_0^{2\pi} d(g_0(e^{i\phi})) \frac{e^{i\phi} + g(x)}{e^{i\phi} - g(x)} d\phi + K_2, \quad x \in R \cup R^+.
\]

(9.53)

The integral (9.43) can be rewritten in a similar way, i.e.

\[
\frac{d}{dx} P(x, 0) = -\frac{i}{\pi} \int_0^{2\pi} d(g_0(e^{i\phi})) \frac{g'(x)e^{i\phi}}{(e^{i\phi} - g(x))^2} d\phi, \quad x \in R \cup R^+.
\]

(9.54)

**Step 2: Numerical evaluation of the integrals (9.53) and (9.54)**
We will evaluate the integrals (9.53) and (9.54) in \( x = 1 \) using the above rewriting and the trapezium rule, for which we split the interval \([0, 2\pi]\) into \( K \) parts of equal length \( 2\pi/K \). From (9.53) and (9.54) we then see that we need to determine the values of the conformal mapping \( g_0(\cdot) \) in the points \( e^{i\phi_k}, \ k = 0, 1, \ldots, K - 1 \), with \( \phi_k = 2\pi k/K \). We further need to determine \( g(1) \) and \( g'(1) \).

**Step 3: Solving Theodorsen’s integral equation (9.33)**
For \( K \) points on the unit circle given by their angles

\[
\{\phi_0, \phi_1, \ldots, \phi_{K-1}\},
\]

we need to solve (9.33) to obtain the corresponding points on \( R \), given by their angles \( \{\theta(\phi_0), \theta(\phi_1), \ldots, \theta(\phi_{K-1})\} \). We determine \( \theta(\phi_k), \ k = 0, 1, \ldots, K - 1 \), iteratively (see Gaier [77], p. 67), from

\[
\theta_0(\phi_k) = \phi_k,
\]

(9.55)

\[
\theta_{n+1}(\phi_k) = \phi_k - \int_0^{2\pi} \log\left\{ \frac{\delta(\theta_n(\omega))}{\cos(\theta_n(\omega))} \right\} \cot\left\{ \frac{1}{2}(\omega - \phi_k) \right\} d\omega,
\]

(9.56)

where \( \delta(\theta_n(\omega)) \) is determined from (see (9.31))

\[
\delta(\theta_n(\omega)) - \cos \theta_n(\omega) \sqrt{m(\delta(\theta_n(\omega)))} = 0,
\]

(9.57)

using the Newton-Raphson root-finding procedure. For each step, the integral in (9.56) is numerically determined by again using the trapezium rule with \( K \) parts of
equal length $2\pi/K$. For the iteration, we have used the following stopping criterion:

$$\max_{k \in \{0, \ldots, K-1\}} |\theta_{n+1}(\phi_k) - \theta_n(\phi_k)| < 10^{-6}. \quad (9.58)$$

Finally, it follows from $\rho(\phi) = \delta(\phi)/\cos \phi$ that the value of $g_0(\cdot)$ in $e^{i\phi_k}$ is given by

$$g_0(e^{i\phi_k}) = \frac{\delta(\theta(\phi_k))}{\cos \theta(\phi_k)} e^{i\phi_k}, \quad k = 0, 1, \ldots, K-1. \quad (9.59)$$

We again set $K$ to 250, although in our experience a far smaller value of $K$ is already sufficient to reach an acceptable level of accuracy.

**Step 4:** Determination of $g(1)$ and $g'(1)$

$g(1)$ is obtained as the unique solution $z$ of $g_0(z) = 1$ on $[0, 1]$, and can be determined using (9.32) and Newton-Raphson. $g'(1)$ is given by (see Boxma & Groenendijk [39])

$$g'(1) = \left[ \frac{1}{g(1)} + \frac{1}{2\pi} \int_0^{2\pi} \log \left\{ \frac{\delta(\theta(\omega))}{\cos \theta(\omega)} \right\} \frac{2e^{i\omega}}{(e^{i\omega} - g(1))^2} d\omega \right]^{-1}. \quad (9.60)$$

We calculate $g'(1)$ by numerically determining the integral (9.60) with the trapezium rule and $K$ set to 250.

### 9.7.4 Conclusions

For both boundary value problems, determining the performance measures comes down to computing real integrals. For boundary value problem I, we have an explicit expression for the conformal mapping $f(y)$, and so computing the real integrals becomes a standard exercise. For boundary value problem II, though, we are not able to derive the required conformal mapping $g(x)$. We therefore choose to numerically determine its inverse conformal mapping in order to compute the integrals. Hence, using boundary value problem II requires some additional effort.

In Subsec. 9.7.1 we saw that both models were useful in computing the performance measures. However, using the approach outlined in Subsec. 9.7.2, boundary value problem I can be applied to determine the performance measures for the complete range of allowed parameter values. Hence, we naturally suggest to use boundary value problem I for computational purposes. If, however, one could derive an explicit expression for the mapping $g(x)$, boundary value problem II would be equally suitable.

### 9.8 Some examples

In this section we present some examples that show the effect of the value of the parameter $p$ on the performance measures. In Sec. 9.7 we have concluded that we can determine the performance measures for the whole set of parameter values $\{\lambda, \nu_1, \nu_2, p\}$ for which the ergodicity condition (9.1) is satisfied. Moreover, we have
9.8 Some examples

seen that part of this set allows for multiple ways to determine the performance measures. Therefore, we cross-checked all results presented in this section whenever possible.

Table 9.1: Performance measures for moderate load \( (P(0, 0) = 1/3) \).

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>( \nu_1 )</th>
<th>( \nu_2 )</th>
<th>( p )</th>
<th>( r_1 )</th>
<th>( r_2 )</th>
<th>( \alpha )</th>
<th>( \beta )</th>
<th>( P(1, 0) )</th>
<th>( P(0, 1) )</th>
<th>( E_{X_1} )</th>
<th>( E_{X_2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0.00</td>
<td>( \infty )</td>
<td>0.33</td>
<td>-</td>
<td>-</td>
<td>0.44</td>
<td>0.67</td>
<td>1.33</td>
<td>0.33</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0.25</td>
<td>1.33</td>
<td>0.44</td>
<td>1.32</td>
<td>0.86</td>
<td>0.47</td>
<td>0.60</td>
<td>1.25</td>
<td>0.50</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0.50</td>
<td>0.67</td>
<td>0.67</td>
<td>1.00</td>
<td>1.15</td>
<td>0.52</td>
<td>0.52</td>
<td>1.02</td>
<td>0.97</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0.75</td>
<td>0.44</td>
<td>1.33</td>
<td>0.44</td>
<td>0.86</td>
<td>0.60</td>
<td>0.47</td>
<td>0.68</td>
<td>1.63</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1.00</td>
<td>0.33</td>
<td>( \infty )</td>
<td>-</td>
<td>-</td>
<td>0.67</td>
<td>0.44</td>
<td>0.50</td>
<td>2.00</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>2</td>
<td>0.00</td>
<td>( \infty )</td>
<td>0.50</td>
<td>-</td>
<td>-</td>
<td>0.50</td>
<td>0.50</td>
<td>1.25</td>
<td>0.50</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>2</td>
<td>0.25</td>
<td>0.67</td>
<td>0.67</td>
<td>1.00</td>
<td>1.15</td>
<td>0.61</td>
<td>0.43</td>
<td>0.76</td>
<td>1.16</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>2</td>
<td>0.50</td>
<td>0.33</td>
<td>1.00</td>
<td>0.46</td>
<td>1.00</td>
<td>0.73</td>
<td>0.40</td>
<td>0.38</td>
<td>1.67</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>2</td>
<td>0.75</td>
<td>0.22</td>
<td>2.00</td>
<td>0.17</td>
<td>0.70</td>
<td>0.80</td>
<td>0.39</td>
<td>0.26</td>
<td>1.83</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>2</td>
<td>1.00</td>
<td>0.17</td>
<td>( \infty )</td>
<td>-</td>
<td>-</td>
<td>0.83</td>
<td>0.39</td>
<td>0.20</td>
<td>1.90</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>6</td>
<td>0.00</td>
<td>( \infty )</td>
<td>0.17</td>
<td>-</td>
<td>-</td>
<td>0.39</td>
<td>0.83</td>
<td>1.58</td>
<td>0.17</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>6</td>
<td>0.25</td>
<td>2.00</td>
<td>0.22</td>
<td>1.57</td>
<td>0.70</td>
<td>0.39</td>
<td>0.80</td>
<td>1.57</td>
<td>0.22</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>6</td>
<td>0.50</td>
<td>1.00</td>
<td>0.33</td>
<td>1.37</td>
<td>1.00</td>
<td>0.40</td>
<td>0.73</td>
<td>1.54</td>
<td>0.34</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>6</td>
<td>0.75</td>
<td>0.67</td>
<td>0.67</td>
<td>1.00</td>
<td>1.15</td>
<td>0.43</td>
<td>0.61</td>
<td>1.42</td>
<td>0.83</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>6</td>
<td>1.00</td>
<td>0.50</td>
<td>( \infty )</td>
<td>-</td>
<td>-</td>
<td>0.50</td>
<td>0.50</td>
<td>1.00</td>
<td>2.50</td>
</tr>
</tbody>
</table>

Table 9.2: Performance measures for high load \( (P(0, 0) = 0.1) \).

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>( \nu_1 )</th>
<th>( \nu_2 )</th>
<th>( p )</th>
<th>( r_1 )</th>
<th>( r_2 )</th>
<th>( \alpha )</th>
<th>( \beta )</th>
<th>( P(1, 0) )</th>
<th>( P(0, 1) )</th>
<th>( E_{X_1} )</th>
<th>( E_{X_2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.8</td>
<td>4</td>
<td>4</td>
<td>0.00</td>
<td>( \infty )</td>
<td>0.45</td>
<td>-</td>
<td>-</td>
<td>0.15</td>
<td>0.55</td>
<td>6.53</td>
<td>0.45</td>
</tr>
<tr>
<td>1.8</td>
<td>4</td>
<td>4</td>
<td>0.25</td>
<td>1.80</td>
<td>0.60</td>
<td>1.22</td>
<td>0.75</td>
<td>0.16</td>
<td>0.42</td>
<td>6.34</td>
<td>0.82</td>
</tr>
<tr>
<td>1.8</td>
<td>4</td>
<td>4</td>
<td>0.50</td>
<td>0.90</td>
<td>0.90</td>
<td>1.00</td>
<td>1.04</td>
<td>0.22</td>
<td>0.22</td>
<td>4.63</td>
<td>4.23</td>
</tr>
<tr>
<td>1.8</td>
<td>4</td>
<td>4</td>
<td>0.75</td>
<td>0.60</td>
<td>1.80</td>
<td>0.41</td>
<td>0.75</td>
<td>0.42</td>
<td>0.16</td>
<td>1.42</td>
<td>10.66</td>
</tr>
<tr>
<td>1.8</td>
<td>4</td>
<td>4</td>
<td>1.00</td>
<td>0.45</td>
<td>( \infty )</td>
<td>-</td>
<td>-</td>
<td>0.55</td>
<td>0.15</td>
<td>0.82</td>
<td>11.86</td>
</tr>
<tr>
<td>1.8</td>
<td>6</td>
<td>3</td>
<td>0.00</td>
<td>( \infty )</td>
<td>0.60</td>
<td>-</td>
<td>-</td>
<td>0.16</td>
<td>0.40</td>
<td>6.60</td>
<td>0.60</td>
</tr>
<tr>
<td>1.8</td>
<td>6</td>
<td>3</td>
<td>0.25</td>
<td>1.20</td>
<td>0.80</td>
<td>1.09</td>
<td>0.90</td>
<td>0.20</td>
<td>0.23</td>
<td>5.78</td>
<td>1.83</td>
</tr>
<tr>
<td>1.8</td>
<td>6</td>
<td>3</td>
<td>0.50</td>
<td>0.60</td>
<td>1.20</td>
<td>0.60</td>
<td>0.90</td>
<td>0.45</td>
<td>0.15</td>
<td>1.31</td>
<td>8.53</td>
</tr>
<tr>
<td>1.8</td>
<td>6</td>
<td>3</td>
<td>0.75</td>
<td>0.40</td>
<td>2.40</td>
<td>0.23</td>
<td>0.65</td>
<td>0.61</td>
<td>0.13</td>
<td>0.64</td>
<td>9.54</td>
</tr>
<tr>
<td>1.8</td>
<td>6</td>
<td>3</td>
<td>1.00</td>
<td>0.30</td>
<td>( \infty )</td>
<td>-</td>
<td>-</td>
<td>0.70</td>
<td>0.13</td>
<td>0.43</td>
<td>9.86</td>
</tr>
<tr>
<td>1.8</td>
<td>3</td>
<td>6</td>
<td>0.25</td>
<td>( \infty )</td>
<td>0.30</td>
<td>-</td>
<td>-</td>
<td>0.13</td>
<td>0.70</td>
<td>6.90</td>
<td>0.30</td>
</tr>
<tr>
<td>1.8</td>
<td>3</td>
<td>6</td>
<td>0.25</td>
<td>2.40</td>
<td>0.40</td>
<td>1.38</td>
<td>0.65</td>
<td>0.13</td>
<td>0.61</td>
<td>6.85</td>
<td>0.82</td>
</tr>
<tr>
<td>1.8</td>
<td>3</td>
<td>6</td>
<td>0.50</td>
<td>1.20</td>
<td>0.60</td>
<td>1.19</td>
<td>0.90</td>
<td>0.15</td>
<td>0.45</td>
<td>6.67</td>
<td>4.23</td>
</tr>
<tr>
<td>1.8</td>
<td>3</td>
<td>6</td>
<td>0.75</td>
<td>0.80</td>
<td>1.20</td>
<td>0.73</td>
<td>0.90</td>
<td>0.23</td>
<td>0.20</td>
<td>3.57</td>
<td>10.26</td>
</tr>
<tr>
<td>1.8</td>
<td>3</td>
<td>6</td>
<td>0.25</td>
<td>0.60</td>
<td>( \infty )</td>
<td>-</td>
<td>-</td>
<td>0.40</td>
<td>0.16</td>
<td>1.50</td>
<td>16.38</td>
</tr>
</tbody>
</table>
Table 9.1 displays the performance measures for a moderate load \((P(0,0) = 1/3)\). The results for the limiting cases \(p = 0\) and \(p = 1\) are obtained with the solutions as given in Resing & Örmecri [139]. Obvious observations are that the fraction of time station 1 is busy, and the mean queue length at station 1, both decrease for higher values of \(p\), and vice versa for station 2. Further note that \(EX_1 + EX_2\) increases as a function of \(p\). Table 9.2 displays the performance measures for a high load \((P(0,0) = 0.1)\), from which similar conclusions can be drawn.

For procedures that require two sequential stages, balancing either the mean queue length or the mean workload might be of interest (see e.g. Andradottir et al. [24]). In Tables 9.1 and 9.2 we see that the difference in mean queue lengths at the two stations is strongly influenced by \(p\). As an example, we have plotted in Fig. 9.5 both mean queue lengths for \(\lambda = 1, \nu_1 = 6, \nu_2 = 2\), and \(p\) running from 0 to 1. The imbalance is minimal when \(EX_1 = EX_2\), i.e. \(p \approx 0.18\). Observe that the optimal value of \(p\) does not correspond to the solution of \(p\nu_1 = (1-p)\nu_2\), which is 0.25.

An example of the influence of \(p\) on the mean workloads \(\frac{1}{\nu_1}EX_1\) and \(\frac{1}{\nu_2}EX_2\) is given by Fig. 9.6, which shows the mean workloads for \(\lambda = 1.8, \nu_1 = 3, \nu_2 = 6\), and \(p\) running from 0 to 1. We observe that the imbalance in workloads is minimal when \(\frac{1}{\nu_1}EX_1 = \frac{1}{\nu_2}EX_2\), i.e. \(p \approx 0.72\).
Chapter 10

Two-station network with coupled processors

In Chapter 9 we have considered the two-stage tandem queue with coupled processors. We have shown that the pgf of the joint stationary queue length distribution can be found using the theory of boundary value problems.

In this chapter we present a more general model of a two-station network with coupled processors. This network consists of two single-server stations, Poisson arrival streams, exponential service times and probabilistic routing. For the network with coupled processors, we will show that the pgf of the joint stationary queue length distribution can again be found using the theory of boundary value problems. The network provides a general framework which, among other things, covers and hence generalizes two key models: The two-stage tandem queue and two parallel queues. The second model has been studied by Fayolle & Iasnogorodski [67]. Their paper has become a classic one, since it introduced the technique of boundary value problems to the field of queueing theory. The chapter is based on Van Leeuwaarden & Resing [P8].
10.1 Introduction

In recent years, the coupled-processors discipline has regained attention. This is due to the fact that the coupled-processors discipline incorporates the generalized processor sharing discipline (GPS). GPS is a popular scheduling discipline in modern communication networks, since it provides a way to achieve service differentiation among different types of customers. The recent work on GPS is focused on deriving characteristics of the queue length and delay distributions, particularly on deriving the asymptotic behavior of tail probabilities (see Van Uitert [156] and the references therein), and less on deriving a transform solution of the stationary distribution (as Fayolle & Iasnogorodski [67] did). It is widely recognized that obtaining transform solutions for GPS (or coupled-processors) models with more than two customer classes is extremely hard. Cohen [52] obtained a partial solution for a three-class GPS model, but up till this day that seems to be as far as it can be taken. The two-station network studied in this chapter is in fact a description of a collection of two-dimensional models with exponential service times for which an analytical solution can be obtained in terms of a transform.

The remainder of this chapter is structured as follows. In Sec. 10.2 we give the model description and derive a functional equation for the joint pgf of the stationary queue length distribution. In Sec. 10.3 we derive expressions and relations for various performance measures. In Sec. 10.4 we treat the case of preemptive priority for one of the stations. We show that from the functional equation the joint pgf of the stationary queue length can be obtained without invoking the theory of boundary value problems. For the GPS case, we do need the theory of boundary value problems, as presented in Secs. 10.5 and 10.6. We end this chapter with some conclusions and suggestions for further research in Sec. 10.7.

