Lindley-type recursions

Lindley-type recursions

Thomas Stieltjes Institute
for Mathematics

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Lindley-type recursions / by Maria Vlasiou. –

Eindhoven: Technische Universiteit Eindhoven, 2006.
Proefschrift. – ISBN 90-386-0784-9 – ISBN 978-90-386-0784-9
NUR 919
Subject headings: recurrence relations / queuing theory
2000 Mathematics Subject Classification: 60K20, 60K25, 90B22

Printed by
Cover design by
Kindly supported by a full scholarship from the legacy of L. Athanasoula

Lindley-type recursions

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 25 september 2006 om 16.00 uur

door

Maria Vlasiou

geboren te Drama, Griekenland

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. J. Wessels
en
prof.dr.ir. O.J. Boxma

Copromotor:
dr.ir. I.J.B.F. Adan

Acknowledgements

The book is finally over and I feel grateful for the faith many people have shown in me throughout all the years I have been studying. This manuscript is the result of a few years of training under the guidance of my supervisor, Ivo Adan. I thank him for his endless patience in his interaction with me, for the fact that he always made time for me and that he constantly stimulated me to keep on learning. Onno Boxma was always available to discuss anything that I would come up with, no matter if it would be the problem that I was stuck with for the past few months, or my opinion on yesterday's tennis game. His opinion, both on maths and on tennis, will always be valued. I truly enjoyed our collaboration. Jaap Wessels has been through numerous drafts of this book and the papers that it is comprised of, and his constructive comments have led to many improvements and additions.

I am blessed to be surrounded by dear friends and family. I often feel that the reason I did not quit early in this project is the constant support and advice of Krishanu Maulik. I am deeply indebted to him for his guidance and unconditional assistance all these years. His friendship is one of the most valuable gifts that I have received in this country. Ioanna Dimitriou and Fanis Matsoukas were always a phone call away, at any our of the day or night, to hear all my mathematical and personal problems. Dimitris Fotiadis and my family where always there to hear me out, even if that meant that they would have to spend several hours on the phone.

A great many colleagues and staff at EURANDOM and TU/e have sustained me over the years in what must have seemed an increasingly testing endeavour. I particularly wish to thank all the support staff in EURANDOM that were always ready to answer questions ranging from peculiarities in the Dutch language to advice on how to fill in my tax form. I wish also to thank Marko Boon and Wil Kortsmit for answering any computer-related question I would come up with. They were both obliging teachers, whose guidance has directly or indirectly produced all graphs appearing in this book. Stef van Eijndhoven is the father of the intriguing differential equation (A.1), which helped me understand better the implications of my own work.

I would not have initiated this project if it were not for the constant encouragement of several professors and friends I have left back in Thessaloniki. Dr. Sotiria Harmousi did not only teach me how to handle my migraines; she also gave me valuable advice that I recall and cherish until today. I am honoured by the trust and confidence that my professors have shown in me. I would like to thank Theodora Theohari-Apostolidi, Lia Kalfa, and Michalis Marias for the constant support throughout all these years.

The joy of my own family life throughout this project is altogether the creation of my partner, Bert. Even a dissertation seems a paltry offering for the daily riches of his companionship.

Notation

A_n	service time of the n-th customer
B_n	preparation time of the n-th customer
W_n	waiting time of the server for the n-th customer
X_{n+1}	$B_{n+1} - A_n$
A, B, W	generic service, preparation, and waiting time respectively
F_Y	distribution function of the random variable Y; e.g., F_A is the service-time distribution
f_Y	density function of the random variable Y
α, β, ω	Laplace-Stieltjes transform of A, B, W respectively
ϕ	Laplace-Stieltjes transform of $A + W$
λ	the scale parameter of the phase-type distribution associated with the service times, whenever relevant
μ	the scale parameter of the phase-type distribution associated with the preparation times, whenever relevant
π_0	$\mathbb{P}[W = 0]$, i.e., the mass of the waiting-time distribution F_W at the origin
$c(k)$	the covariance function $\mathrm{cov}[W_1, W_{1+k}]$
θ	throughput
G_i	the Erlang distribution with i stages
$\mathbb{E}[Y\,;E]$	$\mathbb{E}[Y \cdot \mathbb{1}_{[E]}]$, where Y is a random variable and E is an event
c_Y^2	the squared coefficient of variation of the random variable Y
$f^{(i)}$	the i-th derivative of the function f
$\mathbb{R}^+$	the set of nonnegative real numbers
$\stackrel{\mathcal{D}}{=}$	equal in distribution

Contents

Chapter 1
Introduction

1.1 Scope of the thesis

We have all had the unpleasant experience of waiting for too long at some queue. We seem to lose a significant amount of time waiting for some operator to reply to our call, or for the doctor to be able to see us. We see queues form in practically every aspect of modern life. A typical example of a queue is the phone calls arriving at a call centre. Several operators provide service to the customers that are calling. The phone calls may be diverted to different queues according to the selection a customer makes based on his needs, or they may be randomly assigned to an available operator. Furthermore, there may be some operators with a specific specialisation, such as the knowledge of a foreign language or expertise in financial matters.

Queues are the object of study of *queuing theory*. Under queuing theory we understand the branch of probability theory that studies models that involve a number of *servers* providing service to at least one queue of *customers*. The customers may come one by one or in groups. Neither the customers nor the servers are necessarily individuals. They may be objects, computer programmes, cars, data packets, etc. For example, in a computer we may think of the processes and applications we are currently using as the "customers" who are served by the central processing unit (CPU).

The study of queuing models is usually motivated by a specific application. For example, one may need to study how many operators are necessary in a call centre, so that customers do not have to wait more than a certain amount of time. One may also study a queuing model as a first approach in understanding the operation of a complicated network (for example in industry or telecommunications). Characteristics that determine a queuing model include the number of the servers, the way in which they serve the customers, which is called *service discipline*, the way customers arrive (for example, one by one or in groups), and the way all the involved parts interact with one another. Every change in these parameters, no matter how small it may seem, has the potential to alter the model significantly. The subjects that are studied can usually be traced back to questions concerning the *quality of service* (for example, the waiting time of an arriving customer) or the *performance* of the system (for example, the average queue length).

In this dissertation, we study a specific equation, which originates from a queuing model that emerges from the manufacturing and warehousing world. However, this model also describes other systems met in everyday situations. In the following section, we shall give in detail various examples of real-life problems that are described by this model and that will serve as our working examples for this monograph. The main characteristic of our model is that it is a system that involves two stations alternatingly served by one server. The equation we are interested in describes a

certain relation between the server and the customers.

The scope of this thesis is twofold. First, we study this equation under various settings; second, we apply these results to the queuing model we are interested in. Furthermore, we also study various other aspects of this model and derive performance characteristics, such as the throughput of the system, that provide us with deeper understanding on how such systems work in practice. Moreover, we shall compare the results we derive for this equation to already existing results in queuing theory, we shall compare our model to similar models that have been studied before, and we shall discuss how our results complement the existing literature on queuing theory.

1.2 Background motivation

Consider an ophthalmologist who performs laser surgeries for cataracts. Since the procedure lasts only 10 minutes and is rather simple, he will typically schedule many consecutive surgeries in one day. Before surgery, the patient undergoes a preparation phase, which does not require the surgeon's attendance. In order to optimise the doctor's utilisation, the following strategy is followed. There are two operating rooms that are constantly occupied with patients. While the surgeon works in one of them, the next patient is being prepared in the other one. As soon as the surgeon completes one operation, he moves to the other room and a new patient starts his preparation period in the room that has just been emptied.

In the above example, the surgeon is the server of our model, and the patients waiting for surgery are the customers. The characteristics of the system are perhaps not so evident. Clearly, there is a single server. Furthermore, there are two service points (i.e. the two operating rooms in the above example) that need to be served by the same server. However, the way the customers arrive does not seem to be relevant. In this example we have assumed that the surgeon has scheduled "a sufficient amount of appointments" in one day. This implies that the surgeon will not have to stop working because there are no patients in the waiting room. Therefore, for all practical purposes, we can safely assume that there is an infinite amount of work to be done. Another important feature is that the customers have a special preparation phase before the server can help them. In our example, this preparation time usually includes tasks such as registering the patient, placing him on the operating chair, giving him the local anaesthesia, and waiting for the drugs to take effect. Finally, yet importantly, the server in this model is obliged to alternate. In the example, the surgeon is obliged to operate next on the patient in the other operating room. A reason may be that a patient who has already received anaesthesia during his preparation phase must be operated before the effects of the drug wear out.

Apart from the above example, we may think of a hairdresser who has an assistant to help with the preparation of the customers or of a canteen with one employee and two counters that the employee serves in turns. It is common practice at hairdressers to have two chairs for customers. In one chair, the hairdresser serves a

customer, while some other customer is being prepared at the next chair. For reasons of fairness, the hairdresser has to serve that particular customer next. The key characteristic common in all these examples is that they involve a single server that alternates between two stations. Although the example of the surgeon we have mentioned seems to be a fairly interesting one, not much research has been devoted to alternating service models.

Another application we are interested in comes from the warehousing world. This example originates from a system involving two carousels served by a picker that alternates between them. Contrary to alternating service models, carousel models have received quite a substantial attention in the literature. We shall review some of the main directions of the literature on carousel models in the next section.

Before getting into the details of the system, we describe the basic characteristics of carousels. A carousel is an automated storage and retrieval system, widely used in modern warehouses. It consists of a number of shelves or drawers, rotating in a closed loop. It is operated by a picker that has a fixed position in front of the carousel, which, by rotating, brings the items to the picker. Carousels come in a huge variety of configurations, sizes, and types. They can be horizontal or vertical and rotate in either one or both directions. Carousels are used in many different situations. For example, e-commerce companies use them to store small items and manage small individual orders.

The system we are interested in consists of two identical bidirectional carousels and one picker. At each carousel, there is an infinite supply of pick orders that need to be processed. The picker alternates between the two carousels, picking one order at a time. Each pick order requires exactly one item. The picking process may be visualised as follows. There is one randomly positioned item at each carousel. When the picker is about to pick an item at one of the carousels, he may have to wait until it is rotated in front of him. In the meantime, the other carousel rotates towards the position of the next item. If the other carousel reaches the origin before the picker completes his previous pick, then the carousel stops and waits for the picker. After completion of a pick, the carousel is instantaneously replenished and starts rotating so that the following item to be collected reaches the origin. The picker in the meantime turns to the other carousel, where he may or may not have to wait, and so on.

Clearly, the server in this system is the picker, and the items are the customers. As before, we have again two service points (i.e., the two carousels) that need to be alternately served by the server. Furthermore, the preparation time of a customer evidently is the time it takes for the carousel to rotate until the item is placed in front of the picker. Another important assumption that we have made also for this problem is that there is an infinite supply of pick orders; in other words, there is an infinite amount of work to be done. The analogies of this example to the previous one are evident. Both examples can be described by a single model. In Section 1.4 we shall formulate the model. Before that though, we shall review some of the developments in the literature on carousels.

1.3 The history of carousels

Carousel models have received much attention in the literature and continue to pose interesting problems. There is a rich literature on carousels that dates back to 1980 [177]. We shall review some of the main research topics that have been of interest to the research community so far. The list of references presented is by no means exhaustive; it rather serves the purpose of indicating the continuing interest in carousels.

As we have mentioned, a carousel is an automated warehousing system. It consists of a series of *shelves* which are linked together and are rotating in a closed loop. Each shelf has several *bins* or *drawers* on which the various stocked items are stored. A typical vertical carousel is given in Figure 1.1. An *order* is a set of items that must be picked together (as, for instance, for a single customer). One of the most important benefits of a carousel is that, instead of having the picker, human or robotic, travel to retrieve an item, the item can travel to the picker. Thus, while the item is travelling, the picker has time to package, label, or pick from another carousel. This practice enhances the operational efficiency of the warehouse. Although both *unidirectional* (one-way rotating) or *bidirectional* (two-way rotating) carousels are encountered in practice, the bidirectional types are the most common (as well as being the most efficient) [81]. Usually a carousel is modelled as a circle, either as a *discrete model* [14, 91, 155, 182], where the circle consists of a fixed number of locations, or as a *continuous* one [69, 116, 164, 173], where the circle has unit length and the locations of the required items are represented as arbitrary points on the circle.

In the following, we classify the literature on carousels according to the main theme handled. There are other ways to review the literature (for example, in chronological order). However, this taxonomy allows for a better overview of the variety of the subjects examined. A crucial distinction is made between systems that involve a single carousel and systems with multiple carousels. The first four categories presented relate to single-carousel systems, while systems with multiple carousels are examined later on.

1.3.1 Storage

The performance of a carousel system depends greatly upon the way it is loaded and the demand frequency of the items placed on it. An effective storage scheme may reduce significantly the travel time of the carousel. Several strategies have been followed in practice to store items on a carousel. The simplest strategy is to place the items randomly on the carousel. *Randomised policies* have been examined extensively [85, 116], and various performance characteristics have been derived under the assumption that the items are uniformly distributed on the carousel.

One way to improve the throughput of a carousel system is to adopt a storage policy other than the randomised assignment policy. Ha and Hwang [74] have studied what they call the "two-class-based storage", which is a storage scheme that divides the items in two classes based on their demand frequency. The items with

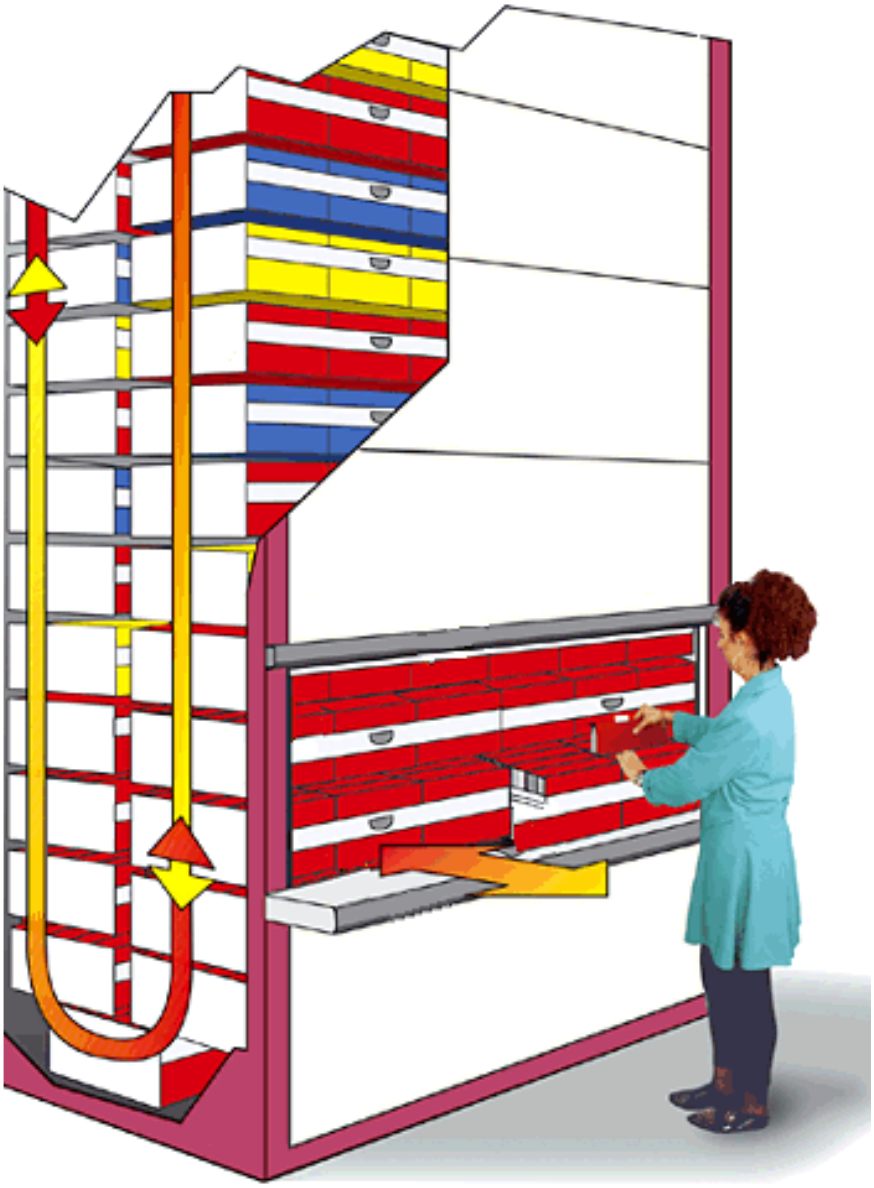

Figure 1.1: A typical vertical carousel.

a higher turnover are randomly assigned to one continuous region of the carousel, while the less frequently asked items occupy the complementary region. The authors show by simulation that the two-class-based storage can reduce significantly the expected cycle time, both in the case where a cycle is a single pick or storage of an item (single-command cycle), and in the case where a cycle consists of the paired operations of storing and retrieving (dual-command cycle). The same authors in [86] examine the effects of the two-class-based storage policy on the throughput of the system, and present a case where there is a 16.29% improvement of this policy over the randomised policy.

Another storage scheme is suggested by Stern [155]. Assignments are made using a *maximal adjacency principle*, that is, two items are placed closely if their probability of appearing in the same order is high. The author evaluates this storage assignment analytically by using a Markov chain model he develops.

The *organ pipe arrangement* for a carousel system is introduced in Lim *et al.* [114] and is proven to be optimal in Bengü [16] and in Vickson and Fujimoto [168] under a wide variety of settings. The organ pipe arrangement has been widely used in storage units, such as magnetic tapes [20] and warehouses [124]. This arrangement is based on the classical mathematical work of Hardy, Littlewood and Polya [77]. Their concept is used to minimise the expected distance travelled by an access head as it travels from one musical record to another. Various optimality properties of this arrangement have been proven; see for example Keane *et al.* [98] and references therein.

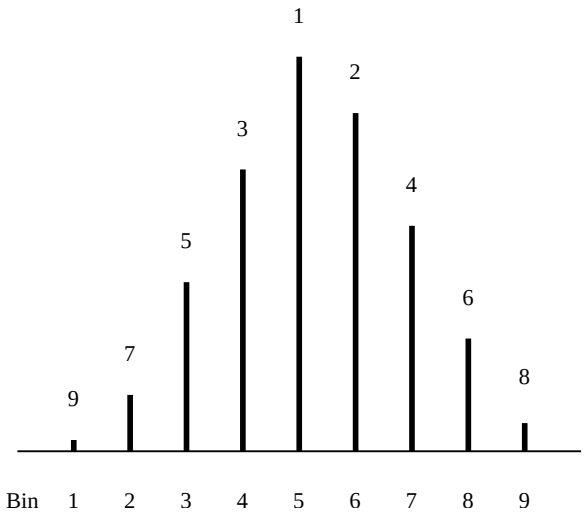

Figure 1.2: Illustration of the organ pipe arrangement, where the upper numbers indicate the frequency ranking of an item.

In carousel systems, this arrangement places the item with the highest demand in an arbitrary bin, the items with the second and third highest demands in the bin next to the first one but from opposite sides, and sequentially all other items next to the previous ones, where the odd-numbered items according to their frequency are grouped together and placed next to one another in a decreasing order from the one side of the most frequent item (and similarly the even-numbered items are grouped together and placed to the other side). Figure 1.2 illustrates the organ pipe arrangement. The numbers at the top indicate the ranking of an item in a decreasing order of frequency.

Another question related to storage is about the number of items of each type that should be stored on the carousel in order to maximise the number of orders that can be retrieved without having to reload. This question is examined in Jacobs *et al.* [91], where the authors propose a heuristic that yields a reasonable solution, the error of which can be bounded. This method has been improved by Yeh [182], where a more accurate solution is obtained, and further on by Kim [101], where it is observed that the heuristic described in [182] does not always lead to the optimal solution. The author constructs an algorithm that yields the optimal solution. This algorithm is further improved in Li and Wan [111].

1.3.2 Picking a single order

One of the most important performance characteristics of a carousel system is the total time to pick an order. The total time to retrieve all items of an order may be expressed as a sum of the total time that the carousel is travelling plus the total time that the carousel is stopped for picking. The latter is effectively the total pick time, and it is not affected by the sequence in which we choose to retrieve the objects. However, the total travelling time greatly depends upon the retrieval sequence. The analysis of the travel time under various strategies is, in general, a non-trivial problem. This problem, however, has been resolved for independent and uniformly distributed item locations [116].

Various picking strategies have been proposed. Bartholdi and Platzman [14] assume a discrete model and study the performance of an algorithm and three heuristics that determine an efficient, but not necessarily optimal, sequence of retrieving all items. A heuristic is a simpler, non-optimal procedure that is based on a specific strategy. The heuristic methods proposed are the *nearest-item heuristic*, where the next item to be picked is always the one that is closer to the picker at any given moment, the *shorter-direction heuristic*, where the carousel chooses the shortest direction between the route that simply rotates clockwise and the route that rotates counterclockwise, and the *monomaniacal heuristic*, that always chooses to rotate to the right and pick items sequentially. The *optimal retrieval* algorithm that is presented enumerates all possible paths; therefore, it is guaranteed to find the quickest sequence in which to retrieve a single order. In [14] the authors prove among other things that the travel time under the nearest-item heuristic is never greater than one rotation of the carousel. Litvak *et al.* [120] improve this upper bound and show that the new upper bound is tight. In [118] Litvak and Adan as-

sume that the positions of the items are independent and uniformly distributed and give a detailed analysis of the distribution of the travel time under this heuristic and its asymptotic behaviour when the number of items tends to infinity.

An interesting picking strategy is the so-called *m-step strategy*, where the carousel chooses the shortest route among the ones that change direction at most once, and only do so after collecting at most m items. Rouwenhorst *et al.* [148] analyse the m-step strategy for $m \leqslant 2$. This means that the carousel changes direction after collecting at most two items. They give stochastic upper bounds for the minimum travel time and study the distribution of the travel time under these assumptions. Their results indicate that this strategy performs very well. Litvak and Adan in [119] compare the nearest-item heuristic with the m-step strategy to conclude that already for $m = 2$, the m-step strategy is very close to optimal, and better than the nearest-item heuristic. Furthermore, they assume that the items are randomly placed on the carousel and derive the distribution and the moments of the travel time, provided that $n > 2m$, where n is the number of items in an order.

Wen and Chang [179] model the carousel as a discrete bidirectional loop and assume that the time to move between the bins of a shelf is *not* negligible. They propose three heuristic solution procedures and compare their performance. An earlier version of this work can be found in Wen [178].

Ghosh and Wells [69] model the carousel as a continuum of *clusters* and *gaps*, where a cluster is a segment on the circle that corresponds to a series of locations that have to be visited for the retrieval of an order, while a gap is the segment of the circle between two clusters. The authors develop two algorithms to find optimal retrieval strategies.

Stern [155] studies properties of the optimal, i.e. minimal, picking sequence both for the *open-loop* strategy, where the carousel remains stationary at the point where the last item was retrieved (awaiting the next order to be fed), and for the *closed-loop* strategy, where the carousel returns to a predefined point after the retrieval of an order is completed. He formally shows that under the open-loop strategy the carousel will change its direction at most once when following the optimal picking sequence, while under the closed-loop strategy the carousel will turn at most twice. A recursive expression for the distribution of the minimal travel time for randomly distributed items is given explicitly by Litvak and Van Zwet [121].

More recent literature includes the work of Wan and Wolff [176] that focuses on minimising the travel time for "clumpy" orders, that is, orders concentrated on a relatively small segment of the carousel, and introduces the nearest-endpoint heuristic for which they obtain conditions for it to be optimal. Under this setting, one can no longer assume that the items' locations are uniformly distributed. The model with non-uniform items' locations reflects a relevant situation when some of the items are required more frequently than others. An interesting work on non-uniformly distributed items is given by Litvak [117], where the focus is on the length of the shortest rotation time needed to collect a single order when the order size is large and the items' locations have a non-uniform continuous distribution with a positive density f on $[0, 1]$.

1.3.3 Picking multiple orders

A popular strategy for reducing the mean travel time per order in carousel storage and retrieval systems is batching together a number of orders and then picking them sequentially. A *batch* is a set of orders that is picked in a single tour. Two consecutively picked items do not necessarily belong to the same order. An excellent literature survey by Van den Berg [165] on planning and control of warehousing systems addresses this issue and the problems that arise if large batches are formed. Apart from the questions mentioned before, Stern [155] also considers the performance of a carousel for a fixed set of order types (for example, big orders with many items, and small ones).

Bartholdi and Platzman [14] are mainly concerned with sequencing batches of requests in a bidirectional carousel. They specify the number of orders to be retrieved (ignoring any new arrivals) and propose three heuristic methods to solve this static problem. Orders may be picked in any sequence (and not necessarily at the order they arrive), but picks within the same order are performed consecutively. They define the *minimum spanning interval*, which is the shortest interval containing all the items of an order and, by assuming that the picker always begins and finishes retrieving an order at one of the endpoints of this interval, they construct the shortest matching chain by ordering the orders accordingly. This procedure may fail to give an uninterrupted sequence in which to pick the orders; therefore, they propose the following heuristics. The first one, called the *hierarchical* heuristic, picks any order that happens to have a common endpoint with another order, and then travels clockwise until an unpicked endpoint is encountered, and repeats the procedure. The *nearest-order* heuristic is practically an extension of the nearest-item heuristic described earlier in the paper, as is the case with the *second monomaniacal* heuristic they propose. Under these heuristics, they obtain upper bounds for the travel time.

Ghosh and Wells [69] assume that the orders have to be picked under a FIFO sequencing restriction, which means that the first order to arrive at the warehouse is the first order that will be picked, and so on. Since the orders are retrieved in a FIFO fashion, the problem is reduced to finding how to retrieve each individual order so that the best overall retrieval is achieved. They develop an algorithm for the optimal retrieval path of n orders via dynamic programming, and show how to update dynamically the solution when new orders arrive.

Rouwenhorst *et al.* [148] model the carousel as an M/G/1 queuing system, where the orders are the "customers" that require service, and the service they get depends on the pick strategy that is followed. This approach permits the derivation of various queuing characteristics such as the mean response time and the waiting time when orders arrive randomly. The authors mention that the tight upper bounds for the mean response time can be further exploited to obtain also good approximations for excess probabilities of the response time.

Van den Berg [164] assumes either a fixed or an arbitrary sequence of orders. When the sequence of the orders is given, he presents an efficient dynamic programming algorithm that finds an optimum path that visits all orders in the specified sequence. Furthermore, when there is no given order sequence, he simplifies the

problem to a *rural postman problem* on a circle and solves this problem to optimality. The rural postman problem is a problem of finding the shortest route in an undirected graph which includes all edges at least one time. Van den Berg [164] concludes that the obtained solution requires at most 1.5 revolutions more than a lower bound of an optimal solution to the original problem. Simulation results suggest that the average rotation time may be reduced up to 25% when allowing a free order sequence.

1.3.4 Design issues

All research papers mentioned so far that deal with travel time models of carousel systems assume average uniform velocity of the carousel. In other words, the main assumption is that the carousel travels with constant speed and the acceleration from the stationary position (when a pick is performed) to its full speed, as well as the deceleration from the maximum speed to zero speed, are negligible factors when computing the travel time of the carousel. Guenov and Raeside [73] give some empirical evidence that the error induced when neglecting acceleration and deceleration of an order picking vehicle is indeed negligible. Thus the problem of minimising retrieval times can be considered to be equivalent to the problem of minimising the average distance travelled by the carousel per retrieval.

Hwang *et al.* [88], however, develop strategies for picking that take into consideration the variation in speed of the carousel. For unit-load automated storage and retrieval systems there are several travel-time models that consider the speed profiles of the storage and retrieval robot. In [88] some relevant references are given. Unlike the unit-load automated storage and retrieval systems, almost all the existing travel-time models for carousel systems assume that the effects of the variation in speed are negligible. In [88] the authors try to bridge this gap in the literature. They assume that the items are randomly distributed on the carousel and derive the expected travel time both in the case of a single command cycle and in the case of a dual command cycle. They verify the accuracy of the proposed models by comparing the results to results directly obtained from discrete racks.

Egbelu and Wu [57] study the problem of pre-positioning the carousel in anticipation of storage or retrieval requests in order to improve the average response time of the system. Choosing the right starting point of a carousel in anticipation of an order is also referred as the *dwell point selection problem.* This strategy becomes relevant when the items are stored under the organ pipe arrangement. In this situation the dwell point should be chosen to be the location of the most popular item; see, e.g., [16].

Spee [153] is concerned with developing design criteria for carousels. He states the basic conditions for designing an automatic order picking system with carousels and comments on the optimal storage design. Namely, he is interested in finding the right number of picking robots and the right number and dimensions of a carousel so that the investment is minimised, provided that the size of the orders that need to be retrieved is given.

McGinnis [127] studies some of the design and control issues relevant to *rotary*

racks. A rotary rack is an automated storage and retrieval system that strongly resembles carousels. In fact, conceptually, a rotary rack is simply a carousel, where the only difference is that each level or shelf of this carousel can rotate independently of the others. The author concludes that, while rotary racks appear to be a simple generalisation of conventional carousels, the control strategies that have been shown effective for carousels do not appear to be as effective for these systems. Rotary racks can be viewed as a multiple-carousel system (where each level is considered as a sub-carousel) with a single picker.

1.3.5 Problems involving multiple carousels

While almost all work mentioned before concerns one-carousel models, real applications have triggered the study of models involving multiple carousels. The study of such models is not as developed yet as the study of models involving a single carousel. The list of references that follows seems to be complete.

Systems with multiple carousels tend to be more complicated. The system cannot be viewed as a number of independently operating carousels, since there may be some interaction between two separate carousels by means of the picker that is assigned to them. Namely, if the number of pickers is less than the number of carousels, then the picking strategy that is chosen for an isolated carousel may affect significantly the waiting time and/or the travel time of another carousel. Thus, one cannot guarantee that minimising the travel time of a single carousel minimises the total travel time of all carousels (and consequently the throughput); the outcome may be quite the contrary because of the system's interdependency. Therefore, multiple-carousel systems merit a special reference.

Multiple-carousel systems tend to have a higher level of throughput; however, they increase the investment cost due to the extra driving and control mechanisms [85, 87]. A natural question is how much the throughput of a standard carousel can be improved by the corresponding multiple-carousel system that has the same number of shelves as the standard carousel.

Perhaps the first academic study that investigates the performance of a system involving several carousels is that of Emerson and Schmatz [60]. The authors simulated the operation of the warehouse of Rockwell's Collins Telecommunications Products. The system consists of twenty-two carousels, where each pair of carousels had a single-operator station (so there are in total eleven operator stations). The questions they are concerned with are how big the batch size of orders should be so as to complete the week's work (which is used as a performance measure) and keep all operators busy, what happens when a carousel or a station is down, and how is an overload or an imbalance (for example, unequal operator performance, unequal carousel loading, or large orders) handled. In order to investigate potential solutions to these three imbalance conditions, the authors investigate two operating rules.

The first operating rule studies six different storage schemes with seven carousel pairs (and thus seven operators). It uses simulation models to study simple storage schemes such as random storage, sequential alternating storage, and storage in the

carousel with the largest number of openings. The aim in [60] is to study the degree of carousel usage. The authors find that there is no significant difference between the carousel loads among the storage schemes. However, they do not treat the problem of optimally assigning items to carousel bins, and do not present any analytical models to help investigate the problem. The second operating rule they investigate is a floating operator. This is an operator who is trained to work at any station, and who is moving to different stations according to specific needs (for example, depending on the size of the queue at a particular station). They conclude that this solution seems advantageous for the purposes of the warehouse they investigate.

Koenigsberg [105] presents analytic solutions for evaluating the performance of a single carousel, and discusses the ways in which his approach can be extended to a system involving two unidirectional carousels both served by a single robotic operator. The carousels are related only through the state of the robot, which means that each carousel is independent of the other except for the time it waits for an operation to commence (such as pick, storage, or repair) because the robot is busy at the other carousel. The author concludes that under some conditions, it is often more advantageous to have two carousels of identical length instead of one carousel of double the length. Furthermore, going to three carousels of equal length (i.e. one third of the length of the single carousel) will offer little further improvement.

Hwang and Ha [85] study the throughput performance both of a single and of a double carousel system. Based on a randomised storage assignment policy, cycle time models are developed for single and dual commands. Furthermore, they examine the value of the information on the succeeding jobs in terms of system efficiency, which may lead to better scheduling of the jobs to be processed.

In a later work, Hwang *et al.* [87] attempt to measure analytically the effects of double shuttles of the storage and retrieval machine (i.e. the robotic picker) on the throughput both of the standard and of the double carousel system. Storage and retrieval machines with double shuttles are machines that have space for two items. Thus, for example, an item can be retrieved from the carousel and stored on one shuttle, while the other shuttle has an item that needs to be stored to the carousel. After this item is stored, a second item can be retrieved from the carousel and placed on the empty shuttle. All these operations occur during a single cycle of the carousel operation. For the double carousel system, a retrieval sequence rule is proposed which utilises the characteristics of the two independently rotating carousels. From the test results, double shuttles are shown to have a substantial improvement over single shuttles. This improvement tends to be more prominent in the double carousel system. Due to cost concerns, the authors note that an economic evaluation will be needed to justify the extra cost of double carousel systems and double shuttles before implementing them in real world situations.

Wen *et al.* [180] consider a system comprised of two carousels and a single retrieval machine. Their main assumption is that every order must be picked in a single tour, i.e., an order cannot be divided into two or more sub-tours. Batching orders together is also not allowed. They analyse the retrieval time and propose four heuristic algorithms for the scheduling sequence of retrieving items from the system to satisfy an order. Their method is an extension of the algorithm presented

in [14] and [155].

Recently, Hassini and Vickson [81] studied storage locations for items, aiming to minimise the long-run expected travel time in a two-carousel setting with a single server. They assume that the products are available at all times (so as to be able to ignore possible delays due to lack of stock), and that orders are not batched; that is, the carousel system processes only single-item orders. This is applicable in situations where individual product orders are processed in a first-come-first-served policy, or when the next item to be retrieved is known only after the present one has been picked. The authors compare the performance of three heuristic storage schemes and a genetic algorithm [70] that for small-sized problems completely enumerates the solution space. They conclude that none of the heuristic approaches leads to a solution that outperforms the algorithmic solution they provide.

The same model is also studied by Park *et al.* [140]. As is the case in [81], in [140] the basic assumptions are that there is an infinite number of items to be picked and that an order consists of a single item. The authors, however, are not interested in storage issues. They further assume that the single operator, the picker, is alternately serving the two carousels. This may cause the picker to have to wait for an amount of time until the item at the carousel he is currently serving is rotated in front of him. They derive the distribution of the waiting time of the picker under specific assumptions for the pick times. This allows them to derive expressions for the system throughput and the picker utilisation.

The models described in Koenigsberg [105], Hassini and Vickson [81], and Park *et al.* [140] are almost identical. However, the questions that are examined differ. This dissertation is motivated by the question investigated by Park *et al.* [140], namely the waiting time of the picker. The waiting time of the picker for each item is given by a very interesting recursion that combines the rotation time of the carousel and the pick time for this item. We shall formally describe the model and derive this recursion in the following section.

1.4 The model

As it is mentioned in Section 1.2, the examples that are given there can be described by a single model. In this section, we shall formulate the model, derive the equation that is the main focus of this dissertation, and introduce the basic notation.

To this end, consider a system consisting of one server and two service points. Only one customer can occupy each service point. The server alternates between the service points, serving one customer at a time. The server is *obliged* to alternate; therefore, he serves all odd-numbered customers at one service point and all even-numbered customers at the other. It is assumed that the system is slightly oversaturated so that at the moment one customer completes his service time another one is immediately available to occupy the same service point. Although in practice the queue at times may be empty, the amount of time it remains so is assumed insignificant. Therefore, we can as well assume that at each service point

there is an infinite queue of customers that needs to be served.

Once a customer enters the service point, his total service is divided into two separate phases. First there is a preparation phase, where the server is *not* involved at all. After the preparation phase is completed, the customer is allowed to start with the second phase, which is the actual service. Figure 1.3 shows a schematic representation of this model.

The customer either has to wait for the server to return from the other service point, where he may be still busy with the previous customer, or he may commence with his actual service immediately after completing his preparation phase. This would be the case only if the server had completed serving the previous customer and was waiting for this customer to complete his preparation phase. Thus the server, after having finished serving a customer at one service point, may have to wait. Once the service is completed, a new customer immediately enters the empty service point and starts his preparation phase without any delay. We are interested in the waiting time of the server.

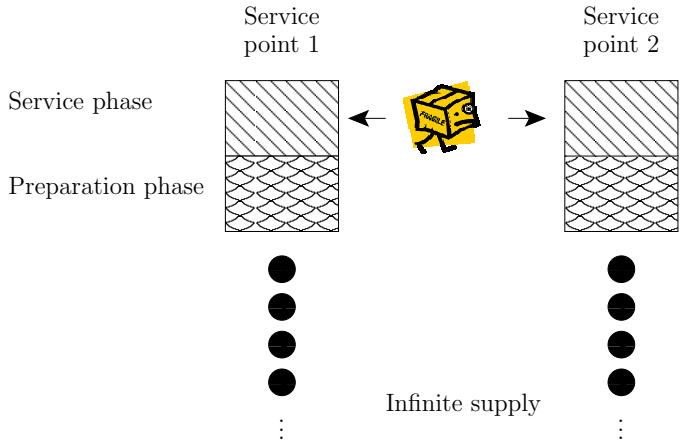

Figure 1.3: The model.

Let B_n denote the preparation time for the n-th customer, and let A_n and W_n be the times the server spends on this customer on service and on waiting for the customer to complete his preparation phase, respectively. Then the server has to wait for the $(n+1)$-st customer at most as long as the preparation time of this customer, which is equal to B_{n+1}. However, when the $(n+1)$-st customer started his preparation time, the server had just moved to serve the n-th customer. Therefore, we need to subtract the time that the customer was busy there, either waiting for the n-th customer, or serving him. Naturally, if $A_n + W_n$ is greater than the preparation of the $(n+1)$-st customer, then the server did not have to wait for this customer and could proceed immediately with the service. Summarising the above, we have that the waiting time W_{n+1} of the server for the $(n+1)$-st customer satisfies the

recursion

$$W_{n+1} = \max\{0, B_{n+1} - A_n - W_n\}, \qquad n \geqslant 1. \tag{1.1}$$

Unless stated otherwise, we assume that $\{A_n\}$ and $\{B_n\}$ are two mutually independent sequences of independent and identically distributed (i.i.d.) non-negative random variables. For the sake of simplicity, from now on we shall use the shorthand $X_{n+1} = B_{n+1} - A_n$, $n \geqslant 1$, unless it is necessary to distinguish between the preparation and the service times. For reasons that will become evident in the sequel, we shall further assume that $\mathbb{P}[X_n < 0] > 0$.

Evidently, all examples given in Section 1.2 are described by this model. The main aim is to calculate explicitly the distribution of the steady-state waiting time of the server, if it exists. In Chapter 2 we prove that such a distribution indeed exists, provided that $\mathbb{P}[X_n < 0] > 0$. For now, it suffices to say that obtaining a closed-form expression of the distribution is, in general, a non-trivial task. When referring to the system in steady state, all subscripts will be suppressed; therefore, the steady-state service, preparation and waiting time are denoted by A, B and W respectively. Naturally, we have that $X = B - A$. So, in steady state, (1.1) becomes

$$W \stackrel{\mathcal{D}}{=} \max\{0, B - A - W\}, \tag{1.2}$$

where A and B are independent. By " $\stackrel{\mathcal{D}}{=}$ " we mean equality in distribution. For this model we shall try to obtain an explicit expression for the distribution of the waiting time W. This will allow us to derive various performance measures, such as the throughput of the system, which is strongly connected to the mean waiting time, and the probability that the server does not have to wait for a customer, which is equal to the mass π_0 of the steady-state waiting time distribution at zero.

For a random variable Y we denote its distribution and its density by F_Y and f_Y respectively. So, for example, the distribution of the steady-state waiting time of the server is denoted by F_W and the density of B is simply f_B. Additionally, the Laplace-Stieltjes transforms of A, B and W are indicated by α, β and ω respectively, so for example,

$$\omega(s) = \int_0^\infty \mathrm{e}^{-sx} \, \mathrm{d}F_W(x).$$

The derivative of order i of ω is denoted by $\omega^{(i)}$ and by definition $\omega^{(0)} = \omega$. Similarly, we define the derivatives of all other distributions, densities, and Laplace transforms that appear. Any further notation will be introduced when it first becomes relevant.

There is a significant difference between the various examples that we have described in Section 1.2. To be precise, in the example inspired by the medical world, the support of the preparation time B is unbounded, while the preparation time in the two-carousel model is limited by the time for a complete rotation of the carousel. We shall observe an important distinction among the techniques that are used and the results that are obtained based on the support of B. Namely, calculations become somewhat cumbersome, although straightforward, when the support of B is bounded. On the other hand, as we shall see in Chapter 5, an unbounded support

does not guarantee that the method followed will necessarily be evident, but it does tend to lead to more simple expressions.

Another important observation that needs to be made is that Equation (1.1) has a striking similarity to Lindley's recursion for the waiting time of a customer in a single-server queue. If only the sign of W_n at the right-hand side of (1.1) were different, then (1.1) would be Lindley's recursion. Lindley's recursion is one of the fundamental and most well-studied equations in queuing theory. We shall discuss this recursion in the following section.

1.5 Lindley's recursion

Lindley's recursion [115] describes the waiting time W_{n+1} of a customer in a single-server queue in terms of the waiting time of the previous customer, his or her service time B_n, and the interarrival time A_n between them. It is assumed that the customers are served in order of arrival. As Figure 1.4 indicates, Lindley's recursion is given by

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

Throughout this thesis, when we refer to "the single-server queue", we imply that the first customer to be served is the first one to come; that is, we refer to a single-server queue where the waiting times of the server are described by the equation above.

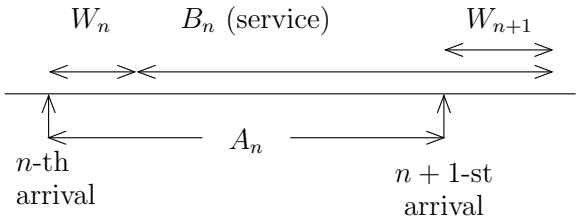

Figure 1.4: Lindley's recursion.

David G. Kendall [100] developed in 1953 a convenient shorthand notation system to classify queuing systems that has to a large extent become standard since then. Kendall's notation has been extended to include various interesting queuing characteristics; a queuing system is denoted by a string of the type $\alpha/\beta/\gamma/\delta/\epsilon$, where α refers to the form of the interarrival distribution, β to the form of the service distribution, γ is the number of servers, δ is the capacity of the waiting room, and ϵ denotes the service discipline. In case only the first three parameters are used, it is implied that the waiting room has infinite capacity and that customers are served in order of arrival. For more details on Kendall's notation see, for example, Asmussen [6, Chapter III].

In this work we adapt this standard notation as follows. We describe the alternating service model we are concerned with by a string of two letters of the fashion α/β, where α now stands for the distribution of the service time A and β refers to the form of the distribution of the preparation time B. The reason, evidently, is that (1.1) differs from the standard Lindley's recursion only in the sign of W that appears at the right-hand side of the equality. Therefore, we can maintain the notation that is widely used for the classic single-server queue. So, for example, M/G refers to the system with exponential service times and generally distributed preparation times, while, in this context, we use "E" when the service or preparation times follow an Erlang distribution. This facilitates the comparison between the two systems.

Some basic queuing models are the M/G/1, the G/M/1, and the G/G/1 queues. Here "G" implies that the interarrival or service times are generally distributed, and "M" (for *Markovian* or *memoryless*) indicates that the underlying distribution is the exponential distribution. Markovian assumptions often greatly simplify the modelling and the solution. They are, therefore, the first approach when faced with a new type of problem, and they will be used here to look into phenomena requiring considerable effort in more general settings. The Markovian set-up has its drawbacks, however. One is that we need to rely on assumptions such as exponential service or preparation times. Phase-type distributions present a partial solution; this class of distributions is dense in the space of distribution functions defined on $[0, \infty)$. Therefore, one can extend the Markovian set-up to a class of models that is also in a certain sense dense. So, motivated by classic queuing theory, we shall analyse (1.2) assuming that the service times are generally distributed and the preparation times follow a phase-type distribution, and vice versa.

Since Lindley's equation is so similar to (1.1), it is reasonable to examine the methods used to derive the steady-state distribution of the waiting time of a customer in a single server queue. One can expect that some of these methods may help determine the distribution of W in (1.2). A wide range of techniques has been used for Lindley's recursion. These techniques usually vary depending on the specific assumptions of the model examined. For the M/G/1 queue, Laplace-Stieltjes transforms yield a straightforward solution (e.g. Cohen [46, Chapter II.4]). Another method used both for the M/G/1 and the G/M/1 queues utilises properties of random walks (e.g. Asmussen [6, Chapter VIII]). Cohen [46, Chapter II.6]) lists another six methods that are often very effective in dealing with special questions regarding the single-server queue. These include Lindley's integral equation, which is in fact a Wiener-Hopf integral equation, and the phase method and its variants. The latter method makes use of the special structure of the phase-type distribution involved by introducing stochastic variables specifying the phase of the service time or of the interarrival time.

Not all of these methods are effective when dealing with the alternating service system described in Section 1.4. For example, the recursion we are dealing with is not a random walk, and therefore no techniques that involve properties of random walks can be used in this case. However, some of the methods mentioned above, or generalisations of them, are effective. For the alternating service model, Laplace-Stieltjes transforms yield simple and straightforward results in case the service time

is generally distributed and the preparation time follows a phase-type distribution; see Chapter 4. The corresponding G/PH/1 single-server queue cannot be solved simply by computing the Laplace-Stieltjes transform of W. In the following section, we shall outline the differences in the analysis and the results of the two models.

In the applied probability literature there has been a considerable amount of interest in generalisations of Lindley's recursion, namely the class of Markov chains described by the recursion $W_{n+1} = g(W_n, X_n)$. For earlier work on such stochastic recursions see for example Bacelli and Brémaud [12], Borovkov and Foss [29] and Brandt *et al.* [38]. Our model is a special case of this general recursion and it is obtained by taking $g(w, x) = \max\{0, x - w\}$. Many structural properties of the recursion $W_{n+1} = g(W_n, X_n)$ have been derived. For example, Asmussen and Sigman [10] develop a duality theory, relating the steady-state distribution to a ruin probability associated with a risk process. More references in this domain can be found for example in Asmussen and Schock Petersen [9], Harrison and Resnick [80], Prabhu [142], and Seal [150].

An important assumption which is frequently made in these studies is that the function $g(w, x)$ is monotone non-decreasing in its main argument w; see for example Borovkov [28] and Kalashnikov [96]. For example, in [10] this assumption is crucial for their duality theory to hold. Clearly, in the model we discuss here, we have that $g(w, x) = \max\{0, x - w\}$, which implies that this assumption does not hold, since g now is a monotone non-increasing function in w. For this reason, a detailed study of (1.1) is of theoretical interest.

1.6 A generalised Wiener-Hopf problem

Since Recursion (1.1) is only a sign apart from Lindley's recursion, it is natural to assess the differences in complexity between these two models. In this section, we present some considerations on this matter. The so-called *generalised Wiener-Hopf problem* will feature prominently in these considerations. Therefore, we formulate this problem.

The Wiener-Hopf problem

First, we briefly consider the ordinary *Wiener-Hopf procedure*. The Wiener-Hopf procedure, also featuring as Wiener-Hopf technique in the literature, was originally invented to solve an integral equation of the type

$$\int_0^\infty f(\xi)K(x - \xi)\,\mathrm{d}\xi = g(x) \tag{1.3}$$

where K and g are given, and f is to be found, cf. Noble [135, p. 49]. As it is shown in [135], this integral equation can be transformed into an equation of the form

$$A(s)\Phi_+(s) + B(s)\Psi_-(s) + C(s) = 0, \tag{1.4}$$

where Φ_+ and Ψ_- are unknown regular functions in the upper and lower half-planes respectively, A, B, and C are known analytic functions, and this equation holds in a

strip of the complex plane. We call (1.3) or (1.4) the Wiener-Hopf integral equation of the first kind.

The fundamental step in the Wiener-Hopf procedure towards solving (1.4) is to find two regular functions K_+ and K_- such that

$$A(s)/B(s) = K_+(s)/K_-(s).$$

Then, by decomposing $K_-(s)C(s)/B(s)$ in the form $C_+(s) + C_-(s)$, (1.4) can be rearranged so as to define a function $J(s)$ by

$$J(s) = K_+(s)\Phi_+(s) + C_+(s) = -K_-(s)\Psi_-(s) - C_-(s). \tag{1.5}$$

By analytic continuation, J can be defined over the whole complex plane, and it can be shown to be regular. Suppose now that it can be shown that

$$|K_+(s)\Phi_+(s) + C_+(s)| < |s|^p$$
$$|K_-(s)\Psi_-(s) + C_-(s)| < |s|^q.$$

Then, by Liouville's theorem we have that J is a polynomial, i.e. both functions $K_+(s)\Phi_+(s) + C_+(s)$ and $K_-(s)\Psi_-(s) + C_-(s)$ are equal to this polynomial, which means that $\Phi_+(s)$ and $\Psi_-(s)$ are determined within a finite number of arbitrary constants which must be determined otherwise. The task of solving equations in the form of (1.3) or (1.4), either by using the Wiener-Hopf procedure or otherwise, is what we call *the Wiener-Hopf problem*.

The crucial step in this technique is finding functions K_+ and K_- such that they satisfy Equation (1.5). All other steps of this technique can be completed by applying general theorems; see Noble [135, pp. 36–38]. For an application, see Chapter 8.

The generalised Wiener-Hopf problem

Let us consider now the generalised Wiener-Hopf equation

$$A(s)\Phi_+(s) + B(s)\Psi_-(s) + C(s) + D(s)\Phi_1(s) + E(s)\Phi_+(-s) = 0, \tag{1.6}$$

where the functions $A, \ldots, E$ are known, Φ_+, Ψ_- are unknown, and Φ_1 is an unknown integral function. Generalised Wiener-Hopf equations differ from the classical Wiener-Hopf equations since the involved plus and minus functions are defined into two different complex planes. Equation (1.6) cannot be solved exactly by the Wiener-Hopf technique. It is remarkable that in some cases a suitable mapping reduces the generalised Wiener-Hopf equations to the classical ones [54].

Usually the Wiener-Hopf formulation involves the factorisation of kernels. However, closed form factorisations of kernels are not available and one often needs to resort to approximate factorisation techniques. Chapter V in [135] discusses some approximate methods which can be used to deal with problems that can be formulated as equations of the form (1.6), or special cases of this equation that do not reduce to equations of the form (1.4). The task of solving equations in the form of (1.6) is what we call *the generalised Wiener-Hopf problem*.

The time-dependent distribution

It is known that for Lindley's recursion one may obtain an explicit expression (see Cohen [46, p. 278]) for the double transform

$$f(r,s) = \sum_{n=1}^{\infty} r^n \mathbb{E}[\mathrm{e}^{-sW_n}],$$

which determines the distribution of W_n given that $W_1 = 0$, by solving a Wiener-Hopf problem. This transform is the generating function of the Laplace-Stieltjes transform of W_n. If one is able to invert (explicitly or numerically) the generating function, then one might obtain an explicit expression for the Laplace-Stieltjes transform, which uniquely defines the time-dependent distribution of the waiting times.

However, the time-dependent distribution of the waiting times for Lindley's recursion is a highly non-trivial problem. An expression is explicitly known only for the M/M/1 queue, while for other settings only an expression for the above mentioned double transform is given. For the Laplace-Stieltjes transform of the weak limit W of W_n one is also called to solve a Wiener-Hopf problem; see, for example, Asmussen [6, Section VIII.3] for more information.

For our model, let H be the generating function of the distribution of W_n, i.e., for $|r| \leqslant 1$,

$$H(r,x) = \sum_{n=0}^{\infty} r^n \mathbb{P}[W_{n+1} \leqslant x], \qquad x \geqslant 0.$$

Assume now that for all n, $X_{n+1} - W_n$ is continuous. Naturally, it suffices to assume that either F_W or F_X is continuous. Then, from (1.1) we have that for $n \geqslant 1$,

$$\mathbb{P}[W_{n+1} \leqslant x] = 1 - \mathbb{P}[X_{n+1} - W_n \geqslant x] = 1 - \int_x^{\infty} \mathbb{P}[W_n \leqslant y - x] \, \mathrm{d}F_X(y).$$

Consequently,

$$\begin{aligned} H(r,x) &= \mathbb{P}[W_1 \leqslant x] + \sum_{n=1}^{\infty} r^n \mathbb{P}[W_{n+1} \leqslant x] \\ &= \mathbb{P}[W_1 \leqslant x] + \frac{r}{1-r} - \sum_{n=1}^{\infty} r^n \int_x^{\infty} \mathbb{P}[W_n \leqslant y - x] \, \mathrm{d}F_X(y) \\ &= \mathbb{P}[W_1 \leqslant x] + \frac{r}{1-r} - r \int_x^{\infty} H(r, y-x) \, \mathrm{d}F_X(y). \end{aligned} \tag{1.7}$$

If one is able to solve the above equation and obtain values for $H(r,x)$ then one can obtain values for $\mathbb{P}[W_n \leqslant x]$ by inverting the generating function H (for example, by applying the Fast Fourier Transform). However, Equation (1.7) can be reduced to a generalised Wiener-Hopf equation which cannot be solved in general. To see this, assume that W_1 and X have densities f_{W_1} and f_X on $(0,\infty)$. Under these

assumptions, we see from (1.7) that H has a derivative h on $(0,\infty)$; therefore, by differentiating with respect to x, (1.7) yields

$$\begin{aligned} h(r,x) &= f_{W_1}(x) + rH(r,0)f_X(x) + r\int_x^\infty h(r,y-x)f_X(y)\,\mathrm{d}y \\ &= f_{W_1}(x) + rH(r,0)f_X(x) + r\int_0^\infty h(r,u)f_X(u+x)\,\mathrm{d}u. \end{aligned} \tag{1.8}$$

Notice that $h(r,x) = \int_0^\infty h(r,y)\delta(x-y)\,\mathrm{d}y$, with δ being the Dirac δ-function. Combining this with (1.8), we obtain that

$$\int_0^\infty h(r,y)\big[\delta(x-y) - rf_X(x+y)\big]\,\mathrm{d}y = rH(r,0)f_X(x) + f_{W_1}(x). \tag{1.9}$$

This equation is equivalent to a generalised Wiener-Hopf equation; see Noble [135, p. 233]. It is shown there that such equations can sometimes be solved, but a general solution, as is possible for the classical Wiener-Hopf problem (arising in Lindley's recursion), seems to be absent; see also a discussion in Section 5.2, where we shall derive a generalised Wiener-Hopf equation for the distribution of W in the M/G model.

The fact that we are dealing with a generalised Wiener-Hopf equation could indicate that deriving the distribution of W_n or W for the alternating service model may be more complicated than for Lindley's recursion. One point we make in this dissertation is that this is not necessarily the case.

The integral equation (1.7) has the following property, which is proven to be valuable in overcoming the difficulties arising by the fact that we are dealing with a generalised Wiener-Hopf equation. Consider the space $\mathcal{L}^\infty\big([0,\infty)\big)$, i.e., the space of measurable and bounded functions on the real line with the norm

$$\|F\| = \sup_{x\geqslant 0}|F(x)|.$$

In this space we define the mapping $\mathcal{T}_r$ by (cf. (1.7))

$$(\mathcal{T}_rF)(x) = \mathbb{P}[W_1 \leqslant x] + \frac{r}{1-r} - r\int_x^\infty F(y-x)\,\mathrm{d}F_X(y).$$

Then for two arbitrary functions F_1 and F_2 in this space we have

$$\begin{aligned} \|(\mathcal{T}_rF_1) - (\mathcal{T}_rF_2)\| &= \sup_{x\geqslant 0}\big|(\mathcal{T}_rF_1)(x) - (\mathcal{T}_rF_2)(x)\big| \\ &= \sup_{x\geqslant 0}\Big|r\int_x^\infty \big[F_2(y-x) - F_1(y-x)\big]\,\mathrm{d}F_X(y)\Big| \\ &\leqslant |r|\sup_{x\geqslant 0}\int_x^\infty \sup_{t\geqslant 0}\big|F_2(t) - F_1(t)\big|\,\mathrm{d}F_X(y) \\ &= |r|\,\|F_1 - F_2\|\sup_{x\geqslant 0}\big(1 - F_X(x)\big) \\ &= |r|\,\|F_1 - F_2\|\,\mathbb{P}[X>0]. \end{aligned}$$

Since $|r| \leqslant 1$ and $\mathbb{P}[X > 0] < 1$, we see that $\mathcal{T}_r$ is a contraction mapping on $\mathcal{L}^\infty\big([0,\infty)\big)$ with contraction coefficient $|r|\mathbb{P}[X > 0]$. Thus, iterating $\mathcal{T}_r$ is an approach in order to obtain H numerically. Summarising the above, we see that we can either try to find H exactly by solving a generalised Wiener-Hopf equation, or numerically by iterating the mapping $\mathcal{T}_r$. If this first step is successful, then one may invert H exactly or numerically (by applying the Fast Fourier Transform) to obtain values for $\mathbb{P}[W_n \leqslant x]$.

Comparison with the G/M/1 queue

There is no explicit expression known so far for the time-dependent distribution of the waiting times in the G/M/1 queue. However, for our model we obtain a very simple and explicit expression for $\mathbb{P}[W_n \leqslant x]$ for the G/M case; see Chapter 4. In our case, this distribution is simply an exponential distribution with mass at the origin. Moreover, it is also very simple to extend these results to the G/PH case. Furthermore, all results are obtained by employing a direct approach, rather than obtaining expressions for the generating function. The reason for this spectacular reduction in complexity is that, if B has a mixed-Erlang distribution, then we can completely describe the system in terms of the evolution of a finite-state Markov chain; for details see Chapter 4.

Not only the time-dependent distribution produces surprising results, but also the steady-state distribution of W yields some striking differences. As we have mentioned, the G/PH model, contrary to the G/PH/1 queue, can indeed be explicitly solved by using Laplace-Stieltjes transforms. The waiting-time distribution in this case is a mixture of Erlang distributions with the same scale parameter for all exponential phases. For the classical G/PH/1 queue, Adan and Zhao [2] show that the waiting-time distribution is a mixture of exponentials with different scale parameters. For a discussion see Chapter 4.

Comparison with the M/G/1 queue

For the M/G/1 queue, although there is no explicit expression for F_W, the Laplace-Stieltjes transform of W can be derived straightforwardly. Moreover, the distribution F_W itself can be rewritten as an infinite sum involving the n-fold convolutions of the residual service time [6]. In our case, however, the M/G model presents the most difficulties. As is the case for the time-dependent distribution with generally distributed service times, also for this case the Laplace-Stieltjes transform of W is the solution of a generalised Wiener-Hopf problem that can not be solved for all distributions of the preparation times. Thus, it is not surprising that not even the transform can be derived explicitly. None of the methods used for Lindley's recursion leads to results in this case. We present a partial solution to this intriguing problem in Chapter 5.

Furthermore, it is also not surprising that, since the derivation of the steady-state distribution of the waiting time is presenting difficulties, it does not seem possible to derive the time-dependent distribution of the waiting times. In comparison, for the M/G/1 queue, the generating function of the Laplace-Stieltjes transform of W_n

conditioned on the number of customers at time zero is completely known; see Cohen [46, Sec. II.4.5].

As will be evident in the sequel, the fact that we are concerned with a generalised Wiener-Hopf problem will naturally generate some additional complexity. However, in specific cases, this model is surprisingly far simpler to analyse than Lindley's recursion. Throughout this monograph, we shall compare the results and the methods used for the alternating service system to the results obtained and the methods used for the analogous case in Lindley's recursion.

1.7 Overview of the thesis

This thesis is organised according to the specific model that is studied. Before making any assumptions on the distribution of the preparation or the service times, we study some general properties of Recursion (1.1) in Chapter 2. The contents of this chapter are based on parts of the work presented in [169] and [174]. In particular, we prove that under some assumptions, there is a unique equilibrium distribution, and the system converges to it. Furthermore, we derive a first crude upper bound for the rate of convergence to the limiting distribution. Moreover, we show that if the limiting distribution of the waiting time of the server is a continuous function on $(0, \infty)$, then it satisfies a contraction mapping, and we retrieve the same upper bound for the rate of convergence to the limiting distribution. We also study the tail asymptotics of W for various classes of distributions of the preparation times, and we derive some properties of the covariance function of the waiting times.

Motivated by the carousel application presented in Section 1.2, in Chapter 3 we review the results obtained by Park *et al.* [140] for the M/U model, and extend these results in two directions. First, we study the PH/U model. In other words, we assume that the service times follow a phase-type distribution, while the preparation times are uniformly distributed and derive the distribution of W. This setting corresponds to uniformly distributed rotation times of the carousels (i.e. the items are randomly located on it) and phase-type distributed pick times. This analysis has already been presented in [173]. Later on, we study the M/P model, where "P" stands for polynomial distributions, and derive the distribution of W. Polynomial distributions are a useful class of distributions since they can approximate any continuous distribution on a bounded support arbitrarily closely. The extension to polynomial distributions is based on the work concluded in [171].

In Chapter 4 we study the G/PH model. This chapter contains almost the complete contents of [170] and [174]. We first discuss the time-dependent distribution of the waiting times and then derive the limiting distribution of W. Furthermore, we drop the assumption that the server is obliged to alternate and derive the time-dependent and the limiting distribution of the waiting times of the server for this model too. We conclude by comparing the two models analytically and numerically.

In Chapter 5 we attempt to derive the steady-state distribution of the M/G model. Although many properties of the M/G/1 queue are known, in our case the M/G model presents the most difficulties. We show again that the derivation of the

steady-state distribution can be formulated as a generalised Wiener-Hopf problem, and we explain why none of the methods presented in the previous chapters or the methods applied to the M/G/1 queue leads to results in this case. In order to obtain reasonable results, we introduce a new class of distributions, in which the derivation of an explicit expression for F_W is possible. The analysis in this chapter is based on parts of [169].

Stimulated by the observation that a wide range of distributions does not belong to the class of distributions introduced in the previous chapter, in Chapter 6 we discuss how we can use the results obtained so far in order to approximate the distribution of W in case that exact computations are not possible. We obtain an upper bound for the incurred approximation error, as presented in [171], and discuss the scenario when the preparation times follow a distribution of a mixed type, i.e., a mixture of a discrete and a continuous distribution.

In Chapter 7 we no longer assume that the preparation times and the service times are two mutually independent sequences of independent and identically distributed random variables. In order to introduce some specific relation between service and preparation times, we turn to the carousel problem and describe an application that leads to dependencies between a consecutive preparation and service time. We analyse this model and study how the different picking strategies affect the waiting time of the picker.

Chapter 8 attempts to bridge the distance between Lindley's recursion and the recursion studied in this thesis. In other words, we study a recursion that has both Lindley's recursion and Recursion (1.1) as special cases. For this system we derive the distribution of the waiting time of the server in case A is generally distributed and B follows a phase-type distribution, and in case A is exponentially distributed and B is deterministic. As we observe in this chapter, the analysis clearly demonstrates the effects both of Lindley's recursion and of Recursion (1.1). The results derived in this chapter can be found in [36].

As mentioned before, we compare all results obtained to the analogous cases for Lindley's recursion. Furthermore, we try to develop an intuitive perception of the results obtained by displaying graphs and numerical results when these are worth mentioning. We present our conclusions and further remarks in Chapter 8.4.

Chapter 2
General properties

2.1 Introduction

In this chapter we study various general properties of Recursion (1.1). These properties form the basis of the more detailed analysis that follows in the thesis. Throughout this chapter we do not need to assume that the sequences of the preparation and the interarrival times are independent of one another. Moreover, we make no assumptions on the exact distribution of A_n or B_n. We see these random variables aggregated in the random variables $X_{n+1} = B_{n+1} - A_n$, for which we assume that they are independent of one another and identically distributed. Sections 2.2 to 2.5 are mainly based on the work presented in [169], while Section 2.3 is presented in [174].

We should first note that, evidently, $\{W_n\}$ is a regenerative process, which in intuitive terms means that the process can be split into cycles that are mutually independent and stochastically equal to one another. The process regenerates every time a zero waiting time occurs. We denote by C the random variable expressing the length of such a cycle. Furthermore, $\{W_n\}$ may be possibly delayed; that is, the process might not start from a regeneration point. Therefore, the time until the first regeneration will have a different distribution than the one of C. We denote the first cycle length by C_1.

In the following, we shall review the subjects we are concerned with in this chapter.

Stability

Naturally, the first property we are concerned with is the stability of the system. In particular, in Section 2.2 we study under which conditions the waiting times W_n converge to a limiting generic waiting time W. We show that there is a unique limiting distribution F_W, and that the system converges to it geometrically fast.

We distinguish between two cases. In the first case, we assume that $\mathbb{P}[X_n < 0] > 0$. This condition implies that, with positive probability, the preparation time will be for some customer less than the service time of the previous customer, which will then cause a zero waiting time for the server. For this case, the process is aperiodic and can be shown to have a finite mean cycle length. Thus, the existence of a limiting distribution and the convergence of the system to it is a direct consequence of standard limiting theorems for regenerative processes.

In a G/G/1 queuing system with arrival rate $1/\mathbb{E}[A]$ and mean service time $\mathbb{E}[B]$ the stability condition that is necessary so that W_n converges in distribution to W is that the *occupation rate* (usually denoted by ρ) is less than the server's capacity, that is taken to be equal to one. In other words, $\mathbb{E}[B] < \mathbb{E}[A]$; see, for example,

Asmussen [6, Section III.6]. As we see, the stability condition for Lindley's recursion is stronger, since $\mathbb{E}[B] < \mathbb{E}[A]$ implies that $\mathbb{P}[B - A < 0] > 0$.

In the second case, we assume that $\mathbb{P}[X_n < 0] = 0$. This case is even more interesting mathematically. The theorems used for the previous case are no longer applicable, since one cannot argue directly that the process is aperiodic with a finite mean cycle length. For this case, we examine the conditions under which a limiting distribution exists and we derive a geometric bound for the rate of convergence of the process to the limiting distribution.

Contraction

As we have seen in Section 1.6, the integral equation for the generating function H that determines the distribution of W_n for our system can be reduced to a generalised Wiener-Hopf equation. Due to this fact, one expects that determining exactly the distribution of W may be a non-trivial task in general. In Section 2.3, however, we prove that, as is the case for H, the distribution F_W is the fixed point of a functional equation that is a contraction, provided that F_X is continuous. The proof is similar to the one we have presented in Section 1.6. The analogous functional equation for Lindley's equation is not a contraction mapping. The analysis of (1.2) is greatly simplified at times because of this property. A direct conclusion is that the distributions of W_n converge to the distribution of W geometrically fast also for the case where $\mathbb{P}[X_n < 0] > 0$.

Tail asymptotics

In Section 2.5 we shall study the tail asymptotics of W for various classes for the distribution of the preparation times. To do so, we first give all necessary definitions and results in Section 2.4. For our problem we show that, for the cases we examine, the tail of W behaves asymptotically like the tail of X.

The tail behaviour of the single-server queue is a well-studied subject in the literature. A comprehensive register of the different theorems on this subject can be found in Korshunov [106]. For Lindley's recursion, the study of the problem is split in three regimes. The tail of W is either exponential when B is light-tailed and satisfies a condition called the "Cramér condition" (cf. Asmussen [6, Section XIII.5]), or it behaves as the tail of the stationary excess distribution of B (cf. Asmussen [6, Section X.9]) if the excess distribution of B is subexponential [59] (which in intuitive terms means that large values of a sum of i.i.d. random variables are most likely caused by a large value of a single term rather than the joint effect of a set of them, as is the case for light-tailed distributions). There exist results also for the case where B is light-tailed but does not satisfy the Cramér condition; see again [106].

Covariance function

If one cannot determine the limiting distribution function of the waiting time analytically, but wants to obtain an estimation from a simulation of the system, then two relevant questions are

1. how long should the simulation run and
2. how big is the variance of the mean of a sample of successive waiting times.

For the determination of the magnitude of this variance it is necessary and sufficient to know the covariance function of the process $\{W_n\}$. We study some properties of the covariance function in Section 2.6. We show that the covariance function between two waiting times is of an alternating sign, and we obtain an upper bound for its absolute value, from which we conclude that the covariance function converges to zero geometrically fast. An explicit expression of the covariance function for the G/PH model will be given in Chapter 4.

The literature on the covariance function of the waiting times for the single-server queue seems to be sporadic. For the M/G/1 queuing system Blomqvist [23] studies the covariance function of the waiting times. Beneš [15] considers a slightly different notion of waiting times. He studies the general properties of the covariance function of the process $W(t)$ of virtual waiting times, where $W(t)$ is the waiting time a customer would encounter if he arrived at time t. For the G/M/1 queue Pakes [139] derives results analogous to those in [23].

Both studies of Blomqvist [23] and Pakes [139] involve the generating functions of the autocorrelations of the waiting times in terms of the probability generating function of the distribution of the number of customers served in a busy period. The latter functions are only implicitly determined as solutions to functional equations. Blanc [22] is concerned with the numerical inversion of the generating functions of the autocorrelations of the waiting times, as they are given in [23] and [139].

For the G/G/1 queue, Daley [53] and Blomqvist [24, 25] give some general properties. In particular, in [53] it is shown that the serial correlation coefficients of a stationary sequence of waiting times are non-negative and decrease monotonically to zero. Further results are obtained for the M/G/1 and M/M/1 queues. In [24] the mean square error of the average waiting times is determined asymptotically by an expression involving the covariance function of the stationary waiting-time process, while in [25] the author gives two heavy-traffic limits for the serial autocorrelation coefficients.

For the single server queue, a few other papers study problems related to the correlation coefficients of various series of random variables. For example, for the M/M/1 queue, Morse [132] investigates the time-continuous $n(t)$-process, where $n(t)$ is the number of customers in the system at time t. For this process, he also obtains the autocorrelation function and studies its properties. From this study, the derivation of an approximation for the relaxation time of the queue length is straightforward, while the standard error of the average queue length is determined by Gebhard [67]. In a recent study, Blanc [21] numerically computes the autocorrelations of interdeparture times.

2.2 Stability

Naturally, the first property we are concerned with is the stability of the system. In this section, we study the convergence of the process $\{W_n\}$ to a limiting waiting time W. We consider two separate cases.

2.2.1 The case $\mathbb{P}[X < 0] > 0$

For the existence of a unique equilibrium distribution, one should note that the stochastic process $\{W_n\}$ is a (possibly delayed) regenerative process with the time points where $W_n = 0$ being the regeneration points. Since $\mathbb{P}[X_n < 0] > 0$, the process is moreover aperiodic. In order to show that the process has a finite mean cycle length define the stopping time $\tau = \inf\{n : W_1 = W_{1+n} = 0\}$, and observe that

$$\mathbb{P}[\tau > n] \leqslant \mathbb{P}[X_k > 0 \text{ for all } k = 2, \dots, n] = \mathbb{P}[X_2 > 0]^{n-1},$$

and $\mathbb{P}[X_2 > 0] < 1$ because of the stability condition we have imposed. Therefore, from the standard theory on regenerative processes it follows that the limiting distribution exists and the process converges to it in total variation; see for example Corollary VI.1.5 or Theorem VII.3.6 in Asmussen [6]. For the application of Theorem VII.3.6 in [6, p. 202] one simply needs to notice that since $\{0\}$ is a regeneration set of the process, $\{W_n\}$ is a Harris chain.

2.2.2 The case $\mathbb{P}[X < 0] = 0$

In the previous case we have examined, the condition that $\mathbb{P}[X < 0] > 0$ guaranteed that the cycle-length distribution is aperiodic and has a finite mean. These statements prove the existence of a total variation limit of the process. However, if we remove this condition, then the above statements do not hold in general, and thus the stability of the system cannot be established by the previously mentioned theorems. In this section, we shall discuss the existence of a limiting distribution and the convergence of the system to it in case $\mathbb{P}[X < 0] = 0$.

In order to prove that there is a unique equilibrium distribution for this case, we need to address three issues: the existence of an invariant distribution, the uniqueness of it and the convergence to it, irrespective of the state of the system at zero. Recall that throughout the thesis we use the following notation: for a random variable Y and an event E we have that $\mathbb{P}[Y \leqslant x\,; E] = \mathbb{E}[\mathbb{1}_{[Y \leqslant x]} \cdot \mathbb{1}_{[E]}]$.

Existence

To prove the existence of an equilibrium distribution, we first give the definition of *tightness*.

Definition 2.1

A sequence ν_n, $n \geqslant 1$, of probability measures on $\mathbb{R}^+$ is said to be *tight* if for every $\epsilon > 0$ there is a number $M < \infty$ such that $\nu_n[0, M] \geqslant 1 - \epsilon$, for all n.

In other words, almost all the mass of each measure is included in a compact set.

Consider now the recursion $W_{n+1} = \max\{0, X_{n+1} - W_n\}$, where $\{X_n\}$ is an i.i.d. sequence of almost surely finite random variables. Let $W_1 = w$ and $M \geqslant w$. Then, since $W_{n+1} \leqslant \max\{0, X_{n+1}\}$ for all $n \geqslant 1$, we have that

$$\mathbb{P}[W_{n+1} \leqslant M] \geqslant \mathbb{P}[\max\{0, X_{n+1}\} \leqslant M] = \mathbb{P}[\max\{0, X_2\} \leqslant M].$$

So we can choose M to be the maximum of w and the $1-\epsilon$ quantile of $\max\{0, X_2\}$. Thus, the sequence $\mathbb{P}[W_n \leqslant x]$ is tight.

Moreover, since the function $g(w, x) = \max\{0, x - w\}$ is continuous in both x and w, the existence of an equilibrium distribution is a direct application of Theorem 4 of Foss and Konstantopoulos [65]. Both the probability space Ω and the Polish space $\mathcal{X}$ that are mentioned in this theorem can be substituted by $\mathbb{R}$ in our case, and the shift (mapping) Θ can be taken to be the function $g(w, x) = \max\{0, x - w\}$ given here. So there exists an almost surely finite random variable W, such that

$$W \stackrel{\mathcal{D}}{=} \max\{0, X_2 - W\}.$$

Uniqueness

Before proving the uniqueness of the equilibrium distribution and the convergence of the process to it, we shall construct a random time that will be useful in proving both results. To do so, along with the assumptions that $\{X_n\}_{n \geqslant 2}$ is an i.i.d. sequence of almost surely finite random variables distributed as X and $\mathbb{P}[X < 0] = 0$ we shall need the additional assumption that X is non-deterministic.

Since X is non-deterministic, we have that there exist constants $\epsilon \in \mathbb{R}^+$, $n \in \mathbb{N}$, such that $\mathbb{P}[X \geqslant (n+1)\epsilon] > 0$ and $\mathbb{P}[X \leqslant n\epsilon] > 0$. Consider now the event

$$E_{n,i} = \{X_i \leqslant n\epsilon; X_{i+1} \geqslant (n+1)\epsilon; X_{i+2} \leqslant n\epsilon; \ldots; X_{i+2n-1} \geqslant (n+1)\epsilon; X_{i+2n} \leqslant n\epsilon\};$$

since the random variables X_i are i.i.d., we shall ignore the second index whenever this is of no consequence. We have that

$$\mathbb{P}[E_n] = \mathbb{P}[X \geqslant (n+1)\epsilon]^n \mathbb{P}[X \leqslant n\epsilon]^{n+1} = q > 0.$$

Consequently, if $E_{n,i}$ occurs, then we have that

$$\begin{aligned} W_i &\leqslant \max\{0, X_i\} \leqslant n\epsilon, \\ W_{i+1} &= X_{i+1} - W_i \geqslant \epsilon, \\ W_{i+2} &= \max\{0, X_{i+2} - W_{i+1}\} \leqslant (n-1)\epsilon, \\ W_{i+3} &= X_{i+3} - W_{i+2} \geqslant 2\epsilon \end{aligned}$$

and so on. That is, for $k = 0, \ldots, n-1$,

$$W_{i+2k} \leqslant (n-k)\epsilon \qquad \text{and} \qquad W_{i+2k+1} \geqslant (k+1)\epsilon.$$

Thus, on $E_{n,i}$ we have that $W_{i+2n} = 0$. Notice that on $E_{n,i}$ this result holds irrespective of the value of W_{i-1}.

Define now the hitting time

$$\tau_{E_n} = \inf\{\ell \geqslant 2 : X_\ell \leqslant n\epsilon\,; X_{\ell+1} \geqslant (n+1)\epsilon\,; \ldots; X_{\ell+2n} \leqslant n\epsilon\}.$$

We shall prove the following proposition.

Proposition 2.1. *For $k \geqslant 1$, $\mathbb{P}[\tau_{E_n} \geqslant (2n+1)k] \leqslant (2n+1)(1-q)^k$.*

Proof. In order for the event $\tau_{E_n} = j$ to happen, we should have that all events $E_{n,i}$ did not occur for all $i = 2, \ldots, j-1$, while $E_{n,j}$ did occur. Let $E^c_{n,i}$ denote the complement of the event $E_{n,i}$. Then, by conditioning we have that

$$\begin{aligned}\mathbb{P}[\tau_{E_n} \geqslant (2n+1)k] &= \sum_{i=0}^{\infty} \mathbb{P}[\tau_{E_n} \geqslant (2n+1)k; E^c_{n,2}; \ldots; E^c_{n,(2n+1)k+i-1}; E_{n,(2n+1)k+i}] \\ &= \sum_{i=0}^{\infty} \mathbb{P}[E^c_{n,2}\,; \ldots; E^c_{n,(2n+1)k+i-1}\,; E_{n,(2n+1)k+i}].\end{aligned}$$

Since $E_{n,i}$ is not independent from $E_{n,j}$ for all $j = i, \ldots, i+2n$, we shall bound the above probability by discarding a number of events so that the remaining ones are independent from one another. Specifically, we keep the event $E^c_{n,2}$, discard the next $2n$ events, keep $E^c_{n,2n+3}$, and so on. In every probability appearing in the sum above, the last two terms we keep are the events

$$E^c_{n,\left[\frac{(2n+1)k+i}{2n+1}\right]-1} \qquad \text{and} \qquad E_{n,(2n+1)k+i},$$

where $[i]$ denotes the integer part of i. Thus,

$$\begin{aligned}&\mathbb{P}[\tau_{E_n} \geqslant (2n+1)k] \\ &\quad\leqslant \sum_{i=0}^{\infty} \mathbb{P}[E^c_{n,2}\,; E^c_{n,(2n+1)+2}\,; \ldots; E^c_{n,(2n+1)\ell+2}, \ell = 0, \ldots, \left[\frac{(2n+1)k+i}{2n+1}\right] - 1; \\ &\qquad\qquad E_{n,(2n+1)k+i}] \\ &\quad= q\sum_{i=0}^{\infty}(1-q)^{\left[\frac{(2n+1)k+i}{2n+1}\right]} = q(1-q)^k \sum_{i=0}^{\infty}(1-q)^{\left[\frac{i}{2n+1}\right]} = (2n+1)(1-q)^k.\end{aligned}$$

□

So far we have that that if X is non-deterministic, there is an event E_n depending only on the sequence $\{X_i\}$, which occurs with positive probability, and which guarantees that the last time associated with this event will produce a zero waiting time. Naturally, the process may reach zero before this time, but the important point here is that we can actually construct such a time. The *coupling time*, that is, the random time after which two processes will coincide, we now use is the

time $\tau = \tau_{E_n} + 2n$. From the above proposition we shall conclude that the rate of convergence to the equilibrium distribution has a geometric bound.

To prove the uniqueness of the equilibrium distribution we assume that there are two solutions W^1, W^2, such that for $i = 1, 2$, we have that $W^i \stackrel{\mathcal{D}}{=} \max\{0, X - W^i\}$. In order to show that W^1 and W^2 have the same distribution, we shall first construct two sequences of waiting times that converge to W^1 and W^2. These sequences are given by $W^i_{n+1} = \max\{0, X_{n+1} - W^i_n\}$ for $i = 1, 2$, where for every n, X_n is equal in distribution to X and $W^i_1 \stackrel{\mathcal{D}}{=} W^i$. Therefore, $\{W^i_n\}$, $i = 1, 2$, is a stationary sequence. Recall that the events $E_{n,i}$ depend only on the sequence $\{X_i\}$. Since the sequences are generated by the same sequence $\{X_i\}$, an event E_n will occur at the same time for both processes. Thus, after some finite time equal to $\tau = \tau_{E_n} + 2n$ both processes simultaneously reach zero, and afterwards they coincide. This implies that they have the same invariant distribution.

Convergence

We need to show that a system that does not start in equilibrium will eventually converge to it. To achieve this, we will compare two systems that are identical, apart from the fact that one of them does not start in equilibrium while the other one does. To this end, for $i = 1, 2$ let the process $\{W^i_n\}$ satisfy the recursion $W^i_{n+1} = \max\{0, X_{n+1} - W^i_n\}$, where where W^1_1 is *not* distributed as W while for every $n \geqslant 1$, $W^2_n \stackrel{\mathcal{D}}{=} W$. As before, we observe that since the events $E_{n,i}$ guarantee that $W_{i+2n} = 0$ irrespective of W_{i-1}, the processes couple after τ. By using this coupling time we readily have from Proposition 2.1 a geometric bound of the rate of convergence to the limiting distribution.

As it is made evident from the above, the conditions that either $\mathbb{P}[X < 0] > 0$ or that X is non-deterministic are crucial in order to prove the uniqueness of the steady-state distribution and the convergence of the process to this distribution. To see the effect of this condition, assume that, for every n, $B_n = 5$ and $A_n = 1$. Therefore, for every n, $X_n = 4$, i.e., strictly positive and deterministic. Furthermore, assume that the first waiting time is equal to the first preparation time, i.e., $W_1 = B_1 (= 4)$. This last assumption is not restrictive at all; it is made only for symmetry reasons and one can start with any other deterministic value for the first waiting time. The process will only reach equilibrium slightly later. In the situation we describe, we have that every odd-numbered waiting time is equal to four and every even-numbered waiting time is equal to zero. Therefore, the limit $\lim_{n\to\infty} \mathbb{P}[W_n \leqslant x]$ is not defined for all values of x.

The above results on uniqueness and convergence are fairly general since we did not have to impose any conditions on the distributions of X or W. Nonetheless, it only proves that there exists a unique invariant distribution F_W; that is, there is only one element of the class of all distribution functions that satisfies Equation (1.2). If we demand though that F_W is continuous, then we can, in fact, prove more. We can expand the class of distributions to the class of measurable bounded functions and

prove that the solution of (1.2) is still unique. The proof of this is the subject of the following section. For the remainder of the thesis we assume that $\mathbb{P}[X < 0] > 0$.

2.3 The recursion revisited

The aim of this section is to examine the set of functions that satisfy (1.2). Note, first, that for $x \geqslant 0$ Equation (1.2) yields that $F_W(x) = \mathbb{P}[W \leqslant x] = \mathbb{P}[X - W \leqslant x]$. Assuming that F_X is continuous, then the last term is equal to $1 - \mathbb{P}[X - W \geqslant x]$, which gives us that

$$F_W(x) = 1 - \int_x^\infty \mathbb{P}[W \leqslant y - x] \, \mathrm{d}F_X(y) = 1 - \int_x^\infty F_W(y - x) \, \mathrm{d}F_X(y).$$

This means that the invariant distribution of W, provided that F_X is continuous, satisfies the functional equation

$$F(x) = 1 - \int_x^\infty F(y - x) \, \mathrm{d}F_X(y). \tag{2.1}$$

Therefore, there exists at least one function that is a solution to (2.1). The question remains though whether there exist other functions, not necessarily distributions, that satisfy (2.1). The following theorem clarifies this matter.

Theorem 2.2. *There is a unique measurable bounded function $F : [0, \infty) \to \mathbb{R}$ that satisfies the functional equation*

$$F(x) = 1 - \int_x^\infty F(y - x) \, \mathrm{d}F_X(y).$$

Proof. Let us consider the space $\mathcal{L}^\infty\big([0, \infty)\big)$, i.e. the space of measurable and bounded functions on the real line with the norm

$$\|F\| = \sup_{t \geqslant 0} |F(t)|.$$

In this space we define the mapping

$$(\mathcal{T}F)(x) = 1 - \int_x^\infty F(y - x) \, \mathrm{d}F_X(y). \tag{2.2}$$

Note that $\mathcal{T}F : \mathcal{L}^\infty\big([0, \infty)\big) \to \mathcal{L}^\infty\big([0, \infty)\big)$, i.e., $\mathcal{T}F$ is measurable and bounded.

For two arbitrary functions F_1 and F_2 in this space we have

$$\begin{aligned}
\|(\mathcal{T}F_1) - (\mathcal{T}F_2)\| &= \sup_{x\geqslant 0}\big|(\mathcal{T}F_1)(x) - (\mathcal{T}F_2)(x)\big| \\
&= \sup_{x\geqslant 0}\Big|\int_x^\infty \big[F_2(y-x) - F_1(y-x)\big]\mathrm{d}F_X(y)\Big| \\
&\leqslant \sup_{x\geqslant 0}\int_x^\infty \sup_{t\geqslant 0}\big|F_2(t) - F_1(t)\big|\,\mathrm{d}F_X(y) \\
&= \|F_1 - F_2\|\,\sup_{x\geqslant 0}\big(1 - F_X(x)\big) \\
&= \|F_1 - F_2\|\,\big(1 - F_X(0)\big).
\end{aligned}$$

Thus,

$$\|(\mathcal{T}F_1) - (\mathcal{T}F_2)\| \leqslant \|F_1 - F_2\|\,\big(1 - F_X(0)\big) = \|F_1 - F_2\|\,\mathbb{P}[B > A].$$

Since $\mathbb{P}[X < 0] > 0$, i.e., $\mathbb{P}[B > A] < 1$ we have a contraction mapping. Furthermore, we know that $\mathcal{L}^\infty\big([0,\infty)\big)$ is a Banach space, therefore by the Banach Fixed Point Theorem [90] we have that (2.1) has a unique solution. □

The set of continuous and bounded functions on $[0,\infty)$ with the norm $\|F\| = \sup_t |F(t)|$ is also a Banach space, since it is a closed subspace of $\mathcal{L}^\infty([0,\infty))$. Since F_W, in case it is continuous, is a solution to Equation (2.1), we have the following corollary.

Corollary 2.3. *The only function satisfying Equation* (2.1) *that is continuous and in* $\mathcal{L}^\infty([0,\infty))$ *is the* unique *limiting distribution* F_W.

One should also note the usefulness of the above result in calculating numerically the invariant distribution. Since we have a contraction mapping, we can evaluate the distribution of W by successive iterations. One can start from some (trivial) distribution and substitute it into the right-hand side of (2.1). This will produce the second term of the iteration, and so on. Furthermore, this iterative approach gives us the distribution of W_n for a given distribution for W_1. Note that we also computed a geometric upper bound for the rate of convergence to the invariant distribution, namely the probability $\mathbb{P}[X > 0]$. We shall utilise this idea of successive iterations in Chapter 7.

So far we know that a limiting distribution of the waiting time exists, and that it can be approximated by successive iterations that will converge to the real distribution geometrically fast. If, however, the knowledge of the full distribution of W is not necessary, one can still estimate the shape of it by examining its tail behaviour. In Section 2.5 we shall discuss the tail behaviour of F_W under various assumptions on the random variable B. To do so though, we will first introduce the notions of regular and rapid variation, and of heavy-, light-, or long-tailed distribution functions in the following section.

2.4 Classification of distribution functions

In this section we introduce some basic definitions concerning (distribution) functions. Although for most of the definitions we present, there are alternative or equivalent statements, here we shall present only one definition for each case. For the scope of this section, let Y be a random variable with distribution function F. We start from the definition of a *heavy-tailed distribution.*

Definition 2.2
A distribution F on $(0, \infty)$ is called *heavy-tailed* if for all $\epsilon > 0$

$$\int_0^\infty \mathrm{e}^{\epsilon x} \, \mathrm{d}F(x) = \infty. \tag{2.3}$$

In other words, a distribution is heavy-tailed if it has no finite exponential moments. Consequently, we call a distribution *light-tailed* if it does have at least one exponential moment. A subclass of the class of heavy-tailed distributions is the class of *long-tailed distributions*, defined as follows.

Definition 2.3
The distribution F of the random variable Y is called *long-tailed* if for any $y > 0$

$$\lim_{x\to\infty} \mathbb{P}[Y > x + y \mid Y > x] = \lim_{x\to\infty} \frac{1 - F(x + y)}{1 - F(x)} = 1.$$

One should note here that there is no universally-accepted definition for heavy-tailed distributions [130]. For example, in Sigman [151] Definition 2.3 is given as the definition of a heavy-tailed distribution, while the term "long-tailed distribution" is not mentioned. As noted in [151], (2.3) is a property of long-tailed distributions, i.e. if F is long tailed, then it satisfies (2.3). The definition chosen for heavy-tailed distributions, as well as all other definitions in this section, can be found in Zwart [183].

Next we give the definition of a *regularly varying function.* A comprehensive account of the theory and applications of regular variation is given by Bingham, Goldie and Teugels [19].

Definition 2.4
A measurable function $f > 0$ is called *regularly varying* of a finite index κ if for all $\ell > 0$ f satisfies

$$\lim_{x\to\infty} \frac{f(\ell x)}{f(x)} = \ell^\kappa.$$

In particular, we call the random variable Y regularly varying, if its tail behaves almost like a power law. In other words, Y is regularly varying if $\mathbb{P}[Y > x]$ is regularly varying with index $-\kappa$, $\kappa > 0$. This implies that

$$\mathbb{P}[Y > x] = L(x)x^{-\kappa},$$

where $L(x)$ is a slowly varying function, i.e. for all $\ell > 0$, $L(\ell x)/L(x) \to 1$, as $x \to \infty$.

Another notion closely related to regular variation is the *domain of attraction of a stable law*. We say that the distribution function F of the random variable Y is in the domain of attraction of a stable law with exponent $\kappa < 2$ if the normalised sum of n independent copies of Y converges in distribution to Y. The following theorem of Breiman [40] gives an equivalent definition that is related to Definition 2.4.

Theorem 2.4 (Breiman [40, p. 207])**.** *The distribution function F is in the domain of attraction of a stable law with exponent $\kappa < 2$ if and only if there are constants $a, b \geqslant 0$, $a + b > 0$, such that*

1. $$\lim_{x\to\infty} \frac{F(-x)}{1 - F(x)} = \frac{a}{b}$$
2. *For all $\ell > 0$,*
$$b > 0 \Rightarrow \lim_{x\to\infty} \frac{1 - F(\ell x)}{1 - F(x)} = \frac{1}{\ell^\kappa}$$
$$a > 0 \Rightarrow \lim_{x\to\infty} \frac{F(-\ell x)}{F(-x)} = \frac{1}{\ell^\kappa}$$

The definition of regular variation demands that the index κ appearing there is finite. This definition can be extended to include the case that κ is equal to infinity, leading to the notion of *rapid variation.* For the purposes of this thesis we shall consider only the case when κ is equal to minus infinity.

Definition 2.5

A measurable function $f : (0, \infty) \to (0, \infty)$ is *rapidly varying* of index $-\infty$ if it satisfies

$$\lim_{x\to\infty} \frac{f(\ell x)}{f(x)} = \begin{cases} 0, & \text{if } \ell > 1; \\ 1, & \text{if } \ell = 1; \\ \infty, & \text{if } 0 < \ell < 1. \end{cases}$$

As before, in case of distributions, the function f in the above definition can be considered to be the distribution tail, which is then extremely light tailed, i.e. lighter than any exponential tail. As it is noted in [19], some of the standard theorems for regularly varying functions have partial analogues for rapid variation.

By convention, we call a random variable regularly varying, heavy-, long-, or light-tailed, if its distribution has the corresponding property. Furthermore, we shall also write "$f \sim g$" when

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1.$$

We can now proceed with studying the tail behaviour of the distribution of W.

2.5 Tail behaviour

We are interested in the tail asymptotics of W. In other words, we would like to know when we can estimate the probability that W exceeds some large value x by using only information from the given distributions of A and B. This information is relevant when, for example, the distribution of W cannot be computed exactly, or when no further knowledge on the distribution is necessary.

Suppose that for some finite constant $\kappa \geqslant 0$

$$\lim_{x\to\infty} \frac{\mathbb{P}[B > x + y]}{\mathbb{P}[B > x]} = \mathrm{e}^{-\kappa y}.$$

Then

$$\lim_{x\to\infty} \frac{\mathbb{P}[\mathrm{e}^B > \mathrm{e}^x \cdot \mathrm{e}^y]}{\mathbb{P}[\mathrm{e}^B > \mathrm{e}^x]} = (\mathrm{e}^y)^{-\kappa},$$

which means that e^B is regularly varying with index $-\kappa$. For the random variable B this means that if $\kappa = 0$, then B is long-tailed, and thus, in particular, heavy-tailed. If $\kappa > 0$, then B is light-tailed, but not lighter than an exponential tail. The above expressions are also useful to demonstrate the connection that exists between regularly-varying and long-tailed or light-tailed random variables.

In order to study the tail behaviour of W we shall need the following proposition that is first shown by Breiman [39].

Proposition 2.5 (Breiman [39, Proposition 3])**.** *Let the random variable $Y \geqslant 0$ be in the domain of attraction of a stable law with exponent $0 < \kappa < 1$ and let Z be a random variable independent of Y with $\mathbb{E}|Z|$ finite. Then YZ is in the domain of attraction of a stable law with the same exponent.*

This result has been further refined by Cline and Samorodnitsky [45] as follows.

Proposition 2.6 (Cline and Samorodnitsky [45, Corollary 3.6])**.** *If $Y > 0$ is a regularly varying random variable with index $-\kappa$, $\kappa \geqslant 0$, and $Z > 0$ is independent of Y with $\mathbb{E}[Z^{\kappa+\epsilon}]$ finite for some $\epsilon > 0$, then YZ is regularly varying with index $-\kappa$. In particular*

$$\mathbb{P}[Y \cdot Z > x] \sim \mathbb{E}[Z^\kappa]\mathbb{P}[Y > x].$$

In order to derive the tail asymptotics of W observe that from Equation (1.2) we have that

$$\mathbb{P}[W > x] = \mathbb{P}[B - (W + A) > x]$$

which implies that

$$\mathbb{P}[\mathrm{e}^W > \mathrm{e}^x] = \mathbb{P}[\mathrm{e}^B \mathrm{e}^{-(W+A)} > \mathrm{e}^x]. \tag{2.4}$$

Applying now Proposition 2.6 to the right-hand side of (2.4) we have that

$$\mathbb{P}[\mathrm{e}^W > \mathrm{e}^x] \sim \mathbb{P}[\mathrm{e}^B > \mathrm{e}^x]\mathbb{E}[\mathrm{e}^{-\kappa(W+A)}]$$

or

$$\mathbb{P}[W > x] \sim \mathbb{P}[B > x]\mathbb{E}[\mathrm{e}^{-\kappa W}]\mathbb{E}[\mathrm{e}^{-\kappa A}].$$

In other words, the tail of W behaves asymptotically as the tail of B, multiplied by a constant. One can write the above result in terms of the tail of X. It suffices to note that

$$\mathbb{P}[X > x] = \mathbb{P}[B - A > x] = \mathbb{P}[\mathrm{e}^{B}\mathrm{e}^{-A} > \mathrm{e}^{x}],$$

and since e^{B} is regularly varying with index $-\kappa$ we have from Proposition 2.6 that the above expression is asymptotically equal to $\mathbb{P}[B > x]\mathbb{E}[\mathrm{e}^{-\kappa A}]$. The above findings are summarised in the following theorem.

Theorem 2.7. *Let e^{B} be regularly varying with index $-\kappa$. Then for the tail of W we have that*

$$\mathbb{P}[W > x] \sim \mathbb{P}[X > x]\,\mathbb{E}[\mathrm{e}^{-\kappa W}].$$

An example of a random variable B that satisfies the conditions of this theorem is the one having asymptotically the tail distribution $\mathbb{P}[B > x] \sim c_0 x^{c_1}\mathrm{e}^{-c_2 x}$, for some real-valued constants c_i, $i = 0, 1, 2$, where $c_0, c_2 > 0$.

Another case that is particularly interesting is when e^{B} is rapidly varying with index $-\infty$, that is

$$\lim_{x\to\infty} \frac{\mathbb{P}[\mathrm{e}^{B} > \mathrm{e}^{x}\cdot\mathrm{e}^{y}]}{\mathbb{P}[\mathrm{e}^{B} > \mathrm{e}^{x}]} = \lim_{x\to\infty} \frac{\mathbb{P}[B > x + y]}{\mathbb{P}[B > x]} = \begin{cases} 0, & \text{if } y > 0; \\ 1, & \text{if } y = 0; \\ \infty, & \text{if } y < 0. \end{cases}$$

This is equivalent to letting the index κ that was given previously go to infinity. For the random variable B this means that B is extremely light tailed. That would be the case if, for example, the tail of B is given by $\mathbb{P}[B > x] = \mathrm{e}^{-x^2}$. As before, we are interested in deriving the asymptotic behaviour of the tail of W in terms of the tail of X. We shall first prove the following lemma.

Lemma 2.1. *If e^{B} is rapidly varying, then also e^{X} is rapidly varying.*

Proof. It suffices to show that for $y > 0$,

$$\lim_{x\to\infty} \frac{\mathbb{P}[X > x + y]}{\mathbb{P}[X > x]} = 0.$$

We have that

$$\frac{\mathbb{P}[X > x + y]}{\mathbb{P}[X > x]} = \frac{\mathbb{P}[B - A > x + y]}{\mathbb{P}[B - A > x]} = \frac{\int_0^\infty \mathbb{P}[B > x + y + z]\,\mathrm{d}F_A(z)}{\int_0^\infty \mathbb{P}[B > x + z]\,\mathrm{d}F_A(z)}. \tag{2.5}$$

Since e^{B} is rapidly varying and $y > 0$, we have that

$$\lim_{x\to\infty} \frac{\mathbb{P}[B > x + y + z]}{\mathbb{P}[B > x + z]} = 0,$$

or in other words, for every $\delta > 0$ there is a finite constant η_δ, such that if $x+z \geqslant \eta_\delta$, then $\mathbb{P}[B > x + y + z] \leqslant \delta \mathbb{P}[B > x + z]$. By taking the limit of (2.5) for x going to infinity, we have that

$$\limsup_{x\to\infty} \frac{\mathbb{P}[X > x + y]}{\mathbb{P}[X > x]} \leqslant \limsup_{x\to\infty} \frac{\delta \int_0^\infty \mathbb{P}[B > x + z] \, \mathrm{d}F_A(z)}{\int_0^\infty \mathbb{P}[B > x + z] \, \mathrm{d}F_A(z)} = \delta,$$

which proves the assertion, since the left-hand side of the above expression is independent of δ, and δ can be chosen to be arbitrarily small. □

To derive the tail asymptotics we shall first decompose the tail of W as follows.

$$\begin{aligned}\mathbb{P}[W > x] &= \mathbb{P}[X - W > x] = \mathbb{P}[X - W > x \,; W = 0] + \mathbb{P}[X - W > x \,; W > 0] \\ &= \mathbb{P}[X > x]\mathbb{P}[W = 0] + \mathbb{P}[X - W > x \,; 0 < W < \epsilon] + \\ &\quad + \mathbb{P}[X - W > x \,; W \geqslant \epsilon], \end{aligned} \tag{2.6}$$

for some $\epsilon > 0$. Since the last two terms of the right-hand side of (2.6) are positive, we can immediately conclude that

$$\liminf_{x\to\infty} \frac{\mathbb{P}[W > x]}{\mathbb{P}[X > x]\mathbb{P}[W = 0]} \geqslant 1.$$

For the upper limit we first observe that

$$\mathbb{P}[X - W > x \,; 0 < W < \epsilon] \leqslant \mathbb{P}[X > x] \, \mathbb{P}[0 < W < \epsilon]$$

and that

$$\mathbb{P}[X - W > x \,; W \geqslant \epsilon] \leqslant \mathbb{P}[X > x + \epsilon] \, \mathbb{P}[W \geqslant \epsilon].$$

Furthermore, since e^X is rapidly varying, we have that for $\epsilon > 0$

$$\lim_{x\to\infty} \frac{\mathbb{P}[X > x + \epsilon]}{\mathbb{P}[X > x]} = 0,$$

or in other words

$$\mathbb{P}[X > x + \epsilon] = o(\mathbb{P}[X > x]).$$

Combining the above arguments we obtain from (2.6) that

$$\limsup_{x\to\infty} \frac{\mathbb{P}[W > x]}{\mathbb{P}[X > x]\mathbb{P}[W = 0]} \leqslant 1 + \frac{\mathbb{P}[0 < W < \epsilon]}{\mathbb{P}[W = 0]}. \tag{2.7}$$

By taking the limit for $\epsilon \to 0$ we have that

$$\limsup_{x\to\infty} \frac{\mathbb{P}[W > x]}{\mathbb{P}[X > x]\mathbb{P}[W = 0]} = 1,$$

since the inequalities in $\mathbb{P}[0 < W < \epsilon]$ are strict and the left-hand side of (2.7) does not depend on ϵ. The above results are summarised in the following proposition.

Theorem 2.8. *Let* e^B *be rapidly varying with index* $-\infty$*. Then for the tail of* W *we have that*

$$\mathbb{P}[W > x] \sim \mathbb{P}[X > x]\,\mathbb{P}[W = 0].$$

In the case when e^B was regularly varying, it was possible to express the tail of W also in terms of the tail of B – instead of the tail of X – simply by applying Breiman's result. In the rapidly-varying case though, this does not seem to be so straightforward. However, in some special situations it is indeed possible to derive the tail of X in terms of the tail of B, and consequently use this form for the tail asymptotics of the waiting time. In the following we shall give one particular example where it is possible to do so.

Assume that A is exponentially distributed with parameter μ and the tail of B is given by $\mathbb{P}[B > x] = \mathrm{e}^{-x^p}$, where $p > 1$. In this example we shall limit ourselves to $p = 2$. However, the extension to the set of natural numbers is almost straightforward. For the tail of X we have that

$$\begin{aligned}\mathbb{P}[X > x] = \mathbb{P}[B - A > x] &= \int_0^\infty \mu \mathrm{e}^{-\mu y} \mathrm{e}^{-(x+y)^2}\,\mathrm{d}y \\ &= \mathrm{e}^{-x^2} \int_0^\infty \mu \mathrm{e}^{-\mu y - y^2} \mathrm{e}^{-2xy}\,\mathrm{d}y \\ &= \mathrm{e}^{-x^2} \frac{1}{x} \int_0^\infty \mu \mathrm{e}^{-\mu \frac{u}{x} - \frac{u^2}{x^2}} \mathrm{e}^{-2u}\,\mathrm{d}u.\end{aligned}$$

Note that the prefactor e^{-x^2} is equal to the tail of B and that the integral at the right-hand side behaves asymptotically like $\mu/2$, as x goes to infinity. In other words, we have that

$$\mathbb{P}[X > x] \sim \mathbb{P}[B > x] \frac{\mu}{2x}.$$

For p being any natural number greater than 1, the procedure is exactly the same. The change of variables that will be the most adequate is $x^{p-1}y = u$. Specifically, we have that

$$\begin{aligned}\mathbb{P}[X > x] &= \int_0^\infty \mu \mathrm{e}^{-\mu y} \mathrm{e}^{-(x+y)^p}\mathrm{d}y \\ &= \mathrm{e}^{-x^p} \int_0^\infty \mu \mathrm{e}^{-\mu y - y^p} \mathrm{e}^{-\sum_{i=1}^{p-1} \binom{p}{i} x^i y^{p-i}}\,\mathrm{d}y \\ &= \mathrm{e}^{-x^p} \frac{1}{x^{p-1}} \int_0^\infty \mu \mathrm{e}^{-\mu \frac{u}{x^{p-1}} - \frac{u^p}{x^{p(p-1)}}} \mathrm{e}^{-\sum_{i=1}^{p-2} \binom{p}{i} \frac{u^{p-i}}{x^{p(p-i-1)}}} \mathrm{e}^{-pu}\,\mathrm{d}u.\end{aligned}$$

Using exactly the same arguments as for the case $p = 2$ we conclude that the asymptotic behaviour of X in this situation is given by

$$\mathbb{P}[X > x] \sim \mathbb{P}[B > x] \frac{\mu}{p\,x^{p-1}}.$$

2.6 The covariance function

In this section we study properties of the covariance function

$$c(k) = \operatorname{cov}[W_1, W_{1+k}].$$

In particular, in Theorem 2.9 we show that the covariance between two waiting times is of alternating sign, while in Theorem 2.10 we bound the absolute value of $c(k)$ to conclude that the covariance function converges to zero geometrically fast.

The result of Theorem 2.9 is an expected effect of the non-standard sign of W_n in Recursion (1.1), while the conclusion we draw from Theorem 2.10 reinforces the results of Section 2.3, where it is shown that $\mathbb{P}[W_n \leqslant x]$ converges geometrically fast to F_W. Recall that $C_1 = \inf\{n \geqslant 1 : W_{n+1} = 0\}$. We proceed with stating Theorem 2.9.

Theorem 2.9. *The covariance function $c(k)$ is non-negative if k is even and non-positive if k is odd. If in addition, X has a strictly positive density on an interval (a, b), $0 < a < b$, and $W \overset{\mathcal{D}}{=} W_1$, then $c(k) > 0$ if k is even, and $c(k) < 0$ if k is odd.*

In order to prove this theorem we shall first prove the following lemma, which is a variation of a result by Angus [4].

Lemma 2.2. *Let Y be a random variable and f a non-decreasing (non-increasing) function defined on the range of Y. Then, provided the expectations exist,*

$$\operatorname{cov}[Y, f(Y)] \geqslant 0 \quad (\operatorname{cov}[Y, f(Y)] \leqslant 0).$$

Furthermore, if

$$\mathbb{P}[Y \in \{y : f(y) \text{ strictly increasing (decreasing) in } y\}] > 0,$$

then

$$\operatorname{cov}[Y, f(Y)] > 0 \quad (\operatorname{cov}[Y, f(Y)] < 0).$$

Proof. We prove this lemma only for f being non-decreasing. The proof for non-increasing f follows analogously. We use the same argument as Angus [4]. Let Z be an i.i.d. copy of Y. So, if f is non-decreasing, then we have that

$$\big(Y - Z\big)\big(f(Y) - f(Z)\big) \geqslant 0.$$

Furthermore, let I_Y be the subset of the domain of f where the function is strictly increasing, i.e., $I_Y = \{y : f(y) \text{ strictly increasing in } y\}$. Then, if $\mathbb{P}[Y \in I_Y] > 0$, we have that

$$\mathbb{P}[\big(Y - Z\big)\big(f(Y) - f(Z)\big) > 0] > 0.$$

By taking expectations, and using the fact that Y and Z are i.i.d. we obtain

$$\mathbb{E}[\big(Y - Z\big)\big(f(Y) - f(Z)\big)] = 2\operatorname{cov}[Y, f(Y)],$$

which is non-negative, and strictly positive if $\mathbb{P}[Y \in I_Y] > 0$. □

Proof of Theorem 2.9. For $k = 0$ the statement is trivial. For any fixed integer $k > 0$ let $X_i = x_i$, $i = 2, \ldots, k+1$, where for all i, $x_i \in \mathbb{R}$. Furthermore, define recursively the functions g_i as follows;

$$g_1(w) = w \quad \text{and} \quad g_{i+1}(w) = \max\{0, x_{i+1} - g_i(w)\}, \quad i = 1, \ldots, k.$$

It can easily be shown now that g_1 is non-decreasing, g_2 is non-increasing and, by iterating, that g_i is non-increasing if i is even, and non-decreasing if i is odd.

Let the first waiting time be fixed; that is, $W_1 = w_1$. Then, it is clear that, for all $i = 1, \ldots, k+1$, the i-th waiting time will be equal to $g_i(w_1)$, cf. Recursion (1.1). Now, write

$$\operatorname{cov}[W_1, W_{1+k}] = \int_{x_2 \in \mathbb{R}, \ldots, x_{k+1} \in \mathbb{R}} \cdots \int \operatorname{cov}[W_1, g_{k+1}(W_1)] \, \mathrm{d}\mathbb{P}[X_2 \leqslant x_2, \ldots, X_{k+1} \leqslant x_{k+1}].$$

From Lemma 2.2, we obtain that $\operatorname{cov}[W_1, g_{k+1}(W_1)] \geqslant 0$ if k is even and that $\operatorname{cov}[W_1, g_{k+1}(W_1)] \leqslant 0$ if k is odd. This concludes the first part of the theorem.

Assume now that X has a strictly positive density on (a, b); therefore, for all $a_1, a_2 \in (a, b)$, with $0 < a_1 < a_2$, we have that $\mathbb{P}[X \in (a_1, a_2)] > 0$. We know already that for any set of fixed constants $\{x_i\}, i = 2, \ldots, k+1$, the functions g_i are monotone (i.e., either non-decreasing or non-increasing). Moreover, observe that if these constants have the property that $x_{k+1} > x_k > \cdots > x_2$, then g_{k+1} is strictly monotone in $(0, x_2)$.

Furthermore, since $\mathbb{P}[X \in (a_1, a_2)] > 0$ and $W \overset{\mathcal{D}}{=} W_1$, we have that

$$\mathbb{P}[W \in (a_1, a_2)] = \mathbb{P}[\max\{0, X - W\} \in (a_1, a_2)] \geqslant \mathbb{P}[W = 0]\mathbb{P}[X \in (a_1, a_2)] > 0,$$

which implies that $\mathbb{P}[W_1 \in (a_1, a_2)] > 0$ for all $a_1, a_2 \in (a, b)$, with $0 < a_1 < a_2$. So we have that if $x_2 > a$, then

$$\mathbb{P}[W \in (0, x_2)] \geqslant \mathbb{P}[W_1 \in (a, x_2)] > 0,$$

which can be rewritten as

$$\mathbb{P}[W \in \{w : g_{k+1}(w) \text{ strictly monotone in } w\}] > 0.$$

Therefore, by Lemma 2.2 we have that $\operatorname{cov}[W_1, g_{k+1}(W_1)] > 0$ (< 0) if k is even (odd).

Now, let S be the subset of $\mathbb{R}^k$ defined as follows,

$$S = \big\{(x_2, x_3, \ldots, x_{k+1}) : x_{k+1} > x_k > \cdots > x_2 > a\big\},$$

and let S^c be its complement. Then

$$\begin{aligned}\operatorname{cov}[W_1, W_{1+k}] = {} & \\ & \int_{(x_2, x_3, \ldots, x_{k+1}) \in S} \cdots \int \operatorname{cov}[W_1, g_{k+1}(W_1)] \, \mathrm{d}\mathbb{P}[X_2 \leqslant x_2, \ldots, X_{k+1} \leqslant x_{k+1}] + \\ & + \int_{(x_2, x_3, \ldots, x_{k+1}) \in S^c} \cdots \int \operatorname{cov}[W_1, g_{k+1}(W_1)] \, \mathrm{d}\mathbb{P}[X_2 \leqslant x_2, \ldots, X_{k+1} \leqslant x_{k+1}]. \quad (2.8)\end{aligned}$$

We know that the second integral at the right-hand side of (2.8) is greater than or equal to zero if k is even and less than or equal to zero if k is odd. It remains to show that the first integral at the right-hand side of (2.8) is strictly positive if k is even and strictly negative if k is odd. Since we integrate over the set S, we have shown that $\mathrm{cov}[W_1, g_{k+1}(W_1)] > 0$ (< 0) if k is even (odd). So it suffices to show that $\mathbb{P}[S'] > 0$, where

$$S' = \big\{(X_2, X_3, \ldots, X_{k+1}) \in S\big\} = \big\{X_{k+1} > X_k > \cdots > X_2 > a\big\}.$$

Indeed, take a partition $\{a_i\}$ of (a, b) such that $a_i = a + [i(b-a)]/k$, $i = 0, \ldots, k$. Then we have that

$$\mathbb{P}\left[X_{k+1} > X_k > \cdots > X_2 > a\right] \geqslant \mathbb{P}\left[X_{k+1} \in (a_{k-1}, b)\,; X_k \in (a_{k-2}, a_{k-1})\,; \cdots ; X_2 \in (a, a_1)\right] = \prod_{i=2}^{k+1} \mathbb{P}\left[X_i \in (a_{i-2}, a_{i-1})\right] > 0,$$

since X has a strictly positive density on (a, b). □

This technique can be also applied to other stochastic recursions; for example, for Lindley's recursion the above argument shows under weak assumptions that the covariance between the waiting time of customer 1 and $k+1$ is strictly positive.

Having seen that the correlations have alternating sign, we now turn to the question of the behaviour of the covariance function $c(k)$ for large k.

Theorem 2.10. *For every value of k we have that*

$$|c(k)| \leqslant 2\mathbb{E}[W_1]\mathbb{E}[X \mid X > 0]\mathbb{P}[X > 0]^k.$$

We see that $c(k)$ converges to zero geometrically fast in k. This is consistent with the fact that the distribution of W_n converges geometrically fast to that of W.

Proof. Write

$$\mathrm{cov}[W_1, W_{1+k}] = \mathrm{cov}[W_1, W_{1+k}\,; C_1 \leqslant k] + \mathrm{cov}[W_1, W_{1+k}\,; C_1 > k]. \tag{2.9}$$

For the first term of the right-hand side of the above equation we have that

$$\begin{aligned}\mathrm{cov}[W_1, W_{1+k}\,; C_1 \leqslant k] &= \sum_{j=1}^{k} \mathrm{cov}[W_1, W_{1+k}\,; C_1 = j] \\ &= \sum_{j=1}^{k} \mathrm{cov}[W_1, W_{1+k} \mid C_1 = j]\ \mathbb{P}[C_1 = j]. \end{aligned} \tag{2.10}$$

Since $C_1 = j$, $j \in \{1, \ldots, k\}$, implies that $W_j = 0$, from the Markov property we immediately conclude that W_{1+k} is independent of W_1. Therefore, for $j \in \{1, \ldots, k\}$, we have that $\mathrm{cov}[W_1, W_{1+k} \mid C_1 = j] = 0$. So (2.10) yields that

$$\mathrm{cov}[W_1, W_{1+k}\,; C_1 \leqslant k] = 0.$$

Thus, from (2.9) and Theorem 2.9 we have that

$$c(k) = \operatorname{cov}[W_1, W_{1+k}\,; C_1 > k] \geqslant 0 \ (\leqslant 0), \quad \text{if } k \text{ is even (odd).}$$

Furthermore, observe that the following bounds hold.

$$\operatorname{cov}[W_1, W_{1+k}\,; C_1 > k] \leqslant \mathbb{E}[W_1 W_{1+k}\,; C_1 > k] + \mathbb{E}[W_1]\mathbb{E}[W_{1+k}]\mathbb{P}[C_1 > k], \quad (2.11)$$

$$\operatorname{cov}[W_1, W_{1+k}\,; C_1 > k] \geqslant -\mathbb{E}[W_{1+k}]\mathbb{E}[W_1\,; C_1 > k] - \mathbb{E}[W_1]\mathbb{E}[W_{1+k}\,; C_1 > k]. \quad (2.12)$$

So, if k is even, then it is bounded from below by zero and above by the right-hand side of (2.11), and similarly for k odd we use (2.12).

Assume now that k is even. In order to bound (2.11) note first that

$$\{C_1 > k\} \subset \{X_2 > 0\,; \cdots ; X_{k+1} > 0\} \quad \text{and} \quad W_{k+1} \leqslant \max\{0, X_{k+1}\}. \quad (2.13)$$

Therefore, for $k \geqslant 1$ we have for the first term of the right-hand side of (2.11) that

$$\begin{aligned}\mathbb{E}[W_1 W_{1+k}\,; C_1 > k] &\leqslant \mathbb{E}[W_1 \max\{0, X_{k+1}\}\,; X_2 > 0\,; \cdots ; X_{k+1} > 0] \\ &= \mathbb{E}[W_1]\mathbb{E}[X_{k+1} \mid X_{k+1} > 0]\mathbb{P}[X > 0]^k.\end{aligned}$$

So (2.11) now yields that

$$\operatorname{cov}[W_1, W_{1+k}\,; C_1 > k] \leqslant \mathbb{E}[W_1]\mathbb{E}[X \mid X > 0]\mathbb{P}[X > 0]^k + \mathbb{E}[W_1]\mathbb{E}[W_{1+k}]\mathbb{P}[C_1 > k].$$

Since

$$\mathbb{P}[C_1 > k] \leqslant \mathbb{P}[X_1 > 0]^k \quad \text{and} \quad \mathbb{E}[W_{1+k}] \leqslant E[X_{k+1}^+] \leqslant \mathbb{E}[X_{k+1} \mid X_{k+1} > 0],$$

we obtain that

$$c(k) \leqslant 2\mathbb{E}[W_1]\mathbb{E}[X \mid X > 0]\mathbb{P}[X > 0]^k,$$

which is exactly what we set to prove for k even.

Assume now that k is odd. So $c(k)$ is now bounded from below by (2.12), which implies that

$$\begin{aligned}|c(k)| &= -\operatorname{cov}[W_1, W_{1+k}\,; C_1 > k] \\ &\leqslant \mathbb{E}[W_{1+k}]\mathbb{E}[W_1\,; C_1 > k] + \mathbb{E}[W_1]\mathbb{E}[W_{1+k}\,; C_1 > k]. \quad (2.14)\end{aligned}$$

All terms in this expression can be straightforwardly bounded, using again (2.13). In particular we obtain as before that $\mathbb{E}[W_{1+k}] \leqslant \mathbb{E}[X \mid X > 0]$ and that

$$\mathbb{E}[W_1\,; C_1 > k] \leqslant \mathbb{E}[W_1]\mathbb{P}[X > 0]^k$$

and

$$\mathbb{E}[W_{1+k}\,; C_1 > k] \leqslant \mathbb{E}[X \mid X > 0]\mathbb{P}[X > 0]^k.$$

Combining these bounds with (2.14) proves the theorem for odd values of k. □

As mentioned before, in [53] it is shown that the serial correlation coefficients of a stationary sequence of waiting times are non-negative and decrease monotonically to zero. Comparing these results to the ones we have obtained in this section, we first observe that the result of Theorem 2.9 is not surprising. It is rather a natural effect of the minus sign that appears in front of W_n in (1.1). However, the condition implied by Theorem 2.10 for the stationary sequence is that $\mathbb{E}[B] < \infty$, which is less restrictive than demanding that the third moment of the service times is finite, as is the case for Lindley's recursion. Furthermore, from Theorem 2.10 we immediately have that the infinite sum of all correlations is finite. For Lindley's recursion, the finiteness of the third moment of B is not sufficient to guarantee this. Even in this case, the series may be converging so slowly to zero, that the sum is infinite. As it is stated in Theorem 2 of [53], what is necessary and sufficient is that the fourth moment of B is also finite.

As we shall see in Chapter 6, to some degree many of the properties mentioned in this chapter are useful to approximate the distribution F_W. We shall first derive, however, an exact expression of the steady-state distribution of the waiting times under different assumptions on the distribution of A and B. In the following chapter we shall focus on preparation times on a bounded support. In the special case that B is uniformly distributed on $[0, 1]$, the results have a direct application to the two-carousel model we have presented in Section 1.2.

Chapter 3

Preparation times on a bounded support

3.1 Introduction

Motivated by the carousel application presented in Section 1.2, in this Chapter we are set to study Equation (1.2)

$$W \stackrel{\mathcal{D}}{=} \max\{0, B - A - W\}$$

under a setting that is relevant to the carousel application we have in mind.

We have already introduced the model in Sections 1.2 and 1.4. As we have explained there, the service points are the two carousels, the preparation time is the time needed until an item rotates in front of the picker, and the service time is the pick time. The pick time is defined as the time from the point that the picker begins the n-th pick until the point he is ready to start picking items from the other carousel. So for example, the time he needs in order to walk between the two carousels – if any – is included in the pick time. Here, we shall reiterate some of the basic characteristics of the model while emphasising various points that are particularly connected to carousels.

To this end, we model the carousel as a continuous loop, rather than consisting of a number of discrete locations. For sake of simplicity, we assume that the maximum time to rotate between any two points on the carousel is one. So we may think of a unidirectional carousel of length one, or of a bidirectional one of length two. Therefore, the mean rotation time for each pick is $1/2$. The carousels rotate at a constant (unit) speed and the time they need to accelerate between zero and the maximum speed is considered to be negligible. Each pick order requires exactly one item, and the orders are processed in the sequence they arrive.

For the first part of this study, we shall assume that the locations of the requested items are independent and uniformly distributed on the carousel. In other words, B follows a uniform distribution on $[0, 1]$. As mentioned in Section 1.3.5, this model has been studied by various authors [81, 105, 140], and in particular, the focus in Park *et al.* [140] is on the waiting-time distribution. In the following section we review the results obtained in [140], we present the results obtained in this chapter, and we discuss how our results contribute to the existing literature.

In the sequel, we shall substitute the words "pick", "rotation", "carousel", and their derivatives, by "service", "preparation", and "service point".