10.2 Model description and functional equation

Consider an open queueing network with two single-server stations, where jobs arrive externally at station \( j \) according to a Poisson process with rate \( \lambda_j \), \( j = 1, 2 \). Every time a job visits station \( j \), \( j = 1, 2 \), it requires an exponential amount of work with parameter \( \nu_j \). The total service capacity of the two stations together is constant. When both stations are nonempty, the service capacity is divided between the two stations according to fixed proportions, \( p_1 \) and \( p_2 \) (\( p_1 + p_2 = 1 \)), and hence the departure rates at station 1 and 2 then equal \( p_1 \nu_1 \) and \( p_2 \nu_2 \), respectively. If one of the stations is empty, the total service capacity of the stations is allocated to the nonempty station. Hence, the departure rate at that station, station \( j \) say, is then temporarily increased to \( \nu_j \).

Figure 10.1 displays the network. We denote by \( r_{ij} \) the probability that a job moves to station \( j \) after receiving service at station \( i \). After receiving service at station 1, a job will join the queue of station 2 w.p. \( r_{12} \), or leave the system w.p. \( 1 - r_{12} \). Similarly, after receiving service at station 2, a job will join the queue of
10.2 Model description and functional equation

\[
\begin{align*}
\lambda_1 & \quad \rightarrow \quad \text{station 1} \quad 1 - r_{21} \quad \rightarrow \quad 1 - r_{12} \\
1 - r_{21} & \quad \leftarrow \quad \text{station 2} \quad r_{21} \quad \leftarrow \quad \lambda_2
\end{align*}
\]

Figure 10.1: Two-station network with probabilistic routing and coupled processors.

station 1 w.p. \( r_{21} \), or leave the system w.p. \( 1 - r_{21} \). We assume that \( r_{11} = 0 \) and \( r_{22} = 0 \), which is for modelling the queue lengths no loss of generality, since \( r_{ii} > 0 \) implies that the service time of a job at station \( i \) is a geometrically distributed sum of exponentially distributed random variables, which is again exponentially distributed. Note that the network model reduces to the tandem queue covered in Chapter 9 by choosing \( \lambda_2 = r_{21} = 0 \) and \( r_{12} = 1 \).

Denoting by \( \gamma_j \) the throughput of jobs at station \( j \), we have that \( \gamma_1 = \lambda_1 + \gamma_2 r_{21} \) and \( \gamma_2 = \lambda_2 + \gamma_1 r_{12} \), which gives

\[
\gamma_1 = \frac{\lambda_1 + r_{21} \lambda_2}{1 - r_{12} r_{21}}, \quad \gamma_2 = \frac{\lambda_2 + r_{12} \lambda_1}{1 - r_{12} r_{21}}. \quad (10.1)
\]

Denote by \( X_j(t) \) the number of jobs at station \( j \) at time \( t \). Under the condition,

\[
\rho_1 + \rho_2 < 1, \quad (10.2)
\]

where \( \rho_j = \gamma_j / \nu_j \), the two-dimensional Markov process

\[
\{(X_1(t), X_2(t)), t \geq 0\} \quad (10.3)
\]

has a unique stationary distribution. This can be explained by the fact that, independent of \( p_1 \) and \( p_2 \), the two stations together always work at capacity 1 (if there is work in the system), and that \( \rho_1 + \rho_2 \) equals the amount of work brought into the system per time unit. Note that in case the servers are not coupled, we have a standard two-station open Jackson network, for which the stationary joint queue length distribution possesses the following product form

\[
\lim_{t \to \infty} \mathbb{P}(X_1(t) = n_1, X_2(t) = n_2) = (1 - \rho_1)\rho_1^{n_1}(1 - \rho_2)\rho_2^{n_2}. \quad (10.4)
\]

For the case with coupled processors, such a product form fails to hold.

Let us denote by \( \pi(n,k) \) the stationary probability of having \( n \) jobs at station 1 and \( k \) jobs at station 2. The following set of balance equations can then be derived (with \( \lambda := \lambda_1 + \lambda_2 \):

\[
\begin{align*}
\lambda \pi(0,0) &= \nu_2(1 - r_{21})\pi(0,1) + \nu_1(1 - r_{12})\pi(1,0), \\
(\lambda + \nu_1)\pi(1,0) &= \lambda_1 \pi(0,0) + \nu_1(1 - r_{12})\pi(2,0) + p_2 \nu_2(1 - r_{21})\pi(1,1) + \nu_2 r_{21} \pi(0,1), \\
(\lambda + \nu_2)\pi(0,1) &= \lambda_2 \pi(0,0) + \nu_2(1 - r_{21})\pi(0,2) + p_1 \nu_1(1 - r_{12})\pi(1,1) + \nu_1 r_{12} \pi(1,0),
\end{align*}
\]
and for $n \geq 1$, $k \geq 1$

\[
\begin{align*}
(\lambda + \nu_1)\pi(n, 0) &= \lambda \pi(n - 1, 0) + \nu_1(1 - r_{12})\pi(n + 1, 0) + p_2\nu_2(1 - r_{21})\pi(n, 1) + p_2\nu_2r_{21}\pi(n - 1, 1), \\
(\lambda + \nu_2)\pi(0, k) &= \lambda \pi(0, k - 1) + \nu_2(1 - r_{21})\pi(0, k + 1) + p_1\nu_1(1 - r_{12})\pi(1, k) + p_1\nu_1r_{12}\pi(1, k - 1), \\
(\lambda + p_1\nu_1 + p_2\nu_2)\pi(1, 1) &= \lambda \pi(0, 1) + \lambda \pi(1, 0) + \nu_1r_{12}\pi(2, 0) + \nu_2r_{21}\pi(0, 2) + p_1\nu_1(1 - r_{12})\pi(2, 1) + p_2\nu_2(1 - r_{21})\pi(1, 2), \\
(\lambda + p_1\nu_1 + p_2\nu_2)\pi(n, 1) &= \lambda \pi(n - 1, 1) + \lambda \pi(n, 0) + \nu_1r_{12}\pi(n + 1, 0) + p_2\nu_2r_{21}\pi(n - 1, 2) + p_1\nu_1(1 - r_{12})\pi(n + 1, 1) + p_2\nu_2(1 - r_{21})\pi(n, 2), \\
(\lambda + p_1\nu_1 + p_2\nu_2)\pi(1, k) &= \lambda \pi(1, k - 1) + \lambda \pi(0, k) + \nu_2r_{21}\pi(0, k + 1) + p_1\nu_1r_{12}\pi(2, k - 1) + p_2\nu_2(1 - r_{12})\pi(1, k + 1) + p_1\nu_1(1 - r_{12})\pi(2, k).
\end{align*}
\]

We define the joint probability generating function

\[
P(x, y) = \sum_{n \geq 0} \sum_{k \geq 0} \pi(n, k)x^n y^k, \quad |x| \leq 1, |y| \leq 1,
\]

which is regular for $|x| < 1$, continuous for $|x| \leq 1$ for every fixed $y$, and similarly for $x$ and $y$ interchanged. From the balance equations it follows that $P(x, y)$ satisfies the following functional equation

\[
h_1(x, y)P(x, y) = h_2(x, y)P(x, 0) + h_3(x, y)P(0, y) + h_4(x, y)P(0, 0), \quad (10.5)
\]

where

\[
\begin{align*}
h_1(x, y) &= (\lambda + p_1\nu_1 + p_2\nu_2)xy - \lambda x^2 y - \lambda_2 xy^2 - p_1\nu_1r_{12}y^2 - p_2\nu_2r_{21}x^2, \\
h_2(x, y) &= p_2(\nu_2 - \nu_1)xy + \nu_1r_{12}y^2 - \nu_2r_{21}x^2 + \nu_1(1 - r_{12})y - \nu_2(1 - r_{21})x, \\
h_3(x, y) &= -p_1(\nu_2 - \nu_1)xy + \nu_1r_{12}y^2 - \nu_2r_{21}x^2 + \nu_1(1 - r_{12})y - \nu_2(1 - r_{21})x, \\
h_4(x, y) &= p_2\nu_1(xy - r_{12}y^2 - (1 - r_{12})y) + p_1\nu_2(xy - r_{21}x^2 - (1 - r_{21})x).
\end{align*}
\]

Observe that the four above functions are all polynomials of degree 2 in both $x$ and $y$, and that $h_2(x, y) = -\frac{p_2}{p_1}h_3(x, y)$.

### 10.3 Performance measures

In this chapter we derive expressions for two performance measures: The fraction of time stations are empty and the mean stationary queue length at the stations.
The fractions of time stations 1 and 2 are empty, given by $P(0,1)$ and $P(1,0)$ respectively, are related as

$$
\gamma_1 = \nu_1(P(1,0) - P(0,0)) + p_1\nu_1(1 - P(1,0) - P(0,1) + P(0,0)),
$$

$$
\gamma_2 = \nu_2(P(0,1) - P(0,0)) + p_2\nu_2(1 - P(1,0) - P(0,1) + P(0,0)).
$$

These equations stem from the following reasoning (similar to the reasoning on p. 181): $P(1,0) - P(0,0)$ is the fraction of time station 1 is nonempty while station 2 is empty, and $1 - P(1,0) - P(0,1) + P(0,0)$ is the fraction of time both stations are nonempty. Thus, the first equation states that, for station 1, the arrival rate equals the departure rate. Similarly, the second equation corresponds to the equality of arrival-departure rates for station 2. Note that the equations are dependent and therefore do not yield an explicit solution for $P(1,0)$ and $P(0,1)$.

The mean stationary queue lengths at both stations, to be denoted by $E_X_1$ and $E_X_2$, are also related. First, we give the following definition:

**Definition 10.3.1** We define by $B_{[\pi_1;\pi_2]}$ a random variable having a two-phase phase-type distribution, starting from phase 1 or phase 2 w.p. $\pi_1$ and $\pi_2$, respectively. The transition rate out of phase 1 (2) equals $\nu_1$ ($\nu_2$). After completing phase 1 (2), the process may continue with phase 2 (1) w.p. $r_{12}$ ($r_{21}$), or enters the absorbing state.

The moments of $B_{[\pi_1;\pi_2]}$ are given by (see Asmussen [25], Proposition 4.1, p. 83)

$$
E(B^k_{[\pi_1;\pi_2]}) = (-1)^k k! \left[ \begin{array}{c} \pi_1 & \pi_2 \\ \nu_1 & \nu_2 \\ r_{12} & r_{21} \\ \end{array} \right]^{-1} \left[ \begin{array}{c} 1 \\ 1 \\ \end{array} \right].
$$

We then have the following result:

**Lemma 10.3.2** The mean queue lengths at both stations are related in the following way:

$$
E(X_1)E(B_{[1;0]}) + E(X_2)E(B_{[0;1]}) = \frac{\rho}{1 - \rho} [2E(B_{[\lambda_1/\lambda;\lambda_2/\lambda]})]^{-1} E(B^2_{[\lambda_1/\lambda;\lambda_2/\lambda]}),
$$

where $\rho = \lambda E(B_{[\lambda_1/\lambda;\lambda_2/\lambda]})$.

**Proof** The left-hand side of (10.7) counts the mean amount of work in the system by multiplying the mean number of jobs by the mean service time they still require before leaving the system. The right-hand side of (10.7) corresponds to the mean amount of work in an $M/G/1$ queue with Poisson arrivals with rate $\lambda$ and service times distributed as $B_{[\lambda_1/\lambda;\lambda_2/\lambda]}$ (see Cohen [51], p. 256).

By (10.7) it suffices to calculate either $E(X_1)$ or $E(X_2)$ to obtain them both. We will show how $E(X_2)$ follows from the solution of the Riemann-Hilbert boundary value problem discussed in Sec. 10.6.
When setting \( x = 1 \) in (10.5), we can divide both sides by \((y - 1)\), and after rewriting we obtain

\[
P(1, y) = \frac{\nu_2 + \nu_1 r_{12} y}{p_2 \nu_2 - (\lambda_2 + p_1 \nu_1 r_{12}) y} (p_2 P(1, 0) - p_1 P(0, y))
+ \frac{p_1 \nu_2 - p_2 \nu_1 r_{12} y}{p_2 \nu_2 - (\lambda_2 + p_1 \nu_1 r_{12}) y} P(0, 0).
\]

Differentiating (10.8) w.r.t. \( y \) yields

\[
\frac{d}{dy} P(1, y) = \frac{\nu_1 \nu_2 r_{12} + \nu_2 \lambda_2}{(p_2 \nu_2 - (\lambda_2 + p_1 \nu_1 r_{12}) y)^2} (p_2 P(1, 0) - p_1 P(0, y))
+ \frac{p_1 \nu_2 \lambda_2 + (2p_1 - 1) \nu_1 \nu_2 r_{12}}{(p_2 \nu_2 - (\lambda_2 + p_1 \nu_1 r_{12}) y)^2} P(0, 0)
- \frac{p_1 (\nu_2 + \nu_1 r_{12}) y}{p_2 \nu_2 - (\lambda_2 + p_1 \nu_1 r_{12}) y} \frac{d}{dy} P(0, y).
\]

Using

\[
p_2 P(1, 0) - p_1 P(0, 1) = \frac{p_2 \gamma_1}{\nu_1} - \frac{p_1 \gamma_2}{\nu_2} + (1 - 2p_1) P(0, 0),
\]

we set \( y = 1 \) in (10.9) and obtain after some rewriting (for \( p_2 \nu_2 \neq \lambda_2 + p_1 \nu_1 r_{12} \))

\[
\text{EX}_2 = \left[ \frac{d}{dy} P(1, y) \right]_{y=1}
= \frac{\gamma_2}{p_2 \nu_2 - (\lambda_2 + p_1 \nu_1 r_{12})}
- \frac{p_1 (\nu_2 + \nu_1 r_{12})}{p_2 \nu_2 - (\lambda_2 + p_1 \nu_1 r_{12})} \left[ \frac{d}{dy} P(0, y) \right]_{y=1}.
\]

Thus, to determine \( \text{EX}_1 \) and \( \text{EX}_2 \), we need to compute \( \left[ \frac{d}{dy} P(0, y) \right]_{y=1} \). Note that for \( p_1 = p, r_{12} = 1, r_{21} = 0 \) and \( \lambda_2 = 0 \), (10.11) reduces to (9.38).

The case that \( p_2 \nu_2 = \lambda_2 + p_1 \nu_1 r_{12} \) is special. Substituting \( y = 1 \) into (10.8) yields after applying l'Hôpital's rule:

\[
1 = \frac{\gamma_1 r_{12} + \lambda_2 - p_2 \nu_2 - (\nu_2 - \lambda_2) \left[ \frac{d}{dy} P(0, y) \right]_{y=1}}{-\lambda_2 - p_1 \nu_1 r_{12}}.
\]

Since \( \gamma_1 r_{12} + \lambda_2 = \gamma_2 \) this gives after some rewriting

\[
\left[ \frac{d}{dy} P(0, y) \right]_{y=1} = \frac{\gamma_2}{\nu_2 - \lambda}.
\]

### 10.4 Preemptive priority at one of the stations

When the proportion of the service capacity \( p_1 \) is set to one, the jobs at station 1 have preemptive priority over the jobs at station 2. For this case the generating
function $P(x, y)$ can be obtained without employing the theory of boundary value problems, as will be shown in this section.

The functional equation (10.5) then reduces to

$$h_1(x, y)P(x, y) = h_3(x, y)P(0, y) + h_4(x, y)P(0, 0).$$

(10.14)

Let $x = \xi(y)$ denote the unique solution of $h_1(x, y) = 0$ within the unit circle. That is

$$\xi(y) = \frac{\nu_1(1 - r_{12} + r_{12}y)}{\lambda_1(1 - \xi(y)) + \nu_1 + \lambda_2(1 - y)}.$$  

(10.15)

For $x = \xi(y)$, the right-hand side of (10.14) should equal zero, yielding

$$P(0, y) = \frac{\nu_2(r_{21}\xi(y) - y + 1 - r_{21})P(0, 0)}{y(\lambda_1\xi(y) - \lambda - \nu_2) + \nu_2(r_{21}\xi(y) + 1 - r_{21}) + \lambda_2y^2}.$$  

(10.16)

Since $P(0, 0) = 1 - \rho_1 - \rho_2$, substituting (10.16) into (10.14) gives an expression for $P(x, y)$ in terms of $\xi(y)$ only, i.e.

$$P(x, y) = \frac{1}{h_1(x, y)} \left[ \frac{h_3(x, y)\nu_2(r_{21}\xi(y) - y + 1 - r_{21})P(0, 0)}{y(\lambda_1\xi(y) - \lambda - \nu_2) + \nu_2(r_{21}\xi(y) + 1 - r_{21}) + \lambda_2y^2} ight.$$  

$$+ h_4(x, y)P(0, 0) \right].$$  

(10.17)

We will now show that (10.15) and (10.17) have probabilistic interpretations.

Let $N$ represent the number of jobs served during a busy period of station 1, and $Y$ the number of external arrivals at station 2 (with rate $\lambda_2$) during a busy period of station 1. Then, whenever station 1 empties, the service at station 2 is restarted, and a certain number of jobs has arrived there. Denote this number by $H$, which may be represented as

$$H = Y + \sum_{i=1}^{N} Z_i,$$  

(10.18)

where $Y$ and $N$ independent, and $Z_i = 1$ w.p. $r_{12}$ and $Z_i = 0$ w.p. $1 - r_{12}$. With $B_1$ the service time of a job at station 1, the pgf of $H$ is given by

$$H(y) = \int_{t=0}^{\infty} \mathbb{E}(y^H|B_1 = t)\nu_1 e^{-\nu_1 t} dt$$  

$$= \int_{t=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\lambda_1 t)^n}{n!} e^{-\lambda_1 t} \sum_{k=0}^{\infty} \frac{(\lambda_2 ty)^k}{k!} e^{-\lambda_2 t}(1 - r_{12} + r_{12}y)$$  

$$\cdot \mathbb{E}(y^{H_1 + \ldots + H_n})\nu_1 e^{-\nu_1 t} dt$$  

$$= (1 - r_{12} + r_{12}y)\nu_1 \int_{t=0}^{\infty} e^{t(\lambda_1 H(y) - \lambda + \lambda_2 y - \nu_1)} dt,$$

which matches (10.15).
To explain (10.17), we introduce the random vector \((Y_1, Y_2)\), where \(Y_1\) and \(Y_2\) represent the stationary number of jobs at stations 1 and 2, respectively, at a point in time during a busy period of station 1. With \(Q(x, y)\) the joint pgf of \((Y_1, Y_2)\), it can be shown that

\[
Q(x, y) = \frac{x(\nu_1 - \lambda_1)(x - H(y))}{(\lambda + \nu_1)x - \lambda_1x^2 - (1 - r_{12})\nu_1 - r_{12}\nu_1 y - \lambda_2xy},
\]

(10.19)

With \(X_2^{(i)}\) the stationary number of jobs at station 2 during idle periods of station 1, we have that

\[
(X_1, X_2) = \begin{cases} 
(0, X_2^{(i)}), & \text{w.p. } q_1, \\
(0, X_2^{(i)}) + (Y_1, Y_2), & \text{w.p. } q_2, \\
(0, X_2^{(i)} - 1|X_2^{(i)} > 0) + (Y_1, Y_2), & \text{w.p. } q_3,
\end{cases}
\]

(10.20)

where \(q_1, q_2, q_3\) denote the probabilities that an arbitrary time point falls within an idle period of station 1, within a busy period of station 1 that is started with an external arrival to station 1, and within a busy period of station 1 that is started with a job coming from station 2, respectively. That is,

\[
q_1 = 1 - \rho_1, \quad q_2 = \rho_1 \frac{\lambda_1(1 - \rho_1)}{\lambda_1(1 - \rho_1) + \nu_2r_{21}\rho_2}, \quad q_3 = \rho_1 \frac{\nu_2r_{21}\rho_2}{\lambda_1(1 - \rho_1) + \nu_2r_{21}\rho_2}.
\]

From (10.20) we see that

\[
P(x, y) = q_1 \frac{P(0, y)}{P(0, 1)} + q_2 \frac{P(0, y)}{P(0, 1)} Q(x, y) + q_3 \frac{1}{y} \left( \frac{P(0, y) - P(0, 0)}{P(0, 1) - P(0, 0)} \right) Q(x, y),
\]

(10.21)

which, after some lengthy calculations, can be shown to be equal to (10.17).