3.2 Previous results

In their work, Park *et al.* [140] show that the random variables W_n form an aperiodic, recurrent Harris chain [128], in order to conclude that the invariant distribution F_W exists, and that the probabilities $\mathbb{P}[W_n \leqslant x]$ converge to it in total variation. The total variation norm is defined as

$$\|\nu_n - \nu\| = \sup\{|\nu_n(S) - \nu(S)| : S \in \mathcal{B}\},$$

where $\mathcal{B}$ is the collection of Borel sets in $\mathbb{R}$. By choosing S to belong to the class $\{(-\infty, x] : x \in \mathbb{R}\}$, we see that convergence in total variation implies weak convergence. Furthermore, they formally prove that the throughput θ of the system is given by

$$\theta = \big(\mathbb{E}[A] + \mathbb{E}[W]\big)^{-1}$$

and they derive two equivalent expressions for the waiting time density. We shall present both of them here.

Since the preparation time B is upper bounded by one, then the maximum waiting time of the server is also upper bounded by one. Furthermore, if A has a density, then one can easily show that W has a density too as follows. From Equation (1.2) we readily have that

$$\mathbb{P}[W \leqslant x] = \int_{-\infty}^{\infty} \mathbb{P}[A \geqslant y - x] \, \mathrm{d}F_{B-W}(y).$$

Since A has a density, we conclude that the integral

$$\int_{-\infty}^{\infty} f_A(y - x) \, \mathrm{d}F_{B-W}(y)$$

exists and is the density of F_W. Recall that $\pi_0 = \mathbb{P}[W = 0]$; then from Equation (1.2) we have that

$$\begin{aligned} F_W(x) &= \mathbb{P}[W \leqslant x] = \mathbb{P}[B - A - W \leqslant x] \\ &= \int_0^{\infty} \int_0^{\infty} \mathbb{P}[B \leqslant x + y + z] \, \mathrm{d}F_A(z) \mathrm{d}F_W(y) \\ &= \pi_0 \int_0^{\infty} \mathbb{P}[B \leqslant x + z] \, \mathrm{d}F_A(z) + \int_0^1 \int_0^{\infty} \mathbb{P}[B \leqslant x + y + z] f_W(y) \, \mathrm{d}F_A(z) \mathrm{d}y. \end{aligned} \tag{3.1}$$

So, for the distribution F_W we have that

$$\begin{aligned} F_W(x) = \pi_0 \bigg(\int_0^{1-x} (x + z) \, \mathrm{d}F_A(z) + \int_{1-x}^{\infty} \mathrm{d}F_A(z) \bigg) + \\ + \int_0^{1-x} \int_0^{1-x-y} (x + y + z) f_W(y) \, \mathrm{d}F_A(z) \mathrm{d}y + \\ + \int_0^{1-x} \int_{1-x-y}^{\infty} f_W(y) \, \mathrm{d}F_A(z) \mathrm{d}y + \int_{1-x}^{1} f_W(y) \, \mathrm{d}y. \end{aligned}$$

Now, by differentiating with respect to x, we have that

$$f_W(x) = \pi_0 F_A(1-x) + \int_0^{1-x} F_A(1-x-y) f_W(y)\,\mathrm{d}y. \tag{3.2}$$

Thus, to determine the invariant distribution F_W it suffices to determine the mass π_0 and the density f_W, provided that $0 \leqslant \pi_0 \leqslant 1$, $f_W \geqslant 0$, and that both quantities satisfy (3.2) and the normalisation equation

$$\pi_0 + \int_0^1 f_W(x)\,\mathrm{d}x = 1. \tag{3.3}$$

Similarly, one can easily prove that (3.2) is equivalent to

$$f_W(x) = \mathbb{P}[A + W \leqslant 1 - x]. \tag{3.4}$$

Park *et al.* [140] consider two specific cases for the service-time distribution. They examine the case of A being deterministic or being exponentially distributed.

Deterministic service times

Assuming that the service times are a constant d might be a reasonable assumption for a robotic server. The main goal in [140] is to determine the throughput of the system. Thus, the non-trivial analysis emerges for $0 \leqslant d \leqslant 1$, since for $d \geqslant 1$, the throughput is equal to d^{-1}.

The authors derive a linear differential equation as follows. Equation (3.4) can now be written as $f_W(x) = F_W(1-x-d)$, so differentiating twice with respect to x yields

$$f_W^{(1)}(x) = -f_W(1-x-d) \quad \text{and} \quad f_W^{(2)}(x) = f_W^{(1)}(1-x-d) = -f_W(x).$$

In other words, they obtain the following second-order homogeneous linear differential equation

$$f_W^{(2)}(x) + f_W(x) = 0, \qquad 0 \leqslant x \leqslant 1-d.$$

The solution to this differential equation is given by

$$f_W(x) = \frac{\cos(x) + \sin(1-d-x)}{1+\sin(1-d)},$$
$$F_W(x) = \frac{\sin(x) + \cos(1-d-x)}{1+\sin(1-d)},$$

which implies that

$$\pi_0 = \frac{\cos(1-d)}{1+\sin(1-d)}.$$

Thus, the throughput is given by

$$\theta = \begin{cases} \frac{1+\sin(1-d)}{\cos(1-d)}, & 0 \leqslant d < 1 \\ \frac{1}{d}, & d \geqslant 1. \end{cases}$$

Exponential service times

Exponentially distributed service times (with rate λ) may be reasonable for human servers. The method followed in [140] in order to derive the density of the waiting times is again by means of differentiation. In particular, they substitute the distribution of A in (3.2) and differentiate two times with respect to x to obtain the equation

$$f_W^{(2)}(x) = \lambda f_W^{(1)}(x) + \lambda f_W(1-x).$$

Then, substitute $f_W(1-x)$ by using (3.2) and differentiate the new equation another two times with respect to x to obtain the linear differential equation

$$f_W^{(4)}(x) - \lambda^2 f_W^{(2)}(x) - \lambda^2 f_W(x) = 0, \qquad 0 \leqslant x \leqslant 1.$$

Notice that this substitution is practically equivalent to a change of variables, where $1-x$ is replaced by x and the procedure is repeated (namely, differentiate again two times). This observation will clarify one of the steps in the analysis that follows in Section 3.4. Now, the solution to the latter equation is given by

$$\begin{aligned} f_W(x) &= c_1 \mathrm{e}^{r_1 x} + c_2 \mathrm{e}^{-r_1 x} + c_3 \cos(r_2 x) + c_4 \sin(r_2 x) \\ F_W(x) &= \frac{c_1}{r_1} \mathrm{e}^{r_1 x} - \frac{c_2}{r_1} \mathrm{e}^{-r_1 x} + \frac{c_3}{r_2} \sin(r_2 x) - \frac{c_4}{r_2} \cos(r_2 x) + c_5 \end{aligned}$$

which implies that

$$\pi_0 = \frac{c_1}{r_1} - \frac{c_2}{r_1} - \frac{c_4}{r_2} + c_5.$$

In the above given expression, the constants c_i are known and the roots r_i are determined by the characteristic equation. Namely,

$$r_1 = \sqrt{\frac{\lambda^2 + \lambda\sqrt{\lambda^2+4}}{2}} \quad \text{and} \quad r_2 = \sqrt{\frac{-\lambda^2 + \lambda\sqrt{\lambda^2+4}}{2}}.$$

The throughput is easy to derive since we know the distribution of W.

In this chapter we complement the above mentioned results in two directions. First, we maintain the assumption that B is uniformly distributed on $[0,1]$ and we analyse (1.2) for more general distributions of the service time A. In particular, in the following section we present an iterative method that is useful for numerical approximations. This method is valid for any distribution F_A. In Sections 3.4 and 3.5 we analytically derive explicit expressions for F_W, when A follows an Erlang distribution or a phase-type distribution respectively. These results are based on [172] and [173].

In the second direction, we extend the results to more general preparation time distributions. Namely, we consider the case where B has a polynomial distribution. Since any continuous distribution on a bounded support can be approximated

arbitrarily close by a polynomial distribution (cf. Chapter 6), this assumption allows for good approximations of distributions on a bounded support. The results of Section 3.6 have been presented in [171].

We conclude the chapter with numerical results and further comments. A statement we shall make there is that the methods presented in this chapter for either direction are closely related. However, before going into details on the matter, we shall proceed with the analysis.

3.3 Iterative approach

Throughout this section we assume that B is uniformly distributed on $[0,1]$. Equation (3.2) is a Fredholm equation [126]; therefore, a natural way to proceed is by successive substitutions. So, substitute f_W at the right-hand side of (3.2) with the left-hand side of the equation to obtain for $0 \leqslant x \leqslant 1$

$$f_W(x) = \pi_0 F_A(1-x) + \pi_0 \int_0^{1-x} F_A(1-x-y)F_A(1-y)\,\mathrm{d}y + \\ + \int_0^{1-x}\int_0^{1-y} F_A(1-x-y)F_A(1-y-z)f_W(z)\,\mathrm{d}z\mathrm{d}y.$$

Define now iteratively the function

$$F_A^{n\,*}(1-x) \stackrel{\mathsf{def}}{=} \int_0^{1-x} F_A(1-x-y)\,F_A^{(n-1)\,*}(1-y)\,\mathrm{d}y, \quad n \geqslant 2, \tag{3.5}$$

while $F_A^{1\,*}(1-x)$ is defined to be equal to $F_A(1-x)$. Then, after n iterations we have that the density f_W is given by

$$f_W(x) = \pi_0 \sum_{i=1}^{n+1} F_A^{i\,*}(1-x) + \int_0^{1-x}\int_0^{1-x_1}\cdots\int_0^{1-x_n} F_A(1-x-x_1)F_A(1-x_1-x_2)\times \\ \times\cdots\times F_A(1-x_n-x_{n+1})f_W(x_{n+1})\,\mathrm{d}x_{n+1}\mathrm{d}x_n\ldots\mathrm{d}x_1.$$

Notice, however, that

$$\lim_{n\to\infty}\int_0^{1-x}\cdots\int_0^{1-x_n} F_A(1-x-x_1)\times\cdots\times F_A(1-x_n-x_{n+1})\times \\ \times f_W(x_{n+1})\,\mathrm{d}x_{n+1}\ldots\mathrm{d}x_1 = 0.$$

To see this, observe that

$$\int_0^{1-x_n} F_A(1-x_n-x_{n+1})f_W(x_{n+1})\,\mathrm{d}x_{n+1} = \mathbb{P}[A+W\leqslant 1-x_n] = p < 1,$$

since A has a support on $[0,\infty)$, and that for the same reason, $q = F_A(1-x-y) < 1$, for $0 \leqslant x, y \leqslant 1$. Therefore, the limit above is less than or equal to

$$\lim_{n\to\infty} p\,q^n \int_0^{1-x}\cdots\int_0^{1-x_{n-1}} \mathrm{d}x_n\ldots\mathrm{d}x_1 = 0,$$

since the multiple integral in the above limit is bounded by one (keep in mind that for all i, $0 \leqslant x_i \leqslant 1$) and q is strictly less than one.

So, by letting n go to infinity, we obtain the formal solution

$$f_W(x) = \pi_0 \sum_{i=1}^{\infty} F_A^{i*}(1-x), \qquad 0 \leqslant x \leqslant 1. \tag{3.6}$$

Since F_A is a distribution, then from (3.5) we have that for $n \geqslant 1$,

$$\begin{aligned} F_A^{(n+2)*}(x) &\leqslant \int_0^x F_A^{(n+1)*}(1-y)\,\mathrm{d}y \\ &\leqslant \int_0^x \int_0^{1-y} F_A^{n*}(1-z)\,\mathrm{d}z\mathrm{d}y \\ &= \int_0^x \int_y^1 F_A^{n*}(z)\,\mathrm{d}z\mathrm{d}y, \end{aligned} \tag{3.7}$$

which implies that

$$F_A^{3*}(x) \leqslant \int_0^x \int_y^1 F_A(z)\,\mathrm{d}z\mathrm{d}y \leqslant \int_0^x (1-y)\,\mathrm{d}y \leqslant \int_0^1 (1-y)\,\mathrm{d}y = \frac{1}{2}.$$

Now, by induction, it can be easily shown that, for $n \geqslant 1$

$$F_A^{2(n+1)*}(x) \leqslant F_A^{(2n+1)*}(x) \leqslant \frac{1}{2^n}, \qquad 0 \leqslant x \leqslant 1. \tag{3.8}$$

One only needs to observe that

$$F_A^{2*}(x) = \int_0^x F_A(x-y)F_A(1-y)\,\mathrm{d}y \leqslant \int_0^x F_A(x)\cdot 1\,\mathrm{d}y = xF_A(x) \leqslant F_A(x),$$

since $0 \leqslant x \leqslant 1$ and that if $F_A^{n*}(x) \leqslant F_A^{(n-1)*}(x)$, then

$$\begin{aligned} F_A^{(n+1)*}(x) &= \int_0^x F_A(x-y)F_A^{n*}(1-y)\,\mathrm{d}y \\ &\leqslant \int_0^x F_A(x-y)F_A^{(n-1)*}(1-y)\,\mathrm{d}y = F_A^{n*}(x). \end{aligned}$$

For the second inequality of (3.8) just notice that for $0 \leqslant x \leqslant 1$

$$F_A^{2(n+1)*}(x) \leqslant \int_0^x F_A^{(2n+1)*}(1-y)\,\mathrm{d}y \leqslant \frac{1}{2^n}\int_0^x \mathrm{d}y \leqslant \frac{1}{2^n}.$$

Furthermore, for $0 \leqslant x \leqslant 1$ we have that (cf. (3.7))

$$F_A^{(2n+3)*}(x) \leqslant \int_0^x \int_y^1 F_A^{(2n+1)*}(z)\,\mathrm{d}z\mathrm{d}y \leqslant \frac{1}{2^n}\Big(x - \frac{x^2}{2}\Big) \leqslant \frac{1}{2^{n+1}}.$$

This means that the infinite sum (3.6) converges (uniformly) for $0 \leqslant x \leqslant 1$; therefore, f_W is well defined. However, for a non-trivial distribution F_A, one cannot easily compute f_W explicitly using (3.6). The difficulty lies in the fact that $F_A^{n\,*}$ is not the n-fold convolution of the distribution function F_A, so all terms have to be computed explicitly.

One should note here though, that if $F_A(1-x-y)$ is a continuous kernel on the compact space $[0,1]^2$, then one can show with a method similar to the one applied in Section 2.3 that the right-hand side of the integral equation (3.2) is a contraction mapping with the contraction constant being equal to $\mathbb{P}[A+U \leqslant 1]$, where U is a uniformly distributed random variable on $[0,1]$ (keep in mind that A has an unbounded support, so this probability is strictly less than one). So the infinite series

$$\pi_0 \sum_{i=1}^{\infty} F_A^{i\,*}(1-x)$$

converges to the steady-state density of the waiting time with at least a geometric rate equal to $\mathbb{P}[A+U \leqslant 1]$. Moreover, the upper bounds of the individual terms of the series, as they are given by Equation (3.8), are not tight. For example, (3.8) suggests that $F_A^{6\,*}(x) \leqslant 1/4$; however, one can easily show by writing out the definition of $F_A^{6\,*}$, using the subsequent definitions for $F_A^{4\,*}$ and $F_A^{5\,*}$, and bounding $F_A^{3\,*}$ by $1/2$, that $F_A^{6\,*}(x) \leqslant 1/6$. Naturally, neither this bound is tight, since the bound for $F_A^{3\,*}$ is not tight. The point is though, that the series converges to the invariant density sufficiently fast, so it is a useful estimate of f_W. The disadvantage of this solution is that it is only a numerical approximation. In the sequel, we shall focus on explicit solutions that lead to more tractable results.

3.4 Erlang service times

Throughout this section we assume that the service times follow an Erlang distribution with scale parameter λ and n stages; that is,

$$F_A(x) = 1 - \mathrm{e}^{-\lambda x} \sum_{i=0}^{n-1} \frac{(\lambda x)^i}{i!}, \qquad x \geqslant 0.$$

As before, B is uniformly distributed on $[0,1]$. We shall present two methods that lead to explicit expressions for the density f_W. The advantage of the first method over the second one is that it gives some insight into the relation of various elements that appear in the solution, and it is computationally more efficient. However, the second method formalises the method used in Park *et al.* [140], and involves simpler calculations. Although the two methods seem different in nature at first, they are connected to one another, and we shall highlight the connection between them.

3.4.1 Laplace transforms approach

In this section we will use Laplace transforms to solve (1.2). Since B has a bounded support, so does W. Therefore, we shall consider Laplace transforms over

a bounded interval. Let ω denote the Laplace transform of f_W over the interval $[0,1]$; that is,

$$\omega(s) = \int_0^1 \mathrm{e}^{-sx} f_W(x)\,\mathrm{d}x.$$

We emphasise that, for the Laplace transform over a bounded interval, the standard properties are no longer valid, in the sense that there are no standard results for calculating the inverse transform over a bounded interval. Note that ω is analytic in the whole complex plane. It is convenient to replace x by $1-x$ in (3.2), yielding

$$f_W(1-x) = \pi_0 F_A(x) + \int_0^x F_A(x-z) f_W(z)\,\mathrm{d}z, \qquad 0 \leqslant x \leqslant 1. \tag{3.9}$$

By taking the Laplace transform of (3.9) and using the normalisation equation (3.3) we obtain

$$\begin{aligned}\mathrm{e}^{-s}\,\omega(-s) = \pi_0\Bigg(\frac{1-\mathrm{e}^{-s}}{s} - \sum_{i=0}^{n-1}\frac{\lambda^i}{(\lambda+s)^{i+1}} + \sum_{i=0}^{n-1}\sum_{j=0}^{i}\frac{\lambda^i}{j!(\lambda+s)^{i+1-j}}\mathrm{e}^{-(\lambda+s)}\Bigg) - \\ - \frac{\mathrm{e}^{-s}}{s}(1-\pi_0) + \frac{1}{s}\,\omega(s) - \sum_{i=0}^{n-1}\frac{\lambda^i}{(\lambda+s)^{i+1}}\,\omega(s) + \\ + \mathrm{e}^{-(\lambda+s)}\sum_{i=0}^{n-1}\sum_{j=0}^{i}\sum_{\ell=0}^{j}\binom{j}{\ell}\frac{\lambda^i}{j!(\lambda+s)^{i+1-j}}\,\omega^{(\ell)}(-\lambda),\end{aligned}$$

which, by rearranging terms and using the identity

$$\sum_{i=0}^{n-1}\frac{\lambda^i}{(\lambda+s)^{i+1}} = \frac{(\lambda+s)^n - \lambda^n}{s(\lambda+s)^n},$$

can be simplified to

$$\begin{aligned}\mathrm{e}^{-s}\,\omega(-s) - \frac{\lambda^n}{s(\lambda+s)^n}\,\omega(s) = \\ \pi_0\Bigg(\frac{\lambda^n}{s(\lambda+s)^n} + \mathrm{e}^{-(\lambda+s)}\sum_{i=0}^{n-1}\sum_{j=0}^{i}\frac{\lambda^i}{j!(\lambda+s)^{i+1-j}}\Bigg) - \\ - \frac{\mathrm{e}^{-s}}{s} + \mathrm{e}^{-(\lambda+s)}\sum_{i=0}^{n-1}\sum_{j=0}^{i}\sum_{\ell=0}^{j}\binom{j}{\ell}\frac{\lambda^i}{j!(\lambda+s)^{i+1-j}}\,\omega^{(\ell)}(-\lambda).\end{aligned} \tag{3.10}$$

In the above expression, $\omega^{(\ell)}$ denotes the ℓ-th derivative of ω. We have used the fact that for a natural number n, the incomplete Gamma function can be written as

$$\Gamma(n,x) = \int_x^\infty y^{n-1}\mathrm{e}^{-y}\,\mathrm{d}y = (n-1)!\,\mathrm{e}^{-x}\sum_{i=0}^{n-1}\frac{x^i}{i!}.$$

A further simplification of (3.10) is possible if we write the distribution of the service time as

$$F_A(x) = \mathrm{e}^{-\lambda x} \sum_{i=n}^{\infty} \frac{(\lambda x)^i}{i!}.$$

However, this form of (3.10) is more practical for numerical computations, since it does not involve infinite sums.

Note that both $\omega(-s)$ and $\omega(s)$ appear in (3.10). To obtain an additional equation we replace s by $-s$ in (3.10) and form a system from which $\omega(s)$ can be solved, yielding the following theorem.

Theorem 3.1. *For all s, the transform ω satisfies*

$$\omega(s)R(s) = -\mathrm{e}^{-s}s(\lambda+s)^n T(-s) - \lambda^n T(s), \tag{3.11}$$

where

$$\begin{aligned} R(s) &= s^2(\lambda^2 - s^2)^n + \lambda^{2n}, \\ T(s) &= \pi_0 \left(\lambda^n + \mathrm{e}^{-(\lambda+s)} \sum_{i=0}^{n-1} \sum_{j=0}^{i} \frac{s\lambda^i(\lambda+s)^{n-i-1+j}}{j!} \right) - \mathrm{e}^{-s}(\lambda+s)^n + \\ &\quad + \mathrm{e}^{-(\lambda+s)} \sum_{i=0}^{n-1} \sum_{j=0}^{i} \sum_{\ell=0}^{j} \binom{j}{\ell} \frac{s\lambda^i(\lambda+s)^{n-i-1+j}}{j!} \omega^{(\ell)}(-\lambda). \end{aligned}$$

In (3.11) we still need to determine the $n+1$ unknowns π_0 and $\omega^{(\ell)}(-\lambda)$ for $\ell = 0, \ldots, n-1$. Note that for any zero of the polynomial R, the left-hand side of (3.11) vanishes (since ω is analytic everywhere). This implies that the right-hand side should also vanish. Hence, the zeros of R provide the equations necessary to determine the unknowns.

Lemma 3.1. *The polynomial R has exactly $2n+2$ simple zeros $r_1, \ldots, r_{2n+2}$ satisfying $r_{2n+3-i} = -r_i$, for $i = 1, \ldots, n+1$.*

Proof. Since $R(s)$ is a polynomial in s^2 of degree $n+1$, it follows that $R(s)$ has exactly $2n+2$ zeros, with the property that each zero s has a companion zero $-s$. Furthermore, we have that the derivative of R is

$$R'(s) = -2s(\lambda^2 - s^2)^{n-1}\big((n+1)s^2 - \lambda^2\big).$$

Therefore, the roots of R' are the following

$$s = 0, s = \pm\lambda, \quad \text{and} \quad s = \pm\frac{\lambda}{\sqrt{1+n}}.$$

Now it is easily verified that for all the above roots of R', we have that $R(s) \neq 0$. In other words, $\gcd[R(s), R'(s)] = 1$. This means that the polynomials $R(s)$ and $R'(s)$ have no common factor of degree greater than zero, or that $R(s)$ has only simple zeros. □

In the following lemma we prove that the $2n+2$ zeros of R produce $n+1$ independent linear equations for the unknowns.

Lemma 3.2. *The probability π_0 and the quantities $\omega^{(\ell)}(-\lambda)$, $\ell = 0, \dots, n-1$ are the unique solution to the $n+1$ linear equations,*

$$\mathrm{e}^{-r_i} r_i (\lambda + r_i)^n T(-r_i) + \lambda^n T(r_i) = 0, \qquad i = 1, \dots, n+1.$$

Proof. For any zero of R the right-hand side of (3.11) should vanish. Hence, for two companion zeros r_i and $r_{2n+3-i} = -r_i$, $i = 1, \dots, n+1$, we have

$$\mathrm{e}^{-r_i} r_i (\lambda + r_i)^n T(-r_i) + \lambda^n T(r_i) = 0, \tag{3.12}$$

$$-\mathrm{e}^{r_i} r_i (\lambda - r_i)^n T(r_i) + \lambda^n T(-r_i) = 0. \tag{3.13}$$

The determinant of (3.12) and (3.13), treated as equations for $T(-r_i)$ and $T(r_i)$, is equal to

$$\begin{vmatrix} \mathrm{e}^{-r_i} r_i (\lambda + r_i)^n & \lambda^n \\ \lambda^n & -\mathrm{e}^{r_i} r_i (\lambda - r_i)^n \end{vmatrix} = R(r_i) = 0.$$

Hence, (3.12) and (3.13) are dependent, and so we may omit one of them. This leaves a system of $n+1$ linear equations for the unknowns π_0 and $\omega^{(\ell)}(-\lambda)$, $\ell = 0, \dots, n-1$. The uniqueness of the solution follows from the general theory of Markov chains that implies that there is a unique equilibrium distribution and thus also a unique solution to (3.10). □

Once π_0 and $\omega^{(\ell)}(-\lambda)$, $\ell = 0, \dots, n-1$ are determined, the transform ω is known. It remains to invert the transform. By collecting the terms that include e^{-s} we can rewrite (3.11) in the form

$$\omega(s) = \frac{P(s)}{R(s)} + \mathrm{e}^{-s} \frac{Q(s)}{R(s)}, \tag{3.14}$$

where $P(s)$ and $Q(s)$ are polynomials of degree $2n+1$ and $n+1$ respectively. Note that, without the last term, the transform is rational so computing the inverse would be straightforward if we had Laplace transforms on $[0, \infty)$. As it is, we must proceed more carefully. Since $\deg[R]$ is greater than $\deg[P]$ and $\deg[Q]$, (3.14) can be decomposed into distinct irreducible fractions. This leads to

$$\omega(s) = \frac{c_1}{s - r_1} + \cdots + \frac{c_{2n+2}}{s - r_{2n+2}} + \mathrm{e}^{-s} \left(\frac{\hat{c}_1}{s - r_1} + \cdots + \frac{\hat{c}_{2n+2}}{s - r_{2n+2}} \right),$$

where the coefficients c_i and $\hat{c}_i$ are given by

$$c_i = \lim_{s \to r_i} \frac{P(s)}{R(s)} (s - r_i) = \frac{P(r_i)}{R'(r_i)}, \qquad \hat{c}_i = \lim_{s \to r_i} \frac{Q(s)}{R(s)} (s - r_i) = \frac{Q(r_i)}{R'(r_i)}. \tag{3.15}$$

Note that the derivative $R'(r_i)$ is non-zero, since r_i is a simple zero. Since $\omega(s)$ is analytic everywhere, we have for every root r_i of $R(s)$ that

$$P(r_i) = -\mathrm{e}^{-r_i} Q(r_i), \qquad i = 1, \dots, 2n+2.$$

Hence, from (3.15) it follows that

$$c_i = -\mathrm{e}^{-r_i}\hat{c}_i, \tag{3.16}$$

and thus

$$\omega(s) = \sum_{i=1}^{2n+2} \frac{c_i}{s - r_i}\big(1 - \mathrm{e}^{r_i - s}\big),$$

which is the transform (over a bounded interval) of a mixture of $2n + 2$ exponentials. Now that the density is known, (3.3) can be used to derive a simple explicit expression for π_0. These findings are summarised in the following theorem.

Theorem 3.2. *The density of W on $[0, 1]$ is given by*

$$f_W(x) = \sum_{i=1}^{2n+2} c_i \mathrm{e}^{r_i x}, \qquad 0 \leqslant x \leqslant 1, \tag{3.17}$$

and

$$\pi_0 = \mathbb{P}[W = 0] = 1 - \sum_{i=1}^{2n+2} \frac{c_i}{r_i}(\mathrm{e}^{r_i} - 1). \tag{3.18}$$

Corollary 3.3. *The throughput θ satisfies*

$$\theta^{-1} = \mathbb{E}[A] + \mathbb{E}[W] = \frac{n}{\lambda} + \sum_{i=1}^{2n+2} \frac{c_i}{r_i^2}[1 + (r_i - 1)\mathrm{e}^{r_i}].$$

Although the roots r_i and coefficients c_i may be complex, the expressions (3.17) and (3.18) will be positive. This follows from the fact that the equilibrium equation (3.2) and the normalisation equation (3.3) have a unique solution. Of course, it is also clear that each root r_i and coefficient c_i have a companion conjugate root and conjugate coefficient, which implies that the imaginary parts in (3.17) and (3.18) cancel.

3.4.2 Differential equations approach

The alternative method to calculate the density f_W is by means of differentiation. This method has been used implicitly in Park *et al.* [140]. The authors there commented that "the approach of deriving a differential equation for each pick-time distribution was rather ad hoc". However, this method can be generalised to include phase-type distributions as well. In this section we shall formalise this method and explain its connection to the Laplace transforms approach we have just developed.

To this end, differentiate (3.9) with respect to x obtaining

$$\frac{\mathrm{d}}{\mathrm{d}x}\big[f_W(1 - x)\big] = \pi_0 \lambda^n \mathrm{e}^{-\lambda x} \frac{x^{n-1}}{(n-1)!} + \int_0^x \lambda^n \mathrm{e}^{-\lambda(x-z)} \frac{(x - z)^{n-1}}{(n-1)!} f_W(z)\,\mathrm{d}z. \tag{3.19}$$

Multiplying with $\mathrm{e}^{\lambda x}$ and differentiating m more times, $m = 1, \ldots, n-1$, with respect to x gives us

$$\frac{\mathrm{d}^m}{\mathrm{d}x^m}\Big[\mathrm{e}^{\lambda x}\frac{\mathrm{d}}{\mathrm{d}x}f_W(1-x)\Big] = \pi_0\lambda^n\frac{x^{n-m-1}}{(n-m-1)!} + \int_0^x \lambda^n \mathrm{e}^{\lambda z}\frac{(x-z)^{n-m-1}}{(n-m-1)!}f_W(z)\,\mathrm{d}z. \tag{3.20}$$

Note that the integral at the right hand side vanishes for $x = 0$. We shall need this remark later in order to derive the initial conditions of the differential equation. Now, by differentiating the equation corresponding to $m = n-1$ we have

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}\Big[\mathrm{e}^{\lambda x}\frac{\mathrm{d}}{\mathrm{d}x}\big[f_W(1-x)\big]\Big] = \mathrm{e}^{\lambda x}f_W(x)\lambda^n \text{ or } \sum_{i=0}^{n}\binom{n}{i}\lambda^{-i}\frac{\mathrm{d}^{i+1}}{\mathrm{d}x^{i+1}}\big[f_W(1-x)\big] = f_W(x).$$

Up to this point we have differentiated with respect to x a total of $n+1$ times. In order to derive a homogeneous linear differential equation we substitute x by $1-y$ and repeat the same procedure. This means that we shall differentiate (with respect to y now) a total of $n+1$ times more. We substitute x by $1-y$ in the last relation to obtain

$$\sum_{i=0}^{n}\binom{n}{i}\lambda^{-i}(-1)^{i+1}\frac{\mathrm{d}^{i+1}}{\mathrm{d}y^{i+1}}f_W(y) = f_W(1-y). \tag{3.21}$$

The change of variables is practically equivalent to the substitution of s by $-s$ that we did in order to obtain equation (3.11). Differentiating once (3.21) with respect to y and combining the result with (3.19) yields

$$\sum_{i=0}^{n}\binom{n}{i}\lambda^{-i}(-1)^{i+1}\frac{\mathrm{d}^{i+2}}{\mathrm{d}y^{i+2}}f_W(y) = \pi_0\lambda^n\mathrm{e}^{-\lambda y}\frac{y^{n-1}}{(n-1)!} + \int_0^y \lambda^n\mathrm{e}^{-\lambda(y-z)}\frac{(y-z)^{n-1}}{(n-1)!}f_W(z)\,\mathrm{d}z.$$

As before, we multiply with $\mathrm{e}^{\lambda y}$ and we differentiate m more times with respect to y, for $m = 1, \ldots, n-1$. Furthermore, the remark we made before is still valid. Namely, all the intermediate steps have a right hand side of the same form as in (3.20); thus, the right-hand side is equal to zero for $x = 0$. One more differentiation gives us

$$\sum_{i=0}^{n}\binom{n}{i}\lambda^{-i}(-1)^{i+1}\frac{\mathrm{d}^n}{\mathrm{d}y^n}\Big[\mathrm{e}^{\lambda y}\frac{\mathrm{d}^{i+2}}{\mathrm{d}y^{i+2}}f_W(y)\Big] = \mathrm{e}^{\lambda y}f_W(y)\lambda^n,$$

which after rewriting the derivatives and arranging the terms becomes

$$\sum_{i=0}^{n}\binom{n}{i}\lambda^{-i}(-1)^{i+1}\sum_{j=0}^{n}\binom{n}{j}\lambda^{-j}\frac{\mathrm{d}^{i+j+2}}{\mathrm{d}x^{i+j+2}}f_W(y) = f_W(y). \tag{3.22}$$

Equation (3.22) is a homogeneous linear differential equation of order $2n+2$. For the solution we need the roots of the characteristic function

$$\sum_{i=0}^{n}\binom{n}{i}\lambda^{-i}(-1)^{i+1}\sum_{j=0}^{n}\binom{n}{j}\lambda^{-j}r^{i+j+2} = 1 \quad \text{or} \quad -r^2\Big(1-\frac{r}{\lambda}\Big)^n\Big(1+\frac{r}{\lambda}\Big)^n = 1,$$

which agrees with $R(r) = 0$. By Lemma 3.1 we know that the roots of this equation are simple, which means that the general solution to (3.22) is given by

$$f_W(x) = \sum_{i=1}^{2n+2} d_i \mathrm{e}^{r_i x}, \qquad 0 \leqslant x \leqslant 1.$$

This proves Theorem 3.2; however, we still need to define the coefficients d_i, for $i = 1, \ldots, 2n+2$.

For the solution we need as many initial conditions as the order of the differential equation. We are going to derive them from the intermediate steps of differentiation. From (3.2) we have that $f_W(1) = 0$ and this will be the first condition. We derive the other $2n+1$ conditions by evaluating at zero each equation that we obtained from the intermediate steps of differentiation. We do not use the last differentiation with respect to y, because this yields Equation (3.22), which is the differential equation. We summarise the above in the following relations.

For $m = 1, \ldots, n-2$ we have

$$\frac{\mathrm{d}}{\mathrm{d}x} f_W(1-x)\bigg|_{x=0} = 0, \qquad \frac{\mathrm{d}^m}{\mathrm{d}x^m}\left[\mathrm{e}^{\lambda x}\frac{\mathrm{d}}{\mathrm{d}x} f_W(1-x)\right]\bigg|_{x=0} = 0,$$

$$\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\left[\mathrm{e}^{\lambda x}\frac{\mathrm{d}}{\mathrm{d}x} f_W(1-x)\right]\bigg|_{x=0} = \pi_0\lambda^n, \qquad \frac{\mathrm{d}^n}{\mathrm{d}x^n}\left[\mathrm{e}^{\lambda x}\frac{\mathrm{d}}{\mathrm{d}x} f_W(1-x)\right]\bigg|_{x=0} = \lambda^n f_W(0),$$

$$\sum_{i=0}^{n}\binom{n}{i}\lambda^{-i}(-1)^{i+1}\frac{\mathrm{d}^{i+2}}{\mathrm{d}y^{i+2}} f_W(y)\bigg|_{y=0} = 0,$$

$$\sum_{i=0}^{n}\binom{n}{i}\lambda^{-i}(-1)^{i+1}\frac{\mathrm{d}^m}{\mathrm{d}y^m}\left[\mathrm{e}^{\lambda y}\frac{\mathrm{d}^{i+2}}{\mathrm{d}y^{i+2}} f_W(y)\right]\bigg|_{y=0} = 0,$$

and

$$\sum_{i=0}^{n}\binom{n}{i}\lambda^{-i}(-1)^{i+1}\frac{\mathrm{d}^{n-1}}{\mathrm{d}y^{n-1}}\left[\mathrm{e}^{\lambda y}\frac{\mathrm{d}^{i+2}}{\mathrm{d}y^{i+2}} f_W(y)\right]\bigg|_{y=0} = \pi_0\lambda^n.$$

Note that all these conditions define uniquely the coefficients d_i, but involve the unknown parameter π_0. We obtain this last parameter by using the normalisation equation (3.3) and this concludes the proof of Theorem 3.2. One observation is that by using this method, one needs to solve a system of linear equations twice as big as the one that appears in Lemma 3.2. Furthermore, we know that the coefficients d_i are equal to the coefficients c_i that appear in (3.17), thus they satisfy (3.16). However, these relations do not become immediately obvious from the analysis. Using Laplace transforms we can derive explicit expressions for the coefficients that appear in the solution, cf. (3.15).

3.5 Phase-type service times

The case we have analysed so far, namely that the service times follow an Erlang distribution with n stages, is a useful preparation for the case we are concerned with in this section. The reason is that now we assume that the service times follow with probability κ_n, $n = 1, \dots, N$ an Erlang distribution of scale parameter λ and n stages. In other words,

$$F_A(x) = \sum_{n=1}^{N} \kappa_n \Big(1 - \mathrm{e}^{-\lambda x} \sum_{i=0}^{n-1} \frac{(\lambda x)^i}{i!} \Big), \qquad x \geqslant 0. \tag{3.23}$$

The class of the phase-type distributions of the above form is dense in the space of distribution functions defined on $[0, \infty)$. This means that for any such distribution function F, there is a sequence F_n of phase-type distributions of this class that converges weakly to F as n goes to infinity; for details see Schassberger [149]. Below we give the result for service time distributions of the form (3.23).

The analysis proceeds along the same lines as in Section 3.4. The formulae in the intermediate steps are simply linear combinations of the ones that appear for Erlang service times. For example, we have that Equation (3.10) now becomes

$$\begin{aligned} \mathrm{e}^{-s}\,\omega(-s) - \omega(s) \sum_{n=1}^{N} \kappa_n \Big(\frac{\lambda^n}{s(\lambda+s)^n} \Big) = \\ \pi_0 \sum_{n=1}^{N} \kappa_n \Big(\frac{\lambda^n}{s(\lambda+s)^n} + \mathrm{e}^{-(\lambda+s)} \sum_{i=0}^{n-1} \sum_{j=0}^{i} \frac{\lambda^i}{j!\,(\lambda+s)^{i+1-j}} \Big) + \\ + \sum_{n=1}^{N} \kappa_n \Big(-\frac{\mathrm{e}^{-s}}{s} + \mathrm{e}^{-(\lambda+s)} \sum_{i=0}^{n-1} \sum_{j=0}^{i} \sum_{\ell=0}^{j} \binom{j}{\ell} \frac{\lambda^i}{j!\,(\lambda+s)^{i+1-j}}\, \omega^{(\ell)}(-\lambda) \Big). \end{aligned} \tag{3.24}$$

In order to obtain the transform ω we form once more a 2×2 system of linear equations by replacing s by $-s$. This leads to the following result.

Theorem 3.4. *For all s, the transform ω satisfies*

$$\omega(s)\widetilde{R}(s) = -\mathrm{e}^{-s} s (\lambda+s)^N \widetilde{T}(-s) - \sum_{n=1}^{N} \kappa_n \lambda^n (\lambda - s)^{N-n} \widetilde{T}(s), \tag{3.25}$$

where

$$\widetilde{R}(s) = s^2(\lambda^2 - s^2)^N + \sum_{n=1}^{N}\sum_{m=1}^{N} \kappa_n\kappa_m\lambda^n\lambda^m(\lambda - s)^{N-n}(\lambda + s)^{N-m},$$

$$\widetilde{T}(s) = \pi_0 \sum_{n=1}^{N} \kappa_n \left(\lambda^n(\lambda + s)^{N-n} + \mathrm{e}^{-(\lambda+s)} \sum_{i=0}^{n-1}\sum_{j=0}^{i} \frac{s\lambda^i(\lambda + s)^{N-i-1+j}}{j!} \right) +$$
$$+ \sum_{n=1}^{N} \kappa_n \left(\mathrm{e}^{-(\lambda+s)} \sum_{i=0}^{n-1}\sum_{j=0}^{i}\sum_{\ell=0}^{j} \binom{j}{\ell} \frac{s\lambda^i(\lambda + s)^{N-i-1+j}}{j!}\, \omega^{(\ell)}(-\lambda) - \right.$$
$$\left. - \mathrm{e}^{-s}(\lambda + s)^N \right).$$

The unknowns π_0 and $\omega^{(\ell)}(-\lambda)$, $\ell = 0, \ldots, n-1$ can be determined in the same way as in Section 3.4. The polynomial $\widetilde{R}$ has exactly $2N+2$ zeros, with the property that each zero s has a companion zero $-s$. We *assume* that all these zeros are simple and label them $\widetilde{r}_1, \ldots, \widetilde{r}_{2N+2}$ such that $\widetilde{r}_{2N+3-i} = -\widetilde{r}_i$ for $i = 1, \ldots, N+1$. Then the following lemma can be readily established.

Lemma 3.3. *The probability π_0 and the quantities $\omega^{(\ell)}(-\lambda)$, $\ell = 0, \ldots, n-1$ are the unique solution to the $N+1$ linear equations,*

$$\mathrm{e}^{-\widetilde{r}_i}\widetilde{r}_i(\lambda + \widetilde{r}_i)^N\widetilde{T}(-\widetilde{r}_i) + \sum_{n=1}^{N} \kappa_n\lambda^n(\lambda - \widetilde{r}_i)^{N-n}\widetilde{T}(\widetilde{r}_i) = 0, \;\; i = 1, \ldots, N+1. \quad (3.26)$$

Given π_0 and $\omega^{(\ell)}(-\lambda)$, $\ell = 0, \ldots, n-1$, the transform ω is completely known. Partial fraction decomposition of the transform yields

$$\omega(s) = \sum_{i=1}^{2N+2} \frac{\widetilde{c}_i}{s - \widetilde{r}_i}\left(1 - \mathrm{e}^{\widetilde{r}_i - s}\right),$$

from which we conclude that the density of the waiting time is a mixture of $2N+2$ exponentials. Hence, as was the case for Erlang service times, the density is given by

$$f_W(x) = \sum_{i=1}^{2N+2} \widetilde{c}_i\mathrm{e}^{\widetilde{r}_i x}.$$

Remark 3.1. When $\widetilde{R}$ has multiple zeros, the analysis proceeds in essentially the same way. For example, if $\widetilde{r}_1 = \widetilde{r}_2$ (so $\widetilde{r}_1$ and, thus, $\widetilde{r}_{2N+2}$ are double zeros), then (3.26) for $i = 1$ is identical to the one for $i = 2$. Nonetheless, an additional equation can be obtained by requiring that the derivative of the right-hand side of (3.25)

should vanish at $s = r_1$. The partial fraction decomposition of ω then becomes

$$\omega(s) = \frac{\widetilde{c}_1}{(s-\widetilde{r}_1)^2}\big(1 - \mathrm{e}^{\widetilde{r}_1 - s} - (s-\widetilde{r}_1)\mathrm{e}^{\widetilde{r}_1 - s}\big) + \sum_{i=2}^{2N+1} \frac{\widetilde{c}_i}{s-\widetilde{r}_i}\big(1 - \mathrm{e}^{\widetilde{r}_i - s}\big) + \\ + \frac{\widetilde{c}_{2N+2}}{(s-\widetilde{r}_{2N+2})^2}\big(1 - \mathrm{e}^{\widetilde{r}_{2N+2} - s} - (s-\widetilde{r}_{2N+2})\mathrm{e}^{\widetilde{r}_{2N+2} - s}\big),$$

the inverse of which is given by

$$f_W(x) = \widetilde{c}_1 x \mathrm{e}^{\widetilde{r}_1 x} + \sum_{i=2}^{2N+1} \widetilde{c}_i \mathrm{e}^{\widetilde{r}_i x} + \widetilde{c}_{2N+2} x \mathrm{e}^{\widetilde{r}_{2N+2} x}.$$

Remark 3.2. In case the pick times follow a phase-type distribution of the form (3.23), we can still obtain the solution by similar methods as in the differential equations approach of Section 3.4.2. If $\widetilde{R}(\cdot)$ has only simple roots, then there are no differences in the analysis. If there are roots with multiplicity greater than one, the differential equation is solved in a similar but not identical manner, involving exponentials multiplied with powers of x (cf. Remark 3.1). For each root r of multiplicity k we need to have k linearly independent solutions, which in this case will be of the form $x^i \, \mathrm{e}^{rx}$, for $i = 0, \ldots, k-1$.

3.6 Polynomial preparation times

In the previous two sections we have maintained the assumption that B is uniformly distributed on $[0, 1]$. In this section we extend the results in Park *et al.* [140] in another direction. Namely, we consider the case where B has a polynomial distribution. This extension, although it does not have a direct application on the two-carousel model, is mathematically interesting, since polynomial distributions can approximate any continuous distribution on a bounded support. As in [140], we assume that A is exponentially distributed with parameter λ. Without loss of generality we can assume that F_B has all its mass on $[0, 1]$. Therefore, let

$$F_B(x) = \begin{cases} \sum_{i=0}^{n} c_i x^i, & \text{for } 0 \leqslant x \leqslant 1, \\ 1, & \text{for } x \geqslant 1, \end{cases} \tag{3.27}$$

where $\sum_{i=0}^{n} c_i = 1$. Substituting the distributions of A and B in (3.1), and differentiating with respect to x, we obtain for $0 \leqslant x \leqslant 1$ that,

$$f_W(x) = \lambda F_W(x) - \lambda \pi_0 \sum_{i=0}^{n} c_i x^i - \lambda \int_0^{1-x} \sum_{i=0}^{n} c_i (x+y)^i f_W(y) \, \mathrm{d}y - \\ - \lambda \int_{1-x}^{1} f_W(y) \, \mathrm{d}y,$$

or that

$$f_W(x) = \lambda F_W(x) - \lambda\pi_0 \sum_{i=0}^{n} c_i x^i - \lambda \sum_{i=0}^{n}\sum_{k=0}^{i} c_i \binom{i}{k} x^{i-k} \int_0^{1-x} y^k f_W(y)\,\mathrm{d}y - \\ - \lambda \int_{1-x}^{1} f_W(y)\,\mathrm{d}y. \quad (3.28)$$

We know from Section 2.2 that (3.28) has a unique solution f_W and π_0, provided that they satisfy the normalisation equation (3.3). To determine f_W and π_0, we shall transform the integral equation (3.28) into a (high order) differential equation for f_W. Differentiating (3.28) with respect to x yields

$$\begin{aligned} f_W^{(1)}(x) = \lambda f_W(x) - \lambda\pi_0 \sum_{i=1}^{n} i c_i x^{i-1} - \\ - \lambda \sum_{i=0}^{n-1}\sum_{j=0}^{i} c_{i+1}(i+1) \binom{i}{j} x^{i-j} \int_0^{1-x} y^j f_W(y)\,\mathrm{d}y + \\ + \lambda \sum_{i=0}^{n}\sum_{j=0}^{i} c_i \binom{i}{j} x^{i-j}(1-x)^j f_W(1-x) - \lambda f_W(1-x), \end{aligned}$$

which can be simplified further to

$$f_W^{(1)}(x) = \lambda f_W(x) - \lambda\pi_0 \sum_{i=1}^{n} i c_i x^{i-1} - \lambda \sum_{i=0}^{n-1}\sum_{j=0}^{i} c_{i+1}(i+1) \binom{i}{j} x^{i-j} \int_0^{1-x} y^j f_W(y)\,\mathrm{d}y,$$

and in general, for $\ell = 1, 2, \ldots, n$,

$$f_W^{(\ell)}(x) = a_\ell(x) + \sum_{i=0}^{\ell-1} \nu_{n-i}(-1)^{\ell-1-i} f_W^{(\ell-1-i)}(1-x), \quad (3.29)$$

where

$$\nu_{n-i} = \lambda \sum_{j=0}^{n-i} \frac{(i+j)!}{j!} c_{i+j}, \quad (3.30)$$

$$\begin{aligned} a_\ell(x) = \lambda f_W^{(\ell-1)}(x) - \lambda\pi_0 \sum_{i=0}^{n-\ell} \frac{(i+\ell)!}{i!} c_{i+\ell}\, x^i - \lambda(-1)^{\ell-1} f_W^{(\ell-1)}(1-x) - \\ - \lambda \sum_{i=0}^{n-\ell}\sum_{j=0}^{i} c_{i+\ell} \frac{(i+\ell)!}{i!} \binom{i}{j} x^{i-j} \int_0^{1-x} y^j f_W(y)\,\mathrm{d}y. \end{aligned} \quad (3.31)$$

From (3.29) we have that the n-th derivative of f_W is given by

$$\begin{aligned} f_W^{(n)}(x) &= \lambda f_W^{(n-1)}(x) - \lambda\pi_0 n! c_n - \lambda(-1)^{n-1} f_W^{(n-1)}(1-x) - \\ &\quad - \lambda n! c_n \int_0^{1-x} f_W(y)\,\mathrm{d}y + \sum_{i=0}^{n-1} \nu_{n-i}(-1)^{n-1-i} f_W^{(n-1-i)}(1-x) \\ &= \lambda f_W^{(n-1)}(x) - \lambda\pi_0 n! c_n - \\ &\quad - \lambda n! c_n \int_0^{1-x} f_W(y)\,\mathrm{d}y + \sum_{i=1}^{n-1} \nu_{n-i}(-1)^{n-1-i} f_W^{(n-1-i)}(1-x), \end{aligned}$$

which implies that for $0 \leqslant x \leqslant 1$,

$$\begin{aligned} f_W^{(n+1)}(x) &= \lambda f_W^{(n)}(x) + \lambda n! c_n f_W(1-x) + \sum_{i=1}^{n-1} \nu_i(-1)^i f_W^{(i)}(1-x) \\ &= \lambda f_W^{(n)}(x) + \sum_{i=0}^{n-1} \nu_i(-1)^i f_W^{(i)}(1-x). \end{aligned} \tag{3.32}$$

Up to this point, we have differentiated Equation (3.28) a total of $n+1$ times. Therefore, we need a total of $n+1$ additional conditions in order to guarantee that any solution to (3.32) is also a solution to (3.28). Since for every value of x in $[0,1]$, Equations (3.28), (3.29), and (3.32) are satisfied, we can evaluate all these equations for a specific x, say $x=0$, which provides us with the $n+1$ initial conditions, for $\ell = 1, 2, \ldots, n$,

$$\begin{aligned} f_W^{(\ell)}(0) &= a_\ell(0) + \sum_{i=0}^{\ell-1} \nu_{n-i}(-1)^{\ell-1-i} f_W^{(\ell-1-i)}(1) \\ \text{and} \quad f_W^{(n+1)}(0) &= \lambda f_W^{(n)}(0) + \sum_{i=0}^{n-1} \nu_i(-1)^i f_W^{(i)}(1). \end{aligned} \tag{3.33}$$

So we now have that Equation (3.32) has a unique solution that satisfies these conditions, along with the normalisation equation (3.3).

Equation (3.32) is a homogeneous linear differential equation, not of a standard form because of the argument $1-x$ that appears at the right-hand side. In the Appendix, we present a differential equation, similar to Equation (3.32), that has some surprising characteristics. Therefore, we need to proceed with caution.

Note that the unknown probability π_0 is not involved in (3.32). We shall solve this equation by transforming it into a differential equation we can handle. To this end, substitute x for $1-x$ in (3.32), to obtain the equation

$$f_W^{(n+1)}(1-x) = \lambda f_W^{(n)}(1-x) + \sum_{i=0}^{n-1} \nu_i(-1)^i f_W^{(i)}(x). \tag{3.34}$$

Equations (3.32) and (3.34) form a system of equations. Now let

$$\boldsymbol{f_W}(x) = \begin{bmatrix} f_W(x) \\ f_W(1-x) \end{bmatrix}, \ \mathbf{A_n} = \begin{bmatrix} 1 & 0 \\ 0 & (-1)^n \end{bmatrix}, \text{ and } \mathbf{J} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Then the system of equations (3.32) and (3.34) can be rewritten as

$$\boldsymbol{f_W^{(n+1)}}(x) = \lambda \mathbf{A_{n+1}} \mathbf{A_n} \boldsymbol{f_W^{(n)}}(x) + \mathbf{A_{n+1}} \mathbf{J} \sum_{i=0}^{n-1} \nu_i (-1)^i \mathbf{A_i} \boldsymbol{f_W^{(i)}}(x). \tag{3.35}$$

In order to derive the characteristic equation of (3.35), we work as follows. We look for solutions of the form $\boldsymbol{\xi} \mathrm{e}^{rx}$, where $\boldsymbol{\xi} = \begin{bmatrix} \zeta \\ \sigma \end{bmatrix}$. Substituting this solution into (3.35) and dividing by e^{rx}, we derive the following linear system that determines $\boldsymbol{\xi}$ and r, which is

$$\begin{aligned} \zeta r^{n+1} &= \lambda \zeta r^n + \sum_{i=0}^{n-1} \nu_i \sigma r^i, \\ \sigma r^{n+1} &= -\lambda \sigma r^n + \sum_{i=0}^{n-1} \nu_i (-1)^{n+1+i} \zeta r^i. \end{aligned} \tag{3.36}$$

In order for a non-trivial solution to exist, the determinant of the coefficients of ζ and σ should be equal to zero. This yields that

$$r^{2n}(r^2 - \lambda^2) + (-1)^n \left(\sum_{i=0}^{n-1} \nu_i r^i \right) \left(\sum_{j=0}^{n-1} \nu_j (-r)^j \right) = 0, \tag{3.37}$$

which is the characteristic equation of (3.35).

Let us *assume* for the moment that the characteristic equation has only simple roots, and label them $r_1, \ldots, r_{2n+2}$. It is interesting to note here that since (3.37) is a polynomial in r^2, for every root r of this polynomial $-r$ is also a root. Therefore, we shall order the roots so that for every i, $r_i = -r_{2n+3-i}$. By substituting each root into the system (3.36), we obtain the corresponding vectors $\boldsymbol{\xi_i}$, $i = 1, \ldots, 2n+2$. Then (3.35) has the $2n+2$ linearly independent solutions $\boldsymbol{\xi_i} \mathrm{e}^{r_i x}$. Thus, the general solution of (3.35) is given by

$$\boldsymbol{f_W}(x) = \sum_{i=1}^{2n+2} d_i \boldsymbol{\xi_i} \mathrm{e}^{r_i x}, \tag{3.38}$$

where d_i are arbitrary constants.

From (3.38) we can immediately conclude that the solution to Equation (3.32) that we are interested in, is of the form

$$f_W(x) = \sum_{i=1}^{2n+2} d_i \zeta_i \mathrm{e}^{r_i x}. \tag{3.39}$$

However, this is not the general solution to (3.32). It does not follow from the derivation of (3.38) that, for any choice of the coefficients d_i, the linear combination (3.39) will satisfy (3.32), since $\zeta_i \mathrm{e}^{r_i x}$ is not a solution to (3.32). Therefore, we substitute (3.39) into (3.32), and by keeping in mind that $r_i = -r_{2n+3-i}$, we have that for every $i = 1, \ldots, 2n+2$,

$$d_i \zeta_i r_i^n (r_i - \lambda) = \mathrm{e}^{-r_i} d_{2n+3-i} \zeta_{2n+3-i} \sum_{j=0}^{n-1} \nu_j r_i^j. \tag{3.40}$$

These are in fact only $n+1$ relations between the unknown coefficients, since it can easily be shown by using the characteristic equation (3.37) that the equations for every i and $2n+3-i$ are identical. Using the relations between the coefficients d_i, one can rewrite (3.39) as the sum of $n+1$ linearly independent solutions to (3.32) as follows

$$f_W(x) = \sum_{i=1}^{n+1} d_i \left(\zeta_i \mathrm{e}^{r_i x} + q_i \ \zeta_{2n+3-i} \mathrm{e}^{-r_i x} \right), \tag{3.41}$$

where q_i follows from (3.40) if we solve for d_{2n+3-i}. Thus, the general solution to (3.32) is given by (3.41). The coefficients d_i, for $i = 1, \ldots, n+1$, and the probability π_0 that we still need to determine, follow now from the initial conditions (3.33) and the normalisation equation (3.3). Namely, by substituting (3.41) to (3.33) and (3.3) we obtain a linear system of $n+2$ equations.

Note that it is not possible to use the same argument in order to determine the coefficients d_i for *any* differential equation of the form (3.32), because of its non-standard form. Here we heavily rely on the fact that we know beforehand that a unique solution exists. We summarise the above in the following theorem.

Theorem 3.5. *Let F_B be a polynomial distribution of the form (3.27). Then the waiting-time distribution F_W has a mass π_0 at the origin, which is given by*

$$\pi_0 = \mathbb{P}[W = 0] = 1 - \sum_{i=1}^{2n+2} \frac{d_i \zeta_i}{r_i} (\mathrm{e}^{r_i} - 1), \tag{3.42}$$

and has a density f_W on $[0, 1]$, given by

$$f_W(x) = \sum_{i=1}^{2n+2} d_i \zeta_i \mathrm{e}^{r_i x}.$$

Although the roots r_i and coefficients d_i may be complex-valued, the density and the probability π_0 that appear in Theorem 3.5 will be non-negative. This follows from the fact that for every distribution F_B of the preparation time, (3.28) has a unique solution which is a distribution. It is also clear that, since the differential equation (3.32) has real coefficients, each root r_i and coefficient d_i have a companion conjugate root and conjugate coefficient, which implies that the imaginary parts cancel.

Remark 3.3. When (3.37) has roots with multiplicity greater than one, the analysis proceeds essentially in the same way. For example assume that $r_1 = r_2$. Then we first look for two solutions to (3.35) of the form $\boldsymbol{\xi} e^{r_1 x}$. If we find only one (that always exists), then we look for a second solution of the form $(x\boldsymbol{\xi} + \boldsymbol{\eta})e^{r_1 x}$, where $\boldsymbol{\eta}$ is again a vector. Substituting this solution into (3.35), we obtain a linear system that determines $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. Thus we can obtain the general solution to the differential equation (3.35). From this point on, by following the same method, we can formulate a linear system that determines the coefficients d_i and π_0, and obtain the solution to (3.32).

Remark 3.4. Another method to derive the solution to the integral equation (3.28) is through Laplace transforms over a bounded interval. We have illustrated this method in Section 3.4.1. However, the method presented here is interesting mainly because of the form of the differential equation (3.32). In the Appendix, we shall illustrate a very interesting and simple case, similar to (3.32), in order to point out the difficulties that arise when treating differential equations of this type.

3.7 Numerical results

This section is devoted to some numerical results. In particular, we want to examine numerically the effects of the squared coefficient of variation of the service time to the throughput of the carousel (in the case in which the items are randomly located on the carousel).

For various values of the mean service time $\mathbb{E}[A]$ we plot in Figure 3.1 the throughput θ versus the squared coefficient of variation of the service time, c_A^2. The mean service time is chosen to be comparable to the mean preparation time, which is 1/2. For each case of $\mathbb{E}[A]$, we fit a mixed Erlang or hyperexponential distribution to $\mathbb{E}[A]$ and c_A^2, depending on whether the squared coefficient of variation is less or greater than one; see, e.g., Tijms [161].

Hyperexponential distributions form another useful class of phase-type distributions. They can be used to model service times with squared coefficient of variation greater than one. Furthermore, hyperexponential distributions are always unimodal, which is not the case for mixed Erlang distributions. The analysis for hyperexponential service times is very similar to the one presented in the Section 3.4.

So, if $1/n \leqslant c_A^2 \leqslant 1/(n-1)$ for some $n = 2, 3, \ldots$, then the mean and squared coefficient of variation of the mixed-Erlang distribution

$$G(x) = p\left(1 - e^{-\lambda x}\sum_{j=0}^{n-2}\frac{(\lambda x)^j}{j!}\right) + (1-p)\left(1 - e^{-\lambda x}\sum_{j=0}^{n-1}\frac{(\lambda x)^j}{j!}\right), \qquad x \geqslant 0,$$

matches with $\mathbb{E}[A]$ and c_A^2, provided the parameters p and λ are chosen as

$$p = \frac{1}{1+c_A^2}\left(nc_A^2 - \sqrt{n(1+c_A^2) - n^2 c_A^2}\,\right), \qquad \lambda = \frac{n-p}{\mathbb{E}[A]}.$$

On the other hand, if $c_A^2 > 1$, then the mean and squared coefficient of variation of the hyperexponential distribution

$$G(x) = p_1(1 - \mathrm{e}^{-\lambda_1 x}) + p_2(1 - \mathrm{e}^{-\lambda_2 x}), \qquad x \geqslant 0,$$

match with $\mathbb{E}[A]$ and c_A^2 provided the parameters $\lambda_1, \lambda_2, p_1$ and p_2 are chosen as

$$p_1 = \frac{1}{2}\left(1 + \sqrt{\frac{c_A^2 - 1}{c_A^2 + 1}}\right), \qquad p_2 = 1 - p_1,$$
$$\lambda_1 = \frac{2p_1}{\mathbb{E}[A]} \qquad \text{and} \quad \lambda_2 = \frac{2p_2}{\mathbb{E}[A]}.$$

For single-server queuing models it is well-known that the mean waiting time depends (approximately linearly) on the squared coefficients of variation of the interarrival (and service) times. The results in Figure 3.1, however, show that for the carousel model, the mean waiting time is not very sensitive to the squared coefficient of variation of the service time and thus neither is the throughput θ; it indeed decreases as c_A^2 increases, but very slowly. This phenomenon may be explained by the fact that the waiting time of the server is bounded by one, that is, the time needed for a full rotation of the carousel.

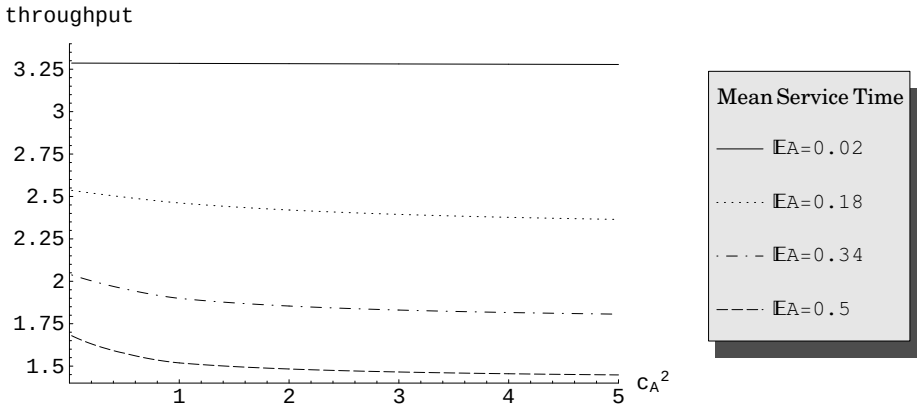

Figure 3.1: The throughput is almost insensitive to c_A^2.

3.8 Concluding remarks

In this chapter, we have considered the PH/U and the M/P models. For Lindley's recursion, there exist results for the more general queues PH/G/1 and M/G/1.

However, the results for these cases are not as explicit as the ones we have derived here.

For the PH/U model we have shown that the steady-state waiting-time distribution has a mass at zero and a density that is a mixture of exponentials. The parameters of these exponentials are the roots of a specific equation and come in groups; for every complex root r, the companion root $-r$ and their conjugates are all roots of the same equation. Thus, the solution is fairly explicitly known.

However, the analogous results for Lindley's equation are not so simple. Consider the PH/G/1 queue. Recall that α and β are the Laplace-Stieltjes transforms of A and B respectively. Since the interarrival times follow a phase-type distribution, α can be written as the ratio α_1/α_2, where α_1 is a *polynomial* of a degree lower than the degree of the *polynomial* α_2, which is taken to be equal to m. Without loss of generality, we can assume that α_1 and α_2 have no common zeros, and that the coefficient of s^m in $\alpha_2(s)$ is equal to one. Furthermore, assume that β and α_2 have no common zero. Then, for $m \geqslant 2$, it is shown in Cohen [46] that

$$\int_0^\infty \mathrm{e}^{-sx}\mathrm{d}F_W(x) = \frac{-\gamma\,\alpha_2(0)s(1-\rho)}{\alpha_2(-s)-\beta(s)\,\alpha_1(-s)}\prod_{i=1}^{m-1}\frac{\delta_i - s}{\delta_i}$$

(recall that $\rho = \mathbb{E}[B]/\mathbb{E}[A]$), where δ_i, $i = 1, \ldots, m-1$ are the roots of the equation

$$\alpha_2(-s) - \beta(s)\,\alpha_1(-s) = 0$$

in the right-half complex plane, and γ is the constant

$$\gamma = \frac{\alpha_2^{(1)}(0) - \alpha_1^{(1)}(0)}{\alpha_2(0)}.$$

Especially for B being uniformly distributed we have that $\beta(s) = (1-\mathrm{e}^{-s})/s$.

The distribution of W for the M/P model is identical to the one we retrieved for PH/U. Namely, we have again that f_W is a mixture of exponentials. Of course now the parameters of these exponentials are the roots to a different equation, but they still present the same characteristics.

The M/P/1 queue is a special case of the M/G/1. For the latter, the distribution of W is known, and it is given as an infinite sum of convolutions of the stationary excess distribution. More specifically, let

$$\widehat{F}_B(x) = \mathbb{E}[B]^{-1}\int_0^x \big(1 - F_B(y)\big)\mathrm{d}y.$$

Then F_W is given by

$$F_W(x) = (1-\rho)\sum_{n=0}^{\infty}\rho^n \widehat{F}_B^{n\,*}(x).$$

Formally, this solution is simple to describe. However, it is not very practical for numerical computations, even if B follows a polynomial distribution.

There are various possible extensions to the results we have presented here. For example, one can retrieve the steady-state distribution of the waiting time for various similar models, such as P/U, PH/P, and H_k/U, where H_k stands for the hyperexponential distribution with k parallel channels. The methods that can be used for these models are identical to the ones described in this chapter. For Lindley's recursion, some of the analogous cases can be explicitly solved. There seems to be though no study of the P/U/1 queuing model.

We have seen that there is a close connection between the Laplace transforms approach and the differential equations approach that we have presented in Section 3.4. Furthermore, we have seen that, similarly, we can derive a solvable differential equation for the M/P case. However, the method becomes more involved as the preparation-time distribution becomes more involved. As we have mentioned in Section 1.4, the techniques used to derive the steady-state distribution of W are classified according to the support of B. This will become more evident in the following chapter, where we consider the case of preparation times that have an unbounded support. More specifically, we shall consider the case that B follows a phase-type distribution.

Chapter 4
The G/PH model

4.1 Introduction

In this chapter we are concerned with the G/PH model. In other words, the service-time distribution F_A is some general distribution on $[0, \infty)$ and the preparation-time distribution F_B is phase-type. Moreover, we assume that for every n, both the service times A_n and the preparation times B_n have finite means. To keep expressions simple, in this chapter, we shall also use the function ϕ defined as $\phi(s) = \omega(s)\,\alpha(s)$. As before, ω is the Laplace-Stieltjes transform of the waiting-time distribution, and α is the Laplace-Stieltjes transform of the service-time distribution. Since the preparation times have an unbounded support, this case is more relevant to the first example mentioned in Section 1.2 than the motivating example of the previous chapter. Namely, for the carousel application, the preparation times are bounded by the time needed for a complete rotation of the carousel, while for the medical application, the preparation times of a patient need not follow a distribution on a bounded support.

As mentioned in Section 1.4, the server is not allowed to serve two consecutive customers at the same service point and must alternate between the service points. This condition is crucial. If we remove this condition, which means that the server will choose to serve the first customer that has completed his preparation time, then the problem turns out to be the classical *machine repair problem*. In that setting, there is a number of machines working in parallel (two in our situation) and one repairman. As soon as a machine fails, it joins the repair queue in order to be served. The machine repair problem, also known as the computer terminal model (see, for example, Bertsekas and Gallager [17]) or as the time sharing system (Kleinrock [103, Section 4.11]), is a well-studied problem in the literature. It is one of the key models to describe problems with a finite input population. A fairly extensive analysis of the machine repair problem can be found in Takács [158, Chapter 5].

The issue that is usually investigated in the machine repair problem is the waiting time of a machine until it becomes again operational. In our situation though we are concerned with the waiting time of the repairman. This question has not been treated in the classical literature, perhaps because in the machine repair problem the operating time of the machine is usually more valuable than the utilisation of the repairman.

Chapter 3 focused on deriving the steady-state distribution of W for two particular cases of Equation (1.2). This chapter, however, focuses both on the *time-dependent* behaviour of the process $\{W_n\}$ and on the derivation of the steady-state distribution of W, both for the model described by (1.1) and for the machine repair model. As we have seen in Section 1.6, the time-dependent waiting-time distribution is determined by the solution of a generalised Wiener-Hopf equation. A simi-

lar observation is valid for the steady-state distribution. Both the time-dependent and the steady-state waiting-time distribution are the fixed point of two distinct contraction mappings, and therefore they can be evaluated numerically. For the time-dependent behaviour, we are particularly interested in the distribution of W_n for any n, in the covariance between W_n and W_{n+k}, and in the distribution of the length of $C = \inf\{n \geqslant 1 : W_{n+1} = 0 \mid W_1 = 0\}$.

An overview of this chapter is as follows. Section 4.2 focuses on the alternating case, namely on Recursion (1.1). In particular, using probabilistic arguments we analyse the time-dependent distribution for exponential preparation times in Section 4.2.1 and for phase-type preparation times in Section 4.2.2. The expressions we obtain are remarkably explicit. Thus, Recursion (1.1) is a rare example of a stochastic model which allows for an explicit time-dependent analysis. The reason is that if B_n has a mixed-Erlang distribution, we can completely describe (1.1) in terms of the evolution of a finite-state Markov chain; for details, see to Section 4.2.2.

We obtain similar explicit results for the distribution of the cycle length C. What is more, we do not need to resort to the usage of generating functions, as is necessary when analysing the corresponding quantity in Lindley's recursion. Note that the interpretation of C for our model is completely different from the one for the corresponding quantity for Lindley's recursion. There, C represents the number of customers that arrived during a busy period. In our setting, C represents the number of pauses a server has until he needs to serve two consecutive customers without any pause. In this sense C can be seen as a "non-busy period".

In Section 4.2.3 we shall derive the steady-state distribution of the waiting time of the server, provided that the preparation time of a customer follows an Erlang distribution, while phase-type distributions are treated in Section 4.2.4.

Continuing with Section 4.3, we shall introduce the machine repair model and analyse the waiting time of the repairman, or in other words we shall remove the restriction that the server alternates between the service points. We compare the two models in Section 4.4. Namely, we compare the steady-state waiting times of the repairman in the classical machine repair problem and our model. We show that the random variables for the waiting time in the two situations are not stochastically ordered. However, on average, the alternating strategy leads to longer waiting times for the server. Furthermore we show that the probability that the server does not have to wait is larger in the alternating service system than in the non-alternating one. This result is perhaps counterintuitive, since the inequality for the mean waiting times of the server in the two situations is reversed. Numerical results related to this comparison are presented in Section 4.5.

Section 4.6 discusses the difference in complexity between the G/M alternating service model and the G/M/1 queue. As it is illustrated in Section 4.2.1, the analysis of the G/M alternating service model is particularly tractable and leads to explicit and fairly simple expressions for the distribution of W_n, the distribution of C, and the covariance between W_n and W_{n+k}. In Section 4.6, however, we see that the G/M/1 queue is rather more complex to analyse. This section also gives a representation for the time-dependent M/M/1 waiting-time distribution, which seems to be a new result. The contents of this chapter are mainly based upon [170] (steady-state

analysis and machine repair problem) and [174] (time-dependent analysis).

4.2 The alternating case

In this section we analyse the alternating service queue under the assumption that the preparation times B_n, $n \geqslant 1$, have a phase-type distribution. In the first part, we derive an explicit expression for the time-dependent distribution of W_n, the distribution of the cycle length C, and the covariance between W_n and W_{n+k} in case B_n is exponentially distributed with rate μ. The phase-type distributions that we will consider are mixtures of Erlang distributions with the same scale parameters. Therefore, in Section 4.2.2 we analyse the time-dependent behaviour of the G/E model, while in Section 4.2.3 we derive the steady-state distribution of the waiting times for this model. In Section 4.2.4 we extend the results derived in Section 4.2.3 to phase-type distributions.

4.2.1 Time-dependent analysis for G/M

The time-dependent waiting-time distribution

Although in the alternating service example it is natural to assume that $W_1 = B_1$, we would like to allow for a more general initial condition. Therefore, we assume that $W_1 = w_1$. Throughout this section, all probabilities are conditioned on this event. We first analyse the distribution of W_2. Write for $x \geqslant 0$,

$$\begin{aligned}\mathbb{P}[W_2 > x] = \mathbb{P}[B_2 > A_1 + w_1 + x] &= \int_0^\infty \mathrm{e}^{-\mu(y+w_1+x)}\,\mathrm{d}F_A(y) \\ &= \mathrm{e}^{-\mu(x+w_1)}\,\alpha(\mu). \end{aligned} \tag{4.1}$$

We see that $\mathbb{P}[W_2 > x \mid W_2 > 0] = \mathrm{e}^{-\mu x}$, that is, the distribution of W_2 is a mixture of a mass at zero and the exponential distribution with rate μ. This is actually the case for the distribution of every W_n. More precisely, the following theorem holds.

Theorem 4.1. *If $W_1 = w_1$, then for every $n \geqslant 1$ the time-dependent distribution of the waiting times is given by*

$$\mathbb{P}[W_{n+1} \leqslant x] = 1 - \mathrm{e}^{-\mu x}\left[\frac{2\,\alpha(\mu)}{2+\alpha(\mu)} + \left(-\frac{\alpha(\mu)}{2}\right)^{n-1}\left(\mathrm{e}^{-\mu w_1}\alpha(\mu) - \frac{2\,\alpha(\mu)}{2+\alpha(\mu)}\right)\right]. \tag{4.2}$$

Note that (4.2) is indeed also valid for $n = 1$, since in this case (4.2) simplifies to $\mathbb{P}[W_2 \leqslant x] = 1 - \mathbb{P}[W_2 > 0]\mathrm{e}^{-\mu x}$, which is consistent with (4.1). Naturally, the term in the brackets at the right-hand side of (4.2) is the probability $\mathbb{P}[W_{n+1} > 0]$.

Proof of Theorem 4.1. In order to compute the distribution of W_{n+1} for $n \geqslant 2$

observe that

$$\mathbb{P}[W_{n+1} > x] = \mathbb{P}[W_{n+1} > x \mid W_n = 0]\,\mathbb{P}[W_n = 0] + \\ + \mathbb{P}[W_{n+1} > x \mid W_n > 0]\,\mathbb{P}[W_n > 0]. \quad (4.3)$$

We shall calculate the first three terms that appear in the right-hand side of (4.3) (as the term $\mathbb{P}[W_n > 0]$ follows immediately from the term $\mathbb{P}[W_n = 0]$). Thus, we need to compute the distribution of W_{n+1} conditioned on the length of the previous waiting time.

Computation of $\mathbb{P}[W_{n+1} > x \mid W_n = 0]$:

We have that, for $n \geqslant 2$,

$$\begin{aligned}\mathbb{P}[W_{n+1} > x \mid W_n = w] &= \mathbb{P}[B_{n+1} - A_n - w > x \mid W_n = w] \\ &= \int_0^\infty \mathbb{P}[B_{n+1} > x + y + w \mid W_n = w]\,\mathrm{d}F_A(y) \\ &= \int_0^\infty \mathrm{e}^{-\mu(x+w)}\mathrm{e}^{-\mu y}\,\mathrm{d}F_A(y) = \mathrm{e}^{-\mu(x+w)}\,\alpha(\mu). \end{aligned} \quad (4.4)$$

For $w = 0$ we readily have the first term at the right-hand side of (4.3), i.e.,

$$\mathbb{P}[W_{n+1} > x \mid W_n = 0] = \mathrm{e}^{-\mu x}\,\alpha(\mu), \quad n \geqslant 2. \quad (4.5)$$

Another implication of Equation (4.4) is that, for $n \geqslant 2$,

$$\mathbb{P}[W_{n+1} > x \mid W_{n+1} > 0, W_n = w] = \frac{\mathbb{P}[W_{n+1} > x \mid W_n = w]}{\mathbb{P}[W_{n+1} > 0 \mid W_n = w]} = \mathrm{e}^{-\mu x}. \quad (4.6)$$

A straightforward conclusion is that

$$\mathbb{P}[W_{n+1} > x \mid W_{n+1} > 0] = \mathrm{e}^{-\mu x}. \quad (4.7)$$

Thus, the distribution of W_{n+1}, provided that W_{n+1} is strictly positive, is exponential and independent of the length of the previous waiting time.

We can extend (4.6) to the following more general property. For any event E of the form $E = \{W_2 \in S_2, \ldots, W_n \in S_n\}$, with $S_k \subseteq [0, \infty)$, $2 \leqslant k \leqslant n$, we have that

$$\mathbb{P}[W_{n+1} > x \mid E, W_{n+1} > 0] = \mathrm{e}^{-\mu x}. \quad (4.8)$$

To see this, write

$$\begin{aligned}\mathbb{P}[W_{n+1} > x \mid E, W_{n+1} > 0] &= \frac{\mathbb{P}[W_{n+1} > x\,; E, W_{n+1} > 0]}{\mathbb{P}[E, W_{n+1} > 0]} \\ &= \frac{\mathbb{P}[W_{n+1} > x\,; E]}{\mathbb{P}[W_{n+1} > 0\,; E]} = \frac{\mathbb{P}[W_{n+1} > x \mid E]}{\mathbb{P}[W_{n+1} > 0 \mid E]}. \end{aligned}$$

Furthermore, for $x \geqslant 0$,

$$\begin{aligned}\mathbb{P}[W_{n+1} > x \mid E] &= \int_0^\infty \mathbb{P}[W_{n+1} > x \mid W_n = w] \, \mathrm{d}\mathbb{P}[W_n \leqslant w \mid E] \\ &= \int_0^\infty \int_0^\infty \mathrm{e}^{-\mu(x+w+y)} \, \mathrm{d}F_A(y) \mathrm{d}\mathbb{P}[W_n \leqslant w \mid E] \\ &= \mathrm{e}^{-\mu x} \, \alpha(\mu) \mathbb{E}[\mathrm{e}^{-\mu W_n} \mid E],\end{aligned}$$

which directly proves (4.8), if we divide by the expression resulting for $x = 0$. Thus, the distribution of W_n is a mixture of a mass at zero and the exponential distribution with rate μ. This property is valid for all $n \geqslant 2$; for $n = 2$ it was shown below (4.1). We now return to the proof of Equation (4.3).

Computation of $\mathbb{P}[W_{n+1} > x \mid W_n > 0]$:

For $n \geqslant 2$ we have that

$$\begin{aligned}\mathbb{P}[W_{n+1} > x \mid W_n > 0] &= \mathbb{P}[W_{n+1} > x, W_{n+1} > 0 \mid W_n > 0] \\ &= \mathbb{P}[W_{n+1} > x \mid W_{n+1} > 0, W_n > 0] \mathbb{P}[W_{n+1} > 0 \mid W_n > 0] \\ &= \mathrm{e}^{-\mu x}(1 - \mathbb{P}[W_{n+1} = 0 \mid W_n > 0]), \qquad (4.9)\end{aligned}$$

where we applied (4.8) with $E = \{W_n > 0\}$ in the final step. We obtain the probability that appears at the right-hand side of (4.9) as follows.

$$\begin{aligned}\mathbb{P}[W_{n+1} = 0 \mid W_n > 0] &= \mathbb{P}[B_{n+1} \leqslant A_n + W_n \mid W_n > 0] \\ &= \int_0^\infty \mathbb{P}[B_{n+1} \leqslant A_n + x \mid W_n > 0] \mu \mathrm{e}^{-\mu x} \, \mathrm{d}x,\end{aligned}$$

where we applied (4.7) to W_n in the last step. Since B_{n+1} is independent of W_n, we have that the previous equation becomes, for $n \geqslant 2$,

$$\begin{aligned}\mathbb{P}[W_{n+1} = 0 \mid W_n > 0] &= \int_0^\infty \int_0^\infty (1 - \mathrm{e}^{-\mu y} \mathrm{e}^{-\mu x}) \mu \mathrm{e}^{-\mu x} \, \mathrm{d}x \mathrm{d}F_A(y) \\ &= 1 - \int_0^\infty \mathrm{e}^{-\mu y} \int_0^\infty \mu \mathrm{e}^{-2\mu x} \, \mathrm{d}x \mathrm{d}F_A(y) = 1 - \frac{1}{2} \, \alpha(\mu). \qquad (4.10)\end{aligned}$$

Combining (4.9) and (4.10) we have that, for $n \geqslant 2$, the third term at the right-hand side of (4.3) is given by

$$\mathbb{P}[W_{n+1} > x \mid W_n > 0] = \frac{1}{2} \, \alpha(\mu) \, \mathrm{e}^{-\mu x}. \qquad (4.11)$$

Computation of $\mathbb{P}[W_n = 0]$:

The last term we need to compute in order to obtain the transient distribution of the waiting times is the probability that the n-th waiting time is equal to zero, cf. (4.3). From (4.1) we readily have that $\mathbb{P}[W_2 = 0] = 1 - \mathrm{e}^{-\mu w_1}\,\alpha(\mu)$. Moreover, for $n \geqslant 2$, we have that

$$\begin{aligned}\mathbb{P}[W_{n+1} = 0] = \mathbb{P}[W_{n+1} = 0 \mid W_n = 0]\,\mathbb{P}[W_n = 0] + \\ + \mathbb{P}[W_{n+1} = 0 \mid W_n > 0]\,\mathbb{P}[W_n > 0],\end{aligned}$$

which implies that (cf. (4.4) and (4.10))

$$\begin{aligned}\mathbb{P}[W_{n+1} = 0] &= \big(1 - \alpha(\mu)\big)\mathbb{P}[W_n = 0] + \Big(1 - \frac{1}{2}\,\alpha(\mu)\Big)\big(1 - \mathbb{P}[W_n = 0]\big) \\ &= 1 - \frac{1}{2}\,\alpha(\mu) - \frac{1}{2}\,\alpha(\mu)\mathbb{P}[W_n = 0].\end{aligned}$$

This gives a first order recursion for $\mathbb{P}[W_{n+1} = 0]$. With simple manipulations it is easy to show that the solution to this recursion for $n \geqslant 2$ is given by

$$\mathbb{P}[W_{n+1} = 0] = \frac{2 - \alpha(\mu)}{2 + \alpha(\mu)} + \Big(-\frac{\alpha(\mu)}{2}\Big)^{n-1}\Big(\mathbb{P}[W_2 = 0] - \frac{2 - \alpha(\mu)}{2 + \alpha(\mu)}\Big). \tag{4.12}$$

So, from Equations (4.3), (4.5), and (4.11) we obtain, for $n \geqslant 2$,

$$\mathbb{P}[W_{n+1} > x] = \mathrm{e}^{-\mu x}\,\alpha(\mu)\mathbb{P}[W_n = 0] + \frac{1}{2}\,\mathrm{e}^{-\mu x}\,\alpha(\mu)\big(1 - \mathbb{P}[W_n = 0]\big).$$

Summing up the results we obtained in Equations (4.3), (4.5), (4.11), (4.12), we have that the distribution of W_{n+1} is given by

$$\mathbb{P}[W_{n+1} \leqslant x] = 1 - \mathrm{e}^{-\mu x}\left[\frac{2\,\alpha(\mu)}{2 + \alpha(\mu)} + \Big(-\frac{\alpha(\mu)}{2}\Big)^{n-1}\Big(\frac{2 - \alpha(\mu)}{2 + \alpha(\mu)} - \mathbb{P}[W_2 = 0]\Big)\right], \tag{4.13}$$

with $\mathbb{P}[W_2 = 0] = 1 - \mathrm{e}^{-\mu w_1}\,\alpha(\mu)$. Substituting $\mathbb{P}[W_2 = 0]$ in this expression gives us Equation (4.2) and completes the proof of the theorem. □

In the medical example given in Section 1.2, it is reasonable to assume that the server has to wait for a full preparation time at the beginning, implying that $W_1 = B_1$. In this case, it is easy to show that

$$\mathbb{P}[W_2 = 0] = \mathbb{P}[B_2 \leqslant A_1 + B_1] = 1 - \frac{1}{2}\,\alpha(\mu),$$

which yields the following corollary, cf. Equation (4.13).

Corollary 4.2. *If $W_1 = B_1$, then for every $n \geqslant 1$ the time-dependent distribution of the waiting times is given by*

$$\mathbb{P}[W_{n+1} \leqslant x] = 1 - \mathrm{e}^{-\mu x}\left[\frac{2\,\alpha(\mu)}{2 + \alpha(\mu)} + \Big(-\frac{\alpha(\mu)}{2}\Big)^{n-1}\Big(\frac{\alpha(\mu)}{2} - \frac{2\,\alpha(\mu)}{2 + \alpha(\mu)}\Big)\right].$$

Another result we can infer from Theorem 4.1 is the speed of convergence of the time-dependent distribution $\mathbb{P}[W_n \leqslant x]$ towards the steady-state distribution $\mathbb{P}[W \leqslant x]$. It is clear from (4.2) that this speed of convergence is geometrically fast at rate $\frac{1}{2}\,\alpha(\mu)$. It is interesting to observe that this rate is twice as fast as predicted by the upper bound in Section 2.3. In Lindley's recursion this speed of converge towards steady state is closely related to the tail behaviour of the cycle length C. In the next part, we see that the same constant $\frac{1}{2}\,\alpha(\mu)$ appears in a crucial way in the distribution of C.

The distribution of the cycle length

As we have mentioned before, $\{W_n\}$ is a regenerative process; regeneration occurs at times when $W_n = 0$. Let C be the random variable describing the length of a generic regeneration cycle, i.e.,

$$C = \inf\{k : W_{n+k} = 0 \mid W_n = 0\} = \inf\{k : W_{1+k} = 0 \mid W_1 = 0\}.$$

We are set to derive the distribution of C. By definition, we have that

$$\mathbb{P}[C = n] = \mathbb{P}[W_{n+1} = 0, W_n > 0, \ldots, W_2 > 0 \mid W_1 = 0].$$

We can now prove the following theorem.

Theorem 4.3. *Let C be the length of a regeneration cycle. Then the distribution of C is given by*

$$\mathbb{P}[C = n] = \begin{cases} 1 - \alpha(\mu) & n = 1 \\ \big(1 - \frac{1}{2}\,\alpha(\mu)\big)\big(\frac{1}{2}\,\alpha(\mu)\big)^{n-2}\alpha(\mu) & n \geqslant 2. \end{cases} \tag{4.14}$$

Proof. For $n = 1$, we readily have from (4.4) that

$$\mathbb{P}[C = 1] = \mathbb{P}[W_2 = 0 \mid W_1 = 0] = 1 - \alpha(\mu). \tag{4.15}$$

For $n \geqslant 2$ we shall prove our assertion by induction. For $n = 2$ we have that

$$\begin{aligned} \mathbb{P}[C = 2] &= \mathbb{P}[W_3 = 0, W_2 > 0 \mid W_1 = 0] \\ &= \mathbb{P}[B_3 \leqslant A_2 + W_2 \mid W_2 > 0, W_1 = 0]\mathbb{P}[W_2 > 0 \mid W_1 = 0]. \end{aligned}$$

Furthermore,

$$\mathbb{P}[B_3 \leqslant A_2 + W_2 \mid W_2 > 0, W_1 = 0] = \int_0^\infty \mathbb{P}[B_3 \leqslant A_2 + x \mid W_2 > 0, W_1 = 0]\,\mu \mathrm{e}^{-\mu x}\,\mathrm{d}x,$$

since (4.6) implies that $\mathbb{P}[W_2 \leqslant x \mid W_2 > 0, W_1 = 0] = 1 - \mathrm{e}^{-\mu x}$. In addition, since B_3 and A_2 are independent of W_2 and W_1, we have now that

$$\begin{aligned} \mathbb{P}[C = 2] &= \mathbb{P}[W_2 > 0 \mid W_1 = 0] \int_0^\infty \int_0^\infty (1 - \mathrm{e}^{-\mu y}\mathrm{e}^{-\mu x})\mu \mathrm{e}^{-\mu x}\,\mathrm{d}x\mathrm{d}F_A(y) \\ &= \alpha(\mu)\Big(1 - \frac{1}{2}\,\alpha(\mu)\Big), \end{aligned}$$

which satisfies (4.14) for $n = 2$.

Next, assume that

$$\mathbb{P}[C = k] = \Big(1 - \frac{1}{2}\,\alpha(\mu)\Big)\Big(\frac{1}{2}\,\alpha(\mu)\Big)^{k-2}\alpha(\mu) \tag{4.16}$$

for all $k \leqslant n$, with $n \geqslant 2$. To complete the proof, we must show that

$$\mathbb{P}[C = n+1] = \Big(1 - \frac{1}{2}\,\alpha(\mu)\Big)\Big(\frac{1}{2}\,\alpha(\mu)\Big)^{n-1}\alpha(\mu).$$

Since $\mathbb{P}[C \geqslant n+1] = 1 - \mathbb{P}[C \leqslant n]$, (4.15) and (4.16) imply that

$$\mathbb{P}[C \geqslant n+1] = \Big(\frac{1}{2}\,\alpha(\mu)\Big)^{n-1}\alpha(\mu).$$

Therefore, we have that

$$\begin{aligned}
\mathbb{P}[C = n+1] &= \mathbb{P}[W_{n+2} = 0, W_{n+1} > 0, \ldots, W_2 > 0 \mid W_1 = 0] \\
&= \mathbb{P}[W_{n+2} = 0 \mid W_{n+1} > 0, \ldots, W_2 > 0, W_1 = 0] \times \\
&\qquad \times \mathbb{P}[W_{n+1} > 0, \ldots, W_2 > 0 \mid W_1 = 0] \\
&= \mathbb{P}[W_{n+2} = 0 \mid W_{n+1} > 0, \ldots, W_2 > 0, W_1 = 0]\,\mathbb{P}[C \geqslant n+1] \\
&= \mathbb{P}[W_{n+2} = 0 \mid W_{n+1} > 0, \ldots, W_2 > 0, W_1 = 0]\Big(\frac{1}{2}\,\alpha(\mu)\Big)^{n-1}\alpha(\mu).
\end{aligned}$$

It suffices to show that

$$\mathbb{P}[W_{n+2} = 0 \mid W_{n+1} > 0, \ldots, W_2 > 0, W_1 = 0] = 1 - \frac{1}{2}\,\alpha(\mu). \tag{4.17}$$

Notice that since $\mathbb{P}[W_{n+2} = 0 \mid W_{n+1} > 0] = 1 - \frac{1}{2}\,\alpha(\mu)$ (cf. (4.10)), Equation (4.17) implies that

$$\mathbb{P}[W_{n+2} = 0 \mid W_{n+1} > 0, \ldots, W_2 > 0, W_1 = 0] = \mathbb{P}[W_{n+2} = 0 \mid W_{n+1} > 0],$$

which does not follow from the Markov property. For this model though, in order to discard conditions regarding previous states, it suffices to know whether the previous waiting time was equal to zero or not.

In order to prove (4.17), observe that

$$\begin{aligned}
&\mathbb{P}[W_{n+2} = 0 \mid W_{n+1} > 0, \ldots, W_2 > 0, W_1 = 0] \\
&\qquad = \mathbb{P}[B_{n+2} \leqslant A_{n+1} + W_{n+1} \mid W_{n+1} > 0, \ldots, W_2 > 0, W_1 = 0] \\
&\qquad = \int_0^\infty \mathbb{P}[B_{n+2} \leqslant A_{n+1} + x \mid W_{n+1} > 0, \ldots, W_2 > 0, W_1 = 0]\mu \mathrm{e}^{-\mu x}\,\mathrm{d}x,
\end{aligned} \tag{4.18}$$

since

$$\mathbb{P}[W_{n+1} \leqslant x \mid W_{n+1} > 0, \ldots, W_2 > 0, W_1 = 0] = 1 - \mathrm{e}^{-\mu x},$$

which follows from (4.8) with $E = \{W_n > 0, \dots, W_2 > 0\}$ and $w_1 = 0$. Equation (4.18) yields

$$\begin{aligned}\mathbb{P}[W_{n+2} = 0 \mid W_{n+1} > 0, \dots, W_2 > 0, W_1 = 0] \\ = \int_0^\infty \int_0^\infty (1 - \mathrm{e}^{-\mu y}\mathrm{e}^{-\mu x})\mu \mathrm{e}^{-\mu x}\,\mathrm{d}x\mathrm{d}F_A(y) = 1 - \frac{1}{2}\,\alpha(\mu),\end{aligned}$$

which is exactly Equation (4.17) that remained to be proven. □

The covariance function

We are interested in the covariance between two waiting times. By definition, we have that

$$\operatorname{cov}[W_n, W_{n+k}] = \mathbb{E}[W_n W_{n+k}] - \mathbb{E}[W_n]\mathbb{E}[W_{n+k}].$$

The terms $\mathbb{E}[W_n]$ and $\mathbb{E}[W_{n+k}]$ can be directly computed, for example, from Theorem 4.1. For the expectation of the product of the two waiting times we have that

$$\mathbb{E}[W_n W_{n+k}] = \int_0^\infty w\,\mathbb{E}[W_{n+k} \mid W_n = w]\,\mathrm{d}\mathbb{P}[W_n \leqslant w].$$

Write

$$\begin{aligned}\mathbb{E}[W_{n+k} \mid W_n = w] &= \mathbb{E}[W_{n+k} \mid W_{n+k} > 0, W_n = w]\,\mathbb{P}[W_{n+k} > 0 \mid W_n = w] \\ &= \frac{1}{\mu}\,\mathbb{P}[W_{1+k} > 0 \mid W_1 = w],\end{aligned}$$

where in the last step, we applied the Markov property as well as the fact that W_{n+k}, given that $W_n = w$ and $W_{n+k} > 0$, is exponentially distributed with rate μ. The latter follows from (4.8). Thus, in order to compute the covariance between W_n and W_{n+k}, we need the distribution of W_{1+k}, conditioned on $W_1 = w$. This distribution has been derived in Theorem 4.1, from which it follows that

$$\mathbb{P}[W_{1+k} > 0 \mid W_1 = w] = \frac{2\,\alpha(\mu)}{2 + \alpha(\mu)}\left[1 - \Big(-\frac{\alpha(\mu)}{2}\Big)^{k-1}\right] + \Big(-\frac{\alpha(\mu)}{2}\Big)^{k-1}\alpha(\mu)\,\mathrm{e}^{-\mu w}.$$

Combining the last three equations, we obtain

$$\begin{aligned}\mathbb{E}[W_n W_{n+k}] &= \frac{1}{\mu}\int_0^\infty w\frac{2\,\alpha(\mu)}{2 + \alpha(\mu)}\left[1 - \Big(-\frac{\alpha(\mu)}{2}\Big)^{k-1}\right]\mathrm{d}\mathbb{P}[W_n \leqslant w] + \\ &\quad + \frac{1}{\mu}\Big(-\frac{\alpha(\mu)}{2}\Big)^{k-1}\alpha(\mu)\int_0^\infty w\mathrm{e}^{-\mu w}\,\mathrm{d}\mathbb{P}[W_n \leqslant w] \\ &= \frac{\mathbb{E}[W_n]}{\mu}\frac{2\,\alpha(\mu)}{2 + \alpha(\mu)}\left[1 - \Big(-\frac{\alpha(\mu)}{2}\Big)^{k-1}\right] + \\ &\quad + \Big(-\frac{\alpha(\mu)}{2}\Big)^{k-1}\alpha(\mu)\int_0^\infty w\mathrm{e}^{-\mu w}\mu\mathrm{e}^{-\mu w}\frac{\mathbb{P}[W_n > 0]}{\mu}\,\mathrm{d}w \\ &= \frac{\mathbb{E}[W_n]}{\mu}\frac{2\,\alpha(\mu)}{2 + \alpha(\mu)}\left[1 - \Big(-\frac{\alpha(\mu)}{2}\Big)^{k-1}\right] - \frac{\mathbb{E}[W_n]}{2\mu}\Big(-\frac{\alpha(\mu)}{2}\Big)^{k}.\end{aligned}$$

Note that the above expression is valid only for $n \geqslant 2$, since we have substituted $\mathrm{d}\mathbb{P}[W_n \leqslant w]$ by $\mu \mathrm{e}^{-\mu w}\mathbb{P}[W_n > 0]\,\mathrm{d}w$, cf. Theorem 4.1. However, if we further assume that W_1, given that $W_1 > 0$, is (like all other W_n's) exponentially distributed with rate μ then the above expression is valid for $n = 1$ too.

All that is left in order to compute the covariance of W_n and W_{n+k} is to compute $\mathbb{E}[W_n]$ and $\mathbb{E}[W_{n+k}]$. To this end, note that for $k \geqslant 1$

$$\begin{aligned}
\mathbb{E}[W_{n+k}] &= \mathbb{E}[W_{n+k} \mid W_{n+k} > 0]\,\mathbb{P}[W_{n+k} > 0] = \frac{1}{\mu}\,\mathbb{P}[W_{n+k} > 0] \\
&= \frac{1}{\mu}\int_0^\infty \mathbb{P}[W_{n+k} > 0 \mid W_n = w]\,\mathrm{d}\mathbb{P}[W_n \leqslant w] \\
&= \frac{1}{\mu}\frac{2\,\alpha(\mu)}{2+\alpha(\mu)}\left[1 - \Big(-\frac{\alpha(\mu)}{2}\Big)^{k-1}\right] + \\
&\quad + \frac{\alpha(\mu)}{\mu}\Big(-\frac{\alpha(\mu)}{2}\Big)^{k-1}\int_0^\infty \mathrm{e}^{-\mu w}\,\mathrm{d}\mathbb{P}[W_n \leqslant w] \\
&= \frac{1}{\mu}\frac{2\,\alpha(\mu)}{2+\alpha(\mu)}\left[1 - \Big(-\frac{\alpha(\mu)}{2}\Big)^{k-1}\right] + \\
&\quad + \frac{\alpha(\mu)}{\mu}\Big(-\frac{\alpha(\mu)}{2}\Big)^{k-1}\Big(\mathbb{P}[W_n = 0] + \frac{\mathbb{P}[W_n > 0]}{2}\Big),
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathbb{E}[W_n]\mathbb{E}[W_{n+k}] = \frac{\mathbb{E}[W_n]}{\mu}\frac{2\,\alpha(\mu)}{2+\alpha(\mu)}\left[1 - \Big(-\frac{\alpha(\mu)}{2}\Big)^{k-1}\right] - \\
- \frac{\mathbb{E}[W_n]}{\mu}\Big(-\frac{\alpha(\mu)}{2}\Big)^{k}\big(2\mathbb{P}[W_n = 0] + \mathbb{P}[W_n > 0]\big).
\end{aligned}$$

Putting everything together we obtain

$$\begin{aligned}
\mathrm{cov}[W_n, W_{n+k}] &= \mathbb{E}[W_n W_{n+k}] - \mathbb{E}[W_n]\mathbb{E}[W_{n+k}] \\
&= \frac{\mathbb{E}[W_n]}{\mu}\big(2\mathbb{P}[W_n = 0] + \mathbb{P}[W_n > 0]\big)\Big(-\frac{\alpha(\mu)}{2}\Big)^{k} - \\
&\quad - \frac{\mathbb{E}[W_n]}{2\mu}\Big(-\frac{\alpha(\mu)}{2}\Big)^{k}.
\end{aligned} \tag{4.19}$$

Simplifying this formula, using $\mathbb{P}[W_n = 0] + \mathbb{P}[W_n > 0] = 1$, we obtain the following theorem, which is the main result of this section.

Theorem 4.4. *For $n \geqslant 2$, the covariance function between W_n and W_{n+k}, $k \geqslant 1$, is given by*

$$\mathrm{cov}[W_n, W_{n+k}] = \frac{\mathbb{E}[W_n]}{\mu}\Big(\mathbb{P}[W_n = 0] + \frac{1}{2}\Big)\Big(-\frac{\alpha(\mu)}{2}\Big)^{k}. \tag{4.20}$$

Furthermore, if W_1, given that $W_1 > 0$, has an exponential distribution with rate μ, then the above expression is valid for $n = 1$ too.

So the covariance function decays geometrically fast. Equation (4.20) is not valid for $k = 0$. The proof fails, for example, when computing $\mathbb{E}[W_{n+k}]$. For $k = 0$ we proceed as follows. Since W_n conditioned on $W_n > 0$ has an exponential distribution with rate μ, we have that

$$\text{var}[W_n] = \frac{\mathbb{P}[W_n > 0](2 - \mathbb{P}[W_n > 0])}{\mu^2}.$$

A particularly tractable case arises when we assume that

$$\mathbb{P}[W_1 > 0] = \frac{2\,\alpha(\mu)}{2 + \alpha(\mu)},$$

which makes $\{W_n\}$ a stationary process. In this case, we obtain the following expression for the covariance function.

Corollary 4.5. *If $\{W_n\}$ is stationary, then for $k \geqslant 1$, we have that*

$$\text{cov}[W_1, W_{1+k}] = \frac{2\,\alpha(\mu)}{2 + \alpha(\mu)}\Big(\frac{3}{2} - \frac{2\,\alpha(\mu)}{2 + \alpha(\mu)}\Big)\frac{1}{\mu^2}\Big(-\frac{\alpha(\mu)}{2}\Big)^k.$$

Theorem 4.4 can be applied to compute the covariance between waiting times in the alternating service example, where $W_1 = B_1$. For this particular case, we plot below the correlation coefficient between W_n and W_{n+k}. More specifically, for the exponentially distributed preparation times μ is chosen equal to one, while the service times A_n are assumed to follow the same Erlang distribution. The parameters of this distribution are chosen so that for two different mean service times, the squared coefficient of variation is equal to three specific values.

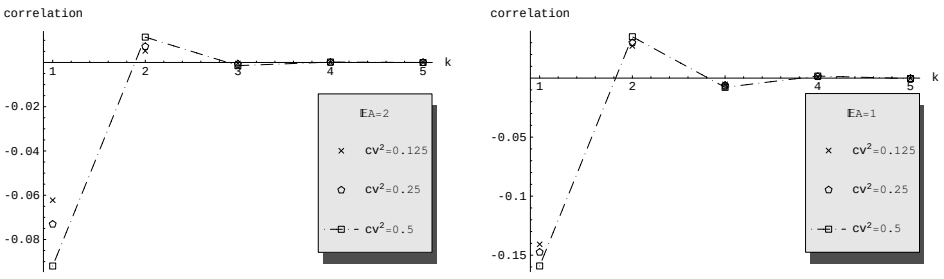

Figure 4.1: The effect of the squared coefficient of variation of A on the waiting-time correlation coefficient.

In Figure 4.1 we have chosen the parameters of the Erlang distribution that the service times follow in such a way, so that for $\mathbb{E}[A] = 1$ or $\mathbb{E}[A] = 2$, the squared coefficient of variation is equal to either 0.125 or 0.25 or 0.5. One can observe that as $\mathbb{E}[B]/\mathbb{E}[A]$ tends to 1, the effect of the squared coefficient of variation of A become negligible. This becomes more apparent in the case where $\mathbb{E}[B]/\mathbb{E}[A] = 2$. There, the

three points for the various cases we examine almost coincide. This is to be expected, since the more dominant the preparation times B_n are, compared to the service times A_n, the more limited is the effect of the distribution of A_n. Furthermore, one can observe that W_n is barely correlated to W_{n+4} or any subsequent waiting times.

However, in Figure 4.2 one can observe that the mean service time strongly affects the correlation coefficient between successive waiting times. Again in this setting, we choose $\mathbb{E}[B] = 1$ and we plot the correlation coefficient between W_n and W_{n+k} for successive values of k. Apparently, as $\mathbb{E}[B]/\mathbb{E}[A]$ tends to 1, or as $\mathbb{E}[A]$ decreases the correlation between waiting times is stronger. However, we have again that waiting times that are four or more steps apart are almost uncorrelated. Furthermore, the squared coefficient of variation of the service times seem to have almost no impact.

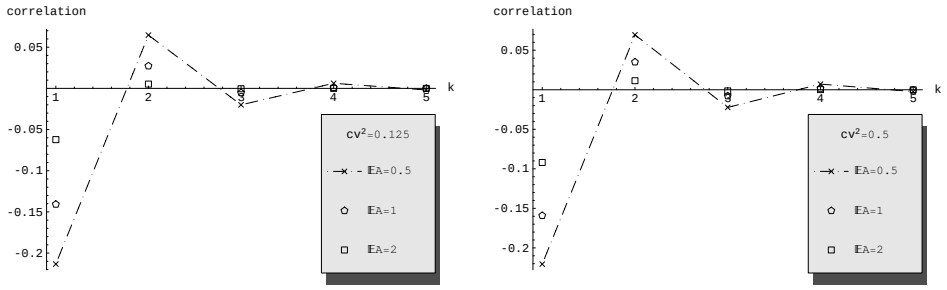

Figure 4.2: The effect of $\mathbb{E}[A]$ on the waiting-time correlation coefficient.

4.2.2 Time-dependent analysis for G/E

In this section we assume that, for all n, the i.i.d. random variables B_n follow an Erlang distribution with N phases and parameter μ. Although the analysis is not as straightforward as in Section 4.2.1, the idea we shall utilise in the following is very simple. Namely, if the random variables B_n follow an Erlang distribution, then we can completely describe the system in terms of a finite-state Markov chain. Thus, it suffices to compute the one-step transition probabilities of this Markov chain. This is done in the following, and is applied to show that W_n has a mixed-Erlang distribution. Subsequently, we derive expressions for the distribution of the cycle length and the covariance.

The time-dependent waiting-time distribution

Let A be a generic service time and E_i be a random variable that follows an Erlang distribution with i phases and parameter μ, which we denote by G_i. Define F_i to be the number of remaining preparation phases that the server sees after his $(i-1)$-th service completion, that is, at the moment he initiates his i-th waiting time. Observe

that $\{F_n\}$ is a Markov chain, and let

$$p_{ij} = \mathbb{P}[F_{n+1} = j \mid F_n = i].$$

Then, for $i, j \in \{1, \dots, N\}$ we have that

$$\begin{aligned} p_{ij} &= \mathbb{P}[\text{exactly } N-j \text{ exponential phases expired during } [0, A+E_i)] \\ &= \int_0^\infty \frac{(\mu x)^{N-j}}{(N-j)!}\, \mathrm{e}^{-\mu x}\, \mathrm{d}\mathbb{P}[A+E_i \le x] \\ &= \frac{(-\mu)^{N-j}}{(N-j)!}\, \mathcal{L}_{A+E_i}^{(N-j)}(\mu), \end{aligned}$$

where $\mathcal{L}_Y^{(N-j)}$ is the $N-j$-th derivative of the Laplace-Stieltjes transform of a random variable Y,

$$\begin{aligned} &= \frac{(-\mu)^{N-j}}{(N-j)!} \sum_{\ell=0}^{N-j} \binom{N-j}{\ell} \alpha^{(N-j-\ell)}(\mu) \left(\left(\frac{\mu}{\mu+s} \right)^i \right)^{(\ell)} \Bigg|_{s=\mu} \\ &= \frac{(-\mu)^{N-j}}{(N-j)!} \sum_{\ell=0}^{N-j} \binom{N-j}{\ell} \alpha^{(N-j-\ell)}(\mu) \left[\left(-\frac{1}{2\mu} \right)^\ell \frac{(i+\ell-1)!}{2^i (i-1)!} \right] \\ &= \frac{(-\mu)^{N-j}}{2^i} \sum_{\ell=0}^{N-j} \binom{i+\ell-1}{i-1} \frac{\alpha^{(N-j-\ell)}(\mu)}{(N-j-\ell)!} \left(-\frac{1}{2\mu} \right)^\ell. \end{aligned}$$

Furthermore, for $j \in \{1, \dots, N\}$ we have that

$$\begin{aligned} p_{0j} &= \mathbb{P}[\text{exactly } N-j \text{ exponential phases expired during } [0, A)] \\ &= \frac{(-\mu)^{N-j}}{(N-j)!}\, \alpha^{(N-j)}(\mu). \end{aligned}$$

The rest of the transition probabilities can be computed by the relations

$$p_{00} = 1 - \sum_{i=0}^{N-1} \frac{(-\mu)^i}{i!}\, \alpha^{(i)}(\mu)$$

and

$$p_{i0} = 1 - \sum_{j=0}^{N-1} \frac{(-\mu)^j}{2^i} \sum_{\ell=0}^{j} \binom{i+\ell-1}{i-1} \frac{\alpha^{(j-\ell)}(\mu)}{(j-\ell)!} \left(-\frac{1}{2\mu} \right)^\ell.$$

Let $\mathbf{P} = (p_{ij})$ be the transition matrix and define $G_0(x) = 1$. Then, the distribution of W_n is given by

$$\begin{aligned} \mathbb{P}[W_n \leqslant x] &= \sum_{i=0}^{N} \mathbb{P}[W_n \leqslant x \mid F_n = i] \mathbb{P}[F_n = i] \\ &= \sum_{i=0}^{N} G_i(x) \mathbb{P}[F_n = i]. \end{aligned}$$

Let $\varpi_{n,i} = \mathbb{P}[F_n = i \mid W_1 = w]$ and $\boldsymbol{\varpi_n}$ be the column-vector $(\varpi_{n,0}, \ldots, \varpi_{n,N})^{\mathrm{T}}$. Then

$$\boldsymbol{\varpi_n} = \mathbf{P}\boldsymbol{\varpi_{n-1}} = \mathbf{P}^{n-2}\boldsymbol{\varpi_2}. \tag{4.21}$$

It remains to compute $\boldsymbol{\varpi_2}$. In the same way we computed p_{ij} we get, for $j \geqslant 1$,

$$\begin{aligned}
\varpi_{2,j} &= \mathbb{P}[\text{exactly } N-j \text{ exponential phases expired during } [0, A+w)] \\
&= \frac{(-\mu)^{N-j}}{(N-j)!}\,\mathcal{L}_{A+w}^{(N-j)}(\mu) \\
&= \frac{(-\mu)^{N-j}}{(N-j)!}\sum_{\ell=0}^{N-j}\binom{N-j}{\ell}\alpha^{(N-j-\ell)}(\mu)\left(\mathrm{e}^{-sw}\right)^{(\ell)}\Big|_{s=\mu} \\
&= (-\mu)^{N-j}\mathrm{e}^{-\mu w}\sum_{\ell=0}^{N-j}\frac{(-w)^{\ell}}{\ell!}\frac{\alpha^{(N-j-\ell)}(\mu)}{(N-j-\ell)!}.
\end{aligned} \tag{4.22}$$

This also characterises $\varpi_{2,0}$. Putting everything together, we obtain the main result of this section.

Theorem 4.6. *For every $n \geqslant 2$, W_n has a mixed-Erlang distribution with parameters μ and $\varpi_{n,0}, \ldots, \varpi_{n,N}$, i.e.,*

$$\mathbb{P}[W_n \leqslant x \mid W_1 = w] = \sum_{i=0}^{N}\varpi_{n,i}G_i(x),$$

with ϖ_n given by Equations (4.21) and (4.22).

It is interesting to note that W_n has a phase-type distribution with $N+1$ phases for all $n \geqslant 2$. This is strikingly different from Lindley's recursion. In Section 4.6, we shall see that for the G/M/1 queue, the waiting time of the n-th customer follows a mixed-Erlang distribution with at most $n+1$ phases; that is, as n grows, the number of phases grows linearly in n.

The distribution of the cycle length

As before, define the cycle length C to be given by

$$C = \inf\{k : W_{1+k} = 0 \mid W_1 = 0\} = \inf\{k : F_{1+k} = 0 \mid F_1 = 0\}.$$

We have that $\mathbb{P}[C = n] = \mathbb{P}[C > n] - \mathbb{P}[C > n+1]$ and

$$\mathbb{P}[C > n] = \sum_{i_0=1}^{N}\mathbb{P}[F_{n+1} = i_0, F_2 \cdot \ldots \cdot F_n > 0 \mid F_1 = 0]. \tag{4.23}$$

Let $t_{0,i_0}^{(n)}$ be the probability that, conditioning on the fact that $F_1 = 0$, we shall go to state i_0 in n steps without passing through state 0 while doing so. Then we have

that

$$\begin{aligned}
t_{0,i_0}^{(n)} &= \mathbb{P}[F_{n+1} = i_0, F_2 \cdot \ldots \cdot F_n > 0 \mid F_1 = 0] \\
&= \sum_{i_1=1}^{N} \mathbb{P}[F_{n+1} = i_0, F_n = i_1, F_2 \cdot \ldots \cdot F_{n-1} > 0 \mid F_1 = 0] \\
&= \sum_{i_1=1}^{N} \mathbb{P}[F_{n+1} = i_0 \mid F_n = i_1, F_2 \cdot \ldots \cdot F_{n-1} > 0, F_1 = 0] \times \\
&\qquad \times \mathbb{P}[F_n = i_1, F_2 \cdot \ldots \cdot F_{n-1} > 0 \mid F_1 = 0] \\
&= \sum_{i_1=1}^{N} p_{i_1 i_0} t_{0,i_0}^{(n-1)} = \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_{n-1}=1}^{N} p_{i_1 i_0}\, p_{i_2 i_1} \cdot \ldots \cdot p_{i_{n-1} i_{n-2}} \times \\
&\qquad \times \mathbb{P}[F_2 = i_{n-1} \mid F_1 = 0]
\end{aligned}$$

Then from (4.23) we have that

$$\mathbb{P}[C > n] = \sum_{i_0=1}^{N} \sum_{i_1=1}^{N} \cdots \sum_{i_{n-1}=1}^{N} p_{i_1 i_0} \cdot \ldots \cdot p_{i_{n-1} i_{n-2}}\, p_{0 i_{n-1}}. \tag{4.24}$$

So, if we define $\mathbf{Q}$ to be the matrix that we obtain if we omit the first line and the first row of the matrix $\mathbf{P}$, $\mathbf{q}$ to be the first row of $\mathbf{P}$ apart from the first element, $\mathbf{I}$ to be the $N \times N$ identity matrix, and $\mathbf{e}$ to be the column vector with all its N entries equal to one, then (4.24) can be rewritten in a more compact form, which is done in the following theorem.

Theorem 4.7. *For every $n \geqslant 1$ we have that the distribution of the cycle length is given by*

$$\mathbb{P}[C = n] = \mathbf{q}\mathbf{Q}^{n-1}(\mathbf{I} - \mathbf{Q})\mathbf{e}.$$

The covariance function

It should be clear by now that extending the results of Section 4.2.1 to Erlang distributed preparation times is feasible, although not as straightforward as before. The results now are given implicitly, in terms of the transition matrix of a finite-state Markov chain. When computing the covariance function between W_n and W_{n+k} the calculations become more complex, and rather long and tedious. In this section we shall only outline the procedure of computing $\operatorname{cov}[W_n, W_{n+k}]$ when, for all n, B_n follows an Erlang distribution.

As before, it suffices to calculate the expectation $\mathbb{E}[W_n W_{n+k}]$, since $\mathbb{E}[W_n]$ and $\mathbb{E}[W_{n+k}]$ can, in principle, be computed directly from the time-dependent distribution. Therefore, it suffices to compute $\mathbb{E}[W_{n+k} \mid W_n = w]$. To this end, we have

that for all $k \geqslant 1$

$$\begin{aligned}\mathbb{E}[W_{n+k} \mid W_n = w] &= \mathbb{E}[W_{1+k} \mid W_1 = w] \\ &= \sum_{i=0}^{N} \mathbb{E}[W_{1+k} \mid F_{1+k} = i, W_1 = w]\, \mathbb{P}[F_{1+k} = i \mid W_1 = w]. \end{aligned} \tag{4.25}$$

Clearly, for any event E depending only on $W_1, \ldots, W_{n-1}$, we have that for $n \geqslant 2$

$$\mathbb{P}[W_n \leqslant x \mid F_n = i, E] = G_i(x).$$

This equation is analogous to Equation (4.8). So (4.25) now becomes

$$\mathbb{E}[W_{n+k} \mid W_n = w] = \sum_{i=0}^{N} \frac{i}{\mu}\, \mathbb{P}[F_{1+k} = i \mid W_1 = w].$$

We have shown that for all $n \geqslant 3$, $\boldsymbol{\varpi_n} = \mathbf{P}^{n-2}\boldsymbol{\varpi_2}$, and the vector $\boldsymbol{\varpi_2}$ is computed, cf. (4.22). From this, we infer that, for all $k \geqslant 1$, $\mathbb{P}[F_{1+k} = i \mid W_1 = w]$ is a polynomial of degree $N - i$ multiplied by $\mathrm{e}^{-\mu w}$. Let $c_{1+k,\,j}$, $j = 0, \ldots, N - i$ be the constants of this polynomial. Then, for $n \geqslant 2$ and $k \geqslant 1$ we have that

$$\mathbb{E}[W_n W_{n+k}] = \sum_{i=0}^{N} \frac{i}{\mu} \sum_{j=0}^{N-i} c_{1+k,\,j} \int_0^\infty w^{j+1} \mathrm{e}^{-\mu w}\, \mathrm{d}\mathbb{P}[W_n \leqslant w].$$

A lengthy but straightforward computation, using Theorem 4.6, shows that

$$\int_0^\infty w^{j+1} \mathrm{e}^{-\mu w}\, \mathrm{d}\mathbb{P}[W_n \leqslant w] = \left(\frac{1}{2}\right)^{j+1} \sum_{\ell=0}^{N} \varpi_{n,\ell} \frac{(\ell + j)!}{2^\ell (\ell - 1)!}.$$

Using Theorem 4.6, we can also compute $\mathbb{E}[W_n]$ and $\mathbb{E}[W_{n+k}]$. Unfortunately, the resulting expression for the covariance is rather complicated and it is therefore omitted.

Remark 4.1. If the preparation times do not have an Erlang distribution, but a mixed-Erlang distribution, the analysis stays completely the same, except for the computation of the transition probabilities p_{ij} and $\varpi_{2,j}$.

4.2.3 Steady-state distribution for G/E

In Section 4.2.2 we have derived the time-dependent distribution of the waiting times in case the preparation times follow an Erlang distribution. The distribution there is given in terms of the equilibrium distribution of a finite-state Markov chain. In this section we study the steady-state distribution of the waiting times for the same setting. Although the steady-state distribution can be seen as the limit for $n \to \infty$ of the expression in Theorem 4.6, the results presented in this section are derived straightforwardly through simple manipulations of Laplace transforms. This approach is simple, direct, and it does not involve the embedded Markov chain, or

its equilibrium distribution. The use of Laplace transforms is a standard approach for the analysis of Lindley's equation. Hence it is natural to try this approach for Equation (1.2).

In the following section, we derive the distribution of the waiting time of the server, assuming that the generic service time A follows some general distribution and the generic preparation time B follows a phase-type distribution. As mentioned before, the phase-type distributions that we consider (cf. Section 4.2.4) are mixtures of Erlang distributions with the same scale parameters. In Section 3.5 we have seen that the study of Erlang service time facilitated the analysis for phase-type service times. The same is valid here; namely, the analysis of phase-type preparation times is a simple expansion of the results derived in this section.

To this end, let B be the sum of N independent random variables $Y_1, \dots, Y_N$ that are exponentially distributed with parameter μ. Then we can readily prove the following.

Theorem 4.8 (Alternating service system)**.** *The waiting-time distribution has a mass π_0 at the origin, which is given by*

$$\pi_0 = \mathbb{P}[B < W + A] = 1 - \sum_{i=0}^{N-1} \frac{(-\mu)^i}{i!}\, \phi^{(i)}(\mu)$$

and has a density f_W on $[0, \infty)$ that is given by

$$f_W(x) = \mu^N \mathrm{e}^{-\mu x} \sum_{i=0}^{N-1} \frac{(-1)^i}{i!}\, \phi^{(i)}(\mu) \frac{x^{N-1-i}}{(N-1-i)!}. \tag{4.26}$$

In the above expression, we have that

$$\phi^{(i)}(\mu) = \sum_{k=0}^{i} \binom{i}{k} \omega^{(k)}(\mu)\, \alpha^{(i-k)}(\mu)$$

and that the parameters $\omega^{(i)}(\mu)$ for $i = 0, \dots, N-1$ are the unique solution to the system of equations

$$\omega(\mu) = 1 - \sum_{i=0}^{N-1} (-\mu)^i \Big(1 - \frac{1}{2^{N-i}}\Big) \sum_{k=0}^{i} \frac{\omega^{(k)}(\mu)\, \alpha^{(i-k)}(\mu)}{k!\,(i-k)!}$$

and for $\ell = 1, \dots, N-1$ (4.27)

$$\omega^{(\ell)}(\mu) = \sum_{i=0}^{N-1} \mu^{i-\ell} \frac{(-1)^{i+\ell}}{2^{N-i+\ell}} \frac{(N-i+\ell-1)!}{(N-i-1)!} \sum_{k=0}^{i} \frac{\omega^{(k)}(\mu)\, \alpha^{(i-k)}(\mu)}{k!\,(i-k)!}.$$

Proof. As before, for a random variable Q and an event E we use the following notation: $\mathbb{E}[Q\,;E] = \mathbb{E}[Q \cdot \mathbb{1}_{[E]}]$. Consider the Laplace transform of (1.2); then we

have that

$$\begin{aligned}\omega(s) = \mathbb{E}[\mathrm{e}^{-sW}] &= \mathbb{P}[B < W + A] + \mathbb{E}[\mathrm{e}^{-s(B-W-A)}\,; B \geqslant W + A] \\ &= \mathbb{P}[B < W + A] + \mathbb{E}[\mathrm{e}^{-s(B-W-A)}\,; Y_1 \geqslant W + A] + \\ &+ \sum_{i=1}^{N-1} \mathbb{E}[\mathrm{e}^{-s(B-W-A)}\,; Y_1 + \cdots + Y_i \leqslant W + A \leqslant Y_1 + \cdots + Y_{i+1}].\end{aligned} \tag{4.28}$$

Using standard techniques and the memoryless property of the exponential distribution one can show that

$$\begin{aligned}&\mathbb{E}[\mathrm{e}^{-s(B-W-A)}\,; Y_1 \geqslant W + A] \\ &\quad= \mathbb{E}[\mathrm{e}^{-s(Y_2+Y_3+\cdots+Y_N)}\mathrm{e}^{-s(Y_1-W-A)}\,; Y_1 \geqslant W + A] \\ &\quad= \Big(\frac{\mu}{\mu+s}\Big)^{N-1} \mathbb{E}[\mathrm{e}^{-s(Y_1-W-A)}\,; Y_1 \geqslant W + A] \\ &\quad= \Big(\frac{\mu}{\mu+s}\Big)^{N-1} \mathbb{E}[\mathrm{e}^{-s(Y_1-W-A)} \mid Y_1 \geqslant W + A]\,\mathbb{P}[Y_1 \geqslant W + A] \\ &\quad= \Big(\frac{\mu}{\mu+s}\Big)^{N} \mathbb{P}[Y_1 \geqslant W + A] \quad \text{(due to the memoryless property)} \\ &\quad= \Big(\frac{\mu}{\mu+s}\Big)^{N} \omega(\mu)\,\alpha(\mu).\end{aligned} \tag{4.29}$$

Additionally, for $Z_i = Y_1 + \cdots + Y_i$ we have that

$$\begin{aligned}&\mathbb{E}[\mathrm{e}^{-s(B-W-A)}\,; Z_i \leqslant W + A \leqslant Z_{i+1}] \\ &= \Big(\frac{\mu}{\mu+s}\Big)^{N-i-1} \mathbb{E}[\mathrm{e}^{-s(Z_{i+1}-W-A)} \mid Z_i \leqslant W + A \leqslant Z_{i+1}]\,\mathbb{P}[Z_i \leqslant W + A \leqslant Z_{i+1}] \\ &\qquad= \Big(\frac{\mu}{\mu+s}\Big)^{N-i} \frac{(-\mu)^i \phi^{(i)}(\mu)}{i!}.\end{aligned} \tag{4.30}$$

Finally, we calculate the probability $\mathbb{P}[B < W + A]$ by substituting $s = 0$ in (4.28) and using equations (4.29) and (4.30). Straightforward calculations give us now that

$$\omega(s) = 1 - \sum_{i=0}^{N-1} \frac{(-\mu)^i}{i!}\,\phi^{(i)}(\mu)\Big(1 - \Big(\frac{\mu}{\mu+s}\Big)^{N-i}\Big). \tag{4.31}$$

Inverting the transform yields the density (4.26).