The above derivation of \(P(x, y)\) can be extended to generally distributed service times, since in that case the decomposition in (10.20) continues to hold.

### 10.5 Analysis of the kernel

From now on we assume that \(p_1 \neq 0, p_2 \neq 0\). In the analysis of the functional equation (10.5) a crucial role is played by the kernel \(h_1(x, y)\). Due to the regularity properties of \(P(x, y)\), for each pair \((x, y)\) on and within the unit circle for which \(h_1(x, y)\) equals zero, the right-hand side of (10.5) must vanish. This provides us with a relation between the unknown functions \(P(0, y)\) and \(P(x, 0)\). Blanc [33] has studied the transient behavior of the two-station network without coupled processors. For this model, the kernel \(h_1(x, y)\) is of the exact same form, and most of the results presented in this section also follow from his work.

Observe that \(h_1(x, y)\) is for each \(x\) a polynomial of degree 2 in \(y\), i.e.

\[
h_1(x, y) = a(x)y^2 + b(x)y + c(x),
\]

(10.22)
For every $x$, there are two possible values of $y$, $y_1(x)$ and $y_2(x)$ say, such that $h_1(x, y_1(x)) = h_1(x, y_2(x)) = 0$. These values can be described by the two-valued function
\[
y(x) = \frac{1}{2a(x)} (-b(x) \pm \sqrt{D(x)}),
\]
where
\[
D(x) = b^2(x) - 4a(x)c(x).
\]

Lemma 10.5.1 The algebraic function $y(x)$, defined by $h_1(x, y(x)) = 0$, has four real branch points $0 \leq x_1 < x_2 \leq 1 < x_3 < x_4$.

Proof The branch points of $y(x)$ are zeros of the discriminant $D(x)$. We have that $D(0) = (p_1 \nu_1 (1 - r_{12}))^2$ and $D(1) = (\lambda_2 + p_1 \nu_1 r_{12} - p_2 \nu_2)^2$. Assuming $D(0)$, $D(1)$ and $\lambda_1$ larger than zero, we have that $D(x) > 0$ as $x \downarrow 0$, $x = 1$ and $x \to \infty$. On the other hand, the function $b(x)$ is negative as $x \downarrow 0$ and $x \to \infty$, but positive at $x = 1$. Hence, $b(x)$ has one zero in the interval $(0, 1)$, and one zero in the interval $(1, \infty)$. At these points, $D(x) = -4a(x)c(x) < 0$. The remaining cases are left to the reader.

For later use, we now study the mapping $y(x)$ for $x \in [x_1, x_2]$ in some more detail. This mapping can be shown to give rise to a smooth and closed contour $L$, as specified in the next lemma.

Lemma 10.5.2 For each $x \in [x_1, x_2]$, $y(x)$ lies on the closed contour $L$, which is symmetric with respect to the real line. For $p_{12} = p_{21} = 0$, $L$ is defined by $|y(x)|^2 = p_2 \nu_2 / \lambda_2$. Otherwise, $|y(x)|^2$ can be written as a function of Re($y$), and
\[
|y|^2 \leq \frac{c(x_2)}{a(x_2)}. \tag{10.24}
\]

Proof For $x \in [x_1, x_2]$, $D(x)$ is negative, so $y_1(x)$ and $y_2(x)$ are complex conjugates. It also follows that $|y(x)|^2 = c(x)/a(x)$, which together with
\[
\frac{d}{dx} \left[ \frac{c(x)}{a(x)} \right] = \frac{p_2 \nu_2 r_{12} (1 - r_{21} + 2r_{21} x + r_{21} \lambda_2 x^2)}{p_1 \nu_1 r_{12} + \lambda_2 x}\]
being nonnegative for $x \in (0, \infty)$ proves (10.24).

We can further solve $|y(x)|^2 = c(x)/a(x)$ as a function of $x$, and denote the solution that lies within $[x_1, x_2]$ by $\tilde{x}(y)$, i.e.
\[
\tilde{x}(y) := \frac{\lambda_2 |y|^2 - p_1 \nu_1 (1 - r_{21}) - \sqrt{(p_1 \nu_1 (1 - r_{21}) - \lambda_2 |y|^2)^2 + 4p_1 \nu_1 r_{12} \nu_2 r_{21} |y|^2}}{2p_2 \nu_2 r_{21}}. \tag{10.26}
\]
So \( \tilde{x}(y) \) is in fact the one-valued inverse function of \( y(x) \). For each \( y \in L \) it then holds that
\[
\text{Re}(y) = \frac{-b(\tilde{x}(y))}{2a(\tilde{x}(y))}
\] (10.27)
Solving (10.27) as a function of \( |y(x)|^2 \) then gives an expression for \( |y(x)|^2 \) in terms of \( \text{Re}(y) \). □

We will henceforth denote the interior of \( L \) by \( L^+ \), and set \( \alpha := y(x_2) = c(x_2)/a(x_2) \), representing the point on \( L \) with the largest modulus.

### 10.6 Boundary value problem

We will now show how the zero-set considered in the previous section leads to a Riemann-Hilbert problem for the function \( P(0, y) \).

**Lemma 10.6.1** The function \( P(0, y) \) is regular in the domain \( L^+ \) and satisfies for \( y \in L \) the condition
\[
\text{Im}[P(0, y)] = \text{Im} \left[ -P(0, 0) \frac{h_4(\tilde{x}(y), y)}{h_3(\tilde{x}(y), y)} \right].
\] (10.28)

**Proof** For zero-pairs \((x, y)\) of the kernel for which \( P(x, y) \) is finite we have
\[
h_2(x, y)P(x, 0) + h_3(x, y)P(0, y) + h_4(x, y)P(0, 0) = 0,
\] (10.29)
from which it follows that
\[
P(0, y) = \frac{1 - p}{p} P(x, 0) - \frac{h_4(x, y)}{h_3(x, y)} P(0, 0).
\] (10.30)
Thus, (10.28) follows from the fact that \( P(x, 0) \) is real for \( x \in [x_1, x_2] \). If \( \alpha \leq 1 \), \( L \) lies entirely within the unit circle. Hence, \( P(0, y) \) is regular in \( L^+ \). If \( \alpha > 1 \), \( P(0, y(x)) \) can be continued analytically over the interval \([x_1, x_2]\) via Equation (10.29), because \( P(x, 0) \) is regular on this interval. Hence, the analytic continuation of \( P(0, y) \) is finite at \( y = y(x_2) \). Because \( P(0, y) \) has a power series expansion at \( y = 0 \) with positive coefficients, this implies that \( P(0, y) \) is regular for \( |y| < y(x_2) \) and hence in \( L^+ \). □

Lemma 10.6.1 shows that determining \( P(0, y) \) reduces to the following Riemann-Hilbert boundary value problem on the contour \( L \): Determine a function \( P(0, y) \) such that
1. \( P(0, y) \) is regular for \( y \in L^+ \) and continuous for \( y \in L^+ \cup L \).
2. \( \text{Re} \ [iP(0, y)] = \chi(y) \), for \( y \in L \),
where

\[ \chi(y) = \operatorname{Im} \left[ \frac{P(0, 0) h_4(x(y), y)}{h_3(x(y), y)} \right]. \]

As done before, we transform the boundary condition (10.28) to a condition on the unit circle (see e.g. Muskhelishvili [121], p. 108). Denote the unit circle by \( C \) and its interior by \( C^+ \). We introduce the conformal mapping:

\[ z = f(y) : L^+ \to C^+, \quad (10.31) \]

and its inverse

\[ y = f_0(z) : C^+ \to L^+. \quad (10.32) \]

Using these mappings, we can reduce the Riemann-Hilbert problem on \( L \) to the following problem: Determine a function \( G(z) \) such that

1. \( G(z) \) is regular for \( z \in C^+ \) and continuous for \( z \in C \cup C^+ \).

2. \( \operatorname{Re} [iG(z)] = \bar{\chi}(z) \), for \( z \in C \), where \( \bar{\chi}(z) = \chi(f_0(z)) \).

The above problem is known as the Dirichlet problem on a circle. Its solution is given by (see Muskhelishvili [121], p. 108)

\[ G(z) = -\frac{1}{2\pi} \oint_C \bar{\chi}(w) \frac{w + z}{w - z} \frac{dw}{w} + K_1, \quad z \in C \cup C^+, \quad (10.33) \]

with \( K_1 \) some real constant. In this way, \( P(0, y) = G(f(y)) \) has been formally determined as

\[ P(0, y) = -\frac{1}{2\pi} \oint_C \bar{\chi}(w) \frac{w + f(y)}{w - f(y)} \frac{dw}{w} + K_1, \quad y \in L \cup L^+. \quad (10.34) \]

In the general case, the mapping \( f_0(z) \) (for evaluating \( \bar{\chi}(w) \)) should be determined using the procedure as described in Sec. 9.5. The procedure consists of finding a fixed number of boundary correspondence points by numerically solving Theodorsen’s integral equation. Exceptions are the case that \( r_{12} = 1, r_{21} = \lambda_2 = 0 \), see Remark 9.4.2, and the case of two parallel queues discussed below.

10.6.1 The case \( r_{12} = r_{21} = 0 \)

In case \( r_{12} = r_{21} = 0 \), there is no routing of customers between the stations, and the network is reduced to two parallel queues. As mentioned before, this model has been analyzed in Fayolle & Iasnogorodski [67]. Since the contour \( L \) is in this case a circle, \( |y(x)|^2 = p_2\nu_2/\lambda_2 \), the mappings \( f(z) \) and \( f_0(z) \) are simply given by

\[ f(y) = \frac{y}{\sqrt{p_2\nu_2/\lambda_2}}, \quad f_0(y) = y\sqrt{p_2\nu_2/\lambda_2}. \quad (10.35) \]
It then readily follows that, for \( y \in L \cup L^+ \),

\[
P(0, y) = -\frac{1}{2\pi} \oint_C \chi(w\sqrt{\frac{p_2\nu_2}{\lambda_2}}) \frac{w + f(y)}{w - f(y)} \, dw + K_1
\]

\[
= -\frac{i}{2\pi} \int_{-\pi}^{\pi} \chi(e^{i\phi}\sqrt{\frac{p_2\nu_2}{\lambda_2}}) \frac{e^{i\phi} + f(y)}{e^{i\phi} - f(y)} \, d\phi + K_1. \tag{10.36}
\]

The case \( r_{12} = r_{21} = 0 \) is the only situation for which we cannot specify the inverse function \( \tilde{x}(y) \) according to (10.26). Instead, we can use the fact that \( L \) is a circle to derive \( \tilde{x}(y) \). We start from the observation that for the zero-pairs \((x, y)\) of the kernel \( h_1(x, y) \) it holds

\[
(\lambda + p_1\nu_1 + p_2\nu_2)x - \lambda_1 x^2 - \lambda_2 xy - p_1\nu_1 - p_2\nu_2 \frac{x}{y} = 0. \tag{10.37}
\]

Also, for each \( y \in L \) we have

\[
y = \frac{p_2\nu_2}{\lambda_2} e^{i\phi} = \frac{p_2\nu_2}{\lambda_2} (\cos \phi + i \sin \phi), \tag{10.38}
\]

for some \( \phi \in [0, 2\pi] \). Plugging (10.38) into (10.37) and solving for \( x \) gives two possible outcomes

\[
x_1(\phi) = \frac{\gamma - \sqrt{\gamma^2 - 4p_1\nu_1 \lambda_1}}{2\lambda_1}, \quad x_2(\phi) = \frac{\gamma + \sqrt{\gamma^2 - 4p_1\nu_1 \lambda_1}}{2\lambda_1}, \tag{10.39}
\]

where \( \gamma := \lambda + p_1\nu_1 + p_2\nu_2 - 2\sqrt{p_2\nu_2 \lambda_2} \cos \phi \). It is straightforward to see that \( x_1(\phi) \in [x_1, x_2] \) for all \( \phi \in [0, 2\pi] \). Therefore, the inverse function \( \tilde{x}(y) \) is for all \( y \in L \) specified by

\[
\tilde{x}(e^{i\phi}\sqrt{p_2\nu_2/\lambda_2}) = x_1(\phi). \tag{10.40}
\]

### 10.7 Conclusions and further research

For the two-station open queuing network with Poisson arrivals, exponential service times, and coupled processors, we have shown that the pgf of the joint stationary queue length distribution can be found using the theory of boundary value problems.

Calculating the performance measures described in Sec. 10.3 involves computational issues as described in Sec. 9.7 for the tandem queue. As for the tandem queue, it is crucial to determine either the mapping \( f(z) \) or its inverse mapping \( f_0(z) \). If no explicit description of either mapping is available, the inverse mapping \( f_0(z) \) can be determined numerically using the procedure described in Sec. 9.5. So far, we have not been able to derive an explicit description for \( f(z) \) or \( f_0(z) \), leaving it as a challenging topic for further research.

Next to coupled processors, we introduced in Subsec. 1.3.1 a scheduling discipline referred to as partial coupling, which, in the more general context of this chapter,
would imply the following. Whenever both queues are nonempty, the capacity is divided among stations 1 and 2 according to fixed fractions. Whenever queue 1 is empty, all capacity goes to station 2. When queue 2 is empty, however, the capacity of station 1 is not increased. Partial coupling is an example of a non-work-conserving scheduling discipline. More generally, we could describe the class of non-work-conserving scheduling disciplines as follows. When both stations are nonempty, the service rate of station $j$ is equal to $\nu_j$. If one of the stations becomes empty, the service rate at the other station changes from $\nu_j$ to $\nu_j^*$, where $\nu_j^* \neq \nu_1 + \nu_2$ for some $j = 1, 2$.

For the two-station open queueing network with Poisson arrivals, exponential service times, and a non-work-conserving scheduling discipline, again a Riemann-Hilbert boundary value problem can be formulated on some contour $M$ (with $M^+$ the interior of $M$): Determine the function $P(0, y)$ such that

1. $P(0, y)$ is regular for $y \in M^+$ and continuous for $y \in M^+ \cup M$.
2. $\text{Re} \{p(y)P(0, y)\} = q(y)$, for $y \in M$.

Note that due to the function $p(y)$, this is a more general Riemann-Hilbert boundary value problem than those we have encountered in Chapters 9 and 10 (in which case $p(y) = 1$). Solving this more general Riemann-Hilbert boundary value problem requires a slightly more complicated technique, as explained in Mushkelisvili [121], p. 100, and Cohen & Boxma [54], p. 56. A detailed study of this more general Riemann-Hilbert boundary value problem is an interesting topic for further research.
Two-station network with coupled processors
In Chapter 9 we considered the tandem queue with coupled processors. In this model the total service capacity is shared between the stations according to fixed proportions, except when one of the stations is empty, in which case the total service capacity is given to the nonempty station. We showed that the problem of finding the generating function of the joint stationary queue length distribution can be reduced to a Riemann-Hilbert boundary value problem.

In this chapter we will be concerned with deriving asymptotic expressions for the stationary queue length distribution at each of the stations. We aim at obtaining functions $f$ of the type

$$P(X_j = n) \sim f(n), \quad j = 1, 2,$$

(11.1)

where the symbol $\sim$ indicates that $P(X_j = n)/f(n) \to 1$ for $n$ tending to infinity. Such asymptotic expressions would be useful in determining the probability of large queue lengths, which is instrumental in calculating quality-of-service measures.
11.1 Introduction

In order to introduce our method for deriving tail asymptotics, we first apply it to the simple case where station 1 gets preemptive priority over station 2. In this case, the first queue will behave as a regular $M/M/1$ queue (as if the second queue does not exist), for which the asymptotic behavior of the queue length distribution is well known. The second queue does depend on the first queue, and so the asymptotic behavior of the queue length distribution at the second queue becomes more complicated. To obtain the asymptotic expressions for the queue length distribution at station 2, we first need to determine the dominant singularity of the pgf of the stationary queue length, defined as the point with smallest modulus at which the pgf fails to be analytic. Next, we expand the pgf around its dominant singularity, leading to asymptotic expressions of the type (11.1). The analysis for the priority case serves as an illustrative example of the steps that should be taken in more complicated cases.

The tandem queue with coupled processors is such a case. The biggest challenge in obtaining asymptotic expressions is finding the dominant singularities. In doing this, we use some results on $P(x, y)$, defined as the joint pgf of the stationary queue lengths at both stations. In Chapter 9 we showed that $P(x, y)$ satisfies a specific type of functional equation. In the same chapter, we exploited some properties of this functional equation thanks to which we could apply the theory of boundary value problems. This resulted in solutions for $P(x, 0), P(x, 1), P(0, y)$ and $P(1, y)$ that are valid only in specific regions of the $x$ and $y$-plane. Therefore, we perform an analytic continuation of the functions in order to determine their dominant singularities. This analytic continuation is based on the functional equation for $P(x, y)$ and is performed in the two-dimensional plane, which makes it harder than for instance in the (one-dimensional) priority case.