Furthermore, the terms $\omega^{(i)}(\mu)$, $i = 0, \ldots, N-1$, that are included in $\phi^{(i)}(\mu)$ still need to be determined. To obtain the values of $\omega^{(i)}(\mu)$, for $i = 0, \ldots, N-1$, we differentiate (4.31) $N-1$ times and we evaluate $\omega^{(i)}(s)$, $i = 0, \ldots, N-1$ at the point $s = \mu$. This gives us the system of equations (4.27). The fact that the solution of the system is unique follows from the general theory of Markov chains that implies that there is a unique equilibrium distribution and thus also a unique solution to (4.27). □

Corollary 4.9. *The throughput θ satisfies*

$$\theta^{-1} = \mathbb{E}[W] + \mathbb{E}[A] = \sum_{i=0}^{N-1} \frac{(-1)^i}{i!}\, \phi^{(i)}(\mu)\mu^{i-1}(N-i) - \alpha'(0).$$

It is quite interesting to note that the density of the waiting time can be rewritten as

$$f_W(x) = \mu \mathrm{e}^{-\mu x} \sum_{i=1}^{N} \pi_i \frac{(\mu x)^{i-1}}{(i-1)!},$$

where

$$\pi_i = \frac{(-\mu)^{N-i}\phi^{(N-i)}(\mu)}{(N-i)!}$$

is the probability that directly after a service completion exactly i exponential phases of B remain. Thus, $\{\pi_i\}$, $i = 0, \ldots, n$, is the equilibrium distribution associated with the matrix $\mathbf{P}$ in Section 4.2.2, i.e. π_i is the limit of $\varpi_{n,i}$ as n goes to infinity.

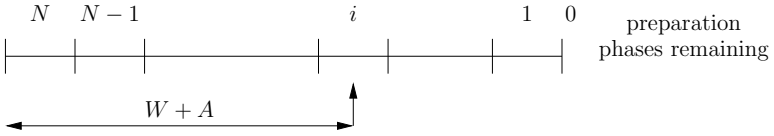

Figure 4.3: The waiting time has a mixed-Erlang distribution.

As it is also clear from Figure 4.3, with probability π_i the distribution of the waiting time is Erlang with i phases, for $i = 1, \ldots, N$. Furthermore, the probability π_0 that the server does not have to wait, or equivalently that at least N exponential phases expired, is $\pi_0 = 1 - \sum_{i=1}^{N} \pi_i$.

Remark 4.2. For $N = 1$ Theorem 4.8 gives the steady-state density of the waiting time in case B is exponentially distributed. In this case, we obtain that

$$\pi_0 = 1 - \frac{2\,\alpha(\mu)}{2 + \alpha(\mu)}$$

and

$$f_W(x) = \mu \mathrm{e}^{-\mu x} \frac{2\,\alpha(\mu)}{2 + \alpha(\mu)},$$

which implies that

$$F_W(x) = 1 - \frac{2\,\alpha(\mu)}{2 + \alpha(\mu)}\, \mathrm{e}^{-\mu x}.$$

This coincides with Theorem 4.1 (or Corollary 4.2 since the distribution of W_1 no longer plays a role in the steady-state distribution of the waiting times) for $n \to \infty$.

Apparently, the same remark is not so straightforward when considering Erlang preparation times, since in this case the time-dependent distribution is derived in terms of the equilibrium distribution of a finite-state Markov chain. This justifies the usage of Laplace transforms for the steady-state distribution, since the result in this case is simple to describe.

So, practically, for the steady-state distribution of the waiting time, the problem is reduced to obtaining the solution of an $N \times N$ linear system. Extending the above result to mixtures of Erlang distributions is simple.

4.2.4 Steady-state distribution for G/PH

For $n = 1, \ldots, N$, let the random variable Y_n follow an Erlang distribution with parameter μ and n phases, and let the random variable B of the preparation times be equal to Y_n with a probability κ_n. In other words the distribution function of B is given by

$$F_B(x) = \sum_{n=1}^{N} \kappa_n \left(1 - \mathrm{e}^{-\mu x} \sum_{j=0}^{n-1} \frac{(\mu x)^j}{j!} \right), \qquad x \geqslant 0. \tag{4.32}$$

So, if $\kappa_n = 0$ for all $n = 1, \ldots, N-1$, we retrieve the Erlang distribution of the previous section. This class of phase-type distributions may be used to approximate any given distribution on $[0, \infty)$ for the preparation times arbitrarily close; see Schassberger [149]. Below we show that Theorem 4.8 can be extended to service distributions of the form (4.32).

By conditioning on the number of phases of B, we find that

$$\omega(s) = \mathbb{E}[\mathrm{e}^{-sW}] = \mathbb{P}[B < W + A] + \sum_{n=1}^{N} \kappa_n \mathbb{E}[\mathrm{e}^{-s(Y_n - W - A)}\,; Y_n \geqslant W + A].$$

Since Y_n follows now an Erlang distribution with n phases, the last equation is practically a linear combination of Equation (4.28), summed over all probabilities κ_n for $n = 1, \ldots, N$. This means that we can directly use the analysis of Section 4.2.3 to calculate the Laplace transform of W in this situation (cf. Equation (4.31)). So we have that

$$\omega(s) = 1 - \sum_{n=1}^{N} \kappa_n \sum_{i=0}^{n-1} \frac{(-\mu)^i}{i!} \phi^{(i)}(\mu) \left(1 - \left(\frac{\mu}{\mu + s} \right)^{n-i} \right), \tag{4.33}$$

where the terms $\phi^{(i)}(\mu)$ can be calculated in a similar fashion as previously. Inverting (4.33) yields the density of the following theorem (cf. Theorem 4.8).

Theorem 4.10. *Let (4.32) be the distribution of the preparation time B. Then the distribution of the server's waiting time has mass π_0 at zero which is given by*

$$\pi_0 = \mathbb{P}[B < W + A] = 1 - \sum_{n=1}^{N} \sum_{i=0}^{n-1} \kappa_n \frac{(-\mu)^i}{i!} \phi^{(i)}(\mu)$$

and has a density on $[0,\infty)$ *that is given by*

$$f_W(x) = \sum_{n=1}^{N} \kappa_n \left(\mu^n \mathrm{e}^{-\mu x} \sum_{i=0}^{n-1} \frac{(-1)^i}{i!} \phi^{(i)}(\mu) \frac{x^{n-1-i}}{(n-1-i)!} \right).$$

One can already see from the above theorem the effect of the different sign in the Lindley-type equation that describes our model. The waiting-time distribution for the GI/PH/1 queue is a mixture of exponentials with different scale parameters (Adan and Zhao [2]) and it is shown to be phase type (Asumssen [5]). In our case we have that the waiting-time distribution is a mixture of Erlang distributions with the same scale parameter for all exponential phases.

As we have mentioned before, the practice of alternating between the service points is inevitably followed in many situations. Still it seems reasonable to argue that it would be more efficient to choose to serve the first customer that has completed his preparation time. If we drop the assumption that the server alternates between the service points then we have the classical machine repair problem, which is the subject of the following section.

4.3 The non-alternating case

There are several variations of the machine repair problem. In the classical machine repair problem , there is a number of machines that are served by a unique repairman when they fail. The machines are working independently and as soon as a machine fails, it joins a queue formed in front of the repairman where it is served in order of arrival. A machine that is repaired is assumed to be as good as new. The model described in Section 1.4 is exactly the classical machine repair problem with two machines after we drop the assumption that the server alternates between the service points. Instead of alternating, the server will serve the first customer that completes his preparation phase, irrespectively of the service point the previous customer occupied. There are two machines that work in parallel (the two service points), the preparation time of the customer is equivalent to the life time of the machine until it fails and the service time of the customer is the time the repairman needs to repair the machine.

What we are interested in is the waiting time of the repairman until a machine breaks down or, in other words, the waiting time of the server until the preparation phase of one of the customers is completed. It is quite surprising that although the machine repair problem under general assumptions is thoroughly treated in the literature, this question remains unanswered. We would like to compare steady-state properties of the alternating service model described in Section 1.4 and the machine repair problem with two machines. Therefore, for this specific machine repair problem, we first need to derive the distribution of the waiting time of the server, when the system is in steady state. The results in this section are limited to the classical machine repair problem with two machines, and do not apply to a greater number of machines. In the following we will refer to the *server* or *customers*

instead of the *repairman* or *machines* in order to illustrate the analogies between the two models.

Let B be the random variable of the time needed for the preparation phase and R be the remaining preparation time just after a service has been completed. Then obviously the waiting time of the server is $W = \min\{B, R\}$. The random variables B and R are independent, so in order to calculate the distribution of W we need the distribution of R. In agreement to the alternating service model, B follows an Erlang distribution with N phases. Note that we do not have a simple Lindley-type recursion for W and therefore this system cannot be easily treated with Laplace transforms. This means that we have to try an alternative approach.

This approach is analogous to the one presented in Section 4.2.2 for the time-dependent analysis of the alternating model. Namely, the system can be fully described by the number of remaining phases of preparation time that a customer has to complete, immediately after a service completion. The state space is finite, since there can be at most N phases remaining and the Markov chain is aperiodic and irreducible, so there is a unique equilibrium distribution $\{\varpi_i\}$, $i = 0, \ldots, N$. After completing a service, the other customer may be already waiting for the server (so the N exponential phases of the Erlang distribution of the preparation have expired) or he is in one of the N phases of the preparation time. This means that the remaining preparation time R that the server sees immediately after completing a service follows the mixed-Erlang distribution $F_R(x) = \varpi_0 + \varpi_1 G_1(x) + \cdots + \varpi_N G_N(x)$ (recall that G_i is the Erlang distribution with i phases).

So in order to derive the distribution of R (and consequently the distribution of W), we need to solve the equilibrium equations $\varpi_i = \sum_k \varpi_k \, p_{ki}$, $i = 0, \ldots, N$, in conjunction with the normalising equation $\sum_k \varpi_k = 1$, where p_{ki} are the one-step transition probabilities. Let us determine the probabilities p_{ij}, for all $i, j \in \{0, \ldots, N\}$. This can be done in the same fashion as in Section 4.2.2.

A transition from state i to state j, for $i, j \in \{1, \ldots, N\}$, can be achieved in two ways: either the customer that has just been served or the other one will finish the preparation phase first. Suppose that the customer that has just been served finishes first. In this case we know that the last event just before the service starts is that the N-th phase of that customer expired. The other customer was in state k and during the service time the other customer reached state j, i.e. $k - j$ phases of that customer have expired. The probability of this event is given by

$$\sum_{k=j}^{N} \left(\frac{1}{2}\right)^{N+i-k} \binom{N+i-k-1}{N-1} \frac{(-\mu)^{k-j}}{(k-j)!} \, \alpha^{(k-j)}(\mu),$$

where α is as before the Laplace transform of the service time. Note that in the above expression we have that

$$\mathbb{P}[\text{exactly } k-j \text{ exponential phases expired during } [0, A)] = \frac{(-\mu)^{k-j}}{(k-j)!} \, \alpha^{(k-j)}(\mu).$$

Similarly we can determine the probability of a transition from state i to state j in

the second case. So in the end we have that for all $i, j \in \{1, \ldots, N\}$,

$$p_{ij} = \sum_{k=j}^{N} \left(\frac{1}{2}\right)^{N+i-k} \left[\binom{N+i-k-1}{N-1} + \binom{N+i-k-1}{N-k} \right] \frac{(-\mu)^{k-j}}{(k-j)!}\, \alpha^{(k-j)}(\mu). \tag{4.34}$$

The transition probabilities from state zero to any state $i = 1, \ldots, N$ are

$$p_{0i} = \frac{(-\mu)^{N-i}}{(N-i)!}\, \alpha^{(N-i)}(\mu), \tag{4.35}$$

since starting from state zero means that the other customer was already waiting when the repairman finished a service and reaching state i means that during the service time, exactly $N-i$ exponential phases expired. For the transition from state zero to state zero we have that during the service time at least N exponential phases expired, so

$$p_{00} = \sum_{i=N}^{\infty} \frac{(-\mu)^{i}}{(i)!}\, \alpha^{(i)}(\mu). \tag{4.36}$$

Similarly, we have that for $i = 1, \ldots, N$

$$p_{i0} = \sum_{k=1}^{N} \left(\frac{1}{2}\right)^{N+i-k} \left[\binom{N+i-k-1}{N-1} + \binom{N+i-k-1}{N-k} \right] \left(\sum_{j=k}^{\infty} \frac{(-\mu)^{j}}{(j)!}\, \alpha^{(j)}(\mu) \right), \tag{4.37}$$

where $\binom{a}{b} = 0$ for $0 \leqslant a < b$.

With the one-step transition probabilities one can determine the equilibrium distribution and thus F_R. Then we have that the distribution of the waiting time of the server, if we drop the assumption that he is alternating between the service points, is given by the following theorem.

Theorem 4.11 (Non-alternating service system)**.** *The waiting-time distribution is*

$$F_W(x) = F_R(x) + F_B(x) - F_R(x)F_B(x),$$

where F_R is the distribution of the remaining preparation time of a customer and is equal to

$$F_R(x) = \varpi_0 + \varpi_1 G_1(x) + \cdots + \varpi_N G_N(x).$$

In the above expression, $\{\varpi_i\}$, $i = 0, \ldots, N$, is the unique solution to the system of equations

$$\varpi_i = \sum_{k=0}^{N} \varpi_k\, p_{ki} \quad \textit{and} \quad \sum_{k=0}^{N} \varpi_k = 1, \quad \textit{for } i = 0, \ldots, N,$$

where p_{ij} are given by the equations (4.34)-(4.37).

Remark 4.3. The above results can be easily extended to phase-type preparation times of the form (4.32). However, this extension does not contribute significantly to the analysis, since it is along the same lines of the analysis in this section.

Remark 4.4. The one-step transition probabilities derived in this section allow for a similar analysis for this model as in Section 4.2.2. We can namely derive the time-dependent distribution, the cycle-length distribution, and the covariance function between two waiting times in the same fashion as before.

This method of defining a Markov chain through the remaining phases of the preparation time after a service has been completed and using the equilibrium distribution in order to calculate the mixing probabilities of R can, of course, also be used in the alternating service system. In that case, the waiting time W is exactly the remaining preparation time R. Furthermore, as we have noted, the probabilities π_i, for $i = 0, \ldots, N$ as defined in Section 4.2.3 form the equilibrium distribution of the underlying Markov chain associated with the transition matrix $\mathbf{P}$ introduced in Section 4.2.2. Furthermore, the system of equations (4.27) can be rewritten as follows:

$$\omega(\mu) = \pi_0 + \sum_{i=1}^{N} \frac{\pi_i}{2^i}$$

$$\omega^{(\ell)}(\mu) = \sum_{i=1}^{N} \frac{\pi_i(-\mu)^{-\ell}}{2^{i+\ell}} \frac{(i+\ell-1)!}{(i-1)!} \quad \text{for } \ell = 1, \ldots, N-1.$$

4.4 Performance comparison

One may wonder if there is any connection between the waiting time of the server in the two models that can help in understanding how the models perform. From this point on in this chapter, we will use the superscript A (NA) for all variables associated with the (non-)alternating service system when we specifically want to distinguish between the two situations. Otherwise the superscript will be suppressed. So, for example, the random variable W^{A} will be the waiting time of the server in the alternating service system.

4.4.1 Stochastic ordering

Suppose that the distributions of the two random variables X and Y have a common support. Then the stochastic ordering $X \geqslant_{st} Y$ is defined as (cf. [109, 125, 157])

$$\mathbb{P}[X \geqslant x] \geqslant \mathbb{P}[Y \geqslant x], \quad \text{for all } x \text{ in the support},$$

and we say that X *dominates* Y.

Intuitively one may argue that $W^{\mathrm{A}} \geqslant_{st} W^{\mathrm{NA}}$ since one expects that large waiting times occur with higher probability in the alternating service system. However this is not true. Let us imagine the situation where the service times are equal to zero. Then in the alternating service system we will have that the waiting time of the server is zero if $B_i \geqslant B_{i+1}$ for some i. So since $\mathbb{P}[B_i \geqslant B_{i+1}] > 0$, we have $\mathbb{P}[W^{\mathrm{A}} = 0] > 0$. In the non-alternating system however, we will have zero waiting time only if both preparation phases finish at exactly the same instant. Since the

preparation times are continuous random variables we have that $\mathbb{P}[B_i = B_{i+1}] = 0$ for every i and thus $\mathbb{P}[W^{\mathrm{NA}} = 0] = 0$. In Figure 4.4 we have plotted the distribution of the waiting time for both situations in the case where the service times are equal to zero and the B follows an Erlang distribution with $\mu = 5$ and five phases.

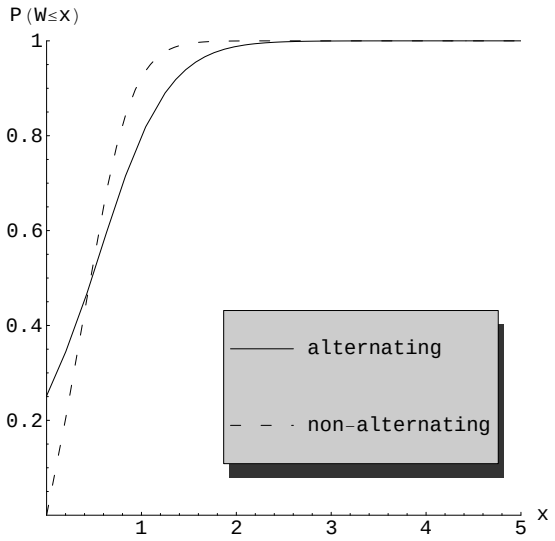

Figure 4.4: W^{A} and W^{NA} are not stochastically ordered.

The situation that we have described above is not a rare example. In fact, the following result holds.

Theorem 4.12. *For any distribution of the preparation and the service time, we have that* $\mathbb{P}[W^{\mathrm{A}} = 0] \geqslant \mathbb{P}[W^{\mathrm{NA}} = 0]$.

Proof. Both processes regenerate when a zero waiting time of the server occurs. Therefore in a cycle there is precisely one customer for whom the server did not have to wait. This means that the fraction of customers for whom the server does not wait is

$$\mathbb{P}[W = 0] = \frac{1}{\mathbb{E}[C]},$$

where $\mathbb{E}[C]$ is the average number of customers in a cycle, i.e. the mean cycle length. So it suffices to show that $\mathbb{E}[C^{\mathrm{A}}] \leqslant \mathbb{E}[C^{\mathrm{NA}}]$.

To prove this, we will couple the two systems and use sample path arguments. We will show that for a given initial state and for any realisation of preparation and service times the number of customers in a cycle is greater in the alternating case than in the non-alternating case. To couple the systems we will use the same

realisations for the preparation and the service times. To this end, let $\{B_i\}$ be a sequence of preparation times and $\{A_i\}$ a sequence of service times. We need to observe the system until the completion of the first cycle. For both systems assume that the server starts servicing the first customer at time zero while at the other service point a customer has just started his preparation phase B_1. Additionally, let R_n be the remaining preparation time for the n-th customer immediately after the service of the $(n-1)$-st customer has finished. As long as $R_n \leqslant B_{n+1}$ both processes are identical, since both servers will alternate between the two service points. In addition, all waiting times until that point will be strictly positive. As soon as $R_n > B_{n+1}$, the alternating service system will regenerate for the first time, since we will have that $W_n^{\mathrm{A}} = R_n$ and $W_{n+1}^{\mathrm{A}} = 0$. The non-alternating system however does not necessarily regenerate. For this system we have that $W_n^{\mathrm{NA}} = B_{n+1}$ and $R_{n+1}^{\mathrm{NA}} = R_n - B_{n+1} - A_n$. Therefore, if $R_{n+1}^{\mathrm{NA}} = 0$ then $W_{n+1}^{\mathrm{NA}} = 0$ and both processes regenerate. Otherwise the non-alternating system will not regenerate. Hence, for each realisation we have that $C^{\mathrm{NA}} \geqslant C^{\mathrm{A}}$ which implies that the mean cycle length $\mathbb{E}[C^{\mathrm{NA}}]$ of the non-alternating system is at least as long as the mean cycle length $\mathbb{E}[C^{\mathrm{A}}]$ of the alternating system. □

4.4.2 Mean waiting times

Although the waiting times in the two situations are not stochastically ordered, we have however that the mean waiting time of the server of the alternating service model $\mathbb{E}[W^{\mathrm{A}}]$ is larger than or equal to the mean waiting time of the server in the non-alternating system $\mathbb{E}[W^{\mathrm{NA}}]$. This is quite natural, since we expect the non-alternating system to perform better in terms of throughput, regardless of the distribution of the preparation phase.

To prove this result for the mean waiting times, we will again couple the two systems. We will make use of the same realisations $\{B_i\}$ and $\{A_i\}$ for the preparation and the service times respectively and we will continue with sample path arguments. We assume that the initial conditions for both systems are the same, i.e. at time zero the server starts servicing the first customer, while at the other service point a customer has just started his preparation phase. Then, for the alternating service system, define:

D_i^{A}: the $i-$th departure time

H_i^{A}: the time the server can start serving the other service point after time D_i^{A}.

Also define in the same way D_i^{NA} and H_i^{NA} for the non-alternating system. We need the following lemma.

Lemma 4.1. *For all i, we have that $D_i^{\mathrm{A}} \geqslant D_i^{\mathrm{NA}}$ and $H_i^{\mathrm{A}} \geqslant H_i^{\mathrm{NA}}$.*

Proof. We will apply induction. For $i = 1$ we have that $D_1^{\mathrm{A}} \geqslant D_1^{\mathrm{NA}}$ and $H_1^{\mathrm{A}} \geqslant H_1^{\mathrm{NA}}$, since

$$D_1^{\mathrm{A}} = A_1 \geqslant A_1 = D_1^{\mathrm{NA}}$$

and thus

$$H_1^{\mathrm{A}} = \max\{D_1^{\mathrm{A}}, B_1\} \geqslant \max\{D_1^{\mathrm{NA}}, B_1\} = H_1^{\mathrm{NA}}.$$

Suppose that for some i we have that $D_{i-1}^{\mathrm{A}} \geqslant D_{i-1}^{\mathrm{NA}}$ and $H_{i-1}^{\mathrm{A}} \geqslant H_{i-1}^{\mathrm{NA}}$. We will prove that $D_i^{\mathrm{A}} \geqslant D_i^{\mathrm{NA}}$ and $H_i^{\mathrm{A}} \geqslant H_i^{\mathrm{NA}}$ and this will conclude the proof.

The first relation is obvious. From the induction hypothesis we have $H_{i-1}^{\mathrm{A}} \geqslant H_{i-1}^{\mathrm{NA}}$, so

$$D_i^{\mathrm{A}} = H_{i-1}^{\mathrm{A}} + A_i \geqslant \min\{H_{i-1}^{\mathrm{NA}}, D_{i-1}^{\mathrm{NA}} + B_i\} + A_i = D_i^{\mathrm{NA}}.$$

For the second inequality, first notice that

$$H_i^{\mathrm{A}} = \max\{D_i^{\mathrm{A}}, D_{i-1}^{\mathrm{A}} + B_i\} \text{ and } H_i^{\mathrm{NA}} = \max\{D_i^{\mathrm{NA}}, \max\{H_{i-1}^{\mathrm{NA}}, D_{i-1}^{\mathrm{NA}} + B_i\}\},$$

because, for example, in the non-alternating case the other service point will either be ready at time D_i^{NA} when the previous customer departs, or it will be ready after the preparation phase at this point is completed, at the time point equal to the maximum of H_{i-1}^{NA} and $D_{i-1}^{\mathrm{NA}} + B_i$.

To prove that

$$H_i^{\mathrm{A}} = \max\{D_i^{\mathrm{A}}, D_{i-1}^{\mathrm{A}} + B_i\} \geqslant \max\{D_i^{\mathrm{NA}}, \max\{H_{i-1}^{\mathrm{NA}}, D_{i-1}^{\mathrm{NA}} + B_i\}\} = H_i^{\mathrm{NA}}, \quad (4.38)$$

we will show that the maximum term of the left-hand side of the inequality (4.38) is greater than or equal to any term of the right-hand side, thus also greater than or equal to the maximum of them.

Assume that $H_i^{\mathrm{A}} = D_i^{\mathrm{A}}$. Then $D_i^{\mathrm{A}} \geqslant D_i^{\mathrm{NA}}$ as we have proven above, furthermore $D_i^{\mathrm{A}} = H_{i-1}^{\mathrm{A}} + A_i \geqslant H_{i-1}^{\mathrm{NA}}$ since $H_{i-1}^{\mathrm{A}} \geqslant H_{i-1}^{\mathrm{NA}}$ and finally since $H_i^{\mathrm{A}} = D_i^{\mathrm{A}}$ then $D_i^{\mathrm{A}} \geqslant D_{i-1}^{\mathrm{A}} + B_i \geqslant D_{i-1}^{\mathrm{NA}} + B_i$. The case for $H_i^{\mathrm{A}} = D_{i-1}^{\mathrm{A}} + B_i$ follows similarly. □

A corollary of the previous result is the following.

Corollary 4.13. *For all i, $\sum_j^i W_j^{\mathrm{A}} \geqslant_{st} \sum_j^i W_j^{\mathrm{NA}}$.*

Proof. The proof is a direct consequence of the fact that for the coupled systems

$$W_1^{\mathrm{A}} + A_1 + \cdots + W_i^{\mathrm{A}} + A_i = D_i^{\mathrm{A}} \geqslant D_i^{\mathrm{NA}} = W_1^{\mathrm{NA}} + A_1 + \cdots + W_i^{\mathrm{NA}} + A_i.$$

□

So, although the random variables W^{A} and W^{NA} are not stochastically ordered, the partial sums of the sequences W_i^{A} and W_i^{NA} are.

It is also interesting to note that Lemma 4.1 immediately implies that the throughput is greater in the non-alternating system than in the alternating system since

$$\theta^{\mathrm{A}} = \lim_{i\to\infty} \frac{i}{D_i^{\mathrm{A}}} \leqslant \lim_{i\to\infty} \frac{i}{D_i^{\mathrm{NA}}} = \theta^{\mathrm{NA}}.$$

Moreover we have that $\theta = \big(\mathbb{E}[W]+\mathbb{E}[A]\big)^{-1}$, so we can readily establish the following result:

Theorem 4.14. *Given any distribution for the preparation and the service time, we have that $\mathbb{E}[W^{\mathrm{A}}] \geqslant \mathbb{E}[W^{\mathrm{NA}}]$.*

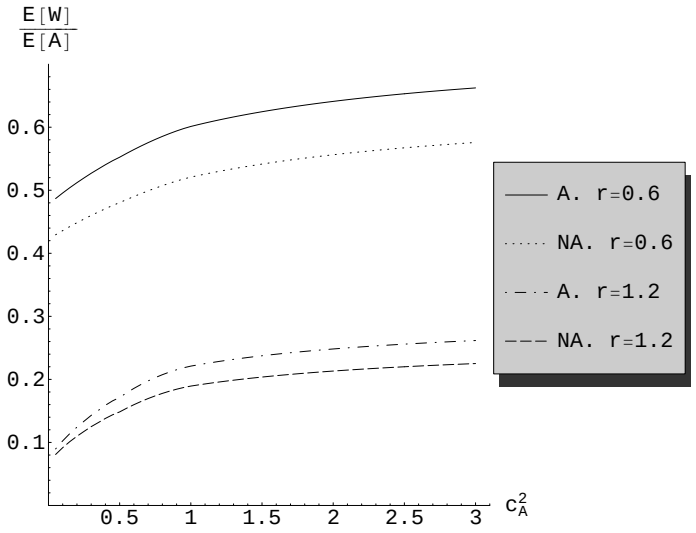

Figure 4.5: $\mathbb{E}[W^{\mathrm{A}}]$ is greater than or equal to $\mathbb{E}[W^{\mathrm{NA}}]$.

Figure 4.5 demonstrates a typical situation. For two values of the ratio

$$r = \frac{\mathbb{E}[A]}{\mathbb{E}[B]}$$

we have plotted the normalised waiting time $\mathbb{E}[W]/\mathbb{E}[A]$ versus the squared coefficient of variation c_A^2 of the service time A. We have chosen the mean service time to be $\mathbb{E}[A] = 1$ and the preparation time to be composed of five exponential phases. As before, A stands for the alternating service system and NA for the non-alternating system. One can see from these two examples that the average waiting time in the alternating service system is larger than in the non-alternating system. As is the case for the GI/G/1 queue, the waiting time depends almost linearly on c_A^2. As c_A^2 increases, the waiting time also increases and for the alternating case the rate of change is bigger. The difference of the mean waiting time in the alternating and the non-alternating case is eventually almost constant and this difference increases as the value of r decreases.

Remark 4.5. From Theorem 4.12 and Theorem 4.14 we can conclude that there is at least one point where the waiting time distributions of both systems intersect. Figure 4.4 suggests though that this point is unique. So, since the mean waiting times are both finite, this implies that W^{NA} is smaller than W^{A} with respect to the *increasing convex ordering*; namely

$$\mathbb{E}[\phi(W^{\mathrm{NA}})] \leqslant \mathbb{E}[\phi(W^{\mathrm{A}})]$$

for all increasing convex functions ϕ, for which the mean exists. This follows as a direct application of the Karlin-Novikoff cut-criterion (cf. Szekli [157]).

4.5 Numerical results

This section is devoted to some numerical results. In Figure 4.5 we have already shown how the normalised waiting time changes when the squared coefficient of variation of the service time is modified. Figure 4.6 shows the normalised waiting time plotted against the squared coefficient of variation of the preparation time. The preparation time is assumed to follow an Erlang distribution. We chose $\mathbb{E}[A] = 1$ and $c_A^2 = 0.2$ and we fitted a mixed Erlang distribution according to the procedure described in Section 3.7.

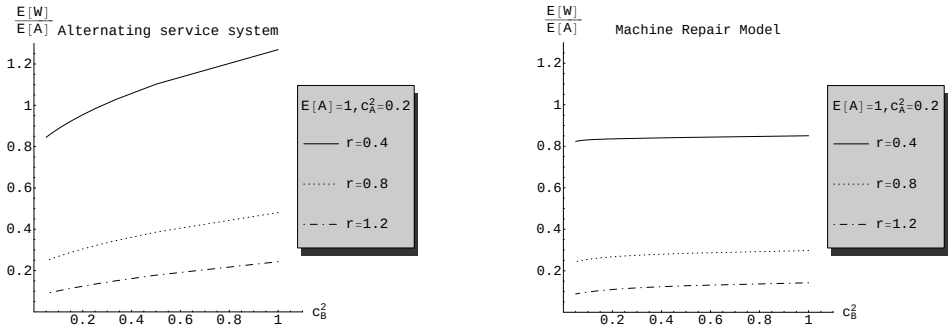

Figure 4.6: The normalised waiting time is almost insensitive to c_B^2 in the non-alternating system.

We have plotted the normalised waiting time for three different values of the ratio r; namely for $r = 0.4$, which implies that the service time is 40% of the preparation time, up to $r = 1.2$. The latter implies that for the alternating service model the server in general does not have to wait much. One can see that the normalised waiting time depends almost linearly on c_B^2 for the alternating service system, but for the non-alternating system it is almost insensitive to c_B^2, and thus to the number of exponential phases of the preparation time. This can be explained by the fact that Erlang loss models are insensitive to the service time distribution apart from its first moment; see for example Kelly [99]. More specifically, one can view the machine repair model that we have described as an $E/G/2/2$ loss system. Here the repairman acts as the Poisson source of an Erlang loss model if B follows an exponential distribution. However, the preparation times are a sum of exponentials and that causes the slight fluctuation in the mean waiting time.

Figure 4.7 shows the normalised waiting time plotted against the mean preparation time. We have chosen c_A^2 to be equal to 0.8 and we have fitted a mixed-Erlang distribution to the mean service time and the squared coefficient of service. As expected, the normalised waiting time $\mathbb{E}[W]/\mathbb{E}[A]$ depends almost linearly on the mean

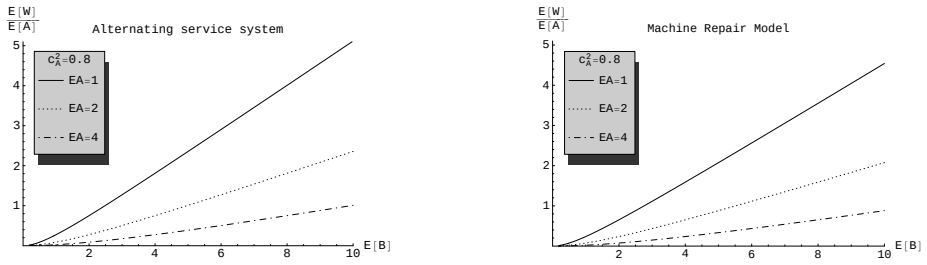

Figure 4.7: The normalised waiting time vs. $\mathbb{E}[B]$.

preparation time. For larger values of the mean preparation time, the normalised waiting time increases.

4.6 A comparison to Lindley's recursion

In Section 4.2.1 we have seen that the distribution of W_n and other characteristics of $\{W_n\}$ are quite explicit if all B_n have an exponential distribution. Since our recursion is, up to a sign, identical to Lindley's recursion, it is interesting to compare the complexity of both recursions. In this section, W_n will denote the waiting time of the n-th *customer* in a G/G/1 queue with first-come-first-served discipline, unless otherwise stated. To this end, let W_n, $n \geqslant 1$ be driven by Lindley's recursion, i.e.,

$$W_{n+1} = \max\{0, X_{n+1} + W_n\},$$

with $X_{n+1} = B_{n+1} - A_n$ as defined before.

In this section we compare the results we have derived so far for the time-dependent behaviour of Recursion (1.1) to the analogous cases for Lindley's recursion. In other words, for the G/M/1 queue we derive the time-dependent distribution of the waiting times in Section 4.6.1, and we review results on the length of the busy cycle in Section 4.6.2. Furthermore, we mention some known results on the covariance function for the G/G/1 queue in Section 4.6.3. For the steady-state distribution of the G/PH model we have already highlighted the differences it presents compared to the steady-state distribution of the G/PH/1 queue in Section 4.2.4.

4.6.1 The time-dependent distribution

The literature on time-dependent properties of Lindley's recursion and other queueing systems usually involves general expressions for the double transform $\sum_{n=0}^{\infty} r^n \mathbb{E}[\mathrm{e}^{-sW_{n+1}}]$, which are derived using Spitzer's identity and the Wiener-Hopf method; see e.g. Asmussen [6] and Cohen [46].

For the distribution of W_n, note that the following representation holds. Let Q_n be the number of customers in the system when the n-th customer arrives, and let $Q_1 = q_1 \geqslant 0$. Then, since all service times (in particular the residual

service times) are exponential with rate μ, W_n has a mixed-Erlang distribution, with mixture probabilities $\mathbb{P}[Q_n = k]$, $k = 0, \ldots, n + q_1$. This result is stated as Equation (3.97) in [46, p. 229]. Although the one-step transition matrix $\mathbf{P}$ can be derived straightforwardly, the probabilities $\mathbb{P}[Q_n = k]$ are fairly complicated, see [46, II.3.4]. However, it is possible to give an expression for the generating function $\sum_{n=0}^{\infty} r^n \mathbb{P}[Q_{1+n} = j]$ if $q_1 = 0$, see the equation below (3.72) in [46, p. 221]. The probabilities $\mathbb{P}[Q_n = k]$ can be computed explicitly for the M/M/1 queue if $q_1 = 0$; see, for example, Equation (2.26) in [46, p. 185]. Therefore, since the mixed-Erlang representation of the distribution of W_n for the G/M/1 case is not very explicit, in the following we give an alternative form of the distribution of W_n for the M/M/1 queue, which we could not find in the literature and which we derive by means of some simple probabilistic arguments.

Let Z be a geometric random variable, independent of everything else, with success probability r, that is, $\mathbb{P}[Z = n] = (1 - r)r^n$. Then, by conditioning on Z, we have that $\mathbb{P}[W_{Z+1} > x] = (1 - r)f(r, x)$, with

$$f(r, x) = \sum_{n=0}^{\infty} r^n \mathbb{P}[W_{n+1} > x]$$

the generating function of $\mathbb{P}[W_{n+1} > x]$. Thus, to get an expression for $f(r, x)$, it suffices to obtain the distribution of W_{Z+1}.

For this, we use two more probabilistic ideas. Assume that $W_1 = 0$. Then we have that $W_{n+1} \stackrel{\mathcal{D}}{=} \max_{k=0,\ldots,n} S_k$, with $S_0 = 0$ and $S_n = X_1 + \cdots + X_n$; see e.g. Asmussen [6, Chapter 8]. Finally, we reduce the problem to computing the distribution of the all-time maximum of a related random walk. For this, we define an i.i.d. sequence of random variables A'_i, $i \geqslant 1$ as follows. For any i we let $A'_i = A_i$ with probability r and $A'_i = \infty$ with probability $1 - r$. We see that we can interpret Z as the first value of i such that $A'_i = \infty$. Define $S'_n = X'_1 + \cdots + X'_n$, with $X'_i = B_{i+1} - A'_i$. Since $S'_n = -\infty$ if $n \geqslant Z$ and $S_n = S'_n$ if $n < Z$, it follows that

$$W_{Z+1} \stackrel{\mathcal{D}}{=} \max_{k=0,\ldots,Z} S_k \stackrel{\mathcal{D}}{=} \max_{k \geqslant 1} S'_k =: M_r.$$

We see that W_{Z+1} has the same distribution as the supremum M_r of a random walk S'_n, $n \geqslant 1$ with exponentially distributed upward jumps. For such random walks it is well known (see Theorem 5.8 in [6, p. 238]) that

$$\mathbb{P}[M_r > x] = \big(1 - \eta(r)/\mu\big)\mathrm{e}^{-\eta(r)x},$$

with $\eta(r)$ the unique positive solution of the equation

$$\frac{\mu}{\mu - \eta(r)} \mathbb{E}[\mathrm{e}^{-\eta(r)A'_1}] = 1. \tag{4.39}$$

We conclude that, for the G/M/1 queue,

$$(1 - r) \sum_{n=0}^{\infty} r^n \mathbb{P}[W_{n+1} > x] = \big(1 - \eta(r)/\mu\big)\mathrm{e}^{-\eta(r)x},$$

and thus, in particular,

$$f(r,x) = \frac{1}{1-r}\big(1-\eta(r)/\mu\big)\mathrm{e}^{-\eta(r)x}. \tag{4.40}$$

Obtaining an explicit expression for $\eta(r)$ is quite complicated. However, if A is exponentially distributed with rate λ, then $\eta(r)$ is given by

$$\eta(r) = \frac{1}{2}\Big(\mu - \lambda + \sqrt{(\lambda-\mu)^2 + 4\lambda\mu(1-r)}\Big). \tag{4.41}$$

Thus, the tail distribution of W_{n+1} for the M/M/1 queue can be derived by substituting (4.41) into the right-hand side of (4.40) and rewriting the resulting expression as a power series in r. This is in general possible, although rather cumbersome. To illustrate this, consider the case where $\rho = \lambda/\mu = 1$. This case results to the most simple expression for $\eta(r)$, as $\eta(r)$ reduces to $\mu\sqrt{1-r}$. Using the power series expansions

$$\mathrm{e}^y = \sum_{i=0}^{\infty} \frac{y^i}{i!}, \quad \text{and} \quad (1-r)^a = \sum_{j=0}^{\infty} \binom{a}{j}(-r)^j,$$

we then see that

$$\begin{aligned} f(r,x) &= \frac{1}{1-r}\big(1-\sqrt{1-r}\big)\mathrm{e}^{-\mu\sqrt{1-r}\,x} \\ &= \sum_{k=0}^{\infty} \frac{(-\mu x)^k}{k!} \sum_{n=0}^{\infty} \left[\binom{\frac{k}{2}-1}{n} - \binom{\frac{k-1}{2}}{n}\right](-r)^n \\ &= \sum_{n=0}^{\infty} r^n (-1)^n \sum_{k=0}^{\infty} \frac{(-\mu x)^k}{k!} \left[\binom{\frac{k}{2}-1}{n} - \binom{\frac{k-1}{2}}{n}\right]. \end{aligned}$$

By identifying this expression as a power series in r we see that, for the M/M/1 queue with $\rho = 1$,

$$\mathbb{P}[W_{n+1} > x \mid W_1 = 0] = (-1)^n \sum_{k=0}^{\infty} \frac{(-\mu x)^k}{k!} \left[\binom{\frac{k}{2}-1}{n} - \binom{\frac{k-1}{2}}{n}\right],$$

since we have assumed that $W_1 = 0$ in order to express W_{n+1} as the maximum of a random walk. We compare this expression to the distribution of the waiting time in the M/M case with $W_1 = 0$ and $\lambda = \mu$. By using that $\alpha(\mu) = \lambda/(\lambda+\mu) = 1/2$, that $\mathbb{P}[W_2 = 0 \mid W_1 = 0] = 1/2$, and Theorem 4.1, we obtain for the $(n+1)$-st waiting time of the *server* in our model that

$$\mathbb{P}[W_{n+1} > x \mid W_1 = 0] = \left[\frac{2}{5} - \frac{1}{10}\Big(-\frac{1}{4}\Big)^n\right]\mathrm{e}^{-\mu x}.$$

We conclude that there is a considerable difference in complexity between the distributions of the waiting time in the two models. We have shown in Theorem 4.1 that the distribution of W_n in the alternating service model is a simple mixture

of an atom at zero and an exponential distribution, while the distribution of W_n in Lindley's recursion can only be represented by its generating function, or by a mixed-Erlang representation of which the mixture probabilities are given by a generating function.

4.6.2 The busy cycle

For Lindley's recursion, the generating function of the busy cycle $C = \inf\{n \geqslant 1 : W_{n+1} = 0 \mid W_1 = 0\}$ can be extracted from Equation (3.89) in Cohen [46, p. 226]. Translated to our notation, we have that

$$\mathbb{E}[r^C] = \frac{r - \lambda(r)}{1 - \lambda(r)},$$

with $\lambda(r)$ the smallest zero in absolute value of the function $z - r\,\alpha\big(\mu(1-z)\big)$. It can be shown that $\lambda(r) = 1 - \eta(r)/\mu$, with $\eta(r)$ determined by (4.39). This implies that

$$\mathbb{E}[r^C] = \frac{\eta(r) - \mu(1-r)}{\eta(r)}.$$

For the M/M/1 queue with load ρ, an explicit expression is available, see for example Equation (2.43) in Cohen [46, p. 190], which states that

$$\mathbb{P}[C = n] = \frac{1}{2n-1}\binom{2n-1}{n}\frac{\rho^{n-1}}{(1+\rho)^{2n-1}}.$$

Again the difference in complexity between Lindley's recursion and our recursion is clear, cf. Theorem 4.3.

4.6.3 The covariance function

For the GI/M/1 queue Pakes [139] studies the covariance function of the waiting times. Theorem 1 of [139] gives the generating function of the correlation coefficients of the waiting times in the stationary GI/M/1 queue in terms of the unique positive solution of a specific functional equation. Furthermore, the correlation coefficients themselves are also given, but now in terms of the probabilities that no other waiting time than W_0 is equal to zero, up to time n. These expressions involve the probability generating function of the distribution of the number of customers served in a busy period, and are not very practical for numerical computations. Blanc [22] is concerned with the numerical inversion of the generating functions of the autocorrelations of the waiting times, as they are given in [139] and in Blomqvist [23], who derives for the M/G/1 queue results analogous to those in [139]. Evidently, these results are far more complicated and implicit than the simple expression for the covariance between W_n and W_{n+k} that is given in Theorem 4.4. Not only is the usage of generating functions not required, but neither is the assumption of stationarity.

To summarise, the time-dependent analysis of (1.1) when A is generally distributed and B is exponential is far more easy, and leads to far more explicit results,

than the analysis and the results obtained for the G/M/1 queue. In the next chapter we study the M/G model. The M/G/1 queue is perhaps the easiest queue to analyse. However, as we shall see in the sequel, the M/G model is the one presenting the most difficulties for the alternating service model we are studying.

Chapter 5
The M/G model

5.1 Introduction

In the previous chapter we have examined various aspects of the G/PH model. Namely, we have assumed that the distribution of the service time F_A is some general distribution on $[0, \infty)$ and the distribution of the preparation time F_B is a mixture of Erlang distributions, and under these assumptions we derived the time-dependent and the steady-state distribution of the waiting times. The methods that we have described there are surprisingly simple. The main idea we have utilised is based on the observation that if the preparation times follow a phase-type distribution, then the analysis of the G/PH model reduces to the analysis of a finite-state Markov chain. Moreover, this observation is valid not only for the examples discussed in Section 1.2, but also for examples where the server is *not* obliged to alternate between the two service stations. This has led us to the study of the waiting time of the server in the machine repair model with two machines, and to the comparison of these two models.

In this chapter we would like to study a more general model than the one analysed in the previous chapter. Specifically, we would like to remove the assumption that the preparation times follow a phase-type distribution, and allow for more general distributions. To keep the analysis simple, we shall first consider exponentially distributed service times. In other words, in this chapter we would like to consider the M/G model. We shall see however in Section 5.2, that the waiting-time distribution for this model is the solution to a generalised Wiener-Hopf equation, as was also the case for the generating function of the waiting times W_n; see Section 1.6. Although this equation can be approximated iteratively, there seems to be no straightforward way to solve it exactly.

Functional analysis has extensively studied various classes of convolution equations. Wiener-Hopf equations are well understood and various methods have been developed to study generalised Wiener-Hopf equations. A standard procedure leading to the solution of generalised Wiener-Hopf equations is to factorise the kernel of the equation (whenever the kernel admits factorisation). Inspired by this technique, in Section 5.3 we shall introduce a class of distributions that possess this property. As we shall see there, this class is strictly bigger than the class of phase-type distributions. Thus, by considering preparation-time distributions belonging to this class we generalise the results obtained in the previous chapter.

For exponential service times, we derive the steady-state waiting-time distribution in Section 5.4, while in Section 5.5 we briefly discuss the procedure for generally distributed service times. In Section 5.6 we work out in detail two examples in order to illustrate how the methods we apply in this chapter evolve, and how they compare to the procedure developed in the previous chapter. We conclude in Section 5.7 with

some remarks and a comparison of this model to the M/G/1 single-server queue. This chapter is based on parts of [169].

5.2 Derivation of the integral equation

In this section we derive the equation that we shall work with later on and we compare this equation with the analogous equation for the M/G/1 single-server queue. Furthermore, we examine various methods that are traditionally used for the single-server queue, but do not seem to be very helpful in our case.

To begin with, consider Equation (1.2)

$$W \stackrel{\mathcal{D}}{=} \max\{0, B - A - W\},$$

where A is the service time and B is the preparation time of a customer, and W is the waiting time of the server in the model described in Section 1.4. As before, we denote by π_0 the mass of the waiting-time distribution at zero. From this equation, we have for the distribution of W that

$$\begin{aligned} F_W(x) &= \mathbb{P}[W \leqslant x] = \mathbb{P}[B - W - A \leqslant x] \\ &= \int_0^\infty \int_0^\infty \mathbb{P}[B \leqslant x + z + y] \, \mathrm{d}F_A(z) \mathrm{d}F_W(y) \\ &= \pi_0 \int_0^\infty \mathbb{P}[B \leqslant x + z] \, \mathrm{d}F_A(z) + \int_{0^+}^\infty \int_0^\infty \mathbb{P}[B \leqslant x + y + z] \, \mathrm{d}F_A(z) \mathrm{d}F_W(y). \end{aligned} \tag{5.1}$$

Assume now that the service time A is exponentially distributed with parameter λ; that is, $f_A(x) = \lambda \mathrm{e}^{-\lambda x}$. One can show that W has a density when A has one in the following way. From Equation (1.2) we readily have that

$$\mathbb{P}[W \leqslant x] = \int_{-\infty}^\infty \mathbb{P}[A \geqslant y - x] \, \mathrm{d}F_{B-W}(y).$$

Since A has a density, the integral

$$\int_{-\infty}^\infty f_A(y - x) \, \mathrm{d}F_{B-W}(y)$$

exists and is the density of F_W. Moreover, since f_A is continuous, it can be shown

that f_W is continuous. Then (5.1) becomes

$$\begin{aligned} F_W(x) &= \pi_0 \int_0^\infty F_B(x+z)\lambda \mathrm{e}^{-\lambda z}\,\mathrm{d}z + \int_0^\infty f_W(y) \int_0^\infty F_B(x+y+z)\lambda \mathrm{e}^{-\lambda z}\,\mathrm{d}z\mathrm{d}y \\ &= \lambda\pi_0 \mathrm{e}^{\lambda x} \int_0^\infty F_B(x+z)\mathrm{e}^{-\lambda(x+z)}\,\mathrm{d}z + \\ &\quad + \int_0^\infty \lambda \mathrm{e}^{\lambda(x+y)} f_W(y) \int_0^\infty F_B(x+z+y)\mathrm{e}^{-\lambda(x+z+y)}\,\mathrm{d}z\mathrm{d}y \\ &= \lambda\pi_0 \mathrm{e}^{\lambda x} \int_x^\infty F_B(u)\mathrm{e}^{-\lambda u}\,\mathrm{d}u + \int_0^\infty \lambda \mathrm{e}^{\lambda(x+y)} f_W(y) \int_{x+y}^\infty F_B(u)\mathrm{e}^{-\lambda u}\,\mathrm{d}u\mathrm{d}y. \end{aligned}$$

For the remainder of this chapter we shall also need to assume that F_B is a continuous function. Therefore, we can differentiate with respect to x using Leibniz's rule to obtain

$$\begin{aligned} f_W(x) &= \lambda^2\pi_0 \mathrm{e}^{\lambda x} \int_x^\infty F_B(u)\mathrm{e}^{-\lambda u}\,\mathrm{d}u - \lambda\pi_0 F_B(x) + \\ &\quad + \lambda^2 \int_0^\infty \mathrm{e}^{\lambda(x+y)} f_W(y) \int_{x+y}^\infty F_B(u)\mathrm{e}^{-\lambda u}\,\mathrm{d}u\mathrm{d}y - \lambda \int_0^\infty F_B(x+y) f_W(y)\,\mathrm{d}y \end{aligned}$$

or

$$f_W(x) = \lambda F_W(x) - \lambda\pi_0 F_B(x) - \lambda \int_0^\infty F_B(x+y) f_W(y)\,\mathrm{d}y. \tag{5.2}$$

What makes this equation troublesome to solve is the plus sign that appears in the integral at the right-hand side. If we were dealing with the classic M/G/1 single-server queue, then the equation for the M/G/1 queue that is analogous to (5.2) would be identical apart from this sign. This difference, nonetheless, is of great importance when we try to derive the waiting-time distribution.

It is not possible to derive a linear differential equation for f_W by differentiating (5.2) as we have done in Chapter 3, since we will not be able to avoid having some integral at the right-hand side.

Taking Laplace transforms, as we did in Section 4.2.3, is also not useful since the integral at (5.2) is not a convolution (as it would be, if only the sign in the argument of F_B were different). In Section 4.2.3 we were able to exploit the memoryless property of the exponential distribution in order to directly derive the Laplace transform of F_W. Unfortunately, if the situation is reversed and we assume that A, instead of B, is exponentially distributed, then this simple calculation fails at its very first step due to the lack of structure of F_B; see Equation (4.28). Therefore, we are not able to directly obtain an expression for the Laplace transform of W.

Additionally, since we no longer consider phase-type distributions F_B, we can no longer define a convenient Markov chain from the equilibrium distribution of which we would be able to deduce F_W. We have used this approach extensively in Chapter 4, both for the alternating service system and the non-alternating service system we considered there. The Markov chain we define there is based on the number of exponential phases that the customer has to complete during his preparation

time when the server returns to that service point. Also this approach is heavily relying on the fact that F_B has some structure, and the memoryless property comes in handy again. Nonetheless, if B follows some general distribution, this method is not applicable either. We would have to incorporate in the Markov chain the – unknown – remaining preparation time, which includes too little information to make any calculations possible.

Moreover, it does not seem possible to factorise the transformed equation into terms that are analytic either in the right half complex plane or in the left half plane (this would allow us to use the Wiener-Hopf technique in order to derive the Laplace transform of W; we shall use this method in Chapter 8). Therefore, the case that we are considering here needs special attention.

One should note here that Equation (5.2) can be reduced to a *generalised Wiener-Hopf* equation. It is known that the following equation

$$\int_0^\infty \big(K(x-y) + F_B(x+y)\big) f_W(y)\,\mathrm{d}y = -\pi_0 F_B(x) \qquad (x>0) \tag{5.3}$$

is equivalent to a generalised Wiener-Hopf equation (see Noble [135, p. 233]). Equation (5.2) reduces to Equation (5.3), if we let the kernel $K(x)$ be the function

$$K(x) = \frac{\delta(x)}{\lambda} - \mathbb{1}_{\{x>0\}} - \frac{F_W(0)}{1-F_W(0)},$$

where $\delta(x)$ is the Dirac δ-function and $\mathbb{1}_{\{x>0\}}$ is the indicator function of the set $\{x>0\}$. This is indeed the case, since we have that

$$\int_0^\infty \Big(\frac{\delta(x-y)}{\lambda} - \mathbb{1}_{\{x>y\}} - \frac{F_W(0)}{1-F_W(0)}\Big) f_W(y)\,\mathrm{d}y = \frac{f_W(x)}{\lambda} - \big[F_W(x) - F_W(0)\big] - F_W(0),$$

which is exactly in the form of Equation (5.2). We were unable though to solve this generalised Wiener-Hopf equation.

Although regular Wiener-Hopf equations can be solved rather straightforwardly, generalised Wiener-Hopf equations cannot be solved by a single technique. For example, Noble [135, Section 5.2] mentions some problems (that lead to special cases of Equation (1.6)) that cannot be solved exactly by the Wiener-Hopf technique (cf. Section 1.6). He then proceeds by giving exact solutions for various special cases of Equation (1.6).

Some generalised Wiener-Hopf equations, however, can be reduced to *Fredholm equations of the second type*, which are equations of the form

$$f(x) - \lambda \int_a^b f(\xi) K(x,\xi)\,\mathrm{d}\xi = g(x),$$

where f is the unknown function and all other terms are given – compare this also to Equation (1.3). It is interesting to note at this point that Equation (5.2) is

a Fredholm integral equation with infinite domain. As is the case for generalised Wiener-Hopf equations, Fredholm equations have been studied also through examining properties of their kernel; textbook references on this domain are Masujima [126], Mikhlin [129], and Tricomi [163].

Under various assumptions, it is well-known that such equations can be solved by the method of successive iterations. We have already observed this in Section 2.3, where it was shown that Equation (5.2) satisfies a contraction mapping. Therefore, successive iterations provide us with a series of functions that converge (geometrically fast) to the unique solution of (5.2).

A particularly tractable case occurs for Fredholm equations with a *multiplicative* (also called *degenerate* [129]) kernel. Such a kernel $K(x, \xi)$ consists of the sum of a finite number of terms, each of which is in its turn the product of two factors, one of which depends only on x, and the other only on ξ. Integral equations with a multiplicative kernel can be simply solved by reduction to a system of algebraic equations.

Since the derivation of a solution to Equation (5.2) for F_B being a general distribution remains challenging, we will limit ourselves to studying under which conditions we can derive an explicit formula for F_W. As discussed in [135], generalised Wiener-Hopf equations can be solved in special cases. In the following section we shall study a class of distribution functions F_B for which such a solution is possible.

5.3 The $\mathcal{M}$ class

First we observe that the kernel of Equation (5.2) is the function $F_B(x + y)$. Motivated by standard techniques regarding the invertibility of generalised Wiener-Hopf operators, we would like to have some multiplicative property of this kernel, in order for us to be able to factorise it. Therefore, before deriving the waiting-time distribution, we first define the class $\mathcal{M}$ as the collection of distribution functions F on $[0, \infty)$ that have the following property. For every $x, y \geqslant 0$, we can decompose the tail of the distribution as follows

$$\overline{F}(x + y) = 1 - F(x + y) = \sum_{i=1}^{n} g_i(x)h_i(y),$$

where for every i, g_i and h_i are arbitrary measurable functions (that can even be constants). Of course, by demanding that F is a distribution we have implicitly made some assumptions on the functions g_i and h_i, but these assumptions are, for the time being, of no real importance.

We have constructed this class of distributions only because of the specific form of the kernel of Equation (5.2), since if F_B belongs to this class, the integral appearing at the right-hand side of (5.2) can be easily computed. A natural question is to investigate how big this class actually is. We shall show that the class $\mathcal{M}$ is particularly rich.

To begin with, one can see that all phase-type distributions are included in $\mathcal{M}$. Moreover, for phase-type distributions all the individual functions g_i and h_i have

a nice interpretation. For the proof, let F be a phase-type distribution. Such a distribution F is defined in terms of a Markov jump proces $J(x)$, $x \geqslant 0$, with finite state space $E \cup \Delta$, such that Δ is the set of absorbing states and E the set of transient states. Then F is the distribution of the time until absorption. It is usually assumed that the process starts in E; see Asmussen [6, Chapter 3]. For our purpose, suppose that we have an $n+1$-state Markov chain, where state 0 is absorbing and states $\{1, \ldots, n\}$ are not. Then

$$\overline{F}(x) = \mathbb{P}[J(x) \text{ is not absorbed}].$$

So we have that

$$\begin{aligned}
\overline{F}(x+y) &= \mathbb{P}[J(x+y) \in \{1, \ldots, n\}] \\
&= \sum_{i=1}^{n} \mathbb{P}[J(x+y) \in \{1, \ldots, n\} \mid J(x) = i] \mathbb{P}[J(x) = i] \\
&= \sum_{i=1}^{n} \mathbb{P}[J(y) \in \{1, \ldots, n\} \mid J(0) = i] \mathbb{P}[J(x) = i] \\
&= \sum_{i=1}^{n} h_i(y) g_i(x),
\end{aligned}$$

with

$$\begin{aligned}
h_i(y) &= \mathbb{P}[J(y) \in \{1, \ldots, n\} \mid J(0) = i] \\
g_i(x) &= \mathbb{P}[J(x) = i].
\end{aligned}$$

So F belongs to $\mathcal{M}$, and the functions h_i and g_i express the probability that the process is in one of the transient states given that it started in state i and the probability that the process is in state i respectively.