The analytic continuation sketched above serves to detect the dominant singularities of the functions $P(x, 0), P(x, 1), P(0, y)$ and $P(1, y)$. By investigating the behavior of these functions in the vicinity of their dominant singularities, we are able to derive asymptotic relationships of the type (11.1). There are some alternative methods for deriving asymptotic expressions for tail probabilities. One alternative method is developed by Foley & McDonald [75] and uses large deviations techniques. The large deviations techniques are accompanied by some intuition (in terms of most probable paths) that can be helpful in understanding why the asymptotic behavior takes on different shapes for different parameter values. Another alternative method is developed by Takahashi et al. [151] for quasi-birth-and-death processes. Indeed, the tandem queue with coupled processors is a quasi-birth-and-death process, and so the method of Takahashi et al. [151], along with the computational framework presented in Haque et al. [84], can be applied to derive asymptotic expressions. We discuss both methods and point out the differences with our analytic approach.

The chapter is structured as follows. After the model description in Sec. 11.2, we first discuss the priority case in Sec. 11.3. Then, for the coupled-processors case, we present the analytic continuation and singularity analysis of $P(x, 0), P(x, 1), P(0, y)$.
11.2 Model description and functional equation

Consider a two-stage tandem queue, where jobs arrive at the first station according to a Poisson process with rate $\lambda$. After receiving service at this station, they move to the second station, and upon completion of service at the second station they leave the system. The service times of a job at each of the stations are exponentially distributed (independent at both stations) with mean 1. Whenever both stations are nonempty, the service rates at each of the stations are $\mu_1$ and $\mu_2$, respectively. Whenever one of the stations is empty, jobs at the nonempty station are served at a rate of $\mu_1 + \mu_2$. Note that the model discussed in Chapter 9 reduces to the present model by setting $\nu_1 = \nu_2 = \mu_1 + \mu_2$ and $p = \mu_1 / (\mu_1 + \mu_2)$. We set the parameters this way for notational convenience, but the theory presented in this chapter can be easily extended to the more general formulation in Chapter 9.

With $X_j(t)$ denoting the number of jobs at station $j$ at time $t$, the two-dimensional process $\{(X_1(t), X_2(t)), t \geq 0\}$ is a Markov process. The ergodicity condition under which this Markov process has a unique stationary distribution is given by

$$\frac{2\lambda}{\mu_1 + \mu_2} < 1.$$  \hfill (11.2)

Without loss of generality, we assume that $\lambda + \mu_1 + \mu_2 = 1$, and (11.2) then reduces to $\lambda < 1/3$. We denote the stationary queue lengths by $X_1$ and $X_2$.

Let $\pi(n, k)$ denote the stationary probability of having $n$ customers at station 1 and $k$ customers at station 2, i.e. $\pi(n, k) = \lim_{t \to \infty} P(X_1(t) = n, X_2(t) = k)$. Also, define the joint pgf

$$P(x, y) = \sum_{n \geq 0} \sum_{k \geq 0} \pi(n, k)x^n y^k, \quad |x| \leq 1, |y| \leq 1,$$

which is, for every fixed $y$, regular for $|x| < 1$ and continuous for $|x| \leq 1$. A similar statement holds for $x$ and $y$ interchanged. The functional equation (9.2) then reduces for this chapter’s parameter setting to

$$h_1(x, y)P(x, y) = h_2(x, y)P(x, 0) + h_3(x, y)P(0, y) + h_4(x, y)P(0, 0),$$  \hfill (11.3)
where

\[ h_1(x, y) = xy - \lambda x^2 y - \mu_1 y^2 - \mu_2 x, \]
\[ h_2(x, y) = \mu_2(y^2 - x), \]
\[ h_3(x, y) = \mu_1(x - y^2), \]
\[ h_4(x, y) = \mu_1 x(y - 1) + \mu_2 y(x - y). \]

### 11.3 Priority for station 1

We now derive asymptotic expressions for the queue length distribution in case station 1 has preemptive priority over station 2. This section serves to illustrate the steps that are needed to obtain asymptotic expressions of type (11.1). It follows from the approach taken in Sec. 10.4 that

\[ P(x, y) = \frac{\rho_1 x(1 - \xi(y)) + x - y}{(\rho_1 + 1)x - \rho_1 x^2 - y} P(0, y), \quad (11.4) \]

where \( \rho_i = \lambda/\mu_i \) and

\[ P(0, y) = \frac{(1 - y) P(0, 0)}{1 - y - \rho_2 y(1 - \xi(y))}, \quad (11.5) \]

\( (P(0, 0) = 1 - \rho_1 - \rho_2) \) and

\[ \xi(y) = \frac{1 + \rho_1}{2\rho_1} (1 - \sqrt{1 - 4\rho_1 y/(1 + \rho_1)^2}). \quad (11.6) \]

From this it follows that

\[ P(1, y) = \frac{1 - y + \rho_1(1 - \xi(y))}{1 - y} P(0, y). \quad (11.7) \]

The function \( \xi(y) \) represents the pgf of the number of customers served in a busy period of an \( M/M/1 \) queue with arrival rate \( \lambda \) and service rate \( \mu_1 \) (see e.g. Cohen [51], p. 190). Denote this random variable by \( \xi \). Then, \( \xi(y) \) allows for an explicit inversion ([51], p. 190)

\[ P(\xi = n) = \frac{1}{n} \binom{2n - 2}{n - 1} \frac{\rho_1^{n-1}}{(1 + \rho_1)^{2n-1}}, \quad n = 1, 2, \ldots. \quad (11.8) \]

Stirling’s approximation \( n! \sim n^n e^{-n} \sqrt{2\pi n} \) yields

\[ \binom{2n}{n} \sim \frac{2^{2n}}{\sqrt{\pi n}}, \quad (11.9) \]
11.3 Priority for station 1

and thus

\[
\mathbb{P}(\xi = n) \sim \frac{1}{n} \frac{2^{2n-2}}{\sqrt{\pi(n-1)}} \frac{1}{\rho_1} \frac{\rho_1^n}{(1 + \rho_1)^{2n}} \\
= \frac{1}{2\rho_1} \frac{1}{2\sqrt{\pi}n^{3/2}} \frac{\sqrt{n}}{\sqrt{n-1}} \left( \frac{4\rho_1}{(1 + \rho_1)^2} \right)^n \\
\sim \frac{1}{2\rho_1} \frac{1}{2\sqrt{\pi}n^{3/2}} \left( \frac{4\rho_1}{(1 + \rho_1)^2} \right)^n.
\] (11.10)

**Remark 11.3.1** Note that (11.4) for \( y = 1 \) yields

\[
P(x, 1) = \frac{1 - \rho_1}{1 - \rho_1 x}.
\] (11.11)

the pgf of the stationary queue length in the \( M/M/1 \) queue. Observe that \( P(x, 1) \) has a pole in \( x = 1/\rho_1 \), and so the asymptotic behavior of the queue length distribution at station 1 satisfies

\[
\mathbb{P}(X_1 = n) \sim c \rho_1^n
\] with

\[
c = 1 - \rho_1
\] and

\[
\mathbb{P}(X_1 = n) = (1 - \rho_1) \rho_1^n.
\]

We are primarily interested in the asymptotic behavior of the queue length at station 2, where we aim at deriving accurate approximations for \( \mathbb{P}(X_2 = n) \) for \( n \) large. For this, we do not have the luxury of an explicit inversion of the pgf, like in case of \( \xi(y) \). Therefore, we use singularity analysis applied to \( \xi(y) \) (as an example) and \( P(1, y) \).

For a function \( f(z) = \sum_{j=0}^{\infty} f_j z^j \), we denote by \( C_z[f(z)] \) the coefficient corresponding to \( z^j \), i.e. \( C_{z^j}[f(z)] = f_j \). For general \( \alpha \) we have that

\[
C_{z^n}[(1 - z)^{-\alpha}] = (-1)^n \binom{n - 1}{\alpha} = \frac{\Gamma(n + \alpha)}{\Gamma(\alpha) \Gamma(n + 1)},
\] (11.12)

where \( \Gamma(z) \) is the Gamma function defined for \( \text{Re}(z) > 0 \) as

\[
\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt.
\] (11.13)

Applying Stirling’s approximation \( \Gamma(n + 1) \sim n^n e^{-n} \sqrt{2\pi n} \) then gives (see e.g. Flajolet & Sedgewick [73])

\[
C_{z^n}[(1 - z)^{-\alpha}] = \frac{n^{\alpha-1}}{\Gamma(\alpha)} \left( 1 + \mathcal{O}(1/n) \right).
\] (11.14)

We now apply the above result to \( \xi(y) \). Denote by \( \mathbb{C}_y \) the complex \( y \)-plane and observe that the function \( \xi(y) \) is analytic in \( \mathbb{C}_y \setminus [(1 + \rho_1)^2/(4\rho_1), \infty) \), i.e. it has a branch point at

\[
y_B = \frac{(1 + \rho_1)^2}{4\rho_1}.
\] (11.15)
This particular case is covered by (11.14) with \( \alpha = -1/2 \), which gives
\[
\mathbb{P}(\xi = n) = C_{y^n}[\xi(y)] = -\frac{1 + \rho_1}{2\rho_1} C_{y^n}\left[\sqrt{1 - 4\rho_1 y/(1 + \rho_1)^2}\right]
\]
\[
= -\frac{1 + \rho_1}{2\rho_1} \left(\frac{4\rho_1}{(1 + \rho_1)^2}\right)^n C_{y^n}\left[\sqrt{1 - y}\right]
\]
\[
= -\frac{1 + \rho_1}{2\rho_1} \left(\frac{4\rho_1}{(1 + \rho_1)^2}\right)^n \frac{1}{\Gamma(-1/2)} n^{-3/2} \left(1 + \mathcal{O}(1/n)\right)
\]
\[
= \frac{1 + \rho_1}{2\rho_1} \left(\frac{4\rho_1}{(1 + \rho_1)^2}\right)^n \frac{1}{2\sqrt{\pi n}} \left(1 + \mathcal{O}(1/n)\right). \tag{11.16}
\]

Note that (11.16) yields (11.10).

We now turn to the function \( P(1, y) \) and study the asymptotic behavior of
\[
C_{y^n}[P(1, y)] = \mathbb{P}(X_2 = n)
\]
by means of singularity analysis. The singularities of \( P(1, y) \) consist of the branch point \( y_B \) and zeros of the denominator of the right-hand side of (11.5):
\[
1 - y - \rho_2 y (1 - \xi(y)) = 0. \tag{11.17}
\]

The question then is which singularity has the smallest modulus, since the singularity of \( P(1, y) \) with the smallest modulus is dominant and determines the asymptotic behavior of the coefficients of \( P(1, y) \), i.e. \( \mathbb{P}(X_2 = n) \), for large values of \( n \).

Because we already know that \( P(1, y) \) has a branch point in \( y_B \), it remains to be investigated whether \( P(1, y) \) has a pole in \( 1 < |y| < y_B \), so whether (11.17) has a solution in \( 1 < |y| < y_B \). A first observation is

**Lemma 11.3.2** If
\[
\rho_2 < \frac{2\rho_1(1 - \rho_1)}{(1 + \rho_1)^2}, \tag{11.18}
\]
then the only solution to (11.17) in the region \( |y| < y_B \) is given by \( y = 1 \).

**Proof** Factorizing \( y = 1 \) from (11.17) yields
\[
\frac{y - 1}{1 - y} - \frac{1}{\rho_2} = 0. \tag{11.19}
\]

For an arbitrary point \( y \) on the circle \( |y| = y_B \) it holds that
\[
\left|\frac{y - 1}{1 - y}\right| \leq |y| \frac{1 - \xi(|y|)}{1 - |y|} = (1 + \rho_1)^2/(4\rho_1) \frac{1 - \frac{1 + \rho_1}{2\rho_1}}{(1 - \rho_1)^2} = \frac{(1 + \rho_1)^2}{2\rho_1(1 - \rho_1)}. \tag{11.20}
\]

Hence, if (11.18) holds, we have by Rouché’s theorem that (11.19) has equally many solutions with \( |y| < y_B \) as \( 1/\rho_2 = 0 \). \( \square \)
11.3 Priority for station 1

Candidate solutions to (11.19) are the solutions to
\[
1 - \frac{4\rho_1}{(1 + \rho_1)^2} y = \left( \frac{2\rho_1}{1 + \rho_1} \left( \frac{1}{\rho_2 y} - \frac{1}{\rho_2} - 1 \right) + 1 \right)^2, \tag{11.21}
\]
given by
\[
y_\pm = \rho_2 - \rho_1 - \rho_1 \rho_2 \pm \frac{\sqrt{4\rho_1 \rho_2^2 + (\rho_2 - \rho_1 \rho_2)^2}}{2\rho_2^2}. \tag{11.22}
\]
Both \(y_+\) and \(y_-\) are real. These real solutions should satisfy \(f(y) = g(y)\) where \(f(y) = 1 + \rho_2 - 1/y\) and \(g(y) = \rho_2 \xi(y)\). For \(y \in [0, y_B]\) both \(f\) and \(g\) are increasing functions with \(\lim_{y \to 0} f(y) = -\infty < g(0) = 0\), \(f(1) = g(1) = \rho_2\), \(f'(1) = 1 > \rho_2/(1 - \rho_1) = g'(1)\). Hence, when \(f(y_B) \leq g(y_B)\), (11.17) has a solution in \((1, y_B]\) equal to \(y_+ =: y_p\). The condition \(f(y_B) \leq g(y_B)\) is equivalent to (see also Fig. 11.1)
\[
\rho_2 \geq \frac{2\rho_1 (1 - \rho_1)}{(1 + \rho_1)^2} =: \rho_c. \tag{11.23}
\]
The negative real solution \(y_-\) can be excluded, since for \(y \in (-\infty, 0]\), \(f(y) > 0\) and \(g(y) < 0\).

This gives the following result:

**Lemma 11.3.3** If \(\rho_2 < \rho_c\), the dominant singularity of the function \(P(1, y)\) is the branch point \(y_B\). If \(\rho_2 > \rho_c\), the dominant singularity of the function \(P(1, y)\) is the pole \(y_p\). If \(\rho_2 = \rho_c\), the dominant singularity of the function \(P(1, y)\) is \(y_B = y_p\).

We may thus conclude the following about the behavior of \(P(1, y)\) near its dominant singularity:

**Lemma 11.3.4**

\[
P(1, y) \approx \begin{cases} 
P(1, y_B) + \gamma_1 \sqrt{1 - y/y_B}, & \rho_2 < \rho_c, \\
\gamma_2 / \sqrt{1 - y/y_B}, & \rho_2 = \rho_c, \\
\gamma_3 / (1 - y/y_p), & \rho_2 > \rho_c,
\end{cases} \tag{11.24}
\]

where \(P(1, y) \approx f(y)\) indicates that \(P(1, y)/f(y) \to 1\) when \(y\) tends to its dominant singularity \(y_B\) or \(y_p\), and

\[
\gamma_1 = \frac{-2P(0, 0)\rho_1 (1 + \rho_1) (\rho_2 + 2\rho_1 \rho_2 + \rho_2^2 (4 + \rho_2))}{(\rho_2 + \rho_2^2 (2 + \rho_2) - 2(1 - \rho_2) \rho_1)^2},
\]
\[
\gamma_2 = \frac{-2P(0, 0)\rho_1 (1 - \rho_1)}{\rho_2 (1 + \rho_1)^2},
\]
\[
\gamma_3 = \frac{P(0, 0)}{y_p} \cdot \frac{1 - y_p + \rho_1 (1 - \xi(y_p))}{-1 - \rho_2 (1 - \xi(y_p)) + \rho_2 y_p \xi'(y_p)}.
\]

**Proof** It immediately follows that
\[
\gamma_3 = \lim_{y \to y_p} (1 - y/y_p) P(1, y). \tag{11.25}
\]
Figure 11.1: Set of allowed parameter values and associated behavior as in Lemma 11.3.4 and Thm. 11.3.5 of the stationary queue length at station 2.

Further, upon substituting $z^2 = 1 - y/y_B$ (and thus $y = y_B(1 - z^2)$) into $\xi(y)$ and $P(1, y)$ we find

$$P(1, z) = P(0, 0) \frac{1 - y_B(1 - z^2) + \rho_1(1 - \bar{\xi}(z))}{1 - y_B(1 - z^2) - \rho_2 y_B(1 - z^2)(1 - \xi(z))},$$  

where

$$\bar{\xi}(z) = \frac{1 + \rho_1}{2\rho_1}(1 - z).$$  

Then $\gamma_1 = \left[\frac{d}{dz} P(1, z)\right]_{z=0}$ and $\gamma_2 = \lim_{z \to 0} z P(1, z)$.

Applying (11.14) for $\alpha = -1/2, 1/2$ and 1 then yields

**Theorem 11.3.5** For the case that station 1 has preemptive priority over station 2, the tail of the probability distribution function of the number $X_2$ of customers in queue 2 is given asymptotically as

(a) If $\rho_2 < \rho_c$,

$$\mathbb{P}(X_2 = n) \sim \gamma_1 \frac{-1}{2\sqrt{\pi n}} \left(\frac{1}{y_B}\right)^n.$$

(b) If $\rho_2 = \rho_c$,

$$\mathbb{P}(X_2 = n) \sim \gamma_2 \frac{1}{2\sqrt{\pi n}} \left(\frac{1}{y_B}\right)^n.$$
If \( \rho_2 > \rho_c \),
\[
\mathbb{P}(X_2 = n) \sim \gamma_3 \left( \frac{1}{y^p} \right)^n,
\]
where the symbol \( \sim \) has been introduced in (11.1).

We conclude the section with a brief summary of the method we used to obtain the tail asymptotics in Thm. 11.3.5. Starting point was the expression for the probability generating function of the probability distribution of interest, in this case (11.7). Then, we had to take three crucial steps: (i) First, we look for the dominant singularity of the probability generating function, defined as the point with smallest modulus outside the unit circle at which the probability generating function fails to be analytic. (ii) Next, we approximate the probability generating function in the vicinity of its dominant singularity, and finally (iii) we use the general result (11.14) to obtain the asymptotic expressions for the tail probabilities. We introduced the steps for the priority case, and we execute them in the next section for the coupled-processors case. There is a crucial difference, though. While for the priority case the probability generating function \( P(0, y) \) is known explicitly, for the coupled-processors case we have only an implicit description of the probability generating function. This makes determining the dominant singularity (step (i)) a bit harder.