However, $\mathcal{M}$ includes more distribution functions apart from the phase-types. A well-known distribution that is not phase-type but has a rational Laplace transform (see, for example, Asmussen [6, p. 87]) is the distribution with a density proportional to $(1 + \sin x)\,\mathrm{e}^{-x}$. So, let the density be $f(x) = c\,(1 + \sin x)\,\mathrm{e}^{-x}$, where

$$c^{-1} = \int_0^\infty (1 + \sin x)\,\mathrm{e}^{-x}\,\mathrm{d}x = \frac{3}{2}.$$

Then the distribution is given by

$$F(x) = 1 - \frac{\mathrm{e}^{-x}(2 + \sin x + \cos x)}{3}$$

and one can easily check now that $\overline{F}(x+y)$ can be decomposed into a finite sum of products of functions of x and of functions of y. In fact, all functions with rational

Laplace transforms are included in this class. To see this, let the function $f(x)$ have the Laplace transform

$$\hat{f}(s) = \frac{P(s)}{Q(s)},$$

where $P(s)$ and $Q(s)$ are polynomials in s with $\deg[P] < \deg[Q]$. Let now the roots of $Q(s)$ be $q_1, \ldots, q_n$ with multiplicities $m_1, \ldots, m_n$ respectively. Then $\hat{f}(s)$ can be decomposed as follows:

$$\hat{f}(s) = \frac{c_1^1}{(s-q_1)} + \frac{c_2^1}{(s-q_1)^2} + \cdots + \frac{c_{m_1}^1}{(s-q_1)^{m_1}} + \frac{c_1^2}{(s-q_2)} + \cdots + \frac{c_{m_n}^n}{(s-q_n)^{m_n}},$$

where the constants c_j^i are given by

$$c_j^i = \frac{1}{(m_i - j)!} \frac{d^{m_i - j}}{ds^{m_i - j}} \left[(s-q_i)^{m_i} \frac{P(s)}{Q(s)} \right] \Bigg|_{s=q_i}.$$

Then $f(x)$ is simply the function

$$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{c_j^i \; x^{j-1}}{(j-1)!} \, \mathrm{e}^{q_i x}.$$

Therefore, the corresponding distribution is given by

$$F(x) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{c_j^i}{(-q_i)^j} \left(1 - \mathrm{e}^{q_i x} \sum_{k=0}^{j-1} \frac{(-q_i x)^k}{k!} \right),$$

which clearly belongs to $\mathcal{M}$. Our conjecture is that $\mathcal{M}$ is exactly the class of distributions with rational Laplace transform.

5.4 Steady-state distribution for M/$\mathcal{M}$

Having defined a class of distributions in which the kernel of Equation (5.2) can be factorised, we now proceed with the derivation of the waiting-time distribution. Denote by $\hat{\beta}$ and γ_i, $i = 1, \ldots, n$, the Laplace transforms of the functions $\overline{F}_B$ and g_i respectively, that is,

$$\hat{\beta}(s) = \int_0^\infty \mathrm{e}^{-sx} \overline{F}_B(x) \, \mathrm{d}x \qquad \text{and} \qquad \gamma_i(s) = \int_0^\infty \mathrm{e}^{-sx} g_i(x) \, \mathrm{d}x.$$

Then the following theorem holds.

Theorem 5.1. *Assume that $F_B \in \mathcal{M}$, is continuous, and that for every $i = 1, \ldots, n$ the functions $h_i(y)$ are bounded on $(0, \infty)$ and*

$$\int_0^\infty |g_i(x)| \, \mathrm{d}x < \infty.$$

Then the distribution of W is given by

$$F_W(x) = 1 - \mathrm{e}^{\lambda x} \int_x^\infty \mathrm{e}^{-\lambda s} \left(\lambda \pi_0 \overline{F}_B(s) + \lambda \sum_{i=1}^n c_i g_i(s) \right) \mathrm{d}s, \tag{5.4}$$

where the constants π_0 and c_i, $i = 1, \ldots, n$, are a solution to the linear system of equations

$$\pi_0 + \lambda \pi_0 \ \hat{\beta}(\lambda) + \lambda \sum_{i=1}^n c_i \, \gamma_i(\lambda) = 1$$

and for $i = 1, \ldots, n$,

$$\begin{aligned} c_i = \lambda \pi_0 \int_0^\infty h_i(x) \left(\overline{F}_B(x) - \lambda \int_x^\infty \mathrm{e}^{-\lambda(s-x)} \overline{F}_B(s) \, \mathrm{d}s \right) \mathrm{d}x + \\ + \lambda \sum_{j=1}^n c_j \int_0^\infty h_i(x) \left(g_j(x) - \lambda \int_x^\infty \mathrm{e}^{-\lambda(s-x)} g_j(s) \, \mathrm{d}s \right) \mathrm{d}x. \end{aligned} \tag{5.5}$$

Proof. Since $F_B \in \mathcal{M}$, (5.2) becomes

$$\begin{aligned} f_W(x) &= \lambda F_W(x) + \lambda \pi_0 \overline{F}_B(x) - \lambda \pi_0 + \lambda \int_0^\infty \overline{F}_B(x+y) f_W(y) \, \mathrm{d}y - \lambda \int_0^\infty f_W(y) \, \mathrm{d}y \\ &= \lambda F_W(x) + \lambda \pi_0 \overline{F}_B(x) - \lambda \pi_0 + \lambda \sum_{i=1}^n g_i(x) \int_0^\infty h_i(y) f_W(y) \, \mathrm{d}y - \lambda(1 - \pi_0), \end{aligned}$$

or

$$f_W(x) = \lambda F_W(x) + \lambda \pi_0 \overline{F}_B(x) + \lambda \sum_{i=1}^n c_i g_i(x) - \lambda, \tag{5.6}$$

where we have defined

$$c_i = \int_0^\infty h_i(y) f_W(y) \, \mathrm{d}y. \tag{5.7}$$

Equation (5.6) is a linear differential equation of first order that satisfies the initial condition $F_W(0) = \pi_0$. Its solution is given by

$$F_W(x) = \mathrm{e}^{\lambda x} \int_0^x \mathrm{e}^{-\lambda s} \left(\lambda \pi_0 \overline{F}_B(s) + \lambda \sum_{i=1}^n c_i g_i(s) - \lambda \right) \mathrm{d}s + \pi_0 \mathrm{e}^{\lambda x}. \tag{5.8}$$

We can rewrite the previous equation as follows.

$$\begin{aligned} F_W(x) &= \mathrm{e}^{\lambda x} \int_0^x \mathrm{e}^{-\lambda s} \left(\lambda \pi_0 \overline{F}_B(s) + \lambda \sum_{i=1}^n c_i g_i(s) \right) \mathrm{d}s + (\pi_0 - 1) \mathrm{e}^{\lambda x} + 1 \\ &= \mathrm{e}^{\lambda x} \left(\pi_0 + \lambda \pi_0 \ \hat{\beta}(\lambda) + \lambda \sum_{i=1}^n c_i \, \gamma_i(\lambda) - 1 \right) - \\ &\quad - \mathrm{e}^{\lambda x} \int_x^\infty \mathrm{e}^{-\lambda s} \left(\lambda \pi_0 \overline{F}_B(s) + \lambda \sum_{i=1}^n c_i g_i(s) \right) \mathrm{d}s + 1. \end{aligned} \tag{5.9}$$

There are $n+1$ unknown terms in the above equation, the probability π_0 and the constants c_i for $i = 1, \ldots, n$. These constants are a solution to a linear system of $n+1$ equations which is formed as follows. The first equation is given by

$$\lim_{x \to \infty} F_W(x) = 1, \tag{5.10}$$

or equivalently,

$$\pi_0 + \lambda \pi_0 \, \hat{\beta}(\lambda) + \lambda \sum_{i=1}^{n} c_i \, \gamma_i(\lambda) = 1. \tag{5.11}$$

For $i = 1, \ldots, n$, we form n additional equations using Equation (5.7) as follows. We substitute f_W by using (5.6). For the distribution F_W that appears in the latter equation we use Equation (5.9), after simplifying this one by using (5.11). With this straightforward calculation we derive linear system for the constants c_i in the form that it appears in (5.5).

For the fact that Equation (5.11) is both necessary and sufficient for (5.10) to hold, one only needs to note that

$$\lim_{x \to \infty} \int_x^{\infty} \mathrm{e}^{-\lambda(s-x)} \left(\lambda \pi_0 \overline{F}_B(s) + \lambda \sum_{i=1}^{n} c_i g_i(s) \right) \mathrm{d}s = 0,$$

since we have that $\int_0^{\infty} |g_i(x)| \, \mathrm{d}x < \infty$.

Denote by Σ the system formed by Equations (5.11) and (5.5). We can show that Σ has at least one solution by constructing one as follows. From Section 2.2.1 we know that there exists at least one invariant distribution for W. This distribution, by definition, satisfies the condition that its limit at infinity equals one and it also satisfies Equation (5.8). Then it is clear that it also satisfies Σ; therefore, Σ has at least one solution.

In Corollary 2.3 we have already seen that if one finds a continuous and bounded solution to (1.2), then this solution is necessarily the limiting distribution. To complete the proof, it remains to show that these conditions apply to (5.8). First of all, (5.8) is clearly a continuous function and since $\lim_{x \to \infty} F_W(x) = 1$ and $0 \leqslant F_W(0) = \pi_0 < \infty$, it is also bounded. Therefore, (5.8) is the limiting distribution. □

Remark 5.1. The conditions that appear in Theorem 5.1 guarantee that all the integrals that appear in the intermediate calculations and in Σ are well defined. In particular, one should note that demanding that

$$\int_0^{\infty} |g_i(x)| \, \mathrm{d}x < \infty$$

implies that the random variable B has a finite mean, $\gamma_i(\lambda)$ and $\hat{\beta}(\lambda)$ exist and are finite numbers, and that

$$\int_0^{\infty} h_i(x) \overline{F}_B(x) \, \mathrm{d}x \qquad \text{and} \qquad \int_0^{\infty} h_i(x) g_j(x) \, \mathrm{d}x$$

exist and are finite; cf. Equation (5.5).

Remark 5.2. We have explained in the proof why Σ has at least one solution, but we have not excluded the possibility that Σ has multiple solutions. In fact, if we choose a decomposition of F_W such that at least one of the functions, say the function h_1, depends linearly on all other functions – in this case the functions h_i –, then we know beforehand that Σ will have multiple solutions. However, the fact that (5.8) is necessarily the *unique* invariant distribution guarantees that the multiple solutions of Σ will make the term $\sum_{i=1}^{n} c_i g_i(s)$ unique, since for each of the solutions of Σ the function F_W appearing in Theorem 5.1 will still be continuous and in $\mathcal{L}([0,\infty))$. Thus, by Corollary 2.3 it will be the unique limiting waiting-time distribution.

Remark 5.3. Equation (5.11) simply states that $\mathbb{P}[W=0]+\mathbb{P}[W>0]=1$. To see that, observe that

$$\lambda\pi_0 \int_0^\infty \mathrm{e}^{-\lambda x}\overline{F}_B(x)\,\mathrm{d}x = \pi_0\mathbb{P}[B>A],$$

and that

$$\begin{aligned}\lambda\sum_{i=1}^{n} c_i\,\gamma_i(\lambda) &= \sum_{i=1}^{n}\int_0^\infty h_i(y)f_W(y)\,\mathrm{d}y\int_0^\infty \lambda\mathrm{e}^{-\lambda x}g_i(x)\,\mathrm{d}x\\ &= \int_0^\infty\int_0^\infty \lambda\mathrm{e}^{-\lambda x}f_W(y)\overline{F}_B(x+y)\,\mathrm{d}x\mathrm{d}y = \mathbb{P}[B-A-W^+>0],\end{aligned}$$

where W^+ is the waiting time, given that it is strictly positive.

5.5 The G/$\mathcal{M}$ model

As one may observe from the proof of Theorem 5.1, the fact that A is exponentially distributed did not have a significant impact on the analysis (apart from keeping expressions simple). One can assume that A follows a mixed-Erlang distribution of the form of Equation (3.23) and the proof remains effectively the same. In fact, the form of the service time distribution is not essential. In the following we shall highlight the basic conclusions one can draw if A follows a general distribution, without being concerned about making rigorous mathematical statements. We simply assume that all functions occurring satisfy conditions which permit us to carry out our operations.

Assume that $F_B \in \mathcal{M}$ and rewrite Equation (5.1) as follows:

$$F_W(x) = \pi_0\int_0^\infty F_B(x+y)\,\mathrm{d}F_A(y) + \sum_{i=1}^{n} g_i(x)\int_{0^+}^\infty\int_0^\infty h_i(y+z)\,\mathrm{d}F_A(z)\mathrm{d}F_W(y). \tag{5.12}$$

Define now the constants

$$c_i = \int_{0^+}^\infty\int_0^\infty h_i(y+z)\,\mathrm{d}F_A(z)\mathrm{d}F_W(y),$$

which we need to express only in terms of the distributions F_A and F_B. To this end, differentiate (5.12) once to obtain

$$f_W(x) = \pi_0 \sum_{k=1}^{n} g_k'(x) \int_0^\infty h_k(y)\,\mathrm{d}F_A(y) + \sum_{k=1}^{n} g_k'(x)c_k.$$

We shall substitute this expression into the definition of the constants c_i in order to form an $n \times n$ linear system for the unknown constants c_i. We have that

$$\begin{aligned}
c_i &= \int_{0^+}^\infty \int_0^\infty h_i(y+z)\,\mathrm{d}F_A(z)\mathrm{d}F_W(y) \\
&= \int_0^\infty \int_0^\infty h_i(y+z) f_W(y)\,\mathrm{d}F_A(z)\mathrm{d}y \\
&= \int_0^\infty \int_0^\infty h_i(y+z)\left(\pi_0 \sum_{k=1}^{n} g_k'(y) \int_0^\infty h_k(u)\,\mathrm{d}F_A(u) + \sum_{k=1}^{n} g_k'(y)c_k\right)\mathrm{d}F_A(z)\mathrm{d}y \\
&= \pi_0 \sum_{k=1}^{n} \int_0^\infty \int_0^\infty \int_0^\infty h_i(y+z) g_k'(y) h_k(u)\,\mathrm{d}F_A(u)\mathrm{d}F_A(z)\mathrm{d}y + \\
&\quad + \sum_{k=1}^{n} c_k \int_0^\infty \int_0^\infty h_i(y+z) g_k'(y)\,\mathrm{d}F_A(z)\mathrm{d}y.
\end{aligned}$$

The probability π_0 will be determined as usual by the normalisation equation

$$\pi_0 + \int_0^\infty f_W(x)\,\mathrm{d}x = 1.$$

As before, one can argue that the linear system determining the constants c_i and the probability π_0 has at least one solution which leads to the unique waiting-time distribution

$$F_W(x) = \pi_0 \int_0^\infty F_B(x+y)\,\mathrm{d}F_A(y) + \sum_{i=1}^{n} c_i g_i(x).$$

This technique is formalised and described in detail in [129, Section I.4].

5.6 Explicit examples

The waiting-time distribution, as it is given by Theorem 5.1, may seem perplexing. It is certainly not straightforward to show even the most basic properties, such as that $\lim_{x\to\infty} F_W(x) = 1$, since the expression involves an exponential term that is unbounded and an integral term that tends to zero as $x \to \infty$. In this section, we shall give the details of the computations for two simple examples.

The first example we shall present is the M/M model. One can derive the steady-state waiting-time distribution for this model either by applying the corresponding theorem in Chapter 4 or by applying the technique developed in this chapter. We

have already seen in Remark 4.2 that the waiting-time distribution for this case is given by

$$F_W(x) = 1 - \frac{2\,\alpha(\mu)}{2+\alpha(\mu)}\,\mathrm{e}^{-\mu x},$$

and the mass of the distribution at the origin is given by

$$\pi_0 = 1 - \frac{2\,\alpha(\mu)}{2+\alpha(\mu)}.$$

In the expressions above we have that $\alpha(\mu) = \lambda/(\lambda+\mu)$. If one wishes to apply Theorem 5.1 however, then the first step is to find a decomposition of the function $\overline{F}_B(x+y)$ into a sum of products of functions either of x or of y. In our case, a decomposition is quite apparent, since $\overline{F}_B(x+y) = \mathrm{e}^{-\mu(x+y)}$; thus, one can simply choose the functions

$$h_1(x) = g_1(x) = \mathrm{e}^{-\mu x}.$$

Obviously, the way to decompose the kernel is not unique, though any decomposition may be used. The next step is to obtain the unique solution of the linear system

$$\begin{aligned}
&\pi_0 + \lambda\pi_0\,\hat{\beta}(\lambda) + \lambda c_1\,\gamma_1(\lambda) = 1,\\
&c_1 = \lambda\pi_0 \int_0^\infty h_1(x)\left(\overline{F}_B(x) - \lambda\int_x^\infty \mathrm{e}^{-\lambda(s-x)}\overline{F}_B(s)\,\mathrm{d}s\right)\mathrm{d}x +\\
&\qquad + \lambda c_1 \int_0^\infty h_1(x)\left(g_1(x) - \lambda\int_x^\infty \mathrm{e}^{-\lambda(s-x)} g_1(s)\,\mathrm{d}s\right)\mathrm{d}x.
\end{aligned}$$

Recall that

$$\hat{\beta}(s) = \int_0^\infty \mathrm{e}^{-sx}\overline{F}_B(x)\,\mathrm{d}x \qquad \text{and} \qquad \gamma_1(s) = \int_0^\infty \mathrm{e}^{-sx} g_1(x)\,\mathrm{d}x.$$

Then the above linear system can be rewritten as

$$\begin{aligned}
\pi_0 + \lambda\pi_0\frac{1}{\lambda+\mu} + \lambda c_1\frac{1}{\lambda+\mu} &= 1,\\
c_1 = \lambda\pi_0\frac{1}{2(\lambda+\mu)} + \lambda c_1\frac{1}{2(\lambda+\mu)},&
\end{aligned}$$

the solution of which is given by

$$\pi_0 = \frac{2\mu+\lambda}{2\mu+3\lambda} \qquad \text{and} \qquad c_1 = \frac{\lambda}{2\mu+3\lambda}.$$

Thus, by Theorem 5.1 we have that the steady-state waiting-time distribution is given by

$$F_W(x) = 1 - \mathrm{e}^{\lambda x}\int_x^\infty \mathrm{e}^{-\lambda s}\left(\lambda\pi_0\mathrm{e}^{-\mu s} + \lambda\frac{\lambda}{2\mu+3\lambda}\mathrm{e}^{-\mu s}\right)\mathrm{d}s,$$

which reduces to the expression we have obtained in Remark 4.2; namely,

$$F_W(x) = 1 - \frac{2\lambda}{2\mu + 3\lambda}\,\mathrm{e}^{-\mu x}.$$

It is quite apparent that if F_B is a phase-type distribution, then the computational effort is greater when applying Theorem 5.1 than Theorem 4.8. In both cases we are called to solve a linear system; however, in order to apply the method developed in this chapter, we also need to find a decomposition of the kernel $F_B(x+y)$ and compute the distribution from Equation (5.4). However, if F_B is not phase type, then this method is the most straightforward one (compared to the option of approximating the given distribution for the preparation times with a phase-type distribution and then applying Theorem 4.8). This leads us to our second example.

The second example we shall present relates to the case where F_B has a rational Laplace transform, but is not phase type. To this end, let the preparation-time distribution be given by

$$F_B(x) = 1 - \frac{\mathrm{e}^{-x}(2 + \sin x + \cos x)}{3}.$$

Since

$$\overline{F}_B(x+y) = \frac{1}{3}\,\mathrm{e}^{-(x+y)}\big(2 + \sin x\cos y + \cos x \sin y + \cos x \cos y - \sin x \sin y\big),$$

we can pick the following functions for the decomposition:

$$g_1(x) = \frac{2}{3}\,\mathrm{e}^{-x}, \qquad h_1(x) = \mathrm{e}^{-x}, \qquad g_2(x) = h_3(x) = g_5(x) = \mathrm{e}^{-x}\sin x,$$
$$g_4(x) = \mathrm{e}^{-x}\cos x, \quad h_5(x) = -\frac{1}{3}\,\mathrm{e}^{-x}\sin x, \quad h_2(x) = g_3(x) = h_4(x) = \frac{1}{3}\,\mathrm{e}^{-x}\cos x.$$

Thus, we have that

$$\hat{\beta}(s) = \frac{6 + 7s + 3s^2}{3(1+s)(2+2s+s^2)}, \qquad \gamma_1(s) = \frac{2}{3(1+s)},$$
$$\gamma_2(s) = \gamma_5(s) = \frac{1}{2+2s+s^2}, \qquad 3\gamma_3(s) = \gamma_4(s) = \frac{1+s}{2+2s+s^2},$$

and the system for the probability π_0 and the constants $c_1, \ldots, c_5$ now becomes

$$\pi_0 + \lambda\pi_0 \frac{6 + 7\lambda + 3\lambda^2}{3(1+\lambda)(2+2\lambda+\lambda^2)} + \\ + \lambda\Big(\frac{2c_1}{3(1+\lambda)} + \frac{c_2 + c_5}{2+2\lambda+\lambda^2} + \frac{(1+\lambda)(c_3+3c_4)}{3(2+2\lambda+\lambda^2)}\Big) = 1,$$

$$c_1 = \lambda\pi_0\Big(\frac{1}{3+3\lambda} + \frac{6+2\lambda}{15(2+2\lambda+\lambda^2)}\Big) + \\ + \frac{1}{15}\lambda\Big(\frac{5c_1}{1+\lambda} + \frac{6c_2 + 4c_3 + 12c_4 + 6c_5 - 3\lambda(c_2 - c_3 - 3c_4 + c_5)}{2+2\lambda+\lambda^2}\Big),$$

$$c_2 = \lambda\pi_0\Big(\frac{4}{45(1+\lambda)} + \frac{4+\lambda}{36(2+2\lambda+\lambda^2)}\Big) + \\ + \lambda\Big(\frac{4c_1}{45(1+\lambda)} + \frac{3(c_2+c_3+3c_4+c_5) - \lambda(3c_2-2c_3-6c_4+3c_5)}{36(2+2\lambda+\lambda^2)}\Big),$$

$$c_3 = \frac{\lambda\pi_0(26+31\lambda+13\lambda^2)}{60(2+4\lambda+3\lambda^2+\lambda^3)} + \\ + \frac{\lambda}{60}\Big(\frac{8c_1}{1+\lambda} + \frac{5(3c_2+c_3+\lambda c_3+3\lambda c_4+3c_4+3c_5)}{2+2\lambda+\lambda^2}\Big),$$

$$c_4 = \lambda\pi_0\Big(\frac{4}{45(1+\lambda)} + \frac{4+\lambda}{36(2+2\lambda+\lambda^2)}\Big) + \\ + \lambda\Big(\frac{4c_1}{45(1+\lambda)} + \frac{3(c_2+c_3+3c_4+c_5) - \lambda(3c_2-2c_3-6c_4+3c_5)}{36(2+2\lambda+\lambda^2)}\Big),$$

$$c_5 = \frac{-\lambda\pi_0(26+31\lambda+13\lambda^2)}{180(2+4\lambda+3\lambda^2+\lambda^3)} - \\ - \lambda\Big(\frac{2c_1}{45(1+\lambda)} + \frac{3c_2+c_3+\lambda c_3+3c_4+3\lambda c_4+3c_5}{36(2+2\lambda+\lambda^2)}\Big).$$

The solution to this system is given by

$$\begin{aligned}
\pi_0 &= \frac{10800+16200\lambda+9753\lambda^2+2542\lambda^3}{10800+27000\lambda+22353\lambda^2+7940\lambda^3},\\
c_1 &= \frac{5760\lambda+6612\lambda^2+2663\lambda^3}{10800+27000\lambda+22353\lambda^2+7940\lambda^3},\\
c_2 = c_4 &= \frac{4680\lambda+5301\lambda^2+2066\lambda^3}{3\,(10800+27000\lambda+22353\lambda^2+7940\lambda^3)},\\
c_3 = -3c_5 &= \frac{2340\lambda+2778\lambda^2+1176\lambda^3}{10800+27000\lambda+22353\lambda^2+7940\lambda^3},
\end{aligned}$$

from which we can compute the waiting-time distribution. For our example, the distribution is given by

$$F_W(x) = 1 - \frac{2\lambda\mathrm{e}^{-x}}{10800+27000\lambda+22353\lambda^2+7940\lambda^3} \times \\ \times \big(5(720+744\lambda+347\lambda^2) + 4(450+645\lambda+241\lambda^2)\cos x + \\ + 2\lambda(255+286\lambda)\sin x\big).$$

As an example, in Figure 5.1 we have plotted the waiting-time distribution for $\lambda = 2$.

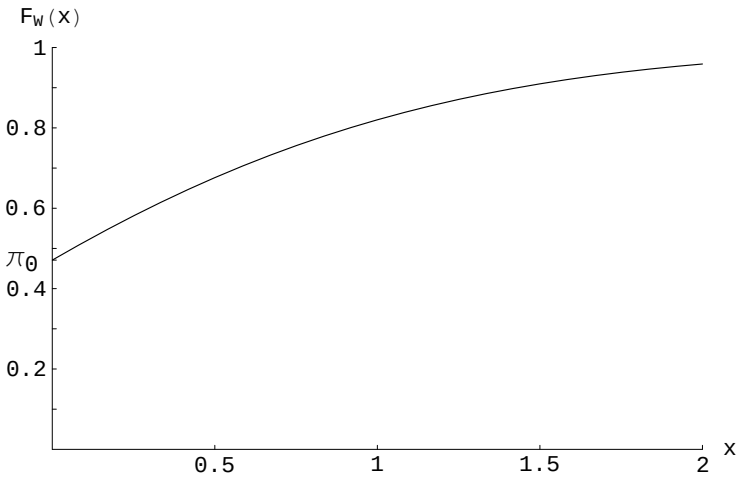

Figure 5.1: The waiting-time distribution for $\lambda = 2$.

A few observations are necessary. As we can see from the above examples, the size of the system cannot be determined before choosing a decomposition of the kernel (for example, even for phase-type distributions it is not necessarily a function of the number of phases of F_B). The technique is, however, simple and can be implemented without any numerical difficulties. Therefore, for preparation-time distributions that are not of the phase-type form considered in Chapter 4 but have a rational Laplace transform, it is more advantageous to apply Theorem 5.1 than approximate the given distribution with a phase-type distribution of the form of Equation (4.32) and then employ Theorem 4.10.

5.7 Concluding remarks

In this chapter, we have reduced the analysis of the M/G model to the solution of a set of linear equations. The technique we have used is a well-known method, frequently used for the solution of Fredholm equations, whenever the kernel of the equation admits factorisation.

A quick comparison of this model to the M/G/1 queue reveals that there is an abundance of differences between them. Consider the M/G/1 queue with occupation rate $\rho < 1$. In steady state we have that the distribution of the waiting time W is given by an infinite sum of convolutions of the residual service time; namely,

$$F_W(x) = (1-\rho)\sum_0^\infty \rho^n \left(\frac{1}{\mathbb{E}[B]}\int_0^x \big(1 - F_B(y)\big)\,\mathrm{d}y\right)^{n*}.$$

It does not seem possible to derive the steady-state waiting-time distribution for

every distribution F_B for our model, while for the single-server queue we have in this case a variety of methods to choose from. The collection of distributions that do not fall under the framework of this chapter includes distributions with discontinuities, distributions without a rational Laplace transform, and distributions on a bounded support. For such distributions, one may approximate F_B by an appropriate distribution, for which we can obtain an exact result for the steady-state waiting-time distribution. We shall consider such approximation schemes in the following chapter.

Since the steady-state distribution of the M/G model is so troublesome, it does not seem surprising that the time-dependent distribution for this model also presents difficulties. Naturally, as we have noted before, the methods developed in the previous chapter cannot be applied here. Due to the lack of structure of the preparation-time distribution, we cannot define a suitable Markov chain without incorporating the – unknown – distribution of the residual preparation time (for example, after a service completion). Again, the difference between this model and the single-server queue is apparent. For the M/G/1 queue the generating function of the Laplace-Stieltjes transform of W_n is explicitly known (see Cohen [46, Section II.4.5]), while even this result seems difficult to derive for our model.

Chapter 6

Approximations

6.1 Introduction

So far we have mainly derived exact formulas for the waiting-time distribution of the server for the alternating-service model described in Section 1.4. These formulas, however, cannot be applied for every possible setting. In particular, if the preparation-time distribution F_B does not belong to the class of distributions described in Section 5.3, then we do not have a closed-form expression for the steady-state waiting-time distribution F_W. In such a case, one has to resort to approximations or simulation.

In Chapter 2 we have derived several results that can be used for various approximation schemes. For example, the covariance function in Section 2.6 can be used to give an approximation of the standard deviation of a series of waiting times obtained by simulation. From the results in Section 2.5 regarding the tail behaviour of the model we can obtain an estimation of the probability that the waiting time exceeds a large value, while from Theorem 2.2 we see that the waiting-time distribution can be approximated by consecutive iterations of a functional equation.

Another possible approach for approximating F_W is to approximate F_B and compute the exact waiting-time distribution for this approximation by applying results derived in the previous chapters. For example, if F_B is approximated by a phase-type distribution, then the resulting steady-state waiting time is given by Theorem 4.10. In this chapter, we discuss such approximation schemes.

An innate question is about the approximation error involved. We treat this subject by giving bounds for the error made in the computation of the waiting-time distribution in case either F_A or F_B are approximated. In order to provide these bounds in Section 6.2, we utilise the fact that the mapping $\mathcal{T}$ given in Equation (2.2) is a contraction mapping. Moreover, in Section 6.3 we discuss how one can approximate a given distribution with a phase-type or a polynomial distribution (in case the given distribution has a bounded support) and we derive a system of differential equations defining f_W for various cases where F_B is a mixed distribution with a continuous and a discrete part. We close the chapter in Section 6.4 with some numerical results and final remarks.

In this chapter, which is partially based on results derived in [171], we use the following notational convention: for a function f we denote by $\widehat{f}$ its approximation both for the case that f is directly approximated and for the case that $\widehat{f}$ is the exact expression derived for some other approximated function.

6.2 Error bounds

Error bounds for queueing models have been studied widely. The main question is to define an upper bound of the distance between the distribution in question and its approximation, that depends on the distance between the governing distributions. These bounds are obtained both in terms of weighted metrics (see, e.g., [96]) and non-weighted metrics (see, e.g., [26, 28] and references therein). An important assumption which is often made in these studies is that the recursion under discussion should be non-decreasing in its main argument. Since this assumption does not hold for the model we discuss, in this section we derive error bounds for the approximated waiting-time distribution in case we approximate either the preparation-time distribution or the service-time distribution. Because of Theorem 2.2, we shall limit ourselves to the uniform norm.

Let $\widehat{F}_B$ be an approximation of F_B and $\widehat{F}_W$ the exact solution that we obtain in that case for the distribution of W. Let $\widehat{B}$ be a random variable that is distributed according to $\widehat{F}_B$, and let $\widehat{X} = \widehat{B} - A$. Define now the mapping (cf. (2.2))

$$(\widehat{\mathcal{T}}F)(x) = 1 - \int_x^\infty F(y-x)\,\mathrm{d}F_{\widehat{X}}(y),$$

which yields that $\widehat{F}_W$ is the solution to $F = \widehat{\mathcal{T}}F$ that can be rewritten in the form (cf. (2.2))

$$(\widehat{\mathcal{T}}F)(x) = \int_0^\infty \int_0^\infty \widehat{F}_B(x+z+y)\,\mathrm{d}F_A(z)\mathrm{d}F(y).$$

Then we can prove the following theorem.

Theorem 6.1. *Let* $\|F_B - \widehat{F}_B\| = \varepsilon$. *Then* $\|F_W - \widehat{F}_W\| \leqslant \varepsilon/(1 - \mathbb{P}[B > A])$.

Proof. We have that

$$\begin{aligned}\|F_W - \widehat{F}_W\| &= \|\mathcal{T}F_W - \widehat{\mathcal{T}}\widehat{F}_W\| = \|\mathcal{T}F_W - \mathcal{T}\widehat{F}_W + \mathcal{T}\widehat{F}_W - \widehat{\mathcal{T}}\widehat{F}_W\| \\ &\leqslant \|\mathcal{T}F_W - \mathcal{T}\widehat{F}_W\| + \|\mathcal{T}\widehat{F}_W - \widehat{\mathcal{T}}\widehat{F}_W\| \\ &\leqslant \mathbb{P}[B > A]\|F_W - \widehat{F}_W\| + \|\mathcal{T}\widehat{F}_W - \widehat{\mathcal{T}}\widehat{F}_W\|,\end{aligned}$$

since $\mathcal{T}$ is a contraction mapping with contraction constant $\mathbb{P}[B > A]$. Furthermore,

$$\begin{aligned}\|\mathcal{T}\widehat{F}_W - \widehat{\mathcal{T}}\widehat{F}_W\| &= \sup_{x\geqslant 0}\Big|\int_0^\infty\int_0^\infty F_B(x+z+y)\,\mathrm{d}F_A(z)\mathrm{d}\widehat{F}_W(y) - \\ &\qquad - \int_0^\infty\int_0^\infty \widehat{F}_B(x+z+y)\,\mathrm{d}F_A(z)\mathrm{d}\widehat{F}_W(y)\Big| \\ &\leqslant \sup_{x\geqslant 0}\int_0^\infty\int_0^\infty \Big|F_B(x+z+y) - \widehat{F}_B(x+z+y)\Big|\,\mathrm{d}F_A(z)\mathrm{d}\widehat{F}_W(y) \\ &\leqslant \sup_{x\geqslant 0}\int_0^\infty\int_0^\infty \sup_{x+y+z\geqslant 0}\Big|F_B(x+z+y) - \widehat{F}_B(x+z+y)\Big|\,\mathrm{d}F_A(z)\mathrm{d}\widehat{F}_W(y) \\ &= \varepsilon\int_0^\infty\int_0^\infty \mathrm{d}F_A(z)\mathrm{d}\widehat{F}_W(y) = \varepsilon.\end{aligned}$$

So $\|F_W - \widehat{F}_W\| \leqslant \mathbb{P}[B > A]\|F_W - \widehat{F}_W\| + \varepsilon$, which is what we wanted to prove. □

An important feature of Equation (1.2) that made the calculation of an error bound straightforward is that the distribution of the waiting time is the fixed point of a contraction mapping. Note that this is not a property of Lindley's recursion.

In case we want to approximate the service-time distribution instead of F_B, then the statement of Theorem 6.1 and its proof remain identical, if we interchange F_A and F_B. Therefore, we omit this proof, and we simply state the following theorem.

Theorem 6.2. *Let* $\|F_A - \widehat{F}_A\| = \varepsilon$. *Then* $\|F_W - \widehat{F}_W\| \leqslant \varepsilon/(1 - \mathbb{P}[B > A])$.

6.3 Approximations of the waiting-time distribution

The result we have obtained in the previous section comes in handy in some cases where it is necessary to resort to approximations of the waiting-time distribution. We have already proven in Chapter 2 that for any distribution of A and B there exists a unique limiting distribution F_W for (1.2), although we may not be able to compute it. As we have seen in the previous chapter, we can compute the waiting-time distribution for F_A being some general distribution and F_B belonging to class $\mathcal{M}$. Futhermore, Chapter 3 covers the case of F_B being a polynomial distribution. Such distributions with bounded support are excluded from class $\mathcal{M}$.

In the present section we propose two approaches to approximating F_W. In Section 6.3.1 we approximate F_B by a phase-type distribution. An important reason is that the class of phase-type distributions is dense; any distribution on $[0, \infty)$ can, in principle, be approximated arbitrarily well by a phase-type distribution (see [149]). In Section 6.3.2 we approximate F_B by a polynomial distribution, which is a more natural choice if F_B has a bounded support. Thus, we can subsequently apply the results of Chapters 5 and 3 respectively.

6.3.1 Fitting phase-type distributions

The fitting techniques available for phase-type distributions are either based on moment matching or on maximum likelihood estimators (MLEs). Moment matching techniques are computationally efficient, but usually apply to somewhat more restrictive models; see Johnson & Taaffe [93, 95, 94], Tijms [160], and recent developments in Osogami [137]. There are various techniques that are based on MLEs. Some examples are:

a) a numerical optimisation method to fit long-tailed distributions into Coxian distributions (Horváth & Telek [84]),

b) a divide-and-conquer technique to fit data sets with non-monotone densities into a mixture of Erlang and hyperexponential distributions and also data sets with completely monotone densities into hyperexponential distributions (Riska *et al.* [143]),

c) the Feldmann-Whitt algorithm [62] which is a heuristic-based approach used for fitting heavy-tailed distributions, such as Weibull or Pareto, into hyperexponential distributions,

d) and the Expectation-Maximisation (EM) algorithm for fitting both data and distributions into general phase-type distributions (Asmussen *et al.* [8]).

Approximation of another distribution by a phase-type distribution, in the sense of minimising the information divergence, can be regarded as an infinite analogue of fitting a phase-type distribution to a sample. Lang & Arthur [107] compared two moment matching techniques and two techniques that are based on MLEs only to conclude that there does not yet exist a single superior parameter approximation method for phase-type distributions.

Regardless of which approximation technique one may choose, the main characteristic that $\widehat{F}_B$ should possess, is that it should minimise the distance from F_B. In other words, one should choose $\widehat{F}_B$ in such a way that $\sup_{x\geqslant 0}|F_B(x)-\widehat{F}_B(x)|$ is as small as possible. Therefore, it may be more reasonable to opt for techniques that match the graph of F_B to a phase-type distribution, since moment matching techniques cannot guarantee that the distance between the two graphs will be minimal. Nonetheless, even if one chooses moment matching techniques (for instance, because they are computationally efficient), then one should compute $\|F_B-\widehat{F}_B\|=\varepsilon$ and decide if the bound of the approximation error of F_W is acceptable. After having chosen an appropriate $\widehat{F}_B$ the computation of $\widehat{F}_W$ is a direct application of the method described in Section 5.5. In the special case where $\widehat{F}_B$ is a mixture of Erlang distributions with the same scale parameter for all exponential phases, then Theorem 4.10 provides us with a simple expression for the density $\widehat{f}_W$. The approximated waiting-time distribution will be in that case a different mixture of Erlang distributions, that will still have the same scale parameter.

6.3.2 Fitting polynomial distributions

If F_B is a continuous distribution on a bounded support, it is reasonable to choose $\widehat{F}_B$ to be a polynomial distribution. The famous *Weierstrass approximation theorem* asserts the possibility of uniform approximation of a continuous, real-valued function on a closed and bounded support by some polynomial. The following theorem is a more precise version of Weierstrass' theorem. It is a special case of the theorem by S. Bernstein that is stated in [63, Section VII.2].

Theorem 6.3. *If F is a continuous distribution on the closed interval $[0,1]$, then as $n\to\infty$*

$$\hat{F}_n(x)=\sum_{k=0}^{n}F(k/n)\binom{n}{k}x^k(1-x)^{n-k}\to F(x)$$

uniformly for $x\in[0,1]$. Furthermore, $\hat{F}_n$ is also a distribution.

Proof. Bernstein's theorem states that if F is a continuous function, then it can be approximated uniformly in x with the polynomial $\hat{F}_n$. In other words, for any given

$\varepsilon > 0$, there is an N independent from x, such that for all $n > N$, $|\hat{F}_n(x) - F(x)| < \varepsilon$, for all x.

It is easy to show that if the function F is a distribution on $[0,1]$, then the approximation $\hat{F}_n(x)$ is also a distribution, since it is continuous, $0 \leqslant \hat{F}_n(x) \leqslant 1$, and, by checking its derivative, we shall show that it is non-decreasing in x. It suffices to note that

$$\begin{aligned}
\hat{F}_n'(x) &= \sum_{k=1}^{n} F(k/n)\binom{n}{k} k x^{k-1}(1-x)^{n-k} - \sum_{k=0}^{n-1} F(k/n)\binom{n}{k} x^k (n-k)(1-x)^{n-k-1} \\
&= \sum_{k=0}^{n-1} x^k (1-x)^{n-k-1} \left[F((k+1)/n)\binom{n}{k+1}(k+1) - F(k/n)\binom{n}{k}(n-k) \right] \\
&= \sum_{k=0}^{n-1} x^k (1-x)^{n-k-1} \frac{n!}{k!(n-k-1)!} \left[F((k+1)/n) - F(k/n) \right].
\end{aligned}$$

The expression in the square brackets at the right hand side is positive since F is a distribution. Therefore, $\hat{F}_n'(x) \geqslant 0$, for $x \in [0,1]$. □

So, given a continuous distribution F_B that has all its mass concentrated on $[0,1]$, one can compute a polynomial distribution $\widehat{F}_B$ that approximates F_B arbitrarily well by using Theorem 6.3. In this sense, the class of polynomial distributions is dense. Then $\widehat{F}_W$ can be computed by using Theorem 3.5, and an error bound for this approximation will be given by Theorem 6.1

6.3.3 Fitting distributions that have one discontinuity

So far, we were concerned with the case where F_B is a continuous distribution that is approximated arbitrarily well by a polynomial or a phase-type distribution $\widehat{F}_B$. However, if F_B is discontinuous, then one may not be able to choose an appropriate approximation that instigates an acceptable error. Nonetheless, one can explicitly compute $\widehat{F}_W$ in the special case that F_B is discontinuous at a single point x_0 of its support as follows. Since F_B is a distribution of a mixed type, it can be decomposed into two parts: a discrete distribution F_D and a continuous one F_C. Now, F_C can be approximated arbitrarily well with either a phase-type or a polynomial distribution, and we obtain that $\widehat{F}_B(x) = p_D F_D(x) + p_C \widehat{F}_C(x)$.

Whether one should choose to approximate F_C with a phase-type or a polynomial distribution depends on the computational efficiency of each method, in combination with the approximation error that occurs. Here, we shall give two specific examples. We assume that F_A is the exponential distribution with parameter λ and we study the case where F_C is (approximated by) either a polynomial distribution on $[0,1]$ or the exponential distribution with parameter μ. In the following, we shall write F_C for the approximation of the continuous part of F_B, as it is not necessary to distinguish between the actual funtion and its approximation.

The exponential case

Since F_A is exponentially distributed, the density of the waiting-time distribution is the unique solution of Equation (5.2). Let now F_B be given by

$$F_B(x) = p_C F_C(x) + p_D F_D(x),$$

where $p_C + p_D = 1$, $F_C(x) = 1 - \mathrm{e}^{-\mu x}$, $x \geqslant 0$, and $F_D(x) = c$ for $0 \leqslant x < x_0$ and $F_D(x) = 1$ for $x \geqslant x_0$. Then from (5.2) we readily have after differentiating that for $0 < x < x_0$

$$f'_W(x) = \lambda f_W(x) - \lambda\mu p_C\big(\pi_0 + \omega(\mu)\big) + \lambda(p_C c - p_D) f_W(x_0 - x)$$

and for $x > x_0$ we have that

$$f_W(x) = \lambda F_W(x) - \lambda + \lambda p_C\big(\pi_0 + \omega(\mu)\big).$$

The latter of these two equations is a simple first-order linear differential equation, while the former equation can be solved for example by using Laplace transforms and tracing back the steps we have followed in Section 3.4.1. These results can be extended to the case where F_C is a usual mixed-Erlang distribution, i.e. F_C is given by the right hand side of Equation (4.32). We shall again have two differential equations as above. For $0 < x < x_0$ the waiting time density will be the sum of two exponentials plus two other terms, each of which is an exponential multiplied by a polynomial; for $x > x_0$, the waiting-time distribution again involves the term $\mathrm{e}^{-\mu x}$ multiplied with a polynomial of degree equal to the number of phases of F_C.

The polynomial case

We now assume that the continuous part of F_B is (or can be approximated by) a polynomial distribution on $[0, 1]$; that is, $F_C(x) = \sum_{i=0}^{n} c_i x^i$, $0 \leqslant x \leqslant 1$, and $\sum c_i = 1$. Then, by differentiating (5.2) a total of $n+1$ times we derive the following two differential equations for the waiting-time density. For the first equation we have that for $0 < x < x_0$

$$f_W^{(n+1)}(x) = \lambda f_W^{(n)}(x) + p_C \sum_{i=0}^{n-1} \nu_i (-1)^i f_W^{(i)}(1 - x) - \lambda p_D (1 - c)(-1)^n f_W^{(n)}(x_0 - x), \tag{6.1}$$

where the constants ν_i are given by Equation (3.30). The second equation is almost identical to Equation (3.32); namely, for $x_0 < x < 1$ we have that

$$f_W^{(n+1)}(x) = \lambda f_W^{(n)}(x) + p_C \sum_{i=0}^{n-1} \nu_i (-1)^i f_W^{(i)}(1 - x), \tag{6.2}$$

where the constants ν_i are defined as above. Notice that (6.1) has one extra term compared to (6.2). Although there exists a unique solution to this system, deriving an exact formula from the above equations seems challenging. For example, a source

of inconvenience is the term $f_W(1-x)$ and all its derivatives that appear in (6.1) and (6.2). One can observe that for $x \in (0, x_0)$ this function is either the solution of Equation (6.2) if $x \in (0, 1-x_0)$ or it is simply the solution of Equation (6.1), in which case the three different arguments appearing in (6.1) make this equation non-standard. A similar observation is valid for Equation (6.2).

However, if we choose $x_0 = 0.5$, i.e. if we limit ourselves to the lattice case for F_D, then both equations given above simplify. The terms involving the argument $1-x$ are now in terms of the solutions of Equations (6.1) and (6.2), since for $x \in (0, 0.5)$, $f_W(1-x)$ is given by the solution of (6.2), and similarly for $x \in (0.5, 1)$. Thus, if F_C has a bounded support, we need to make the extra assumption that the discontinuity occurs exactly in the middle of the support. Then these two differential equations form a system, since each of them involves the density on the other interval, that can be solved explicitly. For the computation of various constants that appear in the solution, apart from all obvious conditions, such as the normalisation equation and the conditions arising by each differentiation, one needs to keep in mind that although F_B is discontinuous, F_W is a continuous distribution.

Remark 6.1. In case that F_D is on a lattice (with more than one discontinuity) one can again derive a system of differential equations that gives the density of the waiting time in each interval. Hence, it seems possible to obtain an explicit solution to this system by following the same method as in Section 3.6. The system of differential equations that arises in the case of multiple discontinuities has an intriguing form. As an example, we shall give this system for a specific case. As before, let F_A and F_C be exponentially distributed with rates λ and μ respectively, and $F_D(x)$ be equal to c_1 if $x \in [0, x_1)$, to c_2 if $x \in [x_1, x_2)$, and to 1 if $x \geqslant x_2$. Then we have that:

For $0 < x < x_1$,

$$f'_W(x) = \lambda f_W(x) - \lambda\mu p_C(\omega(\mu)+\pi_0)\mathrm{e}^{-\mu x} - \lambda(1-c_2)f_W(x_2-x) - \\ - \lambda p_D(c_2-c_1)f_W(x_1-x),$$

for $x_1 < x < x_2$,

$$f'_W(x) = \lambda f_W(x) - \lambda\mu p_C(\omega(\mu)+\pi_0)\mathrm{e}^{-\mu x} - \lambda(1-c_2)f_W(x_2-x),$$

and for $x > x_2$,

$$f_W(x) = \lambda F_W(x) - \lambda + \lambda p_C(\omega(\mu)+\pi_0)\mathrm{e}^{-\mu x}.$$

As we see, each differential equation is identical to the previous one if we remove the last term. The equation in the last interval can always be solved explicitly (up to the normalisation constant), and on a lattice the rest of the equations will involve the density on at least one other interval, thus forming a system that can also be solved explicitly. The usual initial conditions (such as the normalisation equation and the continuity of F_W) will provide enough equations to determine all constants appearing in the computations.

6.4 Numerical results

This section is devoted to some numerical results. For the service-time distribution F_A, we have chosen the exponential distribution with parameter $\lambda = 1$. For a given distribution F_B we calculate from Theorem 6.3 three polynomial distributions (of first, fifth, and tenth order) that approximate F_B, and we plot the resulting densities and distributions of the waiting time. The distribution F_B considered is the piecewise polynomial distribution

$$F_B(x) = (2x^2)\mathbb{1}_{[0\leqslant x\leqslant 1/2]} + (-2x^2 + 4x - 1)\mathbb{1}_{[1/2\leqslant x\leqslant 1]} + \mathbb{1}_{[x\geqslant 1]},$$

where $\mathbb{1}_{[S]}$ is the indicator function of the set S. This distribution is simply the well-known symmetric *triangular* distribution on $[0, 1]$. In Figure 6.1 we plot f_B and the three corresponding densities of the polynomial distributions we have chosen for the approximation of F_B next to the waiting-time densities and its approximations, while in Figure 6.2 we plot the analogous graphs for the distributions. Note that the last plot uses two different scales.

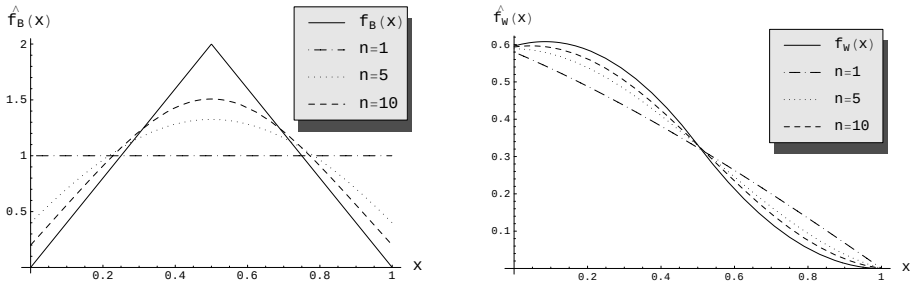

Figure 6.1: The density f_B and its approximations, and the resulting waiting-time densities.

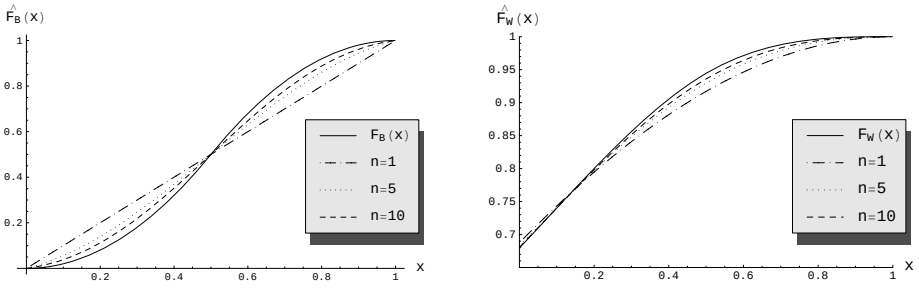

Figure 6.2: The distribution F_B and its approximations, and the resulting waiting-time distributions.

For the above approximations we have computed the distance between F_B and

$\widehat{F}_B$, f_W and $\widehat{f}_W$, F_W and $\widehat{F}_W$, as well as the error bound for F_W and $\widehat{F}_W$ as it is predicted by Theorem 6.1. Evidently, the error bound that is predicted by Theorem 6.1 is rather crude. Furthermore, the resulting error between F_W and $\widehat{F}_W$ in this case is approximately 3 times smaller than the error incurred between the densities (which is an expected consequence of the mass π_0 of the distribution at the origin), and 4.5 times smaller than the initial approximation error between F_B and $\widehat{F}_B$. Evidently, the service-time distribution smoothes out the resulting error for F_W when approximating F_B. The above are summarised in Table 6.1.

	$\|f_B - \widehat{f}_B\|$	$\|F_B - \widehat{F}_B\|$	$\|f_W - \widehat{f}_W\|$	$\|F_W - \widehat{F}_W\|$	bound
$n = 1$	1	0.1250	0.0841	0.0274	0.3283
$n = 5$	0.6750	0.0664	0.0449	0.0147	0.1744
$n = 10$	0.4922	0.0385	0.0264	0.0086	0.1013

Table 6.1: The distances between the real distributions or densities and their approximations, and the error bound given in Theorem 6.1.

Chapter 7
Dependencies

7.1 Introduction

In the previous chapters we have studied Recursion (1.1) under the assumption that $\{A_n\}$ and $\{B_n\}$ are two i.i.d. sequences of random variables that are mutually independent. The assumption that all random variables involved are independent of one another arises naturally in various cases due to the model specifications, and usually simplifies the analysis. However, in many applications there exist dependence structures between the preparation times and the service times, which means that this assumption is simply not correct. In the present chapter we remove the assumption that all random variables involved are independent of one another, and we study two specific dependence structures between $\{A_n\}$ and $\{B_n\}$ that are described below.

In the first example we study, the distributions of the preparation and service times are regulated by an irreducible discrete-time Markov chain. Specifically, we assume that each transition of the Markov chain generates a new preparation time and its corresponding service time. Given the state of the Markov chain at times n and $n+1$, the distributions of A_n and B_{n+1} are independent of one another for all n. However, the distributions of A_n and B_n depend on the state of the Markov chain. Such a dependence structure occurs naturally in many applications. For example, in the application involving two carousels that is described in Section 1.2, one can intuitively see that if an order consists of multiple items on one carousel that need to be picked, then there are strategies for the preparation of the carousel, where a long preparation time B_n implies that the service time A_n (i.e. the time necessary to pick all items on that carousel) will be relatively short, while being independent of all other past or future preparation and service times.

The second dependence structure we study assumes that the random variables A_n and B_{n+1} have a joint distribution. In particular, given the length of the service time A_n, the following preparation time has a Laplace-Stieltjes transform of a specific form. The form we choose is rather general and allows for various specific dependence structures and preparation time distributions. Later on, we shall give a few specific examples. Dependencies between a service time and the *following* preparation time are also possible in applications. Again for the carousel model described in Section 1.2 with orders consisting of multiple items, one can have that a "smart" preparation strategy is followed, which anticipates the expected delay of the server for the previous order. Thus, knowing that the previous service time is relatively long, the other carousel rotates at a starting point that may be further away, but reduces the service time of the following order.

Both examples studied are also motivated by analogous cases studied for Lindley's recursion. In Section 7.2 we derive the steady-state waiting-time distribution

for the first case studied. We assume that the service times follow some general distribution that depends on the state of the Markov chain. For the preparation times, in Section 7.2.1 we assume that they are exponentially distributed (with a rate depending on the state of the Markov chain), while in Section 7.2.2 we extend the analysis to mixed-Erlang distributions. In Section 7.3 we derive F_W for the second case studied. We assume that F_A is the exponential distribution with rate λ, although, as we remark later on, the analysis can be extended to mixed-Erlang distributions. We conclude in Section 7.4, where we compare our results to the analogous results for Lindley's recursion and where we make some final remarks.

7.2 Markov-modulated dependencies

In this section we study the case where the preparation times and the service times depend on a common discrete-time Markov chain. This model allows dependencies between preparation and service times. The waiting time in this case is directly derived by using Laplace transforms. For Lindley's recursion, the analogous model has been analysed in Adan and Kulkarni [1], which we closely follow. In the next section, we analyse this model for exponentially distributed preparation times, and in Section 7.2.2 we generalise this result to phase-type preparation times of the form of Equation (4.32). Recall that for a random variable Y and an event E we have that $\mathbb{P}[Y \leqslant x\,; E] = \mathbb{E}[\mathbb{1}_{[Y \leqslant x]} \cdot \mathbb{1}_{[E]}]$, and likewise for expectations.

We assume that the sequences $\{A_n\}$ and $\{B_n\}$ are both autocorrelated and cross-correlated. The distributions of the preparation and service times are regulated by an irreducible discrete-time Markov chain $\{Z_n\}$, $n \geqslant 1$, with state space $\{1, 2, \ldots, M\}$ and transition probability matrix $\mathbf{P} = (p_{i,j})$. More precisely, we have that

$$\begin{aligned}
&\mathbb{P}[A_n \leqslant x\,; B_{n+1} \leqslant y\,; Z_{n+1} = j \mid Z_n = i\,; B_n\,; (A_\ell, B_\ell, Z_\ell), 1 \leqslant \ell \leqslant n-1] \\
&\qquad = \mathbb{P}[A_1 \leqslant x\,; B_2 \leqslant y\,; Z_2 = j \mid Z_1 = i] \\
&\qquad = p_{i,j}\mathbb{P}[A_1 \leqslant x\,; B_2 \leqslant y \mid Z_1 = i\,; Z_2 = j] \\
&\qquad = p_{i,j}F_{A,i}(x)F_{B,j}(y), \qquad (7.1)
\end{aligned}$$

where $x, y \geqslant 0$ and where $i, j = 1, 2, \ldots, M$. Thus, given Z_n and Z_{n+1}, the distributions of A_n and B_{n+1} are independent of one another for all n. The random variables A_n follow an arbitrary distribution that is independent of the past, given Z_n, while B_n follows in general a phase-type distribution that is depending on the state of Z_n.

Observe that (W_n, Z_n) is a Markov chain. Since for all n, we have assumed that $\mathbb{P}[X_n < 0] > 0$, this Markov chain is stable. To see that, notice that since $\mathbb{P}[X_n < 0] > 0$, there is an i such that $\mathbb{P}[X_n < 0, Z_n = i] > 0$, and without loss of generality we can take $i = 1$. A regeneration point occurs if $W_n = 0$ and $Z_n = 1$; thus if $X_n < 0$ and $Z_n = 1$ the process regenerates and the event $W_n = 0$ is included in the event $X_n < 0$. Since the Markov chain $\{Z_n\}$ reaches state 1 infinitely often, the time between two occurrences of state 1 is finite in expectation, and for each

time this state is reached there is a positive probability that $X_n < 0$, we will have that in a geometric number of steps both events $X_n < 0$ and $Z_n = 1$ will happen at the same step. In other words, we will have a regeneration point. Since the Markov chain (W_n, Z_n) is stable, define for $\mathrm{Re}(s) \geqslant 0$, $n \geqslant 1$, and $j = 1, 2, \ldots, M$ the transforms

$$\omega_j^n(s) = \mathbb{E}[\mathrm{e}^{-sW_n}\,; Z_n = j],$$

and

$$\omega_j(s) = \lim_{n\to\infty} \omega_j^{n+1}(s).$$

Let λ_i^{-1} be the mean and s_i be the second moment of the service time distribution $F_{A,i}$. Analogously, define μ_i^{-1} as the mean of $F_{B,i}$ and σ_i as its second moment. Moreover, denote by $\boldsymbol{\varpi} = (\varpi_1, \varpi_2, \ldots, \varpi_M)$ the stationary distribution of the Markov chain $\{Z_n\}$. Then, in steady state, the autocorrelation between A_m and A_{m+n} is given by

$$\rho[A_m, A_{m+n}] = \rho[A_1, A_{n+1}] = \frac{\sum_{i=1}^M \sum_{j=1}^M \varpi_i \big(p_{i,j}^{(n)} - \varpi_j\big) \lambda_i^{-1} \lambda_j^{-1}}{\sum_{i=1}^M \varpi_i s_i - \Big(\sum_{i=1}^M \varpi_i \lambda_i^{-1}\Big)^2},$$

where

$$p_{i,j}^{(n)} = \mathbb{P}[Z_{n+1} = j \mid Z_1 = i], \qquad n \geqslant 0, \quad 1 \leqslant i, j \leqslant M.$$

A similar expression holds for the autocorrelation between preparation times. Provided $\mathbf{P}$ is aperiodic, $p_{i,j}^{(n)}$ converges to ϖ_j geometrically as n tends to infinity. Hence, the autocorrelation function approaches zero geometrically fast as the lag goes to infinity. For the cross-correlation between A_n and B_n we have that

$$\rho[A_n, B_n] = \rho[A_1, B_1] = \frac{\sum_{i=1}^M \varpi_i \lambda_i^{-1} \mu_i^{-1} - \widehat{\mu}\,\widehat{\lambda}}{\Big(\sum_{i=1}^M \varpi_i s_i - \widehat{\lambda}^2\Big)^{1/2} \Big(\sum_{j=1}^M \varpi_j \sigma_i - \widehat{\mu}^2\Big)^{1/2}},$$

where $\widehat{\lambda} = \sum_{i=1}^M \varpi_i \lambda_i^{-1}$ and $\widehat{\mu} = \sum_{i=1}^M \varpi_i \mu_i^{-1}$.

When the $\{Z_n\}$ is in state j, we denote by $f_{W,j}$ the steady-state waiting-time density, and by α_j the Laplace-Stieltjes transform of F_A. Moreover, recall that the derivative of order i of a function f is denoted by $f^{(i)}$ and by definition $f^{(0)} = f$.