### 11.4 Coupled processors and analytic continuation

From now on we exclude the priority case \((\mu_1 \neq 0 \text{ and } \mu_2 \neq 0)\). Observe that for each \( x \), \( h_1(x, y) \) is a polynomial of degree 2 in \( y \). For every \( x \), there are two possible values
\[
y_+(x) = \frac{x - \lambda x^2 + \sqrt{(x - \lambda x^2)^2 - 4\mu_1 \mu_2 x}}{2\mu_1},
\]
\[
y_-(x) = \frac{x - \lambda x^2 - \sqrt{(x - \lambda x^2)^2 - 4\mu_1 \mu_2 x}}{2\mu_1},
\]
such that \( h_1(x, y_+(x)) = h_1(x, y_-(x)) = 0 \). Similarly, for every \( y \), there are two possible values
\[
x_+(y) = \frac{y - \mu_2 + \sqrt{(y - \mu_2)^2 - 4\lambda \mu_1 y^3}}{2\lambda y},
\]
\[
x_-(y) = \frac{y - \mu_2 - \sqrt{(y - \mu_2)^2 - 4\lambda \mu_1 y^3}}{2\lambda y},
\]
such that \( h_1(x_+(y), y) = h_1(x_-(y), y) = 0 \). The following result holds:

**Lemma 11.4.1** The functions \( x_+(y) \) and \( x_-(y) \) have four real branch points \( 0 < y_1 < y_2 \leq 1 < y_3 < y_4 \). The functions \( y_+(x) \) and \( y_-(x) \) have four real branch points \( 0 = x_1 < x_2 \leq 1 < x_3 < x_4 \).

The proof of the above lemma is straightforward, see Lemma 9.3.1. A far more intricate result is the following:
Theorem 11.4.2  (Fayolle et al. [68], p. 41, Thm. 3.2.3) The function $P(x,0)$ is a meromorphic function in the complex $x$-plane cut along $[x_3, x_4]$. The function $P(0, y)$ is a meromorphic function in the complex $y$-plane cut along $[y_3, y_4]$.

The proof of Thm. 11.4.2 requires notions as Riemann surfaces and Galois automorphisms, which go beyond the scope of this chapter. Therefore, we will apply the result here without providing the proof. Consider now the function $P(x,1)$, specified by (11.3) for $|x| \leq 1$, $|y| \leq 1$ as

$$P(x,1) = \frac{h_2(x,1)P(x,0) + h_3(x,1)P(0,1) + h_4(x,1)P(0,0)}{h_1(x,1)} = \frac{\mu_2P(x,0) - \mu_1P(0,1) - \mu_2P(0,0)}{\lambda x - \mu_1}. \quad (11.32)$$

By Thm. 11.4.2 and (11.32), $P(x,1)$ is a meromorphic function in the complex $x$-plane cut along $[x_3, x_4]$. Note that the singularities of $P(x,1)$ consist of $\hat{x} = \mu_1/\lambda$ and the singularities of $P(x,0)$. $P(x,0)$ has at least one singularity given by the branch point $x_3$ (see Thm. 11.4.2).

Similarly, consider the function $P(1,y)$ specified by (11.3) for $|x| \leq 1$, $|y| \leq 1$ as

$$P(1,y) = \frac{h_2(1,y)P(1,0) + h_3(1,y)P(0,y) + h_4(1,y)P(0,0)}{h_1(1,y)} = \frac{-\mu_2(1+y)P(1,0) + \mu_1(1+y)P(0,y) + (\mu_2y - \mu_1)P(0,0)}{\mu_1 y - \mu_2}. \quad (11.33)$$

By Thm. 11.4.2 and (11.33), $P(1,y)$ is meromorphic function in the complex $y$-plane cut along $[y_3, y_4]$. The singularities of $P(1,y)$ consist of $\hat{y} = \mu_2/\mu_1$ and the singularities of $P(0,y)$ (one of which is the branch point $y_3$, see Thm. 11.4.2).

11.4.1 Singularity analysis for $P(x,0)$ and $P(x,1)$

We want to determine the dominant singularity of the function $P(x,0)$, i.e. the singularity of $P(x,0)$ with the smallest modulus. By Pringsheim’s theorem (see e.g. Flajolet & Sedgewick [73]) we know that the dominant singularity of a Taylor series with non-negative coefficients is real-valued. Since $P(x,0)$ is a Taylor series with non-negative coefficients, whose coefficients add up to a number smaller than one, we know that the dominant singularity of $P(x,0)$ should lie in the interval $(1, \infty)$. Since we know that $P(x,0)$ has a branch point in $x_3$, it remains to be determined whether $P(x,0)$ has a singularity in the interval $(1, x_3)$. This will be done by performing an analytic continuation of $P(x,0)$.

Starting point for this analytic continuation is the following relation that holds for zero-pairs $(x_+(y), y)$ of $h_1(x, y)$:

$$0 = h_2(x_+(y), y)P(x_+(y), 0) + h_3(x_+(y), y)P(0, y) + h_4(x_+(y), y)P(0,0). \quad (11.34)$$
Since \( P(0, y) \) is by Thm. 11.4.2 a meromorphic function in the complex \( y \)-plane cut along \([y_3, y_4]\), we can by (11.34) continue the function \( P(x, 0) \) analytically over the path \([y_2, y_3]\) using the function \( x_+(y) \). Note that such a continuation is not required for expanding the region of analyticity of \( P(x, 0) \), since this is already given by Thm. 11.4.2. The sole purpose of this continuation is to determine whether or not \( P(x, 0) \) has a singularity in \((1, x_3)\). First, we rewrite (11.34) as

\[
P(x_+(y), 0) = \frac{h_3(x_+(y), y)P(0, y) + h_4(x_+(y), y)P(0, 0)}{-h_2(x_+(y), y)}. \tag{11.35}
\]

Then, observe that \( h_2(x, y) = 0 \) implies \( h_3(x, y) = 0 \) and \( x = y^2 \). Since \( h_4(y^2, y) = (\mu_1 + \mu_2)y^2(y - 1) \), it follows from (11.35) that a pole of \( P(x, 0) \) in \((1, x_3)\) is a zero of \( h_2(x_+(y), y) \) with \( x_+(y) \in (1, x_3) \).

If \( x \) is a common root of \( h_3(x, y) = 0 \) and \( h_2(x, y) = h_3(x, y) = 0 \), then the point \((x, y)\) is a point of intersection of the curves \( h_1(x, y) = 0 \) and \( x = y^2 \). Substituting \( x = y^2 \) into \( h_1(x, y) \) yields

\[
h_1(y^2, y) = y^3 - \lambda y^5 - (\mu_1 + \mu_2)y^2 = y^3 - \lambda y^5 - (1 - \lambda)y^2 = y^3(1 - y)(\lambda y^2 + \lambda y + \lambda - 1). \tag{11.36}
\]

Hence, the intersection points of the curves \( h_1(x, y) = 0 \) and \( x = y^2 \) satisfy \( y^2(1 - y)(\lambda y^2 + \lambda y + \lambda - 1) = 0 \). The solutions \( y = 0 \) and \( y = 1 \) lead to trivial intersection points \((0, 0)\) and \((1, 1)\). The last term in (11.36) yields two additional solutions to \( h_1(y^2, y) = 0 \) given by

\[
y^* = \frac{-\lambda + \sqrt{\lambda(4 - 3\lambda)}}{2\lambda}, \quad \bar{y}^* = \frac{-\lambda - \sqrt{\lambda(4 - 3\lambda)}}{2\lambda}. \tag{11.37}
\]

When \( y \) varies from \( y_2 \) to \( y_3 \), the domain delimited by the functions \( x_+(y) \) and \( x_-(y) \) has the form indicated in Fig. 11.2. We denote the egg-shaped curve by \( \Omega \), and its interior by \( \Omega^+ \). Note that \( \Omega \cup \Omega^+ \) is convex and bounded, and the point \((1, 1)\) lies on \( \Omega \). Since the curve \( x = y^2 \) goes through the point \((1, 1)\), there are precisely two intersection points of the curves \( x = y^2 \) and \( \Omega \). The second intersection point can be shown to be \((x^*, y^*)\) with \( x^* = (y^*)^2 \),

\[
x^* = \frac{2(1 - \lambda)^2}{2\lambda - \lambda^2 + \lambda\sqrt{\lambda(4 - 3\lambda)}}. \tag{11.38}
\]

Note that \( y^* \) always lies in the interval \((1, y_3)\) and \( x^* \) always lies in the interval \((1, x_3)\).
Figure 11.2: The curve $\Omega$ for $\lambda = 3/17$, $\mu_1 = 6/17$, $\mu_2 = 8/17$. The intersection point $(x^*, y^*)$ is given by $(2.950, 1.717)$.

Figure 11.3: The analytic continuation of $P(x, 0)$ from $\beta$ to $x_3$ through the function $x_+(y)$, for $y$ increasing from $y_2$ to $\tilde{y}$ for $\lambda = 3/17$, $\mu_1 = 6/17$, $\mu_2 = 8/17$. 
We aim at determining the dominant singularity of \( P(x, 1) \). The dominant singularity can either be \( \hat{x} = \mu_1/\lambda \) or one of the singularities of \( P(x, 0) \) given by the branch point \( x_3 \) and the pole \( x^* \). We have some further knowledge of the branch point \( x_3 \). Evidently, \( x_3 \) belongs to a zero-pair of \( h_1(x, y) = 0 \), \((x_3, \hat{y})\) say, that lies on \( \Omega \), i.e. (see Fig. 11.2)

\[
\hat{y} = y_+(x_3) = y_-(x_3) .
\] (11.39)

Moreover, at \((x_3, \hat{y})\) we have

\[
\frac{d}{dy} h_1(x, y) = x - \lambda x^2 - 2\mu_1 y = 0 .
\] (11.40)

Substituting \( x - \lambda x^2 = 2\mu_1 y \) into \( h_1(x, y) = 0 \) yields \( x = (\mu_1/\mu_2)y^2 \). The point \((x_3, \hat{y})\) is thus an intersection point of the curves \( h_1(x, y) = 0 \) and \( x = (\mu_1/\mu_2)y^2 \), and so \( \hat{y} = h_2^{1/\lambda} (\mu_1/\mu_2)x_3^3 \).

There are two special points on \( \Omega \):

\[
\alpha = y_+(x_2) = y_-(x_2) ,
\]

\[
\beta = x_+(y_2) = x_-(y_2) ,
\]

which have been introduced before in Chapter 9, see (9.9) and (9.13). Since the model in this chapter is a special case of the model in Chapter 9, we know by Lemma 9.5.1 that \( P(x, 0) \) has no singularity in \([0, \beta]\) and \( P(0, y) \) has no singularity in \([0, \alpha]\). Hence, in order to determine the dominant singularity of \( P(x, 0) \), we need to investigate the function \( P(x, 0) \) on the interval \((\max\{\beta, 1\}, x_3]\) by continuing the function \( P(x, 0) \) analytically through \( x_+(y) \) over the path \( [y_2, y_3] \). This is plotted in Fig. 11.3 for \( \lambda = 3/17 \), \( \mu_1 = 6/17 \) and \( \mu_2 = 8/17 \). So we need to check whether or not we encounter a pole, i.e. an intersection point between the curves \( x = y^2 \) and \( x_+(y) \) for values of \( y \) increasing from \( y_2 \) to \( \hat{y} \). If there is no such intersection point, the function \( P(x, 0) \) has no singularity in \((1, x_3]\). This can be formulated as follows:

**Lemma 11.4.3** Consider the points \((x_3, \hat{y})\) and \((x^*, y^*)\) on \( \Omega \). It holds that \( x^* \in (1, x_3] \) and \( y^* \in (1, y_3] \), and \( x^* \) is a pole of the function \( P(x, 0) \) if \( y^* \in (1, \hat{y}] \).

Now use the fact that the point \((x^*, y^*)\) lies on the curve \( y = h_2^{1/\lambda} (\mu_2/\mu_1)x_3^2 \) and the point \((x_3, \hat{y})\) lies on the curve \( y = h_2^{1/\lambda} (\mu_2/\mu_1)x_3^2 \). If \( \mu_1 < \mu_2 \), the analytic continuation of \( P(x, 0) \) by the function \( x_+(y) \) over \([y_2, y_3]\) hits the point \((x^*, y^*)\) before \((x_3, \hat{y})\). The opposite holds if \( \mu_1 > \mu_2 \), and thus

**Lemma 11.4.4** If \( \mu_1 > \mu_2 \), the dominant singularity of the function \( P(x, 0) \) is the branch point \( x_3 \). If \( \mu_1 < \mu_2 \), the dominant singularity of the function \( P(x, 0) \) is the pole \( x^* \). If \( \mu_1 = \mu_2 \), the dominant singularity of the function \( P(x, 0) \) is \( x_3 = x^* \).

We now relate the singularities of \( P(x, 0) \), given by \( x_3 \) and \( x^* \), to \( \hat{x} \), because one of these three points is the dominant singularity of \( P(x, 1) \).

**Lemma 11.4.5** (i) If \( \mu_1 \geq \mu_2 \), the singularity \( \hat{x} \) of the function \( P(x, 1) \) is removable, and the dominant singularity of \( P(x, 1) \) is \( x_3 \). (ii) If \( \mu_1 < \mu_2 \), the singularity \( \hat{x} \) is a pole of the function \( P(x, 1) \), and \( \hat{x} \) is in fact the dominant singularity of \( P(x, 1) \).
Proof We can restrict ourself to the case that \( \mu_1 > \lambda \), since otherwise \( \hat{x} \leq 1 \). Let us then first show that there are some special points on \( \Omega \).

![Figure 11.4: The curve \( \Omega \) for \( \lambda = 3/17 \), \( \mu_1 = 6/17 \), \( \mu_2 = 8/17 \). \( \Omega \) always contains the points indicated in the picture.](image)

When one follows the trajectory indicated in Fig. 11.4 starting from \((1, 1)\), it can be readily seen that the following points (with \( \hat{y} = \mu_2/\mu_1 \)) lie on \( \Omega \):

\[
\begin{array}{c|c|c|c|c|c|c|c|c|c}
 x & 1 & 1 & \frac{\mu_2}{\lambda} & \frac{\mu_2}{\lambda} & \hat{x} & \hat{x} \\
y & 1 & \frac{\hat{y}}{y} & \frac{\hat{y}}{y} & \frac{\mu_2}{\lambda} & \frac{\mu_2}{\lambda} & 1 \\
\end{array}
\]

![Table 11.1: Points on \( \Omega \).](image)

For instance, the points \((\hat{x}, \mu_2/\lambda)\) and \((\hat{x}, 1)\) are connected in Fig. 11.4 because \( y_-(\hat{x}) \) and \( y_+(\hat{x}) \) take on the values \( \mu_2/\lambda \) and 1. Also, the points \((\hat{x}, 1)\) and \((1, 1)\) are connected because \( x_-(1) \) and \( x_+(1) \) take on the values \( \hat{x} \) and 1. Remember that for each point \((x, y)\) on \( \Omega \), if \( P(x, 0) \) and \( P(0, y) \) are finite, we have

\[
h_2(x, y)P(x, 0) + h_3(x, y)P(0, y) + h_4(x, y)P(0, 0) = 0. \tag{11.41}
\]

So the points \((\hat{x}, 1)\), \((1, 1)\) and \((1, \hat{y})\), provided that \( P(x, 0) \) and \( P(0, y) \) are analytic in the \( x \) and \( y \)-values, respectively, yield

\[
-h_2(\hat{x}, 1)P(\hat{x}, 0) = h_3(\hat{x}, 1)P(0, 1) + h_4(\hat{x}, 1)P(0, 0) \tag{11.42}
\]

\[
-h_2(1, 1)P(0, 1) = h_3(1, 1)P(1, 0) + h_4(1, 1)P(0, 0) \tag{11.43}
\]

\[
-h_2(1, \hat{y})P(1, 0) = h_3(1, \hat{y})P(0, \hat{y}) + h_4(1, \hat{y})P(0, 0). \tag{11.44}
\]
From (11.32) it can be easily seen that the \( \hat{x} \) is a removable singularity of \( P(x, 1) \) if (11.42) holds. Moreover, from (11.42)-(11.44) we see that (11.42) holds if and only if (11.44) holds.

Now take \( \mu_1 \geq \mu_2 \) (and \( \mu_1 > \lambda \)), in which case the dominant singularity of \( P(x, 0) \) is \( x_3 \). It obviously holds that \( \hat{x} < x_3 \). Hence, \( \hat{x} \) could be the dominant singularity of \( P(x, 1) \). In case \( \mu_1 \geq \mu_2 \), however, \( \hat{x} \) is a removable singularity, as shown next. If \( \mu_1 \geq \mu_2 \), it can be seen from Fig. 11.4 that \( \hat{y} \leq \alpha \). From this, we may conclude that the function \( P(0, y) \) is analytic in \( \hat{y} \), so (11.44) holds. We therefore conclude that \( P(x, 0) \) is analytic in \( \hat{x} = \mu_1/\lambda \) and (11.42) holds, in which case \( \hat{x} \) is a removable singularity of \( P(x, 1) \). The dominant singularity of \( P(x, 1) \) is then \( x_3 \), which proves Lemma 11.4.5(i).

Now assume \( \mu_1 < \mu_2 \). The dominant singularity of \( P(x, 0) \) is then \( x^* \), which belongs to the intersection point \((x^*, y^*)\) of \( \Omega \) and \( y = |\sqrt{x}| \). Now look at Fig. 11.4 and note that \( \mu_1 < \mu_2 \) implies that \( \mu_2 > \lambda \). It can be easily seen that the curve \( y = |\sqrt{x}| \) goes through the straight line connecting \((\hat{x}, \mu_2/\lambda)\) and \((\hat{x}, 1)\), and so we conclude that \( \hat{x} < x^* \). Hence, if \( \hat{x} \) is not removable, it is the dominant singularity of \( P(x, 1) \). In case \( \mu_1 < \mu_2, \hat{y} > \alpha \), and so \( P(0, y) \) is not analytic in \( \hat{y} \), (11.44) does not hold and \( \hat{x} \) is not removable. This proves Lemma 11.4.5(ii).

\section*{11.4.2 Singularity analysis for \( P(0, y) \) and \( P(1, y) \)}

We now determine, in a similar way as for \( P(x, 0) \), the dominant singularity of the function \( P(0, y) \). We know that the dominant singularity of \( P(0, y) \) lies in the interval \((1, \infty)\) and that \( y_3 \) is a branch point of \( P(0, y) \), so it remains to be determined whether \( P(0, y) \) has a singularity in \((1, y_3)\). This will be done by performing an analytic continuation of \( P(0, y) \). For zero-pairs \((x, y_+(x))\) of the kernel \( h_1(x, y) \) we have that:

\begin{equation}
0 = h_2(x, y_+(x))P(x, 0) + h_3(x, y_+(x))P(0, y_+(x)) + h_4(x, y_+(x))P(0, 0).
\end{equation}

The dominant singularity of \( P(1, y) \) can either be \( \hat{y} = \mu_2/\mu_1 \) or one of the singularities of \( P(0, y) \) given by \( y_3 \) and \( y^* \). For \( y_3 \) belonging to the boundary point \((\hat{x}, y_3)\) with

\begin{equation}
\hat{x} = x_+(y_3) = x_-(y_3),
\end{equation}

it holds that

\begin{equation}
\frac{d}{dx} h_1(x, y) = y - 2\lambda xy - \mu_2 = 0.
\end{equation}

Substituting \( y = 2\lambda xy + \mu_2 \) into \( h_1(x, y) = 0 \) yields that the point \((\hat{x}, y_3)\) is an intersection point of the curve \( y = (\lambda/\mu_1)x^2 \) and \( \Omega \) (so \( \hat{x} = |\sqrt{(\mu_1/\lambda)y_3}| \)). The pole \( y^* \) lies on the curve \( y = |\sqrt{x}| \), which gives the following results:

\begin{lemma}
Consider the points \((\hat{x}, y_3)\) and \((x^*, y^*)\) on \( \Omega \). It holds that \( x^* \in (1, x_3) \) and \( y^* \in (1, y_3) \), and \( y^* \) is a pole of the function \( P(0, y) \) if \( x^* \in (1, \hat{x}) \).
\end{lemma}

\begin{lemma}
If \((\lambda/\mu_1)(x^*)^2 > |\sqrt{x^*}| \), the dominant singularity of the function \( P(0, y) \) is the branch point \( y_3 \). If \((\lambda/\mu_1)(x^*)^2 < |\sqrt{x^*}| \), the dominant singularity of
the function $P(0, y)$ is the pole $y^*$. If $(\lambda/\mu_1)(x^*)^2 = |\sqrt{x^*}|$, the dominant singularity of the function $P(0, y)$ is $y_3 = y^*$.

Now we relate the two singularities of $P(0, y)$, $y_3$ and $y^*$, to $\hat{y} = \mu_2/\mu_1$, which gives:

**Lemma 11.4.8** (i) If $\mu_2 > \mu_1 > \lambda$, the singularity $\hat{y}$ is a pole and the dominant singularity of the function $P(1, y)$. If not, the singularity $\hat{y}$ of the function $P(1, y)$ is removable, and the dominant singularity of $P(1, y)$ is the dominant singularity of $P(0, y)$.

**Proof** We can restrict ourself to the case that $\mu_2 > \mu_1$, since otherwise $\hat{y} \leq 1$ and therefore $\hat{y}$ cannot be the dominant singularity of $P(1, y)$. In case $\mu_2 > \lambda \geq \mu_1$, the singularity $\hat{y}$ of $P(1, y)$ is removable, which can again be seen from (11.42)-(11.44). In case $\lambda \geq \mu_1$, it can be seen from Fig. 11.4 that $\hat{x} \leq \beta$. From this, we may conclude that the function $P(x, 0)$ is analytic in $\hat{x}$, and so by (11.42)-(11.44) the function $P(0, y)$ is analytic in $\hat{y}$, (11.44) holds, in which case we conclude from (11.33) that $\hat{y}$ is a removable singularity of $P(1, y)$. If $\mu_2 > \mu_1 > \lambda$, the singularity $\hat{y}$ cannot be removed, since $\hat{x} > \beta$. It that case, it obviously holds that $\hat{y}$ is the dominant singularity of $P(1, y)$. □

11.5 Tail behavior

By analyzing the behavior of the functions $P(x, 0)$ and $P(0, y)$ in the vicinity of their dominant singularities, we obtain asymptotic expressions for their coefficients $P(X_1 = n, X_2 = 0)$ and $P(X_1 = 0, X_2 = n)$ for large values of $n$. We apply the knowledge on the dominant singularities of $P(x, 0)$ along with (11.14) to obtain the following result:

**Theorem 11.5.1** The tail of the probability distribution of the number $X_1$ of customers in queue 1 while queue 2 is empty, is given asymptotically as

(a) If $\mu_1 < \mu_2$,

$$P(X_1 = n, X_2 = 0) \sim \kappa_1 \left( \frac{1}{x^*} \right)^n.$$  

(b) If $\mu_1 = \mu_2$, it holds that $x^* = x_3$, and

$$P(X_1 = n, X_2 = 0) \sim \kappa_2 \frac{1}{\sqrt{n}} \left( \frac{1}{x_3} \right)^n.$$  

(c) If $\mu_1 > \mu_2$,

$$P(X_1 = n, X_2 = 0) \sim \kappa_3 \frac{1}{n \sqrt{n}} \left( \frac{1}{x_3} \right)^n,$$

where $\kappa_1, \kappa_2, \kappa_3$ are constants.

Similarly, we have
11.5 Tail behavior

**Theorem 11.5.2** The tail of the probability distribution of the number $X_2$ of customers in queue 2 while queue 1 is empty, is given asymptotically as

(a) If $(\lambda/\mu_1)(x^*)^2 < |x^*|$, 

$$P(X_1 = 0, X_2 = n) \sim \kappa_4 \left(\frac{1}{y^*}\right)^n.$$ 

(b) If $(\lambda/\mu_1)(x^*)^2 = |x^*|$, it holds that $y^* = y_3$, and 

$$P(X_1 = 0, X_2 = n) \sim \kappa_5 \frac{1}{\sqrt{n}} \left(\frac{1}{y_3}\right)^n.$$ 

(c) If $(\lambda/\mu_1)(x^*)^2 > |x^*|$, 

$$P(X_1 = 0, X_2 = n) \sim \kappa_6 \frac{1}{\sqrt{n}} \left(\frac{1}{y_3}\right)^n,$$

where $\kappa_4, \kappa_5, \kappa_6$ are constants.

We now apply the knowledge on the dominant singularities of $P(x, 1)$ and $P(1, y)$ along with (11.14) to obtain asymptotic expressions for the coefficients of $P(x, 1)$ and $P(1, y)$. For the coefficients of $P(x, 1)$, we have the following result:

**Theorem 11.5.3** The tail of the probability distribution of the number $X_1$ of customers in queue 1 is given asymptotically as

(a) If $\mu_1 < \mu_2$, 

$$P(X_1 = n) \sim \kappa_7 \left(\frac{1}{x}\right)^n.$$ 

(b) If $\mu_1 = \mu_2$, it holds that $x^* = x_3$, and 

$$P(X_1 = n) \sim \kappa_8 \frac{1}{\sqrt{n}} \left(\frac{1}{x_3}\right)^n.$$ 

(c) If $\mu_1 > \mu_2$, 

$$P(X_1 = n) \sim \kappa_9 \frac{1}{\sqrt{n}} \left(\frac{1}{x_3}\right)^n,$$

where $\kappa_7, \kappa_8, \kappa_9$ are constants.

Similarly, for the coefficients of $P(1, y)$ we find:

**Theorem 11.5.4** The tail of the probability distribution of the number $X_2$ of customers in queue 2 is given asymptotically as

(a) If $(\lambda/\mu_1)(x^*)^2 < |x^*|$ and it does not hold that $\mu_2 > \mu_1 > \lambda$, 

$$P(X_2 = n) \sim \kappa_{10} \left(\frac{1}{y}\right)^n.$$
(b) If \((\lambda / \mu_1)x^* \geq |\sqrt{x^*}|\) and it does not hold that \(\mu_2 > \mu_1 > \lambda\), \(y^* = y_3\), and
\[
P(X_2 = n) \sim \kappa_{11} \frac{1}{\sqrt{n}} \left( \frac{1}{y_3} \right)^n.
\]

(c) If \((\lambda / \mu_1)x^* > |\sqrt{x^*}|\) and it does not hold that \(\mu_2 > \mu_1 > \lambda\),
\[
P(X_2 = n) \sim \kappa_{12} \frac{1}{n \sqrt{n}} \left( \frac{1}{y_3} \right)^n.
\]

(d) If \(\mu_2 > \mu_1 > \lambda\),
\[
P(X_2 = n) \sim \kappa_{13} \left( \frac{1}{y} \right)^n,
\]
where \(\kappa_{10}, \kappa_{11}, \kappa_{12}, \kappa_{13}\) are constants.

### 11.6 Alternative approaches

We mentioned in the introduction that there are some alternative methods for deriving tail asymptotics: A method developed by Foley & McDonald [75] that uses large deviations techniques, and a method developed by Takahashi et al. [151] that considers the model of interest as a quasi-birth-and-death model. We discuss both methods and point out the differences with our analytic approach.

We first describe the method of Foley & McDonald. In Subsec. 11.6.1 we introduce the method by applying it to the priority case, obtaining similar results as in Sec. 11.3. Moreover, the method of Foley & McDonald explains why a certain type of asymptotic behavior occurs. Then, in Subsecs. 11.6.2 and 11.6.3, we apply the method of Foley & McDonald to the tandem queue with coupled processors. As it turns out, some of the results of Sec. 11.5 can be derived, while others cannot. This does not mean that the latter results cannot be proved using large deviations techniques, but they cannot be proved using the standardized approach as presented in Foley & McDonald [75]. Finally, in Subsec. 11.6.4, we discuss the method of Takahashi et al. [151]. This method can only be used to derive part of the results presented in Sec. 11.5.

#### 11.6.1 Priority case

Since the method developed by Foley & McDonald [75] is rather involved, we introduce the method by applying it to the relatively simple case in which station 1 has preemptive priority over station 2 (covered in Sec. 11.3). We will describe the key steps of the method and refer for the details to [75].

The first step is to uniformize the continuous-time Markov process, so that we obtain a discrete-time Markov chain. Then, the discrete-time Markov chain is exponentially twisted as indicated in Fig. 11.5.
Now consider the event that the queue at station 2 gets large. Foley & McDonald distinguish three types of most probable paths of how this event may occur: A jitter path, a bridge path and a cascade path. A jitter path follows the horizontal axis and spends a non-zero proportion of time on the axis. A bridge path travels along the horizontal axis, but instead of jittering along the horizontal axis, it skims above and rarely touches the axis. A cascade path first goes up the vertical axis, and then moves into the horizontal direction while it drops down in the vertical direction at the same time.

Let us return to Fig. 11.5. A jittering path would imply that queue 2 gets larger while queue 1 is more or less empty, where a cascade path would imply that queue 1 builds up until a certain point, after which the customers in queue 1 are dumped into queue 2. The bridge path is essentially something in between these scenarios.

In the standardized approach presented by Foley & McDonald [75], it is assumed that a cascade path does not occur. Further, they introduce the following functions:

**Definition 11.6.1** The functions $R^+$ and $R^-$ are defined as

$$R^+(\theta_1, \theta_2) = \sum_{n', m'} r_{(n,m)-(n',m')} e^{\theta_1(n'-n)} e^{\theta_2(m'-m)},$$

$$R^-(\theta_1, \theta_2) = \sum_{n', m'} r_{(n,0)-(n',m')} e^{\theta_1(n'-n)} e^{\theta_2 m'},$$

where $r_{(n,m)-(n',m')}$ is the transition rate of going from state $(n, m)$ to $(n', m')$ in the original uniformized system and $m \neq 0$.

Let us first investigate the possibility that the most probable path is a bridge path.

**Definition 11.6.2** Let $(\theta_1^b, \theta_2^b)$, $\theta_1^b > 0$ be the solution to

$$R^+(\theta_1^b, \theta_2^b) = 1,$$

$$\frac{dR^+(\theta_1^b, \theta_2^b)}{d\theta_2^b} = 0.$$

*Figure 11.5:* Phase diagram for the queue lengths in case station 1 has preemptive priority over station 2: Normal and twisted version.
Foley & McDonald [75] prove:

**Condition 11.6.3 (bridge)** The bridge path occurs when the following two conditions are satisfied:

1. \( \sum_{n=0}^{\infty} \pi(n,0)e^{\theta_2 n} < \infty, \)
2. \( R^-(\theta_1, \theta_2) \leq R^+(\theta_1, \theta_2). \)

Condition 11.6.3(i) guarantees that no cascade path ever occurs. In the priority case we have that

\[
R^+(\theta_1, \theta_2) = \mu_2 + \lambda e^{\theta_2} + \mu_1 e^{\theta_1 - \theta_2},
\]
\[
R^-(\theta_1, \theta_2) = \mu_1 + \lambda e^{\theta_2} + \mu_2 e^{-\theta_1},
\]

and so from (11.50) and (11.51) it follows that

\[
e^{-\theta_1^0} = \frac{4\rho_1}{(1 + \rho_1)^2}, \quad e^{-\theta_2^0} = \frac{2\rho_1}{1 + \rho_1}.
\]  

Now, Condition 11.6.3(i) can be shown to hold in the following way. Since the marginal distribution of \( X_1 \) is just the stationary queue length distribution of an \( M/M/1 \) queue with load \( \rho_1 \), we have that \( \pi(n,0) \leq \sum_{k=0}^{\infty} \pi(n,k) = (1 - \rho_1)\rho_1^n \), and so

\[
\sum_{n=0}^{\infty} \pi(n,0)e^{\theta_2 n} \leq \sum_{n=0}^{\infty} (1 - \rho_1)\rho_1^n e^{\theta_2 n}
\]
\[
= (1 - \rho_1) \sum_{n=0}^{\infty} \left( \frac{\rho_1(1 + \rho_1)}{2\rho_1} \right)^n \]
\[
= 2 < \infty.
\]  

Condition 11.6.3(ii) is equivalent with

\[
\rho_2 \leq \frac{2\rho_1(1 - \rho_1)}{(1 + \rho_1)^2} =: \rho_c.
\]  

Then by Theorem 5 in Foley & McDonald [75] we have the following result: If Condition 11.6.3(i) is satisfied and Condition 11.6.3(ii) holds with strict inequality (\( \rho_2 < \rho_c \)), we have that

\[
P(X_1 = m, X_2 = n) \sim \zeta_1 \phi_1(m) \frac{1}{2\sqrt{\pi n^3}} e^{-\theta_1^0 n},
\]

with \( \zeta_1 \) some constant and \( \phi_1(m) \) some function of \( m \). If Condition 11.6.3(i) is satisfied and Condition 11.6.3(ii) holds with strict equality (\( \rho_2 = \rho_c \)), then

\[
P(X_1 = m, X_2 = n) \sim \zeta_2 \phi_2(m) \frac{1}{\sqrt{2\pi n}} e^{-\theta_2^0 n},
\]

with \( \zeta_2 \) some constant and \( \phi_2(m) \) some function of \( m \).

Let us now investigate the possibility that the most probable path is a jitter path.
Definition 11.6.4 Let \( (\theta^{j}_1, \theta^{j}_2), \theta^{j}_1 > 0 \) be the solution to
\[
R^{+}(\theta^{j}_1, \theta^{j}_2) = R^{-}(\theta^{j}_1, \theta^{j}_2) = 1. \tag{11.57}
\]

Foley & McDonald [75] prove that (Theorem 5):

Condition 11.6.5 (jitter) The jitter path occurs when the following two conditions are satisfied:

(i) \( \sum_{n=0}^{\infty} \pi(n,0)e^{\theta^{j}_2n} < \infty \),

(ii) \( R^{-}(\theta^{b}_1, \theta^{b}_2) > R^{+}(\theta^{b}_1, \theta^{b}_2) \).

In the priority case Condition 11.6.5(ii) is equivalent with \( \rho^{2} > \rho^{c} \). Furthermore, Foley & McDonald show that when Condition 11.6.5(ii) holds, \( \theta^{j}_2 < \theta^{b}_2 \), and so Condition 11.6.5(i) holds since Condition 11.6.3(i) holds. Furthermore, it follows that
\[
e^{-\theta^{j}_1} = \frac{\rho^{2} - \rho^{1} - \rho^{1}\rho^{2} + \sqrt{4\rho^{1}\rho^{2}^{2} + (\rho^{2} - \rho^{1} - \rho^{1}\rho^{2})^{2}}}{2\rho^{2}},
\]
\[
e^{-\theta^{j}_2} = \frac{\rho^{2} + \rho^{1} + \rho^{1}\rho^{2} - \sqrt{4\rho^{1}\rho^{2}^{2} + (\rho^{2} - \rho^{1} - \rho^{1}\rho^{2})^{2}}}{2\rho^{1}\rho^{2}}.
\]

Then we have by Theorem 5 in Foley & McDonald [75] that if Condition 11.6.5(i) and Condition 11.6.5(ii) hold \( (\rho^{2} > \rho^{c}) \) then
\[
\mathbb{P}(X_1 = m, X_2 = n) \sim \zeta_{3}\phi_{3}(m)e^{-\theta^{j}_1n}, \tag{11.58}
\]
with \( \zeta_{3} \) some constant and \( \phi_{3}(m) \) some function of \( m \). Now, since \( e^{-\theta^{j}_1} = 1/y_B \) and \( e^{-\theta^{j}_2} = 1/y_P \), we see that (11.55), (11.56) and (11.58) give similar tail asymptotics as in Thm. 11.3.5. However, there are two crucial differences. First, (11.55), (11.56) and (11.58) are results for \( \mathbb{P}(X_1 = m, X_2 = n) \), while Thm. 11.3.5 concerns \( \mathbb{P}(X_2 = n) \). Unfortunately, the theory in Foley & McDonald [75] does not lead to asymptotics of \( \mathbb{P}(X_2 = n) \), since it is not clear what the asymptotic behavior of \( \mathbb{P}(X_1 = m, X_2 = n) \) is for \( m \to \infty \). Second, the constants in Thm. 11.3.5 are known, while these cannot be determined with the theory of Foley & McDonald.