7.2.1 Exponential preparation times

In this section we assume that $F_{B,j}(x) = 1 - \mathrm{e}^{-\mu_j x}$. We are interested in the steady-state waiting-time distribution. For the derivation, we shall use Laplace transforms and follow arguments similar to the ones appearing in Section 4.2.3. The next theorem gives the equations satisfied by the waiting-time densities

$$\boldsymbol{f_W}(x) = \big(f_{W,1}(x), f_{W,2}(x), \ldots, f_{W,M}(x)\big).$$

Theorem 7.1. *Let the preparation-time and service-time distributions be governed by a common discrete-time Markov chain according to Equation (7.1), where $p_{i,j}$ is the one-step transition probability of the Markov chain from state i to state j and $F_{B,j}(x) = 1 - \mathrm{e}^{-\mu_j x}$. Given that the Markov chain is in state j, in equilibrium the waiting-time distribution has mass $\pi_{0,j} = \varpi_j - c_j$ at the origin, where ϖ_j is the stationary probability that the Markov chain is in state j and where the constant c_j is given by $c_j = \sum_{i=1}^{M} p_{i,j}\, \omega_i(\mu_j)\, \alpha_i(\mu_j)$. Moreover, the waiting-time density given by*

$$f_{W,j}(x) = \mu_j c_j\, \mathrm{e}^{-\mu_j x}.$$

The M^2 unknown constants $\omega_i(\mu_j)$ that are needed in order to determine the unknown constants c_j are the unique solution to the system of linear equations given by the expression

$$\omega_j(\mu_\ell) = \varpi_j - \frac{\mu_\ell}{\mu_j + \mu_\ell} \sum_{i=1}^{M} p_{i,j}\, \omega_i(\mu_j)\, \alpha_i(\mu_j), \qquad j, \ell = 1, \ldots, M.$$

Proof. From Recursion (1.1) we obtain the following equation for the transforms ω_j^{n+1}, $j = 1, \ldots, M$.

$$\begin{aligned}
\omega_j^{n+1}(s) &= \mathbb{E}[\mathrm{e}^{-sW_{n+1}}\,; Z_{n+1} = j] \\
&= \sum_{i=1}^{M} \mathbb{P}[Z_n = i]\, \mathbb{E}[\mathrm{e}^{-s \max\{0, B_{n+1} - A_n - W_n\}}\,; Z_{n+1} = j \mid Z_n = i] \\
&= \sum_{i=1}^{M} \mathbb{P}[Z_n = i] p_{i,j} \Bigg(\mathbb{E}[\int_0^{A_n + W_n} f_{B_{n+1}}(x)\, \mathrm{d}x \mid Z_n = i\,; Z_{n+1} = j] + \\
&\quad + \mathbb{E}[\int_{A_n + W_n}^{\infty} \mathrm{e}^{-s(x - A_n - W_n)} f_{B_{n+1}}(x)\, \mathrm{d}x \mid Z_n = i\,; Z_{n+1} = j] \Bigg). \qquad (7.2)
\end{aligned}$$

Since $Z_{n+1} = j$, we have that B_{n+1} is now exponentially distributed with rate μ_j. Thus, the above equation becomes

$$\begin{aligned}
\omega_j^{n+1}(s) &= \sum_{i=1}^{M} \mathbb{P}[Z_n = i] p_{i,j} \Bigg(\mathbb{E}[\int_0^{A_n + W_n} \mu_j \mathrm{e}^{-\mu_j x}\, \mathrm{d}x \mid Z_n = i] + \\
&\qquad + \mathbb{E}[\int_{A_n + W_n}^{\infty} \mathrm{e}^{-s(x - A_n - W_n)} \mu_j \mathrm{e}^{-\mu_j x}\, \mathrm{d}x \mid Z_n = i] \Bigg) \\
&= \sum_{i=1}^{M} \mathbb{P}[Z_n = i] p_{i,j}\, \mathbb{E}[1 - \mathrm{e}^{-\mu_j (A_n + W_n)} + \frac{\mu_j}{\mu_j + s}\, \mathrm{e}^{-\mu_j (A_n + W_n)} \mid Z_n = i] \\
&= \sum_{i=1}^{M} \mathbb{P}[Z_n = i] p_{i,j} \Big(1 - \frac{s}{\mu_j + s}\, \mathbb{E}[\mathrm{e}^{-\mu_j (A_n + W_n)} \mid Z_n = i] \Big) \\
&= \sum_{i=1}^{M} p_{i,j} \Big(\mathbb{P}[Z_n = i] - \frac{s}{\mu_j + s}\, \omega_i^n(\mu_j)\, \alpha_i(\mu_j) \Big).
\end{aligned}$$

So for $n \to \infty$ we have that $\omega_j(s)$ is given by

$$\omega_j(s) = \varpi_j - \sum_{i=1}^{M} p_{i,j}\, \omega_i(\mu_j)\, \alpha_i(\mu_j) + \frac{\mu_j}{\mu_j + s} \sum_{i=1}^{M} p_{i,j}\, \omega_i(\mu_j)\, \alpha_i(\mu_j). \tag{7.3}$$

Define the constants

$$c_j = \sum_{i=1}^{M} p_{i,j}\, \omega_i(\mu_j)\, \alpha_i(\mu_j).$$

Inverting the Laplace transform ω_j yields that the density of the waiting time is given by

$$f_{W,j}(x) = \mu_j c_j\, \mathrm{e}^{-\mu_j x},$$

and the corresponding distribution has mass $\pi_{0,j} = \varpi_j - c_j$ at the origin. For $i, j = 1, \ldots, M$, the M^2 unknown constants $\omega_i(\mu_j)$ that are needed in order to determine the unknown constants c_j are the unique solution to the system of linear equations given by the expression

$$\omega_j(\mu_\ell) = \varpi_j - \frac{\mu_\ell}{\mu_j + \mu_\ell} \sum_{i=1}^{M} p_{i,j}\, \omega_i(\mu_j)\, \alpha_i(\mu_j), \qquad j, \ell = 1, \ldots, M; \tag{7.4}$$

see Equation (7.3). The uniqueness of the solution follows from the general theory of Markov chains that there is a unique stationary distribution and thus also a unique solution to the system of equations formed by (7.4) for all $j, \ell = 1, \ldots, M$. □

The result given in Theorem 7.1 is expected. Evidently, since B_n is exponentially distributed (with a rate depending on the state of the Markov chain) and W_n is the residual preparation time, we have that for every state j of the Markov chain, the waiting-time distribution has mass at zero and the conditional waiting time is exponentially distributed with rate μ_j.

Observe that Theorem 7.1 reduces to the statement of Theorem 4.8 which gives the steady-state waiting-time density in case $\{A_n\}$ and $\{B_n\}$ are mutually independent sequences of i.i.d. random variables and B follows an Erlang distribution. Specifically, if the Markov chain in Theorem 7.1 has only one state (and thus there is a unique service-time distribution and a unique rate μ for the exponentially distributed preparation times) and the Erlang distribution F_B in Theorem 4.8 has only one phase (thus $N = 1$, which implies that F_B is exponentially distributed with rate μ) then the statements of these two theorems are identical. Observe, for example, that (7.4) reduces to (4.27) as now $\varpi_j = 1$, $p_{i,j} = 1$, and $\mu_j = \mu_\ell = \mu$.

Moreover, we see that the proofs of both theorems are quite similar. For the Markov-modulated case, the only additional effort we need to make is to keep track of the state of the Markov chain.

7.2.2 Phase-type preparation times

Since for exponentially distributed preparation times, the case where the distributions of A_n and B_n depend on the state of a Markov chain is so similar to the independent case, it is not surprising that, as before, we can extend the analysis of the Markov-modulated dependencies to phase-type distributions for the preparation times.

Assume now that if the Markov chain is in state j, the preparation time is with probability κ_n equal to a random variable Y_n, $n = 1, \ldots, N$, that follows an Erlang distribution with parameter μ_j and n phases. In other words the distribution function of B is given by (cf. (4.32))

$$F_{B,j}(x) = \sum_{n=1}^{N} \kappa_n \left(1 - \mathrm{e}^{-\mu_j x} \sum_{\ell=0}^{n-1} \frac{(\mu_j x)^\ell}{\ell!}\right), \qquad x \geqslant 0. \tag{7.5}$$

As remarked before, this class of phase-type distributions may be used to approximate any given distribution on $[0, \infty)$ for the preparation times arbitrarily close; see Schassberger [149]. The waiting-time density for this case is given by the following theorem.

Theorem 7.2. *Under the conditions of Theorem 7.1, with the modification that $F_{B,j}$ is now given by (7.5) we have that, in equilibrium, the waiting time has mass*

$$\varpi_j - \sum_{i=1}^{M} \sum_{n=1}^{N} \sum_{\ell=0}^{n-1} \sum_{m=0}^{\ell} \kappa_n p_{i,j} \frac{\mu_j^\ell}{\ell!} \binom{\ell}{m} (-1)^\ell \big(\alpha_i(\mu_j)\big)^{(\ell-m)} \big(\omega_i(\mu_j)\big)^{(m)}$$

at the origin, and density given by

$$f_{W,j}(x) = \sum_{i=1}^{M} \sum_{n=1}^{N} \sum_{\ell=0}^{n-1} \sum_{m=0}^{\ell} \kappa_n p_{i,j} \frac{(-1)^\ell}{\ell!} \binom{\ell}{m} \big(\alpha_i(\mu_j)\big)^{(\ell-m)} \times \\ \times \big(\omega_i(\mu_j)\big)^{(m)} \mu_j^n \mathrm{e}^{-\mu_j x} \frac{x^{n-\ell-1}}{(n-\ell-1)!}.$$

Proof. For the proof, we shall refrain from presenting detailed computations, as the analysis is straightforward and similar to the one for the exponential case. We give, however, a few intermediate formulas. From (7.2) and for the preparation time distributions we are considering we have that

$$\omega_j^{n+1}(s) = \sum_{i=1}^{M} \mathbb{P}[Z_n = i] p_{i,j} \Bigg(1 - \\ - \sum_{n=1}^{N} \kappa_n \mathbb{E}[\mathrm{e}^{-\mu_j (A_n + W_n)} \sum_{\ell=0}^{n-1} \frac{\mu_j^\ell (A_n + W_n)^\ell}{\ell!} \mid Z_n = i] + \\ + \sum_{n=1}^{N} \kappa_n \left(\frac{\mu_j}{\mu_j + s}\right)^n \mathbb{E}[\mathrm{e}^{-\mu_j (A_n + W_n)} \sum_{\ell=0}^{n-1} \frac{(\mu_j + s)^\ell (A_n + W_n)^\ell}{\ell!} \mid Z_n = i] \Bigg).$$

So for $n \to \infty$ we have that $\omega_j(s)$ is given by (cf. (7.3))

$$\omega_j(s) = u_j + \sum_{i=1}^{M}\sum_{n=1}^{N}\sum_{\ell=0}^{n-1}\sum_{m=0}^{\ell} \kappa_n p_{i,j} \frac{(-\mu_j)^\ell}{\ell!}\binom{\ell}{m}\big(\alpha_i(\mu_j)\big)^{(\ell-m)} \times \\ \times \big(\omega_i(\mu_j)\big)^{(m)}\Big(\frac{\mu_j}{\mu_j+s}\Big)^{n-\ell}, \quad (7.6)$$

where

$$u_j = \varpi_j - \sum_{i=1}^{M}\sum_{n=1}^{N}\sum_{\ell=0}^{n-1}\sum_{m=0}^{\ell} \kappa_n p_{i,j} \frac{\mu_j^\ell}{\ell!}\binom{\ell}{m}(-1)^\ell\big(\alpha_i(\mu_j)\big)^{(\ell-m)}\big(\omega_i(\mu_j)\big)^{(m)}.$$

Inverting the Laplace transform ω_j yields that the density of the waiting time is given by the expression presented in the theorem and the corresponding distribution has mass u_j at the origin. For $i, j = 1, \ldots, M$, the M^2 unknown constants $\omega_i(\mu_j)$ are the unique solution to the system of linear equations resulting from substituting s for μ_k, $k = 1, \ldots, M$, in (7.6). □

Naturally, the expression for the density appearing in Theorem 7.2 reduces for $N = 1$ to the expression given in Theorem 7.1, and for $M = 1$ it reduces to the expression in Theorem 4.10, which gives the waiting-time density for the independent case.

7.3 Services depending on the previous preparation time

In this section, we study the second dependence structure mentioned in the introduction. We assume that for all n, the service times A_n are distributed as A, which in turn is exponentially distributed with rate λ. Moreover, the Laplace-Stieltjes transform of the preparation time B_{n+1}, given that the previous service time A_n equals t, is of the form

$$\mathbb{E}[\mathrm{e}^{-sB_{n+1}} \mid A_n = t] = \mathbb{E}[\mathrm{e}^{-sB} \mid A = t] = \chi(s)\mathrm{e}^{-\psi(s)t}. \quad (7.7)$$

Observe that now the preparation time B_{n+1} depends only on the previous service time, while in the Markov-modulated case we have examined previously all preparation and service times are correlated between and among one another, since their distributions depend on a common Markov chain.

For the above dependence structure, we further assume that χ and ψ are rational functions; i.e.,

$$\chi(s) = \frac{P_1(s)}{Q_1(s)} \qquad \text{and} \qquad \psi(s) = \frac{P_2(s)}{Q_2(s)}, \quad (7.8)$$

where Q_1 and Q_2 are polynomials of degrees M and N respectively, and P_1 and P_2 are polynomials of degrees less than M and less than N respectively. From the form of Equation (7.7) we see that a number of other assumptions have been implicitly

made. For example, since for $s = 0$ the expectation $\mathbb{E}[\mathrm{e}^{-sB} \mid A = t]$ should be equal to one, we have implicitly assumed that $\psi(0) = 0$ and $\chi(0) = 1$. We shall mention other implications of such type only when necessary.

The preparation time B_{n+1} consists of two parts: a component which depends on the previous service time, represented by $\mathrm{e}^{-\psi(s)t}$, and an 'ordinary' preparation time with Laplace-Stieltjes transform $\chi(s)$, which does not depend on the interarrival time. From (7.7) we have that the bivariate Laplace-Stieltjes transform of the generic preparation and service time is given by

$$\mathbb{E}[\mathrm{e}^{-sB-zA}] = \int_0^\infty \lambda \mathrm{e}^{-\lambda t}\mathrm{e}^{-zt}\chi(s)\mathrm{e}^{-\psi(s)t}\,\mathrm{d}t = \frac{\lambda\chi(s)}{\lambda + \psi(s) + z}, \tag{7.9}$$

for $\mathrm{Re}(\lambda + \psi(s) + z) > 0$. This expression leads to

$$\mathbb{E}[B] = \frac{\psi'(0)}{\lambda} - \chi'(0), \quad \text{and} \quad \mathbb{E}[AB] = \frac{2\psi'(0) - \lambda\chi'(0)}{\lambda^2},$$

from which we have that the covariance function between a preparation time and a service time is given by

$$\mathrm{cov}[A, B] = \frac{\psi'(0)}{\lambda^2}.$$

The correlation between these two variables can be also computed, see Boxma and Combé [33]. Thus, given the covariance (or correlation) between A and B, one can construct a distribution function F_B that has the desired effect.

In order to derive the steady-state waiting-time distribution, we shall first derive the Laplace-Stieltjes transform of F_W. We follow a method based on Wiener-Hopf decomposition. A straightforward calculation, starting from (1.2), yields that

$$\begin{aligned}
\omega(s) &= \mathbb{E}[\mathrm{e}^{-sW}] \\
&= \mathbb{P}[B \leqslant W + A] + \mathbb{E}[\mathrm{e}^{-s(B-W-A)}] - \mathbb{E}[\mathrm{e}^{-s(B-W-A)}\,; B \leqslant W + A] \\
&= \mathbb{P}[B \leqslant W + A] + \mathbb{E}[\mathrm{e}^{sW}]\,\mathbb{E}[\mathrm{e}^{-s(B-A)}] - \mathbb{E}[\mathrm{e}^{-s(B-W-A)}\,; B \leqslant W + A],
\end{aligned}$$

since $B - A$ and W are independent. Therefore, from (7.9) we have for $\mathrm{Re}(s) = 0$ that

$$\begin{aligned}
\omega(s) &= \mathbb{P}[B \leqslant W + A] + \omega(-s)\frac{\lambda\chi(s)}{\lambda - s + \psi(s)} - \\
&\qquad - \mathbb{E}[\mathrm{e}^{-s(B-W-A)} \mid B \leqslant W + A]\mathbb{P}[B \leqslant W + A] \\
&= \omega(-s)\frac{P_1(s)}{Q_1(s)}\frac{\lambda Q_2(s)}{(\lambda - s)Q_2(s) + P_2(s)} + \\
&\qquad + \mathbb{P}[B \leqslant W + A]\big(1 - \mathbb{E}[\mathrm{e}^{-s(B-W-A)} \mid B \leqslant W + A]\big),
\end{aligned} \tag{7.10}$$

which can be rewritten as

$$\begin{aligned}
\omega(s)Q_1(s)\big((\lambda - s)Q_2(s) + P_2(s)\big) = \\
\lambda\omega(-s)P_1(s)Q_2(s) + Q_1(s)\big((\lambda - s)Q_2(s) + P_2(s)\big)\times \\
\times\, \mathbb{P}[B \leqslant W + A]\big(1 - \mathbb{E}[\mathrm{e}^{-s(B-W-A)} \mid B \leqslant W + A]\big).
\end{aligned} \tag{7.11}$$

We can observe that $Q_1(s)\big((\lambda-s)Q_2(s)+P_2(s)\big)$ is a polynomial of degree $M+N+1$ and also that the left-hand side of (7.11) is analytic for $\mathrm{Re}(s) > 0$ and continuous for $\mathrm{Re}(s) \geqslant 0$, and the right-hand side of (7.11) is analytic for $\mathrm{Re}(s) < 0$ and continuous for $\mathrm{Re}(s) \leqslant 0$. Liouville's theorem [162] states that:

> If a function $f(z)$ is analytic for all finite values of z, and as $|z| \to \infty$ we have that $f(z) = O(|z|^k)$, then $f(z)$ is a polynomial of degree less than or equal to k.

Therefore, from Liouville's theorem we conclude that both sides of (7.11) are the same $M + N + 1$-st degree polynomial, say, $\sum_{i=0}^{M+N+1} q_i s^i$. Hence,

$$\omega(s) = \frac{\sum_{i=0}^{M+N+1} q_i s^i}{Q_1(s)\big((\lambda - s)Q_2(s) + P_2(s)\big)}. \tag{7.12}$$

In the expression above, the constants q_i are not determined so far. In order to obtain the transform, observe that ω is a fraction of two polynomials both of degree $M + N + 1$. Let r_i, $i = 1, \ldots, M + N + 1$, be the zeros of the denominator. Ignoring the special case of zeros with multiplicity greater than one, partial fraction decomposition yields that (7.12) can be rewritten as

$$\omega(s) = c_0 + \sum_{i=1}^{M+N+1} \frac{c_i}{s - r_i}, \tag{7.13}$$

which implies that the waiting-time distribution has a mass at the origin that is given by

$$\mathbb{P}[W = 0] = \lim_{s \to \infty} \mathbb{E}[\mathrm{e}^{-sW}] = c_0$$

and has a density that is given by

$$f_W(x) = \sum_{i=1}^{M+N+1} c_i \mathrm{e}^{r_i x}. \tag{7.14}$$

All that remains is to determine the $M + N + 2$ constants c_i. To do so, we work as follows. We express the terms $\mathbb{P}[B \leqslant W + A]$ and $\mathbb{E}[\mathrm{e}^{-s(B-W-A)} \mid B \leqslant W + A]$ that appear at the right-hand side of (7.11) in terms of the constants c_i by using (7.13) and (7.14). Then, we substitute these expressions and (7.13) in the left-hand side of (7.11). Thus we obtain a new equation in terms of the constants c_i that we shall differentiate a total of $M + N + 1$ times. We shall evaluate each of these derivatives for $s = 0$ and thus we obtain a linear system of $M + N + 1$ equations for the constants c_i, $i = 0, \ldots, M + N + 1$. The last equation that is necessary to uniquely determine the constants c_i is the normalisation equation

$$c_0 + \int_0^\infty f_W(x)\,\mathrm{d}x = 1.$$

We summarise the above in the following theorem.

Theorem 7.3. *Let the service time A be exponentially distributed with rate λ. Under the assumptions described by Equations (7.7) and (7.8), we have that the limiting distribution of the waiting time has mass c_0 at the origin and a density on $[0, \infty)$ that is given by*

$$f_W(x) = \sum_{i=1}^{M+N+1} c_i \mathrm{e}^{r_i x}.$$

in the expression above, the constants r_i are the $M + N + 1$ zeros of the equation

$$Q_1(s)\big((\lambda - s)Q_2(s) + P_2(s)\big) = 0$$

and the coefficients c_i are derived as described above.

Remark 7.1. Although the roots r_i and coefficients c_i may be complex, the density and the mass c_0 at zero will be positive. This follows from the fact that there is a unique equilibrium distribution and thus a unique solution to the linear system for the coefficients c_i. Of course, it is also clear that each root r_i and coefficient c_i have a companion conjugate root and conjugate coefficient, which implies that the imaginary parts appearing in the density cancel.

Remark 7.2. When $Q_1(s)\big((\lambda - s)Q_2(s) + P_2(s)\big)$ has multiple zeros, the analysis proceeds in essentially the same way. For example, if $r_1 = r_2$, then the partial fraction decomposition of ω becomes

$$\omega(s) = c_0 + \frac{c_1}{(s - r_1)^2} + \sum_{i=2}^{M+N+1} \frac{c_i}{s - r_i},$$

the inverse of which is given by

$$f_W(x) = c_1 x \mathrm{e}^{r_1 x} + \sum_{i=2}^{M+N+1} c_i \mathrm{e}^{r_i x}.$$

Remark 7.3. For the service time A we have considered only the exponential distribution, mainly because we can illustrate the technique we use without complicating the analysis. However, we can extend this class by considering distributions with a mixed-Erlang distribution of the form of Equation (3.23) and the proof remains essentially the same. The resulting density of the waiting time is again of the form

$$f_W(x) = \sum_{i=1}^{K} c_i \mathrm{e}^{r_i x}.$$

In the expression above, we have that $K = M + n(N + 1)$, where n is the number of phases of the Erlang distribution with the most phases that A follows with a certain probability. The constants r_i are the zeros to the equation

$$Q_1(s)\big((\lambda - s)Q_2(s) + P_2(s)\big)^n = 0,$$

and the coefficients c_i are determined in a similar fashion as before. Naturally, if any of the zeros r_i has multiplicity greater than one, the form the the density changes analogously; see also the previous remark.

We now present a few examples where we show how some classic dependence structures fit into this class. In the examples below, we only discuss how to derive the function ψ.

The independent case

The dependence structure described by Equation (7.7) includes a great variety of dependence structures, including the independent case. If for all s we have that $\psi(s) = 0$, then the Laplace-Stieltjes transform of B is independent of the length of the service time, and thus the function χ appearing in (7.7) is in fact the Laplace-Stieltjes transform of B, i.e. the function β. Observe that we have assumed that B has a rational Laplace-Stieltjes transform, which is necessary in order to decompose Equation (7.10) into functions that are analytic either at the left-half plane or at the right-half plane; see also Equation (1.5).

Linear Dependence

Assume that the service time A and the preparation time B are linearly dependent; that is, $B = cA$. Then,

$$\mathbb{E}[\mathrm{e}^{-sB} \mid A = t] = \mathbb{E}[\mathrm{e}^{-scA} \mid A = t] = \mathrm{e}^{-sct}.$$

Thus we have that $\chi(s) = 1$, and $\psi(s) = cs$, and both functions satisfy our assumptions.

The Compound Poisson Process

In this case we assume that given that $A = t$, the preparation time B is equal to $\sum_{i=1}^{N(t)} C_i$, where $N(t)$ is a Poisson process with rate γ, and $\{C_i\}$ is a sequence of i.i.d. random variables, where each of them is distributed like C, and where C has a rational Laplace-Stieltjes transform. Under this assumption, we have that

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{-sB} \mid A = t] = \mathbb{E}[\mathrm{e}^{-s\sum_{i=1}^{N(t)} C_i} \mid A = t] &= \sum_{k=0}^{\infty} \mathbb{E}[\mathrm{e}^{-s\sum_{i=1}^{k} C_i} \mid A = t]\, \mathrm{e}^{-\gamma t} \frac{(\gamma t)^k}{k!} \\ &= \sum_{k=0}^{\infty} \big(\mathbb{E}[\mathrm{e}^{-sC}]\big)^k \mathrm{e}^{-\gamma t} \frac{(\gamma t)^k}{k!} = \mathrm{e}^{-\psi(s)t},\end{aligned}$$

where $\psi(s) = \gamma\big(1 - \mathbb{E}[\mathrm{e}^{-sC}]\big)$. As before, in this case we have that $\chi(s) = 1$.

Brownian Motion

In this case we assume that given that $A = t$, the random variable B is normally distributed with mean μt and variance $\sigma^2 t$. Then we have that

$$\mathbb{E}[\mathrm{e}^{-sB} \mid A = t] = \int_{-\infty}^{\infty} \mathrm{e}^{-sx} \frac{\mathrm{e}^{-(x-\mu t)^2/(2\sigma^2 t)}}{\sigma\sqrt{2\pi t}}\, \mathrm{d}x = \mathrm{e}^{-\psi(s)t},$$

where $\psi(s) = \mu s - s^2\sigma^2/2$, and $\chi(s) = 1$. Naturally, if B is interpreted as the preparation time of a customer in the models described in Section 1.2, then assuming that B is normally distributed is not a natural assumption, since the preparation time of a customer is non-negative. However, in the analysis we do not need the condition of B being non-negative; therefore, it is mathematically possible to consider this case. For further examples of distributions satisfying the condition described by (7.7), see Boxma and Combé [33].

7.4 A comparison to Lindley's recursion

Models with dependencies between interarrival and service time have been studied by several authors. A review of the early literature can be found in Bhat [18]. Such dependencies arise naturally in various applications. For example, the phenomenon of dependence among the interarrival times in the packet streams of voice and data traffic is well known; see, e.g., [82, 83, 154]. However, in [64] the authors argue that in packet communication networks one should also expect two additional forms of dependence: between successive service times and among interarrival times and service times. These forms of dependence occur because of the presence of bursty arrivals and multiple sources with different mean service times (due to different packet lengths), and they may have a dominant effect on waiting times and queue lengths. In the following, we give an overview of some results derived for Lindley's equation that are related to the results presented in this chapter.

Markov-modulated dependencies for Lindley's recursion

Dependence structures of the form of Equation (7.1), and several generalisations, have been studied extensively for Lindley's recursion. The basic model is described in Adan and Kulkarni [1]. The authors study a single-server queue where the interarrival times and the service times depend on a common discrete-time Markov chain in a similar way as the one described by Equation (7.1). This model generalises the well-known MAP/G/1 queue by allowing dependencies between interarrival and service times. The waiting-time process is directly analysed in a similar method to the one described in this chapter, thus deriving the Laplace-Stieltjes transform of the steady-state waiting-time distribution. By exploiting a well-known relation between the waiting time of a customer and the number of customers left behind by a departing customer, they also derive the Laplace-Stieltjes transform of the queue length distribution at departure epochs and at arbitrary time points. The process analysed in [1] is a special case of the class of processes considered in Asmussen and Kella [7], where the results for the Markov-modulated M/G/1 queue have already been sketched. In [1], however, all results are given explicitly and the analysis extends to the study of the queue length distribution.

Although the analysis for Lindley's case in [1] is quite similar to the one followed in Section 7.2, the resulting Laplace transform for the waiting-time density is rather more complicated and it does not seem straightforward to invert it directly. Moreover, in order to determine the quantities analogous to the coefficients c_i appearing

in Theorem 7.1, one needs to study the solutions of a certain equation –a step which is not required for this model.

In the literature much attention has been devoted to single-server queues with Markovian Arrival Processes (MAP), see, e.g., [123, and the references therein]. The MAP/G/1 queue provides a powerful framework to model dependencies between successive interarrival times, but typically the service times are assumed to be i.i.d. and independent of the arrival process. The model considered [1] is a generalisation of the MAP/G/1 queue, by also allowing dependencies between successive service times and between interarrival times and service times.

In [123] the MAP construction is generalised to the Batch Markovian Arrival Process (BMAP) to allow batch arrivals. The framework of the BMAP/G/1 queue can also be used to model dependence between interarrival times and service times. This is described in [49]. The important observation is that the arrival of a batch (the size of which depends on the state of the underlying Markov chain) can be viewed as the arrival of a super customer, whose service time is distributed as the sum of the service requests of the customers in the batch.

A special case of Markov-dependent interarrival and service times is the model with strictly periodic arrivals. These models arise, for example, in the modelling of inventory systems using periodic ordering policies; see [152] and [175]. Queueing models with periodic arrival processes have also been studied extensively; see for example [47, 110, 144, 145].

Markov-modulated dependencies of this form have also been considered in insurance mathematics. For example, Albrecher and Boxma [3] consider the same semi-Markovian dependence structure for W_n, X_n and Z_n, where now W_i denotes the interarrival time between two claims, X_n is the size of the n-th claim and $\{Z_n\}$ is the regulating Markov chain. This work unifies and generalises various other models considered in insurance mathematics, see [11, 68, 112, 113].

Dependencies of the form of (7.7) for Lindley's recursion

The dependence structure (7.7) that we have presented occurs in simple queuing models. Consider the following situation. Work arrives at a single server queue according to a process with stationary, non-negative independent increments. This work, however, does not enter immediately the queue of the server facility; instead it is accumulated behind a gate. At exponential interarivals the gate is opened and – after the addition of an independent component – the work is collected and delivered as a single customer at the queue of the service facility. The additional component may be viewed as a set-up time.

Due to the exponentially distributed interarrival times of customers, we can view the service facility as an M/G/1 queue in which the interarrival and service time for each customer are positively correlated. Indeed, if the interval between two consecutive openings of the gate is relatively long (short), it is likely that a relatively large (small) amount of work has accumulated during that interval. This model is a unification and generalisation of the M/G/1 queue with a positive correlation between interarrival and service times [30, 31, 44, 51] and has been analysed in

Boxma and Combé [33], which we closely followed in Section 7.3. In Combé and Boxma [49] it is shown that the collect system can also be modelled by using the BMAP framework.

Models with a linear dependence between the service time and the preceding interarrival time have been studied in [44, 50, 52]; see also [35]. Other papers [51, 75, 76, 131] analyse the M/M/1 queue where the service time and the preceding interarrival time have a bivariate exponential density with a positive correlation. The linear and bivariate exponential cases are both contained in the correlated M/G/1 queue studied by [30, 31, 33].

In particular, in our notation, the goal in [33] is to extend the analysis of [30, 31, 44, 51] to M/G/1 queues with arrival rate λ in which the Laplace-Stieltjes transform of the service time B_{n+1}, given that the interarrival time A_n equals t is of the form of (7.7). The authors give a few examples and show how previously analysed models for the M/G/1 queues with dependence fit into this structure. They analyse the joint distribution of the waiting and service time of an arbitrary customer in steady state, and they also present a vacation-type workload decomposition for the M/G/1 queue with the exponential gating mechanism described above.

One additional assumption made in [33] is that the function ψ appearing in (7.7) has a completely monotone increasing derivative. This implies that the function $\mathrm{e}^{-\psi(s)}$ is the Laplace-Stieltjes transform of an infinitely divisible probability distribution, which in its turn is a distribution of increments in processes with stationary independent increments, i.e. Lévy processes. Although the monotonicity of the derivative of ψ was not required for our case, most of the examples presented in Section 7.3 are examples of Lévy processes.

As a final remark we should add that various other types of dependencies can be analysed numerically, if not analytically. As an example of a dependence structure arising from a specific application, consider the carousel model described in Section 1.2 with the modification that each pick order requires now more than one item. Under the service time A of an order, we understand both the actual time to pick all items and the time the carousel spends in rotating from the moment we complete the first pick until the moment we are about to start the last pick of an order. The preparation time B in this case is the time needed until the carousel rotates to a specified point before the first pick takes place. For example, we can take as preparation strategy the following: upon arrival of a new order to a carousel, the carousel rotates to the nearest item to the origin (and waits there, should the server still be occupied with the previous order). We see that a relatively long preparation time implies that all items of the order are placed in a relatively short interval. Thus, for every n, the random variables A_n and B_n depend on one another and are negatively correlated. For a given strategy, one can compute the distribution of B given the length of A. However, the exact computation of the waiting-time distribution is usually cumbersome. Naturally, if the waiting-time distribution is the fixed point of a contraction mapping, one can numerically approximate F_W by successive iterations.

Chapter 8
A more general Lindley-type recursion

8.1 Introduction

In the previous chapters we have studied various aspects of the stochastic recursion (1.1), mainly focusing on the derivation of the steady-state waiting-time distribution. In this chapter, we generalise some of the results derived so far by considering the Lindley-type recursion

$$W_{n+1} = \max\{0, B_n - A_n + Y_n W_n\}, \tag{8.1}$$

where for every n, the random variable Y_n is equal to plus or minus one according to the probabilities $\mathbb{P}[Y_n = 1] = p$ and $\mathbb{P}[Y_n = -1] = 1 - p$, $0 \leqslant p \leqslant 1$. Recursion (8.1) reduces to the classical Lindley recursion [115] when $\mathbb{P}[Y_n = 1] = 1$ for every n; see Section 1.5. Furthermore, if $\mathbb{P}[Y_n = -1] = 1$, then (8.1) reduces to the recursion studied in Chapters 2–7. Evidently, this recursion, like Recursion (1.1), is also not monotone increasing, while this is usually assumed for generalisations of Lindley's recursion; see again Section 1.5. The material presented in this chapter is based on work already carried out in [36].

Recursion (8.1) is a special case of more general recursions that have been studied in the literature, see for example Diaconis and Freedman [56], Goldie [71], Borovkov [27], and references cited in these studies. Studying a recursion that contains both Lindley's classical recursion and Recursion (1.1) as special cases seems of interest in its own right. Additional motivation for studying this recursion is supplied by the fact that, for $0 < p < 1$, the resulting model can be interpreted as a special case of a queuing model in which service and interarrival times depend on waiting times. We shall now discuss the latter model.

Consider an extension of the standard G/G/1 queue in which the service times and the interarrival times depend linearly and randomly on the waiting times. Namely, the model is specified by a stationary and ergodic sequence of four-tuples of non-negative random variables $\{(A_n, B_n, \widehat{A}_n, \widehat{B}_n)\}$, $n \geqslant 0$. The sequence $\{W_n\}$ is defined recursively by

$$W_{n+1} = \max\{0, \overline{B}_n - \overline{A}_n + W_n\},$$

where

$$\overline{A}_n = A_n + \widehat{A}_n W_n,$$
$$\overline{B}_n = B_n + \widehat{B}_n W_n.$$

We interpret W_n as the waiting time and $\overline{B}_n$ as the service time of customer n. Furthermore, we take $\overline{A}_n$ to be the interarrival time between customers n and $n+1$.

We call B_n the *nominal service time* of customer n and A_n the *nominal interarrival time* between customers n and $n+1$, because these would be the actual times if the additional shift were omitted, that is, if $\mathbb{P}[\widehat{A}_n = \widehat{B}_n = 0] = 1$. Evidently, the waiting times satisfy the generalised Lindley recursion (8.1), where we have written $Y_n = 1 + \widehat{B}_n - \widehat{A}_n$.

Queues with state-dependent service and arrival processes have been studied extensively; see for example Brill [41], Callahan [43], Harris [78, 79], Laslett *et al.* [108], Mudrov [133], Posner [141], Rosenshine [146], and Sugawara and Takahashi [156]. Such queues arise, for example, whenever the first customer in a busy period requires a setup time from the server. This situation may arise among bank tellers, machine repairmen, hospital emergency teams, copying systems, computer terminals, iron and steel processing [156], etc. For various other applications and further results see also Brill and Posner [42], and references therein.

This model – for generally distributed random variables Y_n – has been introduced in Whitt [181], where the focus is on conditions for the process to converge to a proper steady-state limit, and on approximations for this limit. Whitt [181] builds upon previous results by Vervaat [167] and Brandt [37] for the unrestricted recursion $W_{n+1} = Y_n W_n + X_n$, where $X_n = B_n - A_n$. There has been considerable previous work on this unrestricted recursion, due to its close connection to the problem of the ruin of an insurer who is exposed to a stochastic economic environment. Such an environment has two kinds of risk, which were called by Norberg [136] insurance risk and financial risk. Indicatively, we mention the work by Tang and Tsitsiashvili [159], and by Kalashnikov and Norberg [97]. In the more general framework, W_n may represent an inventory in time period n (e.g. cash), Y_n may represent a multiplicative, possibly random, decay or growth factor between times n and $n+1$ (e.g. interest rate) and $B_n - A_n$ may represent a quantity that is added or subtracted between times n and $n+1$ (e.g. deposit minus withdrawal). Obviously, the positive-part operator is appropriate for many applications [181].

In this chapter, we present an exact analysis of the steady-state distribution of $\{W_n\}$, $n = 1, 2, \ldots$, as given by (8.1) with $\mathbb{P}[Y_n = 1] = p$ and $\mathbb{P}[Y_n = -1] = 1 - p$. For $0 < p < 1$, this amounts to analysing the above-described G/G/1 extension where $\widehat{A}_n = \widehat{B}_n$ with probability p, and $\widehat{A}_n = 2 + \widehat{B}_n$ with probability $1 - p$. This problem, and state-dependent queuing processes in general, is connected to LaPalice queuing models, introduced by Jacquet [92], where customers are scheduled in such a way that the period between two consecutively scheduled customers is greater than or equal to the service time of the first customer.

This chapter is organised in the following way. In Section 8.2 we comment on the stability of the process $\{W_n\}$, as it is defined by Recursion (8.1). In the remainder of the chapter it is assumed that the steady-state distribution of $\{W_n\}$ exists. Section 8.3 is devoted to the determination of the distribution of W when A is generally distributed and B has a phase-type distribution. In Section 8.4 we determine the distribution of W when A is exponentially distributed and B is deterministic. At the end of each section we compare the results that we derive to the already known results for Lindley's recursion (i.e. for $p = 1$) and to the equivalent results for the Lindley-type recursion arising for $p = 0$, that is, for Recursion (1.1).

The notation used here is identical to the conventions made so far. As before, for a random variable X we denote its distribution by F_X and its density by f_X. Furthermore, we denote by $f^{(i)}$ the i-th derivative of the function f. The Laplace-Stieltjes transforms of F_A and F_W are respectively denoted by α and ω. Moreover, we also use the function ϕ defined as $\phi(s) = \omega(s)\,\alpha(s)$. Finally, in order to distinguish between the various special cases of this model, we shall use Kendall's adapted terminology that we have already established in Section 1.5. Thus, e.g., M/G refers now to the model described by Recursion (8.1), unless otherwise stated.

8.2 Stability

The following result on the convergence of the process $\{W_n\}$ to a proper limit W is shown in Whitt [181]. It is included here only for completeness.

From Recursion (8.1), it is obvious that if we replace Y_n by $\max\{0, Y_n\}$ and $B_n - A_n$ by $\max\{0, B_n - A_n\}$, then the resulting waiting times will be at least as large as the ones given by (8.1). Moreover, when we make this change, the positive-part operator is not necessary anymore.

Lemma 8.1 (Whitt [181, Lemma 1]). *If W_n satisfies (8.1), then with probability 1, $W_n \leqslant Z_n$ for all n, where*

$$Z_{n+1} = \max\{0, Y_n\}Z_n + \max\{0, B_n - A_n\}, \quad n \geqslant 0, \tag{8.2}$$

and $Z_0 = W_0 \geqslant 0$.

So if W_n satisfies (8.1), Z_n satisfies (8.2), and Z_n converges to the proper limit Z, then $\{W_n\}$ is tight and $\mathbb{P}[W > x] \leqslant \mathbb{P}[Z > x]$ for all x, where W is the limit in distribution of any convergent subsequence of $\{W_n\}$. This observation, combined with Theorem 1 of Brandt [37], which implies that Z_n satisfying (8.2) converges to a proper limit if $\mathbb{P}[\max\{0, Y_n\} = 0] = \mathbb{P}[Y_n \leqslant 0] > 0$, leads to the following theorem.

Theorem 8.1 (Whitt [181, Theorem 1]). *The series $\{W_n\}$ is tight for all values of $\rho = \mathbb{E}[B_0]/\mathbb{E}[A_0]$ and W_0. If, in addition, $0 \leqslant p < 1$ and $\{(Y_n, B_n - A_n)\}$ is a sequence of independent vectors with*

$$\mathbb{P}[Y_0 \leqslant 0, B_0 - A_0 \leqslant 0] > 0,$$

then the events $\{W_n = 0\}$ are regeneration points with finite mean time and $\{W_n\}$ converges in distribution to a proper limit W as $n \to \infty$ for all ρ and W_0.

Naturally, for $p = 1$, i.e. for the classical Lindley recursion, we need the additional condition that $\rho < 1$.

Therefore, assume that the sequences $\widehat{B}_n - \widehat{A}_n$ and $B_n - A_n$ are independent stationary sequences, that are also independent of one another, and that for all n, A_n and B_n are non-negative. Then the conditions of Theorem 8.1 hold, so there exists a proper limit W, and for the system in steady-state we write

$$W \stackrel{\mathcal{D}}{=} \max\{0, B - A + YW\}, \tag{8.3}$$

where "$\stackrel{\mathcal{D}}{=}$" denotes equality in distribution, where A, B are generic random variables distributed like A_n, B_n, and where $\mathbb{P}[Y=1]=p$ and $\mathbb{P}[Y=-1]=1-p$.

The sequences $\{A_n\}$ and $\{B_n\}$ are assumed to be independent sequences of i.i.d. non-negative random variables. In the next two sections, our main goal is to determine the distribution of W for the two cases when A is generally distributed and B follows a phase-type distribution, and when A is exponentially distributed and B deterministic.

Since this model behaves like an ordinary G/G/1 queue when $p=1$, and like an alternating service model when $p=0$, it is reasonable to assume that a way to proceed with the analysis is by choosing a method that works well for both special cases. For the G/PH case, the method we follow is based on a Wiener-Hopf decomposition. Namely, we derive an expression for the Laplace-Stieltjes transform of the distribution F_W and we manipulate this expression accordingly until we can determine two expressions for the same function involving the transform, that are both defined on the imaginary axis of the complex plane and approach their values there continuously. Then, by using standard theorems from calculus, we shall be able to determine ω. For the M/D case, the analysis combines elements from the analysis we have seen in Section 3.6 and from the analysis of the M/D/1 queue.

8.3 The G/PH model

In this section we assume that A is generally distributed, while B follows the mixed-Erlang distribution we have encountered before; see for example Section 3.5 and Section 4.2.4. Specifically, we assume that with probability κ_n the nominal service time B follows an Erlang distribution with parameter μ and n phases, i.e.,

$$F_B(x)=\sum_{n=1}^{N}\kappa_n\left(1-\mathrm{e}^{-\mu x}\sum_{j=0}^{n-1}\frac{(\mu x)^j}{j!}\right)=\sum_{n=1}^{N}\kappa_n\sum_{j=n}^{\infty}\mathrm{e}^{-\mu x}\frac{(\mu x)^j}{j!},\qquad x\geqslant 0,\quad (8.4)$$

with Laplace-Stieltjes transform $\sum_{n=1}^{N}\kappa_n\big(\mu/(\mu+s)\big)^n$. These distributions, i.e. mixtures of Erlang distributions, are special cases of Coxian or phase-type distributions. It is sufficient to consider only this class, since it may be used to approximate any given continuous distribution on $[0,\infty)$ arbitrarily close; see Schassberger [149]. We are interested in the distribution of W.

Derivation of the waiting-time distribution

In order to derive the distribution of W, we shall first derive the Laplace-Stieltjes transform of F_W. A straightforward calculation yields for values of s such that $\mathrm{Re}(s)=0$:

$$\begin{aligned}\omega(s)&=\mathbb{E}[\mathrm{e}^{-sW}]=p\,\mathbb{E}[\mathrm{e}^{-s\max\{0,B-A+W\}}]+(1-p)\,\mathbb{E}[\mathrm{e}^{-s\max\{0,B-A-W\}}]\\&=p\,\mathbb{P}[W+B\leqslant A]+p\,\mathbb{E}[\mathrm{e}^{-s(B-A+W)}]-p\,\mathbb{E}[\mathrm{e}^{-s(B-A+W)}\,;W+B\leqslant A]+\\&\quad+(1-p)\,\mathbb{P}[B\leqslant W+A]+(1-p)\,\mathbb{E}[\mathrm{e}^{-s(B-A-W)}\,;B\geqslant W+A];\end{aligned}\qquad (8.5)$$

here A, B and W are independent random variables. The G/PH case for the Lindley-type equation $W \stackrel{\mathcal{D}}{=} \max\{0, B - A - W\}$ has already been analysed in Chapter 4, and the Laplace-Stieltjes transform of the corresponding W is given there. From Equation (4.33) we can readily copy an expression for the last two terms appearing in (8.5), so ω can now be written as

$$\begin{aligned}\omega(s) = p\,\mathbb{P}[W + B \leqslant A] + p\,\alpha(-s)\,\omega(s)\sum_{n=1}^{N}\kappa_n\Big(\frac{\mu}{\mu+s}\Big)^n - \\ - p\,\mathbb{E}[\mathrm{e}^{-s(B-A+W)}\,; W + B \leqslant A] + \\ + (1-p)\left[1 - \sum_{n=1}^{N}\sum_{i=0}^{n-1}\kappa_n\frac{(-\mu)^i}{i!}\,\phi^{(i)}(\mu)\left(1 - \Big(\frac{\mu}{\mu+s}\Big)^{n-i}\right)\right].\end{aligned}$$

So, for $\mathrm{Re}(s) = 0$ we have that

$$\begin{aligned}\omega(s)\left[1 - p\,\alpha(-s)\sum_{n=1}^{N}\kappa_n\Big(\frac{\mu}{\mu+s}\Big)^n\right] = \\ p\,\mathbb{P}[W + B \leqslant A] - p\,\mathbb{E}[\mathrm{e}^{-s(B-A+W)}\,; W + B \leqslant A] + \\ + (1-p)\left[1 - \sum_{n=1}^{N}\sum_{i=0}^{n-1}\kappa_n\frac{(-\mu)^i}{i!}\,\phi^{(i)}(\mu)\left(1 - \Big(\frac{\mu}{\mu+s}\Big)^{n-i}\right)\right]. \quad (8.6)\end{aligned}$$

Cohen [46, p. 322–323] shows by applying Rouché's theorem that the function

$$1 - p\,\alpha(-s)\sum_{n=1}^{N}\kappa_n\Big(\frac{\mu}{\mu+s}\Big)^n \equiv \frac{1}{(\mu+s)^N}\left[(\mu+s)^N - p\,\alpha(-s)\sum_{n=1}^{N}\kappa_n\mu^n(\mu+s)^{N-n}\right]$$

has exactly N zeros $\xi_i(p)$ in the left-half plane if $0 < p < 1$ (it is assumed that $\alpha(\mu) \neq 0$, which is not an essential restriction) or if $p = 1$ and $\mathbb{E}[B] < \mathbb{E}[A]$. Naturally, this statement is not valid if $p = 0$; therefore, this case needs to be excluded from this point on. So we rewrite (8.6) as follows

$$\begin{aligned}\omega(s)\prod_{i=1}^{N}\big(s - \xi_i(p)\big) = \frac{\prod_{i=1}^{N}\big(s - \xi_i(p)\big)}{(\mu+s)^N - p\,\alpha(-s)\sum_{n=1}^{N}\kappa_n\mu^n(\mu+s)^{N-n}} \times \\ \times \Bigg[p\,(\mu+s)^N\mathbb{P}[W + B \leqslant A] - p\,(\mu+s)^N\mathbb{E}[\mathrm{e}^{-s(B-A+W)}\,; W + B \leqslant A] + \\ + (1-p)\left[(\mu+s)^N - \sum_{n=1}^{N}\sum_{i=0}^{n-1}\kappa_n\frac{(-\mu)^i}{i!}\,\phi^{(i)}(\mu)\big((\mu+s)^N - \mu^{n-i}(\mu+s)^{N-n+i}\big)\right]\Bigg]. \quad (8.7)\end{aligned}$$

Thus, we have conveniently eliminated all poles from the right-hand side of the equation, since the ones remaining in the denominator of the fraction are removed by the zeros of the numerator of this fraction.

We can observe that the left-hand side of (8.7) is analytic for $\mathrm{Re}(s) > 0$ and continuous for $\mathrm{Re}(s) \geqslant 0$, and the right-hand side of (8.7) is analytic for $\mathrm{Re}(s) < 0$ and continuous for $\mathrm{Re}(s) \leqslant 0$. So from Liouville's theorem [162] (see also Section 7.3) we have that both sides of (8.7) are the same N-th degree polynomial, say, $\sum_{i=0}^{N} q_i s^i$. Hence,

$$\omega(s) = \frac{\sum_{i=0}^{N} q_i s^i}{\prod_{i=1}^{N}\big(s - \xi_i(p)\big)}. \tag{8.8}$$

In the expression above, the constants q_i are not determined so far, while the roots $\xi_i(p)$ are known. In order to obtain the transform, observe that ω is a fraction of two polynomials of degree N. So, ignoring the special case of multiple zeros $\xi_i(p)$, partial fraction decomposition yields that (8.8) can be rewritten as

$$\omega(s) = c_0 + \sum_{i=1}^{N} \frac{c_i}{s - \xi_i(p)}, \tag{8.9}$$

which implies that the waiting-time distribution has a mass at the origin that is given by

$$\mathbb{P}[W = 0] = \lim_{s\to\infty} \mathbb{E}[\mathrm{e}^{-sW}] = c_0$$

and has a density that is given by

$$f_W(x) = \sum_{i=1}^{N} c_i \mathrm{e}^{\xi_i(p)x}.$$

All that remains is to determine the $N + 1$ constants c_i. To do so, we work as follows. We shall substitute (8.9) in the left-hand side of (8.7), and express the terms $\mathbb{P}[W+B \leqslant A]$ and $\mathbb{E}[\mathrm{e}^{-s(B-A+W)}\,; W+B \leqslant A]$ that appear at the right-hand side of (8.7) in terms of the constants c_i. Note that the terms $\phi^{(i)}(\mu)$ that appear at the right-hand side of (8.7) can also be expressed in terms of the constants c_i. Thus we obtain a new equation that we shall differentiate a total of N times. We shall evaluate each of these derivatives for $s = 0$ and thus we obtain a linear system of N equations for the constants c_i, $i = 0, \ldots, N$. The last equation that is necessary to uniquely determine the constants c_i is the normalisation equation

$$c_0 + \int_0^\infty f_W(x)\,\mathrm{d}x = 1. \tag{8.10}$$

To begin with, note that

$$\mathbb{P}[W + B \leqslant A] = \mathbb{P}[W = 0]\mathbb{P}[B \leqslant A] + \int_0^\infty \mathbb{P}[B \leqslant A - x] \sum_{i=1}^{N} c_i \mathrm{e}^{\xi_i(p)x}\,\mathrm{d}x, \tag{8.11}$$

with

$$\begin{aligned}\mathbb{P}[B \leqslant A] &= \int_0^\infty \sum_{n=1}^N \kappa_n \left(\mathrm{e}^{-\mu x} \sum_{j=n}^\infty \frac{(\mu x)^j}{j!} \right) \mathrm{d}F_A(x) \\ &= \sum_{n=1}^N \sum_{i=n}^\infty \kappa_n \frac{(-\mu)^i}{i!}\, \alpha^{(i)}(\mu), \end{aligned} \tag{8.12}$$

and

$$\begin{aligned}&\int_0^\infty \mathbb{P}[B \leqslant A - x] \sum_{i=1}^N c_i \mathrm{e}^{\xi_i(p)x}\, \mathrm{d}x \\ &\quad = \int_0^\infty \int_0^\infty \mathbb{P}[B \leqslant y - x] \sum_{i=1}^N c_i \mathrm{e}^{\xi_i(p)x}\, \mathrm{d}x \mathrm{d}F_A(y) \\ &\quad = \int_0^\infty \int_0^y \mathrm{e}^{-\mu(y-x)} \sum_{n=1}^N \sum_{j=n}^\infty \kappa_n \frac{\big(\mu(y-x)\big)^j}{j!} \sum_{i=1}^N c_i \mathrm{e}^{\xi_i(p)x}\, \mathrm{d}x \mathrm{d}F_A(y) \\ &\quad = \sum_{n=1}^N \sum_{j=n}^\infty \sum_{i=1}^N \sum_{k=j+1}^\infty \kappa_n c_i \frac{\mu^j \big(\mu + \xi_i(p)\big)^{k-j-1}}{k!(-1)^k}\, \alpha^{(k)}(\mu). \end{aligned} \tag{8.13}$$

Likewise, we have that

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{-s(B-A+W)}\,; W + B \leqslant A] &= \mathbb{P}[W = 0]\, \mathbb{E}[\mathrm{e}^{-s(B-A)}\,; B \leqslant A] + \\ &\quad + \int_0^\infty \mathbb{E}[\mathrm{e}^{-s(B-A+x)}\,; x + B \leqslant A] \sum_{i=1}^N c_i \mathrm{e}^{\xi_i(p)x}\, \mathrm{d}x, \end{aligned} \tag{8.14}$$

with

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{-s(B-A)}\,; B \leqslant A] &= \int_0^\infty \int_0^x \mathrm{e}^{-s(y-x)} \sum_{n=1}^N \kappa_n \mu \mathrm{e}^{-\mu y} \frac{(\mu y)^{n-1}}{(n-1)!}\, \mathrm{d}y \mathrm{d}F_A(x) \\ &= \int_0^\infty \mathrm{e}^{xs} \sum_{n=1}^N \kappa_n \Big(\frac{\mu}{\mu + s}\Big)^n \sum_{i=n}^\infty \mathrm{e}^{-x(\mu+s)} \frac{x^i (\mu+s)^i}{i!}\, \mathrm{d}F_A(x) \\ &= \sum_{n=1}^N \sum_{i=n}^\infty \kappa_n\, \mu^n \frac{(-1)^i}{i!}\, (\mu + s)^{i-n}\, \alpha^{(i)}(\mu), \end{aligned} \tag{8.15}$$

and

$$\begin{aligned}
&\int_0^\infty \mathbb{E}[\mathrm{e}^{-s(B-A+x)}\,; x+B\leqslant A]\sum_{i=1}^N c_i \mathrm{e}^{\xi_i(p)x}\,\mathrm{d}x \\
&\quad= \int_0^\infty \int_0^y \int_0^{y-x} \mathrm{e}^{-s(z-y+x)} \sum_{n=1}^N \kappa_n \mu \mathrm{e}^{-\mu z}\frac{(\mu z)^{n-1}}{(n-1)!}\sum_{i=1}^N c_i \mathrm{e}^{\xi_i(p)x}\,\mathrm{d}z\mathrm{d}x\mathrm{d}F_A(y) \\
&\quad= \int_0^\infty \int_0^y \sum_{n=1}^N \kappa_n \Big(\frac{\mu}{\mu+s}\Big)^n \mathrm{e}^{-s(x-y)} \sum_{j=n}^\infty \mathrm{e}^{-(\mu+s)(y-x)}\frac{(\mu+s)^j(y-x)^j}{j!}\times \\
&\qquad\qquad \times \sum_{i=1}^N c_i \mathrm{e}^{\xi_i(p)x}\,\mathrm{d}x\mathrm{d}F_A(y) \\
&\quad= \sum_{n=1}^N \sum_{j=n}^\infty \sum_{i=1}^N \sum_{k=j+1}^\infty \kappa_n c_i \Big(\frac{\mu}{\mu+s}\Big)^n \frac{(\mu+s)^j\big(\mu+\xi_i(p)\big)^{k-j-1}}{k!(-1)^k}\,\alpha^{(k)}(\mu). \qquad (8.16)
\end{aligned}$$

So, using (8.12) and (8.13), substitute (8.11) in the right-hand side of (8.7), and likewise for (8.14). Furthermore, as mentioned before, substitute (8.9) into the left-hand side of (8.7) to obtain an expression, where both sides can be reduced to an N-th degree polynomial in s. By evaluating this polynomial and all its derivatives for $s=0$ we obtain N equations binding the constants c_i. These equations, and the normalisation equation (8.10), form a linear system for the constants c_i, $i=0,\ldots,N$, that uniquely determines them (see also Remark 8.1 below). For example, the first equation, evaluated at $s=0$, yields that

$$c_0 - \sum_{i=1}^N \frac{c_i}{\xi_i(p)} = \frac{1-p}{1-p\,\alpha(0)} = 1,$$

since $\alpha(0)=1$. We summarise the above in the following theorem.

Theorem 8.2. *Consider the recursion given by (8.1), and assume that $0<p<1$. Let (8.4) be the distribution of the random variable B. Then the limiting distribution of the waiting time has mass c_0 at the origin and a density on $[0,\infty)$ that is given by*

$$f_W(x) = \sum_{i=1}^N c_i \mathrm{e}^{\xi_i(p)x}.$$

In the above equation, the constants $\xi_i(p)$, with $\mathrm{Re}(\xi_i(p))<0$, are the N roots of

$$(\mu+s)^N - p\,\alpha(-s)\sum_{n=1}^N \kappa_n \mu^n (\mu+s)^{N-n} = 0,$$

and the $N+1$ constants c_i are the unique solution to the linear system described above.

Remark 8.1. Although the roots $\xi_i(p)$ and coefficients c_i may be complex, the density and the mass c_0 at zero will be positive. This follows from the fact that there is a unique equilibrium distribution and thus a unique solution to the linear system for the coefficients c_i. Of course, it is also clear that each root $\xi_i(p)$ and coefficient c_i have a companion conjugate root and conjugate coefficient, which implies that the imaginary parts appearing in the density cancel.

Remark 8.2. In case that $\xi_i(p)$ has multiplicity greater than one for one or more values of i, the analysis proceeds in essentially the same way. For example, if $\xi_1(p) = \xi_2(p)$, then the partial fraction decomposition of ω becomes

$$\omega(s) = c_0 + \frac{c_1}{\big(s - \xi_1(p)\big)^2} + \sum_{i=2}^{N} \frac{c_i}{s - \xi_i(p)},$$

the inverse of which is given by

$$f_W(x) = c_1 x \mathrm{e}^{\xi_1(p)x} + \sum_{i=2}^{N} c_i \mathrm{e}^{\xi_i(p)x}.$$

Remark 8.3. For the nominal service time B we have considered only mixtures of Erlang distributions, mainly because this class approximates well any continuous distribution on $[0, \infty)$ and because we can illustrate the techniques we use without complicating the analysis. However, we can extend this class by considering distributions with a rational Laplace transform. The analysis in [170] can be extended to such distributions, and the analysis in Cohen [46, Section II.5.10] is already given for such distributions, so the results given there can be implemented directly.

Remark 8.4. The analysis we have presented so far can be directly extended to the case where Y takes any finite number of negative values. In other words, let the distribution of Y be given by $\mathbb{P}[Y = 1] = p$, and for $i = 1, \ldots, n$, $\mathbb{P}[Y = -u_i] = p_i$, where $u_i > 0$ and $\sum_i p_i = 1 - p$. Then, for example, Equation (8.6) becomes

$$\omega(s)\left[1 - p\,\alpha(-s) \sum_{n=1}^{N} \kappa_n \Big(\frac{\mu}{\mu + s}\Big)^n\right] =$$

$$p\,\mathbb{P}[W + B \leqslant A] - p\,\mathbb{E}[\mathrm{e}^{-s(B-A+W)}\,; W + B \leqslant A] + \sum_{i=1}^{n} p_i\,\mathbb{P}[B \leqslant u_i W + A] +$$

$$+ \sum_{i=1}^{n} p_i \left(\sum_{n=1}^{N} \kappa_n \Big(\frac{\mu}{\mu + s}\Big)^n \alpha(-s)\,\omega(-u_i s) - \mathbb{E}[\mathrm{e}^{-s(B-A-u_i W)}\,; B \leqslant u_i W + A]\right).$$

Following the same steps as below (8.6), we can conclude that the waiting time density is again given by a mixture of exponentials of the form

$$f_W(x) = \sum_{i=1}^{N} \hat{c}_i \mathrm{e}^{\xi_i(p)x},$$

where the new constants $\hat{c}_i$ (and the mass of the distribution at zero, given by $\hat{c}_0$) are to be determined as the unique solution to a linear system of equations. The only additional remark necessary when forming this linear system is to observe that both the probability $\mathbb{P}[B \leqslant u_i W + A]$ and the expectation $\mathbb{E}[\mathrm{e}^{-s(B-A-u_i W)}\,; B \leqslant u_i W + A]$ can be expressed linearly in terms of the constants $\hat{c}_i$.