Remark 11.6.6 We mentioned earlier that the theory of Foley & McDonald can provide some intuition that can be helpful in understanding why the asymptotic behavior takes on different shapes in different situations. Now, the outcome of the analysis is that we have either a bridge path or a jitter path. The bridge path corresponds to the case in Thm. 11.3.5 where the dominant singularity of \( P(1,y) \) is a singularity of \( \xi(y) \), the pgf of the number of customers served in a busy period of an \( M/M/1 \) queue with arrival rate \( \lambda \) and service rate \( \mu_1 \). One can intuitively understand how a large busy period at queue 1 causes queue 2 to get large, but how does this correspond to a bridge path? Take another look at Fig. 11.5, and assume
that the horizontal and vertical axis represent the number of jobs served and the number of jobs in the queue, respectively, both during one busy period of an \( M/M/1 \) queue with arrival rate \( \lambda \) and service rate \( \mu_1 \). Then, a busy period is represented by the process moving from the vertical axis into the interior of the state space and at one point hitting the horizontal axis (the end of the busy period). So, by definition, this process skims above and does not touch the horizontal axis, which is the exact description of a bridge path. The jitter path corresponds to the case in Thm. 11.3.5 where the dominant singularity of \( P(1, y) \) is a singularity of \( P(0, y) \), which suggests that queue 2 gets large while queue 1 stays empty. Remember that this indeed is the description of a jitter path.

### 11.6.2 Coupled processors: Queue 1

We now apply the method of Foley & McDonald to the coupled-processors case, to derive tail asymptotics for the queue length distribution at station 1. Note that in this case the continuous-time Markov chain is already uniformized, due to the work-conserving property and the assumption \( \lambda + \mu_1 + \mu_2 = 1 \). The exponential twisting is then displayed in Fig. 11.6.

![Figure 11.6: Phase diagram for the coupled-processors case with the stationary queue length at station 1 at the horizontal axis: Normal and twisted version.](image)

The functions \( R^+ \) and \( R^- \) are given by

\[
R^+(\theta_1, \theta_2) = \lambda e^{\theta_1} + \mu_1 e^{\theta_2 - \theta_1} + \mu_2 e^{-\theta_2},
\]

\[
R^-(\theta_1, \theta_2) = \lambda e^{\theta_1} + (\mu_1 + \mu_2) e^{\theta_2 - \theta_1}.
\]

Further, the solution \((\theta_1^b, \theta_2^b), \theta_1^b > 0\) to (11.50) and (11.51) is implicitly defined as

\[
e^{-\theta_1^b} = \frac{\mu_2}{\mu_1} (e^{-\theta_2^b})^2. \tag{11.59}\]

Explicit expressions for \( e^{-\theta_1^b} \) and \( e^{-\theta_2^b} \) can be found, but these are not presented here because of their lengthiness.
11.6 Alternative approaches

**Condition 11.6.7 (bridge)** The bridge path occurs when the following two conditions are satisfied:

(i) \( \sum_{n=0}^{\infty} \pi(0,n)e^{\theta_2 n} < \infty \),

(ii) \( R^-(\theta_1^b, \theta_2^b) \leq R^+(\theta_1^b, \theta_2^b) \).

Note that Condition 11.6.7(i) is in terms of \( \pi(0,n) \), whereas Condition 11.6.3(i) is in terms of \( \pi(n,0) \). This is because in the current subsection, the queue length at station 1 is plotted on the horizontal axis. Condition 11.6.7(ii) is equivalent with \( \mu_1 \geq \mu_2 \).

Then by Theorem 5 in Foley & McDonald [75] we have: If Condition 11.6.7(i) is satisfied and \( \mu_1 > \mu_2 \) then

\[
P(X_1 = n, X_2 = m) \sim \zeta_4 \phi_4(m) \frac{1}{\sqrt{n}} e^{-\theta_1^b n},
\]

with \( \zeta_4 \) some constant and \( \phi_4(m) \) some function of \( m \). If Condition 11.6.7(i) is satisfied and \( \mu_1 = \mu_2 \), then

\[
P(X_1 = n, X_2 = m) \sim \zeta_5 \phi_5(m) \frac{1}{\sqrt{n}} e^{-\theta_1^b n},
\]

with \( \zeta_5 \) some constant and \( \phi_5(m) \) some function of \( m \).

The conditions for a jitter path read:

**Condition 11.6.8 (jitter)** The jitter path occurs when the following two conditions are satisfied:

(i) \( \sum_{n=0}^{\infty} \pi(0,n)e^{\theta_2 n} < \infty \),

(ii) \( R^-(\theta_1^j, \theta_2^j) > R^+(\theta_1^j, \theta_2^j) \).

Condition 11.6.8(ii) is equivalent with \( \mu_1 < \mu_2 \). With \( (\theta_1^j, \theta_2^j), \theta_1^j > 0 \) the solution to (11.57), it follows that

\[
e^{-\theta_1^j} = \frac{2\lambda - \lambda^2 + \lambda \sqrt{\lambda(4 - 3\lambda)}}{2(1 - \lambda)^2},
\]

\[
e^{-\theta_2^j} = \frac{\lambda + \sqrt{\lambda(4 - 3\lambda)}}{2(1 - \lambda)}.
\]

Then we have by Theorem 5 in Foley & McDonald [75] that if Condition 11.6.8(i) is satisfied and \( \mu_1 < \mu_2 \) then

\[
P(X_1 = n, X_2 = m) \sim \zeta_6 \phi_6(m)e^{-\theta_1^j n},
\]

with \( \zeta_6 \) some constant and \( \phi_6(m) \) some function of \( m \).
So, the remaining problem is how to check Conditions 11.6.7(i) and 11.6.8(i)? This is far from straightforward. In [74], Foley & McDonald have presented an approach for checking these conditions for a modified Jackson network. This approach is highly problem-dependent and is based on large deviations and specific optimization techniques. At this point, we shall not give an analysis for checking 11.6.7(i) and 11.6.8(i). Instead, we compare the results (11.61)-(11.63) for $m = 0$ with those in Thm. 11.5.1.

First note that $e^{-\theta_1} = 1/x^*$. Hence, in case $\mu_1 < \mu_2$, Thm. 11.5.1(a) corresponds to the jitter case (11.63), which suggests that for $\mu_1 < \mu_2$ Condition 11.6.8(i) is satisfied. It can be shown that $e^{-\theta_1}$ equals $1/x_3$. Hence, in case $\mu_1 = \mu_2$, Thm. 11.5.1(b) corresponds to the case (11.62), which suggests that for $\mu_1 = \mu_2$ Condition 11.6.7(i) is satisfied. For $\mu_1 > \mu_2$, Thm. 11.5.1(c) obviously corresponds to (11.61), which suggests that for $\mu_1 > \mu_2$ Condition 11.6.7(i) is satisfied. This brings us to the following conjecture:

**Conjecture 11.6.9** Condition 11.6.7(i) holds if $\mu_1 \geq \mu_2$. Condition 11.6.8(i) holds if $\mu_1 < \mu_2$.

### 11.6.3 Coupled processors: Queue 2

We now apply the method of Foley & McDonald to the coupled-processors case to derive tail asymptotics for the queue length distribution at station 2. The twisted uniformized Markov chain is displayed in Fig. 11.7.

![Phase diagram for the coupled-processors case with the stationary queue length at station 2 at the horizontal axis: Normal and twisted version.](image)

We then get that

$$R^+(\theta_1, \theta_2) = \lambda e^{\theta_2} + \mu_1 e^{\theta_1 - \theta_2} + \mu_2 e^{-\theta_1},$$
$$R^-(\theta_1, \theta_2) = \lambda e^{\theta_2} + (\mu_1 + \mu_2) e^{-\theta_1}.$$

For $(\theta_1^b, \theta_2^b), \theta_1^b > 0$ the solution to (11.50) and (11.51), the following relation holds

$$e^{-\theta_1^b} = \frac{\mu_1}{\lambda} (e^{-\theta_2^b})^2.$$  \hspace{1cm} (11.64)
Again, the bridge case holds when Condition 11.6.3(ii) is equivalent with
\[(e^{-\theta_1})^2 \leq e^{-\theta_2} \iff \left(\frac{\mu_1}{\lambda}\right)^2 (e^{-\theta_2})^3 \leq 1. \tag{11.65}\]

Then by Theorem 5 in Foley & McDonald [75] we have: If Condition 11.6.3(i) holds and \((e^{-\theta_1})^2 < e^{-\theta_2}\), then
\[P(X_1 = m, X_2 = n) \sim \zeta_7 \phi_7(m) \frac{1}{\sqrt{n^3}} e^{-\theta_1 n}, \tag{11.66}\]
with \(\zeta_7\) some constant and \(\phi_7(m)\) some function of \(m\). If Condition 11.6.3(i) holds and \((e^{-\theta_1})^2 = e^{-\theta_2}\), then
\[P(X_1 = m, X_2 = n) \sim \zeta_8 \phi_8(m) \frac{1}{\sqrt{n}} e^{-\theta_1 n}, \tag{11.67}\]
with \(\zeta_8\) some constant and \(\phi_8(m)\) some function of \(m\).

Condition 11.6.5(ii) is equivalent with \((e^{-\theta_1})^2 > e^{-\theta_2}\). The solution \((\theta_1^j, \theta_2^j)\) of (11.57) can be shown to be
\[e^{-\theta_1^j} = \frac{\lambda + \sqrt{\lambda(4 - 3\lambda)}}{2(1 - \lambda)}, \]
\[e^{-\theta_2^j} = \frac{2\lambda^2 + \lambda \sqrt{\lambda(4 - 3\lambda)}}{2(1 - \lambda)^2}. \]

Then we have by Theorem 5 in Foley & McDonald [75] that if Condition 11.6.5(i) holds and \((e^{-\theta_1})^2 > e^{-\theta_2}\), then
\[P(X_1 = m, X_2 = n) \sim \zeta_9 \phi_9(m) e^{-\theta_1 n}, \tag{11.68}\]
with \(\zeta_9\) some constant and \(\phi_9(m)\) some function of \(m\).

Again we have the problem that we cannot check Conditions 11.6.3(i) and 11.6.5(i) in a straightforward way. Let us compare the results (11.66)-(11.68) for \(m = 0\) with those in Thm. 11.5.2. First note that \(e^{-\theta_1^j} = 1/y_3\) and \(e^{-\theta_2^j} = 1/y^*\). Also, using (11.64), it can be seen that
\[\left(\frac{\mu_1}{\lambda}\right)^2 (e^{-\theta_2})^3 \leq 1 \iff (\lambda/\mu_1)(x^*)^2 \geq |\sqrt{x^*}|, \tag{11.69}\]
which leads us to the following conjecture

**Conjecture 11.6.10** Condition 11.6.3(i) holds if \((\lambda/\mu_1)(x^*)^2 \geq |\sqrt{x^*}|\). Condition 11.6.5(i) holds if \((\lambda/\mu_1)(x^*)^2 < |\sqrt{x^*}|\).
11.6.4 QBD method

We now discuss the method developed by Takahashi et al. [151]. Consider a Markov process on the two-dimensional state space \( \{(n, k) | n \geq 0, 0 \leq k \leq H\} \). The first component is referred to as *level*, the second as *phase*. So we refer by level \( n \) to the set of states \( \{(n, 0), (n, 1), \ldots, (n, H)\} \). Such a Markov process is called a homogeneous quasi-birth-and-death (QBD) process when one-step transitions are restricted to states in the same level or in the two adjacent levels, and the transition rates are assumed to be level-independent.

Now order the states as

\[
\{(0, 0), \ldots, (0, H), (1, 0), \ldots, (1, H), \ldots, (n, 0), \ldots, (n, H)\},
\]

and assume that the infinitesimal generator \( Q \) has the following block tridiagonal structure:

\[
Q = \begin{pmatrix}
B_1 & B_0 & 0 & 0 & 0 & \ldots \\
B_2 & A_1 & A_0 & 0 & 0 & \ldots \\
0 & A_2 & A_1 & A_0 & 0 & \ldots \\
0 & 0 & A_2 & A_1 & A_0 & \ldots \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots \\
\end{pmatrix},
\]

(11.70)

where \( 0, A_0, A_1 \) and \( A_2 \) are square matrices of order \( H + 1 \). The matrix \( 0 \) has all zero entries. The matrices \( A_0, A_2, B_0 \) and \( B_2 \) are nonnegative and the matrices \( B_1 \) and \( A_1 \) have nonnegative off-diagonal elements and strictly negative diagonals.

The QBD process driven by \( Q \) is ergodic if and only if it satisfies the mean drift condition (see Neuts [124])

\[
w A_0 e < w A_2 e,
\]

(11.71)

where \( w = (w_0, \ldots, w_H) \) is the equilibrium distribution of the generator \( A_0 + A_1 + A_2 \) and \( e \) the unity vector. When (11.71) is satisfied, the stationary distribution of the QBD process exists. Denoting by \( \pi(n, k) \) the stationary probability of the process being in state \( (n, k) \), and using the vector notation \( \pi_n = (\pi(n, 0), \ldots, \pi(n, H)) \), the balance equations of the QBD process are given by

\[
\pi_{n-1} A_0 + \pi_n A_1 + \pi_{n+1} A_2 = 0, \quad n \geq 2,
\]

(11.72)

and

\[
\pi_0 B_1 + \pi_1 B_2 = 0,
\]

(11.73)

\[
\pi_0 B_0 + \pi_1 A_1 + \pi_2 A_2 = 0.
\]

(11.74)

Introducing the rate matrix \( R \) as the minimal nonnegative solution of the nonlinear matrix equation

\[
A_0 + R A_1 + R^2 A_2 = 0,
\]

(11.75)

it can be proved that the equilibrium probabilities satisfy (see e.g. Neuts [124])

\[
\pi_{n+1} = \pi_n R, \quad n \geq 1.
\]

(11.76)
It is known that positive-recurrent QBD processes with finite phase-space have a stationary distribution that decays geometrically in the level. The decay parameter then equals the spectral radius of $R$. For a comprehensive treatment of QBD processes with a finite phase-space we refer to Neuts [124] and Latouche & Ramaswami [109].

The two-stage tandem queue covered in this chapter is a QBD processes with infinite phase-space ($H = \infty$). For instance, if $(n, k)$ denotes the state that in equilibrium queue 1 is of length $n$ and queue 2 is of length $k$, the coupled-processors case is described by

$$A_0 = B_0 = \begin{pmatrix} \lambda^2 & \lambda & \ddots \\ \lambda & \lambda^2 & \ddots \\ & \ddots & \ddots & \ddots \end{pmatrix}, \quad A_1 = \begin{pmatrix} \Delta & \Delta & \ddots \\ \mu_2 & \Delta & \ddots \\ & \ddots & \ddots & \ddots \end{pmatrix},$$

and

$$A_2 = B_2 = \begin{pmatrix} 0 & \mu_1 + \mu_2 & \mu_1 & \ddots \\ \mu_1 & 0 & \mu_1 & \ddots \\ \mu_2 & \mu_1 & 0 & \ddots \\ \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \quad B_1 = \begin{pmatrix} \Delta & \Delta & \ddots \\ \mu_1 + \mu_2 & \Delta & \ddots \\ & \ddots & \ddots & \ddots \end{pmatrix}.$$

Here, we have denoted the diagonal elements of $B_1$ and $A_1$ by $\Delta$, which are such that the row sums of $Q$ equal zero.

Determining the decay rate for QBD processes with infinite phase-space is less straightforward than in the finite case, since the $R$-matrix is infinite-dimensional, and its spectral properties are not obvious. Takahashi et al. [151] presented conditions on QBD processes with infinite phase-space such that the stationary distributions of these processes decay geometrically in the level. This geometric tail behavior in the level direction can be described as

$$\pi(n, k) \sim c_1 \eta^n b_k, \quad n \to \infty,$$  \hspace{1cm} (11.77)

where $c$ is a constant, $b_k$ is a constant that depends on the phase, and $\eta$ is the decay rate. In this chapter, we are not interested in the $b_k$, and so we use instead of (11.77) the formulation

$$\mathbb{P}(X_1 = n) = \sum_{k=0}^{\infty} \pi(n, k) \sim \bar{c}_1 \eta^n, \quad n \to \infty,$$  \hspace{1cm} (11.78)

where $\bar{c} = c \sum_{k=0}^{\infty} b_k$ is a constant.

A rigorous treatment of the conditions under which there is a decay of the above type has been presented in Haque et al. [84]. They provide a computational framework for verifying the conditions under which the QBD process has a geometric tail in the level direction. Using this framework, we can derive the following result in the coupled-processors case:
If $\mu_1 < \mu_2$, we have that
\[ P(X_1 = n) \sim \bar{c}n^n, \quad n \to \infty, \quad (11.79) \]
where
\[ \eta = \frac{(2\lambda - \lambda^2) + \lambda \sqrt{\lambda(4 - 3\lambda)}}{2(1 - \lambda)^2}. \quad (11.80) \]

Now note that $\eta = 1/x^* = 1/(y^*)^2$ (see (11.37)), and so we obtain the exact same result as in Thm. 11.5.3(a).

Unfortunately, the QBD method cannot be used to prove Lemma 11.5.3(b). That is, if the conditions in Haque et al. [84] are satisfied, there is a geometric tail behavior as in (11.77), but if these conditions are not satisfied, not much is known about the tail behavior. This is a major drawback of the QBD method. In the terminology of Foley & McDonald, one could say that the QBD method is suited only for determining jitter paths. Indeed, with the QBD method, we are able to prove next to Thm. 11.5.3(a), Thms. 11.3.5(c) and 11.5.4(a). All three correspond to jitter paths.

As the results presented in this chapter indicate, there are many regimes in which the conditions for the geometric tail behavior are not satisfied. In that respect, see also Kroese et al. [107], who give a detailed treatment of a two-station Jackson network and show that it might exhibit some non-trivial tail behavior that cannot be captured by the QBD method.

### 11.7 Conclusions and further research

For the two-station tandem queue with either preemptive priority for station 1 or coupled processors, we have derived asymptotic expressions for the stationary queue length distribution at each of the stations. For the preemptive-priority case, we have derived the asymptotic expressions by investigating the functions $P(x, 1)$ and $P(1, y)$ in the vicinity of their dominant singularities. For the coupled-processors case, determining the dominant singularities required an analytic continuation of the functions $P(x, 0)$ and $P(0, y)$.

There are two alternative methods for deriving tail asymptotics in the above model: The method of Foley & McDonald [75] and the QBD method. The QBD method is only suited for determining the conditions under which a specific type of geometric tail behavior occurs. The method of Foley & McDonald is based on large deviations techniques. Using their standardized approach, the results presented in this chapter can only partially be obtained. It is possible, though, to derive all the results using large deviations techniques, but this requires a problem-dependent analysis which is far from straightforward. Foley & McDonald have presented such an analysis for a modified Jackson network in [74]. A similar analysis that would lead to the proof of Conjectures 11.6.9 and 11.6.10 is a topic for future research.

It should be stressed that both the method of Foley & McDonald and the QBD method can be applied to higher-dimensional models, while the approach presented...
in this chapter is a strictly two-dimensional one. For two-dimensional models in a more general context, a further comparison between our approach, the method of Foley & McDonald and the QBD method is an interesting topic for further research.

In this chapter, all asymptotic expressions contain some multiplicative constant. Where both the method of Foley & McDonald and the QBD method do not lead to explicit expressions for these multiplicative constants, our approach does. Since we essentially evaluate a function in the vicinity of its dominant singularity, the constant is just the residue of this function at its dominant singularity. For the priority case in Sec. 11.3, we considered the function \( P(1, y) \) and determined its dominant singularity, \( y_D \) say. Since the function \( P(1, y) \) was fully known, the residue \( \lim_{y \to y_D} P(1, y) \) (and so the multiplicative constant in the asymptotic expression, see Thm. 11.3.5) could be determined explicitly.

For the coupled-processors case, things get more complicated. Again consider the function \( P(1, y) \) and assume that its dominant singularity is \( y_D \). Now, for determining the residue \( \lim_{y \to y_D} P(1, y) \), we need to have an explicit expression for \( P(1, y) \). Remember that we have presented an expression for \( P(1, y) \) in Chapter 9 by providing an expression for \( P(0, y) \) (which in turn yields the expression for \( (P(1, y)) \) as the solution of a Riemann-Hilbert boundary value problem. This solution is valid on and within a certain contour, which we have denoted in Chapter 9 by \( L \). Now, if \( y_D \) lies on or within this contour, we can evaluate the function \( P(0, y) \) in \( y = y_D \), and so we can provide an explicit expression for \( \lim_{y \to y_D} P(1, y) \) and thus for the multiplicative constant.

If \( y_D \) lies outside \( L \), the result in Chapter 9 does not provide us with the means to determine \( \lim_{y \to y_D} P(1, y) \). In this case, we need to obtain a solution for \( P(0, y) \) that is valid not only for values of \( y \) on and inside \( L \), but also for the region that contains \( y_D \). For two independent \( M/M/1 \) queues with coupled processors, Guillemin & Pinchon [82] have shown how this can be done. They also solve a boundary value problem, and provide an analytic continuation of the solution to the boundary value problem such that the residue of the function of interest at its dominant singularity can be determined explicitly. We believe that the same approach can be chosen for the models in Chapters 9-11, and this is a challenging topic for further research.
Bibliography

Self-references


---

**References**


of multidimensional queues. In Advances in Queueing. Theory, Methods and


Probability Letters 31: 169-175.

Monthly 56: 325-327.

[149] Syski, R. (1967). Pollaczek’s methods in queueing theory. In Queueing The-
tory, Recent Developments and Applications (ed. R. Cruon), Engl. Univ. Press,
Londen, 33-60.

der Berliner Math. Gesellschaft 22: 50-64; also in Collected Papers, Vol. 1:
645-662, Birkhäuser, Boston.

steady-state probabilities in a quasi-birth-and-death process with a countable


Sons, New York.

versity Press, New York.


sis, Eindhoven University of Technology.


Index

Abel’s theorem, 27, 85
adaptive scheduling, 161
ALOHA, 8
argument principle, 34
bidirectional data transfer, 1
busy period $M/M/1$ queue, 208
cable access network, 1
Cauchy’s theorem, 49, 78
complementary error function, 90, 97
conformal mapping, 178
contention tree, 8
corresponding-boundaries theorem, 180
coupled processors, 11
Cramer’s rule, 39
cyclic behavior, 155, 160
delayed flexible boundary model, 9
Dirichlet problem, 177, 201
discrete $D/G/1$ queue, 84
discrete Pareto distribution, 38
discrete standard lognormal distribution, 38
distinctness of roots, 45
downstream channel, 2
elementary symmetric function, 23
fixed boundary model, 7
fixed-cycle traffic light queue, 152, 157
forced request slots, 8
Fourier sampling, 46
full coupling, 11
functional equation, 172
Gamma function, 51, 209
generalized principle of the argument, 34
generalized processor sharing, 11, 192
generalized Szegő curve, 123
generating function technique, 137
Haight’s zeta distribution, 39
homotopy, 49, 55
implicit framing, 143
implicit function theorem, 34, 186
incomplete Gamma function, 99
increasing failure rate, 110
injective mapping, 45, 48
Jackson network, 11, 193
Jensen’s inequality, 114, 122
Jordan curve, 45, 49, 180
Kakeya’s theorem, 55
Kingman’s upper bound, 110, 122
Lagrange’s inversion theorem, 44, 58
large deviations, 13, 206, 222
Lerch’s transcendent, 96
Liouville’s theorem, 30
Little’s law, 147, 148
machine repair model, 6
maximum principle, 72
Mills’ ratio, 98
multi-access communication, 1, 2
Newton-Raphson, 187
partial coupling, 11, 202
partial fraction expansion, 26
phase-type distribution, 195
preemptive priority, 196
principal value, 44
Pringsheim’s theorem, 56, 214
product form solution, 11, 193

QROOT, 56, 81
quasi-birth-and-death process, 206

random access protocol, 2
regularization, 46
relaxation time, 33, 83, 85
request-grant mechanism, 2
residual lifetime, 143, 144
Riemann zeta distribution, 38
Riemann zeta function, 38
Riemann-Hilbert problem, 176, 200
right-continuous random walk, 78
Rouché’s theorem, 35, 70, 71, 78, 186, 210

saddle point method, 90
Spitzer’s identity, 27, 78
starshaped, 72, 180
Stirling’s formula, 51, 209
Szegö curve, 47

Taylor series expansion, 65, 183
Theodorsen’s integral equation, 180, 187, 201
transient behavior, 27
transmission delay, 4
trapezium rule, 186

uniform convergence, 36
univalent functions, 55, 72

Vandermonde matrix, 23

Weibull distribution, 38
Wiener-Hopf factorization, 30

Zipf distribution, 38
Samenvatting

Kabelnetwerken zijn oorspronkelijk aangelegd voor het verzenden van beelden vanuit een centraal punt naar televisietoestellen van huishoudens. Het laatste decennium zijn kabelnetwerken geschikt gemaakt voor interactieve toepassingen als het internet. Waar televisies slechts signalen ontvangen, zijn computers ook bronnen die data versturen. Er is dan sprake van dataverkeer in twee richtingen: van het centrale punt naar televisies/computers en van computers naar het centrale punt. Een dergelijk kabelnetwerk vormt daarmee een systeem waarvan de capaciteit gedeeld moet worden door vele computers. Bovendien gaat data verloren wanneer meerdere computers tegelijkertijd data versturen. Dit vormt vooral een probleem wanneer het netwerk behoorlijk belast wordt. In die situatie is er dan ook behoefte aan het reguleren van het dataverkeer. Belangrijk daarbij is dat de computers slechts met elkaar verbonden zijn door het centrale punt en derhalve niet rechtstreeks met elkaar kunnen communiceren.

Een gangbare manier van reguleren is een reserveringsmechanisme. Hierbij moet een computer eerst capaciteit reserveren alvorens data te kunnen versturen. Een computer stuurt daartoe een verzoek naar het centrale punt. Dit verzoek kan nog steeds verloren gaan wanneer meerdere computers tegelijkertijd verzoeken sturen, maar de verzoeken zijn zeer klein en vragen weinig capaciteit. Wanneer het verzoek wordt gelanceerd krijgt de gebruiker, gedurende een voorgeschreven periode, het exclusieve gebruik van het netwerk.

Het reserveringsmechanisme vereist een extra inspanning in de vorm van het versturen van verzoeken, maar in ruil daarvoor komt alle relevante informatie binnen bij het centrale punt, waardoor een efficiënte capaciteitstoewijzing kan worden bepaald.

Vanuit het perspectief van de gebruiker is het versturen van data dan een tweestapsprocedure: het versturen van het verzoek en het verzenden van de feitelijke data. Bij het centrale punt moet dan worden beslist hoe de capaciteit van het netwerk wordt verdeeld over deze twee procedures. We duiden deze capaciteitsverdeling aan met toewijzingsmechanisme. Dit proefschrift richt zich op het modelleren van de vertraging die gebruikers in deze twee procedures ondervinden. Meer specifiek richten wij ons op het onderzoeken van de invloed van het toewijzingsmechanisme op de vertraging. We vertalen de tweestapsprocedure in verschillende wachtrijmodellen. Deze modellen worden vervolgens wiskundig geanalyseerd en leiden tot uitdrukkingen voor prestatiematen als bijvoorbeeld het gemiddelde en de variantie van de vertraging die een gebruiker ervaart. Dergelijke prestatiematen
Samenvatting

zijn zeer relevant voor het bepalen van de kwaliteit van de service. De kwaliteit van de aangeboden dienst is cruciaal voor een kabelnetwerk dat commercieel wordt geëxploiteerd voor het verlenen van toegang tot het internet. De modellen zijn dan ook zeer geschikt voor een scenario-analyse, waarbij de invloed van verschillende factoren op de prestatiematen kan worden onderzocht. Hierbij moet men denken aan factoren als de capaciteit van het netwerk, het gedrag van gebruikers en het toewijzingsmechanisme.

De meeste modellen zoals ontwikkeld en geanalyseerd in dit proefschrift zijn toepasbaar op een breder gebied dan dat van kabelnetwerken. Karakteristieken als de gedeelde capaciteit en het reserveringsmechanisme zijn ook van toepassing op bijvoorbeeld draadloos internetverkeer en mobiele telefonie over satellieten. De modellen zijn feitelijk één-, twee- of hoger-dimensionale Markov processen en de technieken die worden gebruikt zijn vanuit een wiskundig oogpunt interessant. Ook streven we in dit proefschrift naar een zo volledig mogelijke beschrijving van het implementeren van de gevonden resultaten.

Dit proefschrift bestaat uit twee delen. Deel 1 beslaat hoofdstuk 2 t/m 6 en handelt over de discrete bulk service queue (DBSQ). Dit model is één van de standaardmodellen voor het analyseren van datacommunicatie. Deel 2 beslaat hoofdstuk 7 t/m 11 en hierin worden modellen gepresenteerd die specifiek zijn ontwikkeld voor de prestatie-analyse van datacommunicatie over kabelnetwerken.

- De DBSQ is een wachtrij met groepsbediening, waarbij tijd is ingedeeld in sloten. Nieuwe klanten sluiten aan het eind van ieder slot aan in de rij en per slot kan een groep van maximaal s klanten worden bediend. Noteer met $X_n$ de rijlengte aan het begin van slot $n$. De relatie tussen de rijlengten aan het begin van twee opeenvolgende sloten kan dan worden beschreven als

$$X_{n+1} = \max\{0, X_n - s\} + A_n, \quad n = 1, 2, \ldots$$

waarbij $A_n$ staat voor het aantal nieuwe klanten dat aansluit aan het eind van slot $n$. Neem vervolgens aan dat $A_n, n = 1, 2, \ldots$ identiek en onafhankelijk verdeeld zijn volgens een discrete, niet-negatieve stochast $A$ met verwachting $\mathbb{E}A < s$. De stationaire verdeling $\mathbb{P}(X = j) = \lim_{n \to \infty} \mathbb{P}(X_n = j), j = 0, 1, 2, \ldots$ bestaat dan.

Het vinden van (karakteristieken van) deze stationaire verdeling staat centraal in hoofdstuk 2 t/m 6.

In hoofdstuk 2 wordt een overzicht gegeven van de bestaande literatuur. Tevens worden de meest gangbare technieken voor het bepalen van de stationaire verdeling geëxplorieerd. We laten zien dat de stationaire verdeling kan worden uitgedrukt ofwel (i) in termen van de nulpunten van $z^s - A(z)$, met $A(z)$ de kansgenererende functie van $A$, ofwel (ii) in termen van een oneindige som met daarin convoluties van de verdeling van $A$. Beide manieren hebben hun eigen beperking: de nulpunten moeten doorgaans numeriek worden bepaald en de oneindige som zal moeten worden afgeknot.

In hoofdstuk 3 wordt een nieuwe techniek geïntroduceerd, waarmee de resultaten van type (ii) verkregen worden uit de resultaten van type (i). Eerst wordt
de observatie gemaakt dat nulpunten van $z^s - A(z)$ met $|z| \leq 1$ op een specifieke curve liggen. Door deze curve vervolgens op een bepaalde manier te beschrijven, wordt een expliciete uitdrukking voor de nulpunten gevonden. Deze uitdrukking is een $2\pi$-periodieke functie waarvan de Fouriercoëfficiënten analytisch bepaald kunnen worden. Met behulp van *Fourier sampling* worden vervolgens de resultaten van type (ii) afgeleid voor het gemiddelde, de variantie en de gehele stationaire verdeling.

In hoofdstuk 4 wordt de expliciete uitdrukking voor de nulpunten nader bekeken. In het bijzonder wordt de uitdrukking getoetst op bruikbaarheid en vergeleken met gangbare numerieke methoden. De uitdrukking blijkt numeriek stabiel en gemakkelijk te implementeren.

In hoofdstuk 5 wordt de oneindige som onderzocht. Omdat we vanwege praktische redenen de som moeten afknotten, zoeken we naar een karakterisering voor de snelheid waarmee de som convergeert. Met behulp van de zogeheten *zadelpunt-methode* wordt een asymptotische uitdrukking gevonden voor de onnauwkeurigheid die wordt veroorzaakt door het afknotten. Deze asymptotische uitdrukking dient als ondersteuning voor het bepalen van het niveau waarop de oneindige som kan worden afgeknot.

In hoofdstuk 6 worden onder- en bovengrenzen afgeleid voor de eerste twee momenten van de stationaire verdeling. De grenzen zijn simpel en bevatten hooguit de eerste drie momenten van de verdeling van $A$. Het wordt aangetoond dat de grenzen scherp zijn en aldus een goed alternatief vormen voor (i) en/of (ii).

In deel 2 komen we toe aan de modellen die zeer specifiek zijn gericht op de toepassing in kabelnetwerken met reserveringen. Een kenmerk van deze modellen is dat ze rekening houden met het feit dat er een vertraging optreedt tussen het toekennen van een reservering (bij het centrale punt) en het op de hoogte stellen van de desbetreffende gebruiker.

In hoofdstuk 7 staat de *data queue* centraal, gedefinieerd als het aantal nog te versturen data-pakketten van gebruikers met een reservering. Tijd wordt ingedeeld in sloten en ieder slot kan worden gebruikt als reserveringsslot, voor het behandelen van nieuwe reserveringen, of als data slot, voor het verzenden van data-pakketten van gebruikers met een reservering. Een reserveringsslot vergroot de data queue en een data slot verkleint de data queue.

De beslissing of een slot wordt gebruikt als reserverings- of data slot wordt voor een vast aantal sloten (een frame) genomen. We bekijken dan twee modellen. Het eerste model wijst per frame een vast aantal reserveringssloten toe en de overige sloten in het frame zijn datasloten. We nemen hierbij aan dat een frame dusdanig lang is dat de gebruikers wiens reserveringen worden toegekend in frame $t$, geïnformeerd zijn aan het begin van frame $t + 1$. De transmissievertraging is dan hooguit één frame. Maar wanneer er geen data is om te versturen gaat de capaciteit van de datasloten verloren. Een tweede model gebruikt de ongebruikte datasloten als reserveringssloten en is daarmee per definitie efficiënter. Voor beide modellen wordt de stationaire verdeling van de grootte van de data queue en de vertraging van een willekeurig datapakketje afgeleid.
Als gevolg van de transmissievertraging kan een gebruiker die een verzoek krijgt toegekend pas na een zekere tijd de data versturen. Wanneer de vertraging groter is dan één frame resulteert dit in een cyclisch gedrag van het dataverkeer over het netwerk, hetgeen een zeer negatief effect kan hebben op de prestatiematen en daarmee op de kwaliteit van de dienstverlening. In hoofdstuk 8 worden toewijzingsmechanismen ontwikkeld die rekening houden met de transmissievertraging. Het wordt aangetoond dat deze toewijzingsmechanismen de kwaliteit van de service substantieel verhogen. De patenteerbaarheid van deze mechanismen wordt nog onderzocht.

In hoofdstuk 9 bekijken we naast de data queue tevens de rij met reserveringaanvragen, aangeduid met request queue. We veronderstellen dat de capaciteit van het netwerk wordt gesplitst tussen de twee wachtrijen volgens een vaste verhouding. Bovendien nemen we aan dat wanneer één van beide wachtrijen leeg is, de volledige capaciteit naar de andere wachtrij gaat. Dit principe staat in de literatuur bekend als gekoppelde processoren. Het systeem van de request queue en data queue met gekoppelde processoren wordt gemodelleerd als een twee-dimensionaal Markov proces. We laten zien dat het bepalen van de twee-dimensionale kansgenererende functie van de stationaire rijlengten gelijk staat aan het oplossen van een speciaal type randwaardeprobleem. We geven de oplossing van het randwaardeprobleem en bespreken tevens de numerieke aspecten van het implementeren van deze oplossing.

In hoofdstuk 10 wordt het model van hoofdstuk 9 uitgebreid tot een netwerk van twee stations met probabilistische routering van klanten. We laten wederom zien dat het bepalen van de twee-dimensionale kansgenererende functie gelijk staat aan het oplossen van een randwaardeprobleem.

In hoofdstuk 11 worden voor het model in hoofdstuk 9 asymptotische uitdrukkingen voor de stationaire verdelingen van de rijlengten afgeleid. Deze uitdrukkingen vormen scherpe benaderingen voor grote waarden van de rijlengten.
About the author

Johan van Leeuwaarden was born in Eindhoven, The Netherlands (NL), on September 14, 1978. He completed grammar school at the Bisschop Beckers College, Eindhoven, in June 1995, and went on to study econometrics at Tilburg University (NL). After an internship at the asset management department of Aegon Insurances, The Hague (NL), he received a master’s degree in econometrics, specialization quantitative finance, in May 2000. Right after, he started with the post-graduate program Mathematics for Industry at the Stan Ackermans Institute in Eindhoven. He spent two months at the Helsinki University of Technology, Finland, working on models for mobile communication over satellites, and eight months at Philips Research, Eindhoven, developing models for Internet provisioning over cable networks. He received the PDEng (Professional Doctorate in Engineering) degree in industrial mathematics in August 2002. In September 2002 he started with a PhD project (funded by Philips Research) in the stochastic networks group at the research institute EURANDOM. Under the supervision of Onno Boxma and Jacques Resing, his work led to several papers in international journals, a patent application, and this thesis. Johan defends his thesis at the Eindhoven University of Technology on June 13, 2005. As of June 1, 2005, he joins the credit risk department of ING, Amsterdam (NL).