The case $p = 0$

We have seen that the case where $Y_n = -1$ for all n, or in other words the case $p = 0$, had to be excluded from the analysis. Equation (8.7) is still valid if we take the constants $\xi_i(0)$ to be defined as in Theorem 8.2. However, one cannot apply Liouville's theorem to the resulting equation. The transform can be inverted directly. As it is shown in Chapter 4, the terms $\phi^{(i)}(\mu)$ that remain to be determined follow by differentiating Equation (8.7) a total of $N-1$ times and evaluating $\omega^i(s)$ at $s = \mu$ for $i = 0, \ldots, N-1$. The density in this case is a mixture of Erlang distributions with the same scale parameter μ for all exponential phases. As we can see, for $p = 0$ the resulting density is intrinsically different from the one described in Theorem 8.2.

The case $p = 1$

If $p = 1$ and $\mathbb{E}[B] < \mathbb{E}[A]$, then we are analysing the steady-state waiting-time distribution of a G/PH/1 queue. Equation (8.7) now reduces to

$$\omega(s) \prod_{i=1}^{N} \big(s - \xi_i(1)\big) = \frac{\prod_{i=1}^{N} \big(s - \xi_i(1)\big)}{(\mu+s)^N - \alpha(-s) \sum_{n=1}^{N} \kappa_n \mu^n (\mu+s)^{N-n}} \times \\ \times \big[(\mu+s)^N \mathbb{P}[W + B \leqslant A] - (\mu+s)^N \mathbb{E}[\mathrm{e}^{-s(B-A+W)}\,; W + B \leqslant A]\big]. \quad (8.17)$$

Earlier we have already observed that the right-hand side of (8.17) is equal to an N-th degree polynomial $\sum_{i=0}^{N} q_i s^i$. Inspection of the right-hand side of (8.17) reveals that it has an N-fold zero in $s = -\mu$. Indeed, all zeros of the numerator of the quotient in the right-hand side cancel against zeros of the denominator, and the term

$$\mathbb{P}[W + B \leqslant A] - \mathbb{E}[\mathrm{e}^{-s(B-A+W)}\,; W + B \leqslant A]$$

is finite for $s = -\mu$. Hence,

$$\sum_{i=0}^{N} q_i s^i = q_N \, (\mu + s)^N. \quad (8.18)$$

Combining (8.17) and (8.18), we conclude that

$$\omega(s) \prod_{i=1}^{N} \big(s - \xi_i(1)\big) = q_N \, (\mu + s)^N,$$

and since $\omega(0) = 1$, the last equation gives us that

$$q_N = \frac{\prod_{i=1}^{N}\big(-\xi_i(1)\big)}{\mu^N}.$$

Thus, we have that

$$\omega(s) = \Big(\frac{\mu + s}{\mu}\Big)^N \prod_{i=1}^{N} \frac{\xi_i(1)}{\xi_i(1) - s},$$

which is in agreement with Equation II.5.190 in [46, p. 324].

8.4 The M/D model

We have examined so far the case where the nominal interarrival time A is generally distributed and the nominal service time B follows a phase-type distribution. In other words, we have studied the case which is in a sense analogous to the ordinary G/PH/1 queue. We now would like to study the reversed situation; namely, the case analogous to the M/G/1 queue. As before, a reasonable assumption is to utilise a method that can be used to derive the steady-state waiting-time distribution both for the M/G/1 queue and for the M/D case of the alternating service model given by Recursion (1.1).

The M/G/1 queue has been studied in much detail. However, the analogous alternating service model – i.e., take $\mathbb{P}(Y = -1) = 1$ in (8.1), so $p = 0$ – seems to be more complicated to analyse, see Chapter 5. As shown in Section 1.6, if $p = 0$, the density of W satisfies a generalised Wiener-Hopf equation, for which no solution is known in general. The presently available results for the distribution of W with $p = 0$ are developed in Chapter 5, where B is assumed to belong to the class $\mathcal{M}$, which strictly bigger than the class of functions with rational Laplace transforms, but not completely general. Moreover, the method developed in Chapter 5 breaks down when applied to (8.3) with Y *not* identically equal to -1. One cannot exploit in this case the special form of the distributions belonging to $\mathcal{M}$, since the "Lindley-part" of the recursion does not allow for a convenient definition of *constants* c_i, as they were defined in Chapter 5. Everything considered, it seems that there is no method available so far that can be used both for the M/G/1 queue and for the M/D case of the alternating service model.

We shall refrain from trying to develop an alternative approach for the M/G case with a more general distribution for B than the one treated in Section 8.3. Instead, we give a detailed analysis of the M/D case: A is exponentially distributed and B is deterministic. This case is neither contained in the G/PH case of the previous section nor has it been treated (for the special choice of $p = 0$) in Chapter 3. Its analysis is of interest for various reasons. To start with, the model generalises the classical M/D/1 queue; additionally, the analysis illustrates the difficulties that arise when studying (8.3) in case A is exponentially distributed and B is generally distributed; finally, the different effects of Lindley's classical recursion and of the Lindley-type recursion (1.1) are clearly exposed. As we shall see in the following, the

analysis can be practically split into two parts, where each part follows the analysis of the corresponding model with $Y \equiv 1$, or $Y \equiv -1$.

Derivation of the waiting-time distribution

As before, consider Equation (8.3), and assume that $Y = 1$ with probability p and $Y = -1$ with probability $1 - p$. Let A be exponentially distributed with rate λ and B be equal to b, where $b > 0$. Furthermore, we shall denote by π_0 the mass of the distribution of W at zero; that is, $\pi_0 = \mathbb{P}[W = 0]$.

For this setting, we have from (8.3) that for $x \geqslant 0$,

$$\begin{aligned} F_W(x) &= \mathbb{P}[\max\{0, b - A + YW\} \leqslant x] = \mathbb{P}[b - A + YW \leqslant x] \\ &= p\,\mathbb{P}[b - A + W \leqslant x] + (1-p)\,\mathbb{P}[b - A - W \leqslant x] \\ &= p\,\pi_0 \mathbb{P}[b - A \leqslant x] + p \int_0^\infty \mathbb{P}[b - A \leqslant x - y] f_W(y)\,\mathrm{d}y + \\ &\quad + (1-p)\,\pi_0 \mathbb{P}[b - A \leqslant x] + (1-p) \int_0^\infty \mathbb{P}[b - A \leqslant x + y] f_W(y)\,\mathrm{d}y \\ &= \pi_0 \mathbb{P}[A \geqslant b - x] + p \int_0^\infty \mathbb{P}[A \geqslant b - x + y] f_W(y)\,\mathrm{d}y + \\ &\quad + (1-p) \int_0^\infty \mathbb{P}[A \geqslant b - x - y] f_W(y)\,\mathrm{d}y. \end{aligned} \tag{8.19}$$

Evidently, in order to rewrite the probabilities appearing in the above expression, we need to consider two separate cases. So, for $0 \leqslant x < b$ the above equation reduces to

$$\begin{aligned} F_W(x) = \pi_0\,\mathrm{e}^{-\lambda(b-x)} &+ p \int_0^\infty \mathrm{e}^{-\lambda(b-x+y)} f_W(y)\,\mathrm{d}y + \\ &+ (1-p) \int_0^{b-x} \mathrm{e}^{-\lambda(b-x-y)} f_W(y)\,\mathrm{d}y + (1-p) \int_{b-x}^\infty f_W(y)\,\mathrm{d}y, \end{aligned} \tag{8.20}$$

and for $x \geqslant b$, Equation (8.19) reduces to

$$F_W(x) = \pi_0 + p \int_0^{x-b} f_W(y)\,\mathrm{d}y + p \int_{x-b}^\infty \mathrm{e}^{-\lambda(b-x+y)} f_W(y)\,\mathrm{d}y + (1-p)(1-\pi_0), \tag{8.21}$$

where we have utilised the normalisation equation

$$\pi_0 + \int_0^\infty f_W(y)\,\mathrm{d}y = 1. \tag{8.22}$$

In the following, we shall derive the distribution on the interval $[0, b)$ and on the interval $[b, \infty)$ separately. At this point though, one should note that from Equation (8.3) it is apparent that for A exponentially distributed and $B = b$, the

distribution of W is continuous on $(0,\infty)$. Also, one can verify that Equation (8.20) for $x = b$ reduces to Equation (8.21) for $x = b$. The fact that F_W is continuous on $(0,\infty)$ will be used extensively in the sequel. Notice also that from Equations (8.20) and (8.21) we can immediately see that we can differentiate $F_W(x)$ for $x \in (0,b)$ and $x \in (b,\infty)$; see, for example, Titchmarsh [162, p. 59].

The distribution on $[0,b)$:

In all subsequent equations it is assumed that $x \in (0,b)$. In order to derive the distribution of W on $[0,b]$, we differentiate (8.20) once to obtain

$$\begin{aligned} f_W(x) = \lambda\pi_0\, \mathrm{e}^{-\lambda(b-x)} + \lambda p \int_0^{\infty} \mathrm{e}^{-\lambda(b-x+y)} f_W(y)\,\mathrm{d}y + \\ + \lambda(1-p)\int_0^{b-x} \mathrm{e}^{-\lambda(b-x-y)} f_W(y)\,\mathrm{d}y - \\ - (1-p)\mathrm{e}^{-\lambda(b-x)}\mathrm{e}^{\lambda(b-x)} f_W(b-x) + (1-p) f_W(b-x). \end{aligned}$$

We rewrite this equation after noticing that the third line is equal to zero, while the sum of the integrals in the first two lines can be rewritten by using (8.20). Thus, we have that

$$\begin{aligned} f_W(x) &= \lambda\pi_0\, \mathrm{e}^{-\lambda(b-x)} + \lambda\left(F_W(x) - \pi_0 \mathrm{e}^{-\lambda(b-x)} - (1-p)\int_{b-x}^{\infty} f_W(y)\,\mathrm{d}y \right) \\ &= \lambda F_W(x) - \lambda(1-p)\int_{b-x}^{\infty} f_W(y)\,\mathrm{d}y. \end{aligned} \tag{8.23}$$

In order to obtain a linear differential equation, differentiate (8.23) once more, which leads to

$$f_W^{'}(x) = \lambda f_W(x) - \lambda(1-p) f_W(b-x). \tag{8.24}$$

Equation (8.24) is a homogeneous linear differential equation, not of a standard form because of the argument $b-x$ that appears at the right-hand side. We have already encountered a similar differential equation of this type in Section 3.6. For the solution, one could follow exactly the same steps. However, Equation (8.24) is significantly less complicated, as it is only of first order, so here we shall present a simpler approach. To solve it, we substitute x for $b-x$ in (8.24) to obtain

$$f_W^{'}(b-x) = \lambda f_W(b-x) - \lambda(1-p) f_W(x). \tag{8.25}$$

Then, we differentiate (8.24) once more to obtain

$$f_W^{''}(x) = \lambda f_W^{'}(x) + \lambda(1-p) f_W^{'}(b-x),$$

and we eliminate the term $f_W^{'}(b-x)$ by using (8.25). Thus, we conclude that

$$f_W^{''}(x) = \lambda^2 p\,(2-p) f_W(x). \tag{8.26}$$

For $p \neq 0$, the solution to this differential equation is given by

$$f_W(x) = d_1 \mathrm{e}^{r_1 x} + d_2 \mathrm{e}^{r_2 x}, \tag{8.27}$$

where r_1 and r_2 are given by

$$r_{1,2} = \pm\lambda\sqrt{p\,(2-p)}, \tag{8.28}$$

and the constants d_1 and d_2 will be determined by the initial conditions. Namely, the solution needs to satisfy (8.24) and the condition $F_W(0) = \pi_0$. Thus, for the first equation, substitute the general solution we have derived into (8.24) and for the second equation, rewrite (8.23) as follows:

$$f_W(x) = \lambda F_W(x) - \lambda(1-p)\left(1 - \pi_0 - \int_0^{b-x} f_W(y)\,\mathrm{d}y\right),$$

substitute $f_W(x)$ from (8.27), and evaluate the resulting equation for $x = 0$. So the first equation we derive is

$$r_1 d_1 \mathrm{e}^{r_1 x} + r_2 d_2 \mathrm{e}^{r_2 x} = \lambda(d_1 \mathrm{e}^{r_1 x} + d_2 \mathrm{e}^{r_2 x}) - \lambda(1-p)(d_1 \mathrm{e}^{r_1(b-x)} + d_2 \mathrm{e}^{r_2(b-x)}),$$

from which we conclude that

$$r_1 d_1 = \lambda d_1 - \lambda(1-p) d_2 \mathrm{e}^{r_2 b},$$

and the second equation we derive is

$$d_1 + d_2 = \lambda\pi_0 - \lambda(1-p)\left(1 - \pi_0 - \left(\frac{d_1}{r_1}\left(\mathrm{e}^{r_1 b} - 1\right) + \frac{d_2}{r_2}\left(\mathrm{e}^{r_2 b} - 1\right)\right)\right).$$

This system uniquely determines d_1 and d_2. Specifically, we have that

$$d_1 = \frac{\lambda^2(1-p)\big(1 - p\,(1-\pi_0) - 2\pi_0\big)r_1}{(\mathrm{e}^{br_1} - 1)\lambda^2(2-p)(1-p) + \mathrm{e}^{br_1} r_1\big(r_1 - \lambda(2-p)\big)},$$
$$d_2 = \frac{\mathrm{e}^{br_1}\lambda(1 - p\,(1-\pi_0) - 2\pi_0)r_1\,(\lambda - r_1)}{(\mathrm{e}^{br_1} - 1)\lambda^2(2-p)(1-p) + \mathrm{e}^{br_1} r_1\big(r_1 - \lambda(2-p)\big)},$$

where in the process we have assumed that $p \neq 1$. Up to this point we have that the waiting-time distribution on $[0, b]$ is given by

$$F_W(x) = \frac{d_1}{r_1}(\mathrm{e}^{r_1 x} - 1) + \frac{d_2}{r_2}(\mathrm{e}^{r_2 x} - 1) + \pi_0, \tag{8.29}$$

where d_1 and d_2 are known up to the probability π_0. The cases for $p = 0$ and $p = 1$ follow directly from Equation (8.26) and will be handled separately in the sequel.

The distribution on $[b, \infty)$:

As before, we obtain a differential equation by differentiating (8.21) once, and substituting the resulting integrals by using (8.21) once more. Thus, we obtain the equation

$$f_W(x) = \lambda\bigg(F_W(x) - \pi_0 - (1-p)(1-\pi_0) - p\int_0^{x-b} f_W(y)\,\mathrm{d}y\bigg),$$

which can be reduced to

$$f_W(x) = \lambda\big(F_W(x) - 1 + p - p\,F_W(x-b)\big). \tag{8.30}$$

Equation (8.30) is a delay differential equation that can be solved recursively. Observe that for $x \in (b, 2b)$, the term $F_W(x-b)$ has been derived in the previous step, so for $x \in (b, 2b)$, Equation (8.30) reduces to an ordinary linear differential equation from which we can easily derive the distribution of W in the interval $(b, 2b)$.

For simplicity, denote by $F_i(x)$ the distribution of W when $x \in [ib, (i+1)b]$, and analogously denote by $f_i(x)$ the density of W, when $x \in (ib, (i+1)b)$. Then (8.30) states that

$$f_i(x) = \lambda\big(F_i(x) - 1 + p - p\,F_{i-1}(x-b)\big),$$

which leads to an expression for F_i that is given in terms of an indefinite integral that is a function of x, that is,

$$F_i(x) = \mathrm{e}^{\lambda x}\bigg(\int \lambda\big(-1 + p - p\,F_{i-1}(x-b)\big)\mathrm{e}^{-\lambda x}\,\mathrm{d}x + \gamma_i\bigg), \quad i \geqslant 1. \tag{8.31}$$

The constants γ_i can be determined by exploiting the fact that the waiting-time distribution is continuous. In particular, every γ_i is determined by the equation

$$F_i(ib) = F_{i-1}(ib). \tag{8.32}$$

Solving Equation (8.31) recursively, we obtain that

$$F_i(x) = 1 - p^i(1-\pi_0) - p^i\Big(\frac{d_1}{r_1} + \frac{d_2}{r_2}\Big) + \sum_{j=1}^{2}\Big(\frac{\lambda p}{\lambda - r_j}\Big)^i \frac{d_j}{r_j}\mathrm{e}^{r_j(x-ib)} + \\ + x\sum_{j=0}^{i-1}(-\lambda p)^j \gamma_{i-j}\,\frac{(x-jb)^{j-1}}{j!}\,\mathrm{e}^{\lambda(x-jb)}. \tag{8.33}$$

Observe that for $i = 0$, if we define the empty sum at the right-hand side to be equal to zero, then the above expression is satisfied. Notice that, since we have made use of the distribution on $[0, b)$ as it is given by (8.29), Equation (8.33) is not valid for

$p = 0$ or $p = 1$. From Equation (8.32) we now have that for every $i \geqslant 1$,

$$\begin{aligned}\gamma_i = \gamma_{i-1} + \mathrm{e}^{-\lambda i b}(1-p)p^{i-1}\Big(\pi_0 - 1 - \frac{d_1 - d_2}{r_1}\Big) - \\ - \sum_{j=1}^{2} \frac{\mathrm{e}^{-\lambda i b} d_j}{r_j}\Big(\frac{\lambda p}{\lambda - r_j}\Big)^i \Big(1 - \frac{(\lambda - r_j)\,\mathrm{e}^{b r_j}}{\lambda p}\Big) + \\ + i \sum_{j=1}^{i-1} \frac{\mathrm{e}^{-\lambda j b}(i-j)^{j-1}(-\lambda p b)^j\,(\gamma_{i-1-j} - \gamma_{i-j})}{j!}, \end{aligned} \tag{8.34}$$

where we have assumed that $\gamma_0 = 0$, and that for $i = 1$, the second sum is equal to zero. These expressions can be simplified further by observing that

$$1 - p^i(1-\pi_0) - p^i\Big(\frac{d_1}{r_1} + \frac{d_2}{r_2}\Big) = 1 - \frac{p^i}{2-p}.$$

Recall that d_1 and d_2, and thus also all constants γ_i, are known in terms of π_0. The probability π_0 that still remains to be determined will be given by the normalisation equation (8.22). Notice though, that since the waiting-time distribution is determined recursively for every interval $[ib, (i+1)b]$, Equation (8.22) yields an infinite sum. The sum is well defined, since a unique density exists. The above findings are summarised in the following theorem.

Theorem 8.3. *Consider the recursion given by (8.1), and assume that $0 < p < 1$. Let A be exponentially distributed with rate λ and B be equal to b, where $b > 0$. Then for $x \in [ib, (i+1)b]$, $i = 0, 1, \ldots$, the limiting distribution of the waiting time is given by*

$$\begin{aligned} F_W(x) = 1 - \frac{p^i}{2-p} + \sum_{j=1}^{2}\Big(\frac{\lambda p}{\lambda - r_j}\Big)^i \frac{d_j}{r_j}\,\mathrm{e}^{r_j(x-ib)} + \\ + x\sum_{j=0}^{i-1}(-\lambda p)^j \gamma_{i-j}\frac{(x-jb)^{j-1}}{j!}\,\mathrm{e}^{\lambda(x-jb)}, \end{aligned}$$

where the constants γ_i are given by Equation (8.34) and the probability π_0 is given by the normalisation equation (8.22).

One might expect though that Equation (8.22) may not be suitable for numerically determining π_0. However, if the probability p is not too close to one, or in other words, if the system does not almost behave like an M/D/1 queue, then one can numerically approximate π_0 from the normalisation equation. As an example, in Figure 8.1 we display a typical plot of the waiting-time distribution. We have chosen $b = 1$, $\lambda = 2$, and $p = 1/3$.

For p close to one, we can see from the expressions for d_1 and d_2 that both the numerators and the denominators of these two constants approach zero. Furthermore,

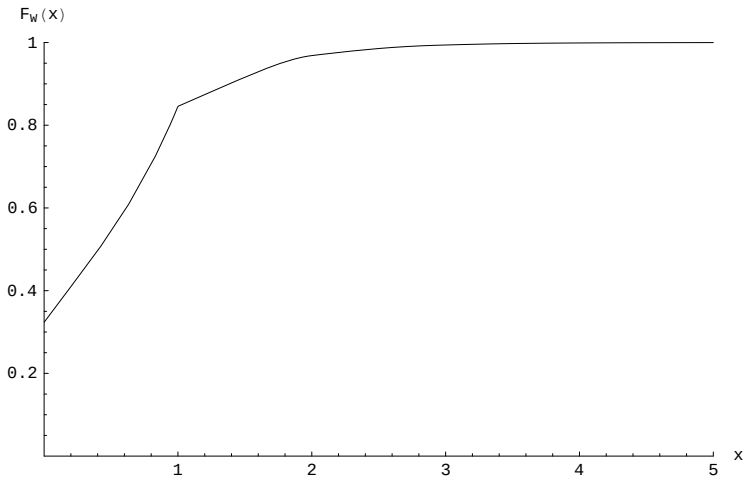

Figure 8.1: The waiting-time distribution for $b = 1$, $\lambda = 2$, and $p = 1/3$.

the denominators $\lambda - r_j$, $j = 1, 2$, that appear in the waiting-time distribution also approach zero, which makes Theorem 8.3 unsuitable for numerical computations for values of p close to one. Moreover, we also see that very large values of the parameter λ may also lead to numerical problems, since λ is involved in the exponent of almost all exponential terms that appear in the waiting-time distribution.

As one can observe from Figure 8.1, and show from Theorem 8.3, F_W is not differentiable for $x = b$. This is not surprising, as the waiting-time distribution is defined by two different equations; namely Equation (8.20) for $x < b$ and Equation (8.21) for $x \geqslant b$. Furthermore, from Equation (8.23) we have that

$$f_W(b^-) = \lambda F_W(b) - \lambda(1-p)(1-\pi_0),$$

and from Equation (8.30) we have that

$$f_W(b^+) = \lambda(F_W(b) - 1 + p - p\,\pi_0).$$

That is, $f_W(b^-) - f_W(b^+) = \lambda\pi_0$.

The case $p = 0$

Observe that if $p = 0$ then the support of W is the interval $[0, b]$. To determine the density of the waiting time, we insert $p = 0$ into Equation (8.26). Thus, we obtain that

$$f_W''(x) = 0,$$

from which we immediately have that

$$f_W(x) = \nu_1 x + \nu_2,$$

for some constants ν_1 and ν_2 such that (8.24) is satisfied. The latter condition implies that for every $x \in (0, b)$ the following equation must hold:

$$\nu_1 = \lambda(\nu_1 x + \nu_2) - \lambda(\nu_1(b - x) + \nu_2).$$

From this we conclude that ν_1 is equal to zero, i.e. the waiting time has a mass at zero and is uniformly distributed on $(0, b)$. To determine the mass π_0 and the constant ν_2 we evaluate (8.23) at $x = 0$ and we use the normalisation equation (8.22), keeping in mind that $f_W(x) = 0$ for $x \in [b, \infty)$. These two equations give the system

$$\begin{aligned} \nu_2 &= \lambda\pi_0 - \lambda(1 - \pi_0 - b\nu_2) \\ \pi_0 + b\nu_2 &= 1, \end{aligned}$$

which yields that if $p = 0$, then

$$f_W(x) = \frac{\lambda}{1 + \lambda b}, \quad 0 < x < b, \qquad \text{and} \qquad \pi_0 = \frac{1}{1 + \lambda b}. \tag{8.35}$$

Evidently, the density in this case is quite different from the density for $p \neq 0$, which is on $(0, b)$ a mixture of two exponentials; see (8.27). However, our numerical experiments verify that for $0 < p = \epsilon << 1$, the waiting-time distribution, which is given by (8.33), is close to the uniform distribution on $(0, b]$ with a mass at zero close to the value of π_0 as it is given by Equation (8.35). This is to be expected, since $p = 0$ is just a special case of Equation (8.26), and thus its solution can be seen as the limiting case of (8.27) for p approaching zero.

Another way to see that $f_W(x) = \lambda\pi_0$, $0 < x < b$, is as follows. Recall that for $p = 0$ and $x \geqslant b$ we have that $f_W(x) = 0$. Equation (8.23) can now be written as

$$f_W(x) = \lambda\pi_0 + \lambda\mathbb{P}[W \in (0, x)] - \lambda\mathbb{P}[W \in (b - x, b)].$$

Replacing x by $b - x$ shows that $f_W(x) = f_W(b - x)$, from which we derive that $\mathbb{P}[W \in (0, x)] = \mathbb{P}[W \in (b - x, b)]$ and finally that $f_W(x) = \lambda\pi_0$, $0 < x < b$. It seems less straightforward to explain *probabilistically* that W, given that $W > 0$, is uniformly distributed. With a view towards the recursion $W \stackrel{\mathcal{D}}{=} \max\{0, b - A - W\}$, we believe that this property is related to the fact that, if n Poisson arrivals occur in some interval, then they are distributed like the n order statistics of the uniform distribution on that interval [147, Section 2.3].

The case $p = 1$

For the M/D/1 queue, Erlang [61] derived the following expression for the waiting-time distribution:

$$\mathbb{P}[W \leqslant x] = (1 - \rho) \sum_{j=0}^{i} \frac{\big(-\lambda(x - jb)\big)^j}{j!} \, \mathrm{e}^{\lambda(x - jb)}, \quad ib \leqslant x < (i + 1)b,$$

where ρ is the traffic intensity. Recall that for the M/D/1 queue we have that $F_W(0) = 1 - \rho$. We see that for $p = 1$ Equation (8.23) indeed leads to the waiting-time distribution $(1-\rho)\,\mathrm{e}^{\lambda x}$, as it is given by Erlang's expression for the first interval

$[0, b)$. For $x \geqslant b$, one needs to recursively solve Equation (8.31) in order to obtain Erlang's expression. However, since the recursive solution we have obtained for our model makes use of $F_W(x)$ as it is given by (8.29), which is not valid for $p = 1$, the waiting-time distribution we have obtained in Theorem 8.3 cannot be extended to the case for $p = 1$.

The terms both in Erlang's expression for the waiting-time distribution of an M/D/1 queue and in Theorem 8.3 alternate in sign and in general are much larger than their sum. Thus, the numerical evaluation of the sum may be hampered by roundoff errors due to the loss of significant digits, in particular under heavy traffic. For the M/D/1 queue, however, a satisfactory solution has been given by Franx [66]. The author uses a probabilistic approach leading to a simple formula for the waiting-time distribution that involves only a finite sum of positive terms; thus, this expression presents no numerical complications, not even for high traffic intensities. For our model, extending Franx's approach is a challenging problem as the representation of various quantities appearing in [66] which are related to the queue length at service initiations is not straightforward.

As we see, the waiting-time distribution in Theorem 8.3 is quite similar to Erlang's expression, so we expect that eventually the solution will suffer from roundoff errors. Furthermore, a significant difference in the numerical computation between the M/D/1 queue and the model described by Recursion (8.1) arises when computing π_0. For any single server queue we know a priori that $\mathbb{P}[W = 0] = 1 - \rho$. In our model, π_0 has to be computed from the normalisation equation, where the numerical complications when calculating the waiting-time distribution become apparent. In particular, as p tends to 1, i.e. as the system behaves almost like an M/D/1 queue, the computation of π_0 becomes more problematic.

As an additional observation, we note that the effects of Lindley's classical recursion and of the Lindley-type recursion (1.1) are quite apparent. The analysis for our model is in a sense separated into two parts: the derivation of the waiting-time distribution in $[0, b)$ and in $[b, \infty)$. In the first part, we see that Equation (8.24) is quite similar to the differential equation (3.32) for the derivation of the waiting-time distribution in case $p = 0$ and B follows a polynomial distribution. Moreover, one could use the same technique to derive a solution, but Equation (8.24) is too simple to call for such means. In the second part, we see the effects of the M/D/1 queue, as we eventually derive F_W in a recursive manner. Furthermore, this model inherits all the numerical difficulties appearing in the classical solution for the M/D/1 queue, plus the additional difficulties of computing π_0. For Lindley's recursion, π_0 is known beforehand, while for the Lindley-type recursion (1.1) π_0 is derived by the normalisation equation. However, for values of p not too close to 1, one can numerically approximate the waiting-time distribution fairly easily, and the results have also been verified by means of simulation. Finally, one could argue that the challenges arising in the analysis of (8.1) are demonstrated also by the fact that the approach for the G/PH case differs significantly from the one followed for the M/D case. The method used in each case does not seem to be successful in handling the other case.

Final remarks

Minor extensions and observations

In this dissertation, we have mainly studied various aspects of the Lindley-type recursion (1.1) and a generalisation of it given by Recursion (8.1). The basic characteristic of both of these recursions is that they do not satisfy the usual assumption made when studying general stochastic recursions. Namely, they are not stochastically monotone increasing in their main argument, which does not allow us to exploit the known duality results between Markov processes and monotone increasing continuous processes (see Asmussen [6, Section IX.4] and Loynes [122]). From the analysis presented in the previous chapters, however, one can draw a few simple conclusions.

As we have observed in various points before (cf. Sections 5.2 and 5.7), the structure of the preparation time distribution F_B (or the lack thereof) is significant in the analysis of this model. If F_A is generally distributed and F_B has some known structure, then we can analyse exactly both the time-dependent and the steady-state behaviour of the system; see Chapter 4. Moreover, we observe that in this case, the methods that can be applied are quite similar to the methods applicable to the standard M/G/1 queuing system. Thus, we observe that we can often learn from the results derived for the M/G/1 queue, since frequently these results can also be derived for the G/M case of our model. In other words, although there seems to be no directly established duality between Lindley's recursion and the Lindley-type recursion (1.1), the experience gathered from Lindley's recursion gives insight on the possibilities arising for (1.1).

Apart from the general observation that Lindley's recursion is potentially didactic when studying this model or trying to extend the results presented in this thesis, there are minor extensions that can be made in almost all chapters we have seen so far. For sake of completeness, in the following we give a non-exhaustive list of such minor extensions. It should be stressed though that most of these minor points are interesting only to the extent that they satisfy one's mathematical curiosity and do not constitute a significant advance of the theory. Key extensions will be presented in the following section.

In Chapter 2, we have studied the stability of the system under the assumption that $\{X_n\}$ is an i.i.d. sequence. If one wishes, however, to introduce any kind of dependencies between preparation and service times, then one should first study the stability of the system under the more general assumption that $\{X_n\}$ is a stationary sequence. For a particular case in Chapter 7 we have seen that the proof is quite straightforward. Moreover, from Foss and Konstantopoulos [65] we know that for any sequence $\{X_n\}$ there exists an equilibrium distribution; see also Section 2.2.2. However, the uniqueness of this distribution and the convergence to it remain to be studied.

The results of Section 3.6 can be extended to mixed-Erlang service times, and

those of Section 4.2.2 regarding the time-dependent waiting-time distribution can be extended to preparation time distributions that belong to class $\mathcal{M}$ (cf. Section 5.3). Another interesting point is the conjecture made in Remark 4.5. Namely, after comparing the performance of the alternating-service system to the one of the non-alternating model, we have observed that the waiting times are not stochastically ordered, but numerical examples suggest that they are ordered with respect to the increasing convex ordering.

In Section 6.2, we have obtained error bounds for the waiting-time distribution in case we approximate either the service-time or the preparation-time distribution. These bounds though, as is observed also in Section 3.3, are not tight, and are limited to the uniform norm. For numerical applications though, one might be interested in deriving tighter bounds and extending these results to weighted norms. This will allow for a greater variety of methods when fitting a distribution to either F_A or F_B. For example many techniques that are used for fitting phase-type distributions to given distributions are based on moment matching. The distance between the two distributions with respect to the uniform norm is not necessarily reduced as the number of moments that are matched increases. However, if we derive error bounds with respect to a norm that involves only the distance between the moments of two distributions, then moment matching techniques are a reasonable choice when fitting distributions.

An obvious extension of the results presented in Chapter 7 is to consider a wide range of other dependence structures than the ones described there. Various kinds of dependencies are application-specific (see, for example, page 142) while others are mathematically interesting, see Nelsen [134]. What is also of interest is to study numerically the effects of the auto- and cross-correlation between the sequences $\{A_n\}$ and $\{B_n\}$ to the waiting-time distribution.

In Chapter 8, we have studied a more general non-increasing Lindley-type recursion, which generalises Recursion (1.1). For this recursion, one could consider the case where the additional sequence $\{Y_n\}$ is i.i.d. and generally distributed. However, the analysis in this case is far from trivial.

For Recursion (1.1), one could also consider applying the moment iteration method for approximating the waiting-time distribution described in De Kok [55]. Some advantages of this method are that it is easy to implement and that it is quite accurate for the G/G/1 queue when the initial distributions have all moments finite. This method eventually matches the two distributions, i.e. the real waiting-time distribution and its approximation, by matching the first two or three moments (although, in principle, the algorithm can be modified to yield better results through iteration of higher moments and fitting distributions to all these moments). This does not guarantee, however, that the approximated waiting-time distribution will be as close as desired (for example, with respect to the uniform norm) to the real waiting-time distribution. Since for our model we know that the waiting-time distribution satisfies a contraction mapping, it seems more advantageous to derive an approximation of F_W by successive iterations of the functional equation (2.1). However, a possible disadvantage of this technique may be that every iteration might increase the complexity of the computation for certain service-time or preparation-

time distributions. The solution to this problem is to approximate one of the two distributions with a distribution that allows for explicit computations. As we have seen in Chapter 6, the approximation error in this case can be controlled.

More interesting extensions usually involve a modification of the model, and will be presented in the following section.

Further research

As we have seen in Section 1.2, the model we have considered in this dissertation applies to a two-carousel system that is operated by a single picker. Two-carousel systems have received some attention in the literature (cf. Section 1.3.5) but many questions remain open. A line of research is directed towards studying the performance of two-carousel systems under various storage-assignment policies (randomised or not), for various pick/travel time strategies and heuristics (sequential picking, nearest-item heuristic, m-step strategies, etc.), for single- or dual-command cycles, and for open- and closed-loop strategies. As explained in Section 1.3.5, two-carousel systems differ in nature and in analysis from the corresponding one-carousel problems even when studied under the same assumptions on the various storage, pick, cycle, and starting-point strategies that are followed. Since two-carousel systems perform in broad terms better than single-carousel systems [87], studying the expected increase of the throughput of the system can help answer questions of financial nature, such as whether the benefits from the increased throughput justify the increased cost of building and operating a two-carousel system.

Another question that arises naturally when considering the two-carousel system, is: "why limit the study to two carousels and not extend the model to multiple carousels?" Consider the situation where a single server (picker) operates three stations (carousels). Apart from the number of stations, all other characteristics of the model remain the same. That is, we again consider an infinite queue of customers (orders) that need to be served, we have again a preparation phase and a service phase in the same station for each customer, and as before, the server is needed only in the service phase and serves all stations cyclically. For three stations, this leads to the recursion

$$W_{n+2} = \max\{0, B_{n+2} - W_{n+1} - A_{n+1} - W_n - A_n\},$$

where now the variables appearing at the right-hand side are not independent of one another, as was the case for all variables appearing at the right-hand side of Recursion (1.1). Although we may assume for convenience that the sequences $\{A_n\}$ and $\{B_n\}$ are independent among them and between them, the waiting times W_n and W_{n+1} are not independent. The state of the system can also be modelled as a two-dimensional Markov chain, where apart from the waiting time of the server for the n-th customer we also need to incorporate the remaining preparation time of the next carousel to be served. Evidently, if the preparation times are assumed to be exponentially distributed, the system (for three or more stations) can be analysed explicitly by similar techniques as the ones applied in Chapter 4.

Naturally, if one considers a system with multiple carousels or stations, one can think about optimisation questions. Namely, as the number of carousels increases, the waiting time of the picker is expected to decrease. After serving a long series of carousels cyclically, when you return to the beginning of the cycle, with high probability the item to be picked will have reached the origin. This implies that an item will have to wait for the picker at the origin more frequently than in the two-carousel system, which means that the throughput of the system decreases. Intuitively, as the number of carousels increases to infinity, the utilisation of the server increases to one, while the throughput of each individual carousel decreases to zero. Given a setting, one might wonder how many carousels a single picker can operate so that we maximise both the throughput of the carousels and the utilisation of the server simultaneously.

It is also interesting to study if the model can be analysed in case there is an arrival process according to which the customers (orders) arrive. For example, if customers arrive according to a Poisson process in front of the service station, where they form an infinite-buffer queue, then undergo a preparation phase as before, and only then receive service from the server, then what can be said for the waiting time of the server? This question can also be combined with the non-alternating system, where the server serves the first customer that has completed the preparation phase, or with Bernoulli-type requests, where the server has to serve with a certain probability at the "first" station and with the complementary probability at the "other" station (potentially waiting for a customer if none is present at the designated station). For each case, one should also consider the stability of the system in case the arrival rate of the customers is less than the throughput of the system with an infinite queue of customers.

In the literature on polling systems, the polling system with two queues where at each queue the server serves exactly one customer before switching to the other queue is often referred to as the *1-limited alternating-service* model. Extending the model of Section 1.4 by introducing an arrival process of the customers as suggested above, is equivalent to studying an 1-limited alternating-service model with switch-over times between the stations (which can be seen as being equivalent to the preparation phase of a customer). The model with two queues, Poisson arrivals, and no switch-over times has first been studied by Eisenberg [58], where the main question studied (as is often the case in the literature on polling systems) is the queue-length distribution. Eisenberg [58] gives the generating function for the stationary joint distribution of the two queue sizes. Cohen and Boxma [48] study the single server queue with two Poissonian arrival streams and no switch-over times. The server handles alternatingly a customer of each queue if the queues are not empty and it is assumed that customers of the same arrival stream have the same service time distribution. It is shown that the determination of the joint queue-length distribution at the departure epoch can be formulated as a Riemann-Hilbert boundary problem that can be completely solved for general service time distributions. Introducing switch-over times increases the complexity of the problem. In Boxma [32] the analysis is extended to include switch-over times of the server between queues, under the restriction that both queues have identical characteristics.

This work is further extended in Boxma and Groenendijk [34], where the authors no longer request that both queues have identical characteristics. It is assumed that service times and switch-over times are generally distributed.

The literature on polling systems with alternating service is not limited to the references above but is rather extensive; see [72, 89, 138] for some references. It seems though, that the question regarding the waiting time of the server for the 1-limited alternating-service polling system with two stations has not been considered outside the scope of this thesis. Thus, introducing an arrival process for the customers in our model complements the existing literature on polling systems and forms a challenging problem. The interesting feature then is that the switch-over time between two queues depends on the current service time.

An extension considered in polling systems is the *k-limited* service policy, where the server switches queues after having served at most k customers in one queue. For an extensive list of references on k-limited polling systems see Van Vuuren and Winands [166]. The main focus of the existing literature is again on the queue-length distribution of all stations. As the authors note in [166], "to this very day, not only hardly any exact results for polling systems with the k-limited service policy have been obtained, but also their derivations give little hope for extensions to more realistic systems". It is worth considering the k-limited service discipline under the exact setting we have established in Section 1.4, where now the focus is on the distribution of the waiting time of the server.

A final open question, not connected directly to applications in carousel systems, has to do with the study of class $\mathcal{M}$ introduced in Section 5.3. We have shown there that this class is at least as big as the class of distributions with rational Laplace transforms. Due to the multiplicative decomposition structure of $\mathcal{M}$, one can easily solve various types of generalised Wiener-Hopf equations within this class. Such equations arise in various areas of research. As an example, we mention the work by Bansal [13] concerning processor sharing queues with bulk arrivals. The processor-sharing queueing system, first introduced by Kleinrock [102], has been of considerable interest and is used extensively to study computer and communication systems. Under this policy, each job receives an equal share of the processor, i.e. if there are n jobs at some time, then each job gets serviced at $1/n$ times the speed of the processor. The remarkable features of processor sharing are its simplicity, the fact that there is no requirement of knowledge of job sizes and its fairness properties (in particular, the expected response time of a job is directly proportional to its size). Bulk arrivals are often used to model the burstiness in the arrival process. Bursty arrivals often occur in modern systems, for example in a web server, usually multiple embedded objects within a web page are requested simultaneously. In [13] the author extends the work of Kleinrock *et al.* [104] on processor sharing queues with bulk arrivals by solving the generalised Wiener-Hopf equation describing the dynamics of the system for distributions with rational Laplace transforms. Should class $\mathcal{M}$ be strictly greater than the class of distributions with rational Laplace transforms, then this work, and many other studies where generalised Wiener-Hopf equations arise, can be directly extended.

Appendix
A remark on Equation (3.32)

Let P and Q be two arbitrary functions on the real numbers. We know that if the linear differential equation of first order

$$\frac{\mathrm{d}}{\mathrm{d}x}f(x) + P(x)f(x) = Q(x)$$

has a solution, then there is a unique solution that satisfies the initial condition $f(x_0) = y_0$, provided that x_0 belongs to the domain of f. However, the above statement is not necessarily true when dealing with differential equations where the argument varies. We have encountered such an equation in Section 3.6; also see Remark 3.4. Here we shall present a simple, yet illustrative, example.

Consider the differential equation

$$f'(x) = \frac{\pi}{2}f(1-x), \qquad x \in \mathbb{R}, \tag{A.1}$$

with the initial condition $f(0) = 0$. Substitute x with $1 - x$ to obtain

$$f'(1-x) = \frac{\pi}{2}f(x),$$

so we readily have that

$$f''(x) = -\frac{\pi}{2}f'(1-x) = -\left(\frac{\pi}{2}\right)^2 f(x).$$

The characteristic equation of this linear differential equation of second order is

$$x^2 + \left(\frac{\pi}{2}\right)^2 = 0,$$

and it has the roots $x = \pm i\pi/2$. Therefore, the general solution to the equation is given by

$$f(x) = c_1 \cos(\frac{\pi x}{2}) + c_2 \sin(\frac{\pi x}{2}).$$

So, from (A.1) we have that

$$-c_1 \sin(\frac{\pi x}{2}) + c_2 \cos(\frac{\pi x}{2}) = c_1 \cos(\frac{\pi(1-x)}{2}) + c_2 \sin(\frac{\pi(1-x)}{2}),$$

or $c_1 = 0$ since the coefficient of c_2 is

$$\cos(\frac{\pi x}{2}) - \sin(\frac{\pi(1-x)}{2}) = 0.$$

Therefore, we have that

$$f(x) = c_2 \sin(\frac{\pi x}{2}),$$

and the initial condition given does not lead towards a *unique* solution for this equation. The conclusion is that with such particular differential equations, a "sufficient" amount of initial conditions does not necessarily guarantee that a unique solution can be obtained. The additional knowledge we had in Section 3.6 that helped overcome such difficulties was that we knew beforehand that a unique solution (i.e. the waiting-time density) *does* exist, and the conditions we derive do not disturb this fact, but arise naturally from the intermediate steps of differentiation.

Samenvatting

In dit proefschrift staat de volgende Lindley-achtige recursie centraal:

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

Deze "niet-stijgende" recursie is belangrijk in de analyse van systemen waarbij een bediende alterneert tussen twee bedieningsstations. Een station biedt ruimte voor één klant. De bediende alterneert tussen beide stations en bedient één klant per keer. Aangenomen wordt dat voortdurend bij beide stations klanten staan te wachten. Zodra een wachtende klant een station betreed, begint de eerste fase van zijn bediening, die bestaat uit een voorbereidende fase. De bediende is hier *niet* bij betrokken: pas nadat de voorbereidende fase is afgerond kan een klant aan de tweede fase van zijn bediening beginnen, welke wordt uitgevoerd door de bediende. Dus de eigenlijke bediening bestaat alleen uit de tweede fase. Het kan voorkomen dat de bediende moet wachten totdat de voorbereiding van de volgende klant is afgelopen. We zijn dan ook geïnteresseerd in de wachttijd van de bediende. Als B_n de voorbereidingstijd is voor de n-de klant en A_n de bedieningstijd is van de n-de klant, dan kan de wachttijd van de bediende voor de $(n+1)$-ste klant beschreven worden door middel van Recursie (1). Een belangrijke observatie is dat deze recursie vrijwel identiek is aan Lindley's recursie. Het enige verschil is het min-teken voor W_n.

Dit model is gemotiveerd door diverse toepassingen waarvan er twee worden besproken in Hoofdstuk 1. De eerste toepassing betreft oog-operaties. De tweede toepassing is gerelateerd aan carousel systemen. Dit soort systemen zijn uitgebreid bestudeerd; Sectie 1.3 geeft een literatuuroverzicht. Verderop in dit hoofdstuk geven we een gedetailleerde modelbeschrijving en noemen we enkele verschillen tussen de analyse van dit model en het standaard wachtrijmodel.

Hoofdstuk 2 bestudeert enkele algemene eigenschappen van Recursie (1), zoals de stabiliteit van het systeem, existentie van een evenwichtsverdeling, convergentie naar deze verdeling als n naar oneindig gaat en het staartgedrag en de covariantie functie van de verdeling van de wachttijd van de bediende.

Een rode draad in dit proefschrift is de afleiding van de evenwichtsverdeling van de wachttijd van de bediende. In de volgende drie hoofdstukken leiden we deze verdeling af onder diverse aannames over de verdeling van de voorbereidingstijd en bedieningstijd van een generieke klant. We bestuderen gevallen die analoog zijn aan de klassieke M/G/1, G/PH/1 en PH/P/1 wachtrijmodellen, waarbij "P" staat voor polynomiale verdelingen. Geïnspireerd door de toepassingen van ons model, bekijken we enkele prestatiematen voor dit systeem, zoals de doorzet. Dit maakt een vergelijk met de prestatie van niet-alternerende systemen mogelijk.

In Hoofdstuk 6 onderzoeken we methoden om de wachttijdverdeling te benaderen door de verdeling van de voorbereidingstijd of bedieningstijd te benaderen met een verdeling die exacte berekeningen mogelijk maakt. We beschrijven hoe zo'n verdel-

ing kan worden gevonden en we geven een bovengrens voor de fout tussen de werkelijke wachttijdverdeling en zijn benadering.

In alle voorgaande hoofdstukken hebben we aangenomen dat alle voorbereidingstijden en bedieningstijden onafhankelijk van elkaar zijn. In Hoofdstuk 7 laten we deze aanname vallen. We onderzoeken twee specifieke vormen van afhankelijkheid tussen deze variabelen. Voor beide vormen leiden we opnieuw de limietverdeling af van de wachttijd van de bediende.

Hoofdstuk 8 analyseert een recursie welke een uitbreiding is van zowel Lindley's recursie als (1). We bekijken, namelijk, de recursie

$$W_{n+1} = \max\{0, B_{n+1} - A_n + Y_n W_n\},$$

met Y_n een stochastische variabele die zowel de waarde 1 als -1 kan aannemen. Voor deze recursie onderzoeken we stabiliteit, en we berekenen de limietverdeling in twee specifieke gevallen, waarmee we de bestaande theorie voor Lindley's recursie en Recursie (1) generaliseren. De analyse maakt duidelijk dat de technieken voor het analyseren van (1) en voor het analyseren Lindley's recursie moeten worden gecombineerd.

Diverse methoden om Lindley's recursie te analyseren zijn ook nuttig voor de analyse van (1). Wanneer we aannemen dat de voorbereidingstijd een fase-type verdeling heeft, dan reduceert de analyse van (1) tot de analyse van een Markovketen met eindige toestandsruimte. Ook kunnen Laplace-transformaties of Wiener-Hopf technieken in diverse gevallen worden toegepast (cf. Sectie 1.6). In andere gevallen moet een niet-standaard differentiaalvergelijking worden opgelost, of moet uitgeweken worden naar een iteratieve benadering van de wachttijdverdeling. In Hoofdstuk 5 dient ook een speciale klasse van verdelingen geïntroduceerd te worden die het mogelijk maakt om een Fredholm vergelijking op te lossen. In de meeste gevallen zijn de resultaten expliciet of kunnen worden weergegeven in termen van de oplossing van een lineair stelsel vergelijkingen, zie bijvoorbeeld Stelling 4.8.

Het proefschrift wordt afgesloten met enkele afsluitende opmerkingen en diverse suggesties voor verder onderzoek.

Bibliography

[1] I. J.-B. F. Adan and V. G. Kulkarni. Single-server queue with Markov-dependent inter-arrival and service times. *Queueing Systems. Theory and Applications*, 45(2):113–134, October 2003. *[130, 140, and 141]*

[2] I. J.-B. F. Adan and Y. Q. Zhao. Analyzing GI/E_r/1 queues. *Operations Research Letters*, 19(4):183–190, October 1996. *[22 and 89]*

[3] H. Albrecher and O. J. Boxma. On the discounted penalty function in a Markov-dependent risk model. *Insurance: Mathematics & Economics*, 37(3):650–672, December 2005. *[141]*

[4] J. E. Angus. Classroom note: The inspection paradox inequality. *SIAM Review*, 39(1):95–97, March 1997. *[40]*

[5] S. Asmussen. Phase-type representations in random walk and queueing problems. *Annals of Probability*, 20(2):772–789, April 1992. *[89]*

[6] S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, New York, 2003. *[16, 17, 20, 22, 26, 28, 98, 99, 108, and 163]*

[7] S. Asmussen and O. Kella. A multi-dimensional martingale for Markov additive processes and its applications. *Advances in Applied Probability*, 32(2):376–393, June 2000. *[140]*

[8] S. Asmussen, O. Nerman, and M. Olson. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23(4):419–441, 1996. *[122]*

[9] S. Asmussen and S. Schock Petersen. Ruin probabilities expressed in terms of storage processes. *Advances in Applied Probability*, 20(4):913–916, December 1989. *[18]*

[10] S. Asmussen and K. Sigman. Monotone stochastic recursions and their duals. *Probability in the Engineering and Informational Sciences*, 10(1):1–20, January 1996. *[18]*

[11] F. Avram and M. Usábel. Ruin probabilities and deficit for the renewal risk model with phase-type interarrival times. *Astin Bulletin*, 34(2), November 2004. *[141]*

[12] F. L. Baccelli and P. Brémaud. *Elements of Queueing Theory: Palm-Martingale Calculus and Stochastic Recurrences*. Springer-Verlag, New York, 1994. *[18]*

[13] N. Bansal. Analysis of the M/G/1 processor-sharing queue with bulk arrivals. *Operations Research Letters*, 31(5):401–405, September 2003. *[167]*

[14] J. J. Bartholdi, III and L. K. Platzman. Retrieval strategies for a carousel conveyor. *IIE Transactions*, 18:166–173, June 1986. *[4, 7, 9, and 13]*

[15] V. E. Beneš. On queues with Poisson arrivals. *The Annals of Mathematical Statistics*, 28(3):670–677, September 1957. *[27]*

[16] G. Bengü. An optimal storage assignment for automated rotating carousels. *IIE Transactions*, 27(1):105–107, February 1995. *[6 and 10]*

[17] D. Bertsekas and R. Gallager. *Data Networks*. Prentice-Hall, Englewood Cliffs, New Jersey, 1992. *[69]*

[18] U. N. Bhat. Queueing systems with first-order dependence. *Opsearch. The Journal of the Operational Research Society of India*, 6:1–24, 1969. *[140]*

[19] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*, volume 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1987. *[34 and 35]*

[20] J. R. Bitner and C. K. Wong. Optimal and near-optimal scheduling algorithms for batched processing in linear storage. *SIAM Journal on Computing*, 8(4):479–498, November 1979. *[6]*

[21] J. P. C. Blanc. Computation of autocorrelations of interdeparture times by numerical transform inversion. *Annals of Operations Research*, 112(1–4):83–100, April 2002. *[27]*

[22] J. P. C. Blanc. Numerical transform inversion for autocorrelations of waiting times. In H. Fleuren, D. den Hertog, and P. Kort, editors, *Operations Research Proceedings 2004*, pages 297–304, Tilburg, 1–3 September 2004. Springer. *[27 and 101]*

[23] N. Blomqvist. The covariance function of the M/G/1 queuing system. *Skandinavisk Aktuarietidskrift*, 50:157–174, 1967. *[27 and 101]*

[24] N. Blomqvist. Estimation of waiting time parameters in the GI/G/1 queueing system, part I: General results. *Skandinavisk Aktuarietidskrift*, 51:178–197, 1968. *[27]*

[25] N. Blomqvist. Estimation of waiting time parameters in the GI/G/1 queueing system, part II: Heavy traffic approximations. *Skandinavisk Aktuarietidskrift*, 52:125–136, 1969. *[27]*

[26] A. A. Borovkov. *Stochastic Processes in Queueing Theory*. Number 4 in Applications of Mathematics. Springer-Verlag, New York, 1976. *[120]*

[27] A. A. Borovkov. *Asymptotic Methods in Queueing Theory*. Wiley, Chichester, 1984. *[143]*

[28] A. A. Borovkov. *Ergodicity and Stability of Stochastic Processes*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 1998. *[18 and 120]*

[29] A. A. Borovkov and S. Foss. Stochastically recursive sequences. *Siberian Advances in Mathematics*, 2(1):16–81, 1992. *[18]*

[30] S. C. Borst, O. J. Boxma, and M. B. Combé. Collection of customers: A correlated M/G/1 queue. *Performance Evaluation Review*, 20(1):47–59, June 1992. *[141 and 142]*

[31] S. C. Borst, O. J. Boxma, and M. B. Combé. An M/G/1 queue with customer collection. *Communications in Statistics. Stochastic Models*, 9(3):341–371, 1993. *[141 and 142]*

[32] O. J. Boxma. Two symmetric queues with alternating service and switching times. In *Performance '84 (Paris, 1984)*, pages 409–431. North-Holland, Amsterdam, 1985. *[166]*

[33] O. J. Boxma and M. B. Combé. The correlated M/G/1 queue. *Archiv für Elektronik und Übertragungstechnik*, 47(5/6):330–335, 1993. *[136, 140, and 142]*

[34] O. J. Boxma and W. P. Groenendijk. Two queues with alternating service and switching times. In O. Boxma and R. Syski, editors, *Queueing Theory and its Applications – Liber Amicorum for J.W. Cohen*, volume 7 of *CWI monographs*, pages 261–282. North-Holland, Amsterdam, 1988. *[167]*

[35] O. J. Boxma and D. Perry. A queueing model with dependence between service and interarrival times. *European Journal of Operational Research*, 128(3):611–624, February 2001. *[142]*

[36] O. J. Boxma and M. Vlasiou. On queues with service and interarrival times depending on waiting times. Technical Report 2006-008, Eurandom, Eindhoven, The Netherlands, 2006. Available at `http://www.eurandom.nl`. *[24 and 143]*

[37] A. Brandt. The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Advances in Applied Probability*, 18(1):211–220, March 1986. *[144 and 145]*

[38] A. Brandt, P. Franken, and B. Lisek. *Stationary Stochastic Models*, volume 78 of *Mathematische Lehrbücher und Monographien, II. Abteilung: Mathematische Monographien*. Akademie-Verlag, Berlin, 1990. *[18]*

[39] L. Breiman. On some limit theorems similar to the arc-sin law. *Theory of Probability and its Applications*, 10(2):323–331, 1965. *[36]*

[40] L. Breiman. *Probability*, volume 7 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, 1992. *[35]*

[41] P. H. Brill. Single-server queues with delay-dependent arrival streams. *Probability in the Engineering and Informational Sciences*, 2:231–247, 1988. *[144]*

[42] P. H. Brill and M. J. M. Posner. A two server queue with nonwaiting customers receiving specialized service. *Management Science*, 27(8):914–925, August 1981. *[144]*

[43] J. R. Callahan. A queue with waiting time dependent service times. *Naval Research Logistics Quarterly*, 20:321–324, 1973. *[144]*

[44] I. Cidon, R. Guérin, A. Khamisy, and M. Sidi. On queues with interarrival times proportional to service times. *Probability in the Engineering and Informational Sciences*, 10(1):87–107, 1996. *[141 and 142]*

[45] D. B. H. Cline and G. Samorodnitsky. Subexponentiality of the product of independent random variables. *Stochastic Processes and their Applications*, 49(1):75–98, January 1994. *[36]*

[46] J. W. Cohen. *The Single Server Queue*. North-Holland Publishing Co., Amsterdam, 1982. *[17, 20, 23, 67, 98, 99, 101, 118, 147, 151, and 153]*

[47] J. W. Cohen. On periodic Pollaczek waiting time processes. In *Athens Conference on Applied Probability and Time Series Analysis, Vol. I (1995)*, volume 114 of *Lecture Notes in Statististics*, pages 361–378. Springer, New York, 1996. *[141]*

[48] J. W. Cohen and O. J. Boxma. The M/G/1 queue with alternating service formulated as a Riemann-Hilbert problem. In *Performance '81 (Amsterdam, 1981)*, pages 181–199. North-Holland, Amsterdam, 1981. *[166]*

[49] M. B. Combé and O. J. Boxma. BMAP modelling of a correlated queue. In J. Walrand, K. Bagchi, and G. W. Zobrist, editors, *Network Performance Modeling and Simulation*, pages 177–196. CRC, first edition, April 1998. *[141 and 142]*

[50] B. W. Conolly. The waiting time process for a certain correlated queue. *Operations Research*, 16(5):1006–1015, September 1968. *[142]*

[51] B. W. Conolly and Q. H. Choo. The waiting time process for a generalized correlated queue with exponential demand and service. *SIAM Journal on Applied Mathematics*, 37(2):263–275, October 1979. *[141 and 142]*

[52] B. W. Conolly and N. Hadidi. A correlated queue. *Journal of Applied Probability*, 6(1):122–136, April 1969. *[142]*

[53] D. J. Daley. The serial correlation coefficients of waiting times in a stationary single server queue. *Journal of the Australian Mathematical Society*, 8:683–699, 1968. *[27 and 44]*

[54] V. G. Daniele. The Wiener-Hopf technique for impenetrable wedges having arbitrary aperture angle. *SIAM Journal on Applied Mathematics*, 63(4):1442–1460, 2003. *[19]*

[55] A. G. De Kok. A moment-iteration method for approximating the waiting-time characteristics of the GI/G/1 queue. *Probability in the Engineering and Informational Sciences*, 3:273–287, 1989. *[164]*

[56] P. Diaconis and D. Freedman. Iterated random functions. *SIAM Review*, 41(1):45–76, March 1999. *[143]*

[57] P. J. Egbelu and C.-T. Wu. Relative positioning of a load extractor for a storage carousel. *IIE Transactions*, 30(4):301–317, April 1998. *[10]*

[58] M. Eisenberg. Two queues with alternating service. *SIAM Journal on Applied Mathematics*, 36(2):287–303, April 1979. *[166]*

[59] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*, volume 33 of *Applications of Mathematics*. Springer Verlag, Berlin, 1997. *[26]*

[60] C. R. Emerson and D. S. Schmatz. Results of modeling an automated warehouse system. *Industrial Engineering*, 13(8):28–32, cont. on p. 90, August 1981. *[11 and 12]*

[61] A. K. Erlang. The theory of probabilities and telephone conversations. In E. Brockmeyer, H. L. Halstrøm, and A. Jensen, editors, *The Life and Works of A. K. Erlang*, number 6 in Applied Mathematics and Computing Machinery Series, pages 131–137. Acta Polytechnica Scandinavica, second edition, 1960. English translation. First published in "Nyt Tidsskrift for Matematik" B, Vol. 20 (1909), p. 33. *[160]*

[62] A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31(3–4):245–279, January 1998. *[122]*

[63] W. Feller. *An Introduction to Probability Theory and its Applications, Vol. II*. John Wiley & Sons Inc., New York, second edition, 1971. *[122]*

[64] K. W. Fendick, V. R. Saksena, and W. Whitt. Dependence in packet queues. *IEEE Transactions on Communications*, 37(11):1173–1183, 1989. *[140]*

[65] S. Foss and T. Konstantopoulos. An overview of some stochastic stability methods. *Journal of the Operations Research Society of Japan*, 47(4):275–303, 2004. *[29 and 163]*

[66] G. J. Franx. A simple solution for the M/D/c waiting time distribution. *Operations Research Letters*, 29(5):221–229, December 2001. *[161]*

[67] R. F. Gebhard. A limiting distribution of an estimate of mean queue length. *Operations Research*, 11(6):1000–1003, November–December 1963. *[27]*

[68] H. U. Gerber and E. S. W. Shiu. The time value of ruin in a Sparre Andersen model. *North American Actuarial Journal*, 9(2):49–84, April 2005. *[141]*

[69] J. B. Ghosh and C. E. Wells. Optimal retrieval strategies for carousel conveyors. *Mathematical and Computer Modelling*, 16(10):59–70, October 1992. *[4, 8, and 9]*

[70] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, Reading, MA, January 1989. *[13]*

[71] C. M. Goldie. Implicit renewal theory and tails of solutions of random equations. *The Annals of Applied Probability*, 1(1):126–166, February 1991. *[143]*

[72] W. P. Groenendijk. *Conservation Laws in Polling Systems*. Ph. D. Thesis, Utrecht University, 1990. *[167]*

[73] M. Guenov and R. Raeside. Real time optimization of man on board order picking. In J. A. White, editor, *Proceedings of the 10th International Conference on Automation in Warehousing*, pages 87–93, Dallas, Texas, October 1989. IFS(Publications). *[10]*

[74] J.-W. Ha and H. Hwang. Class-based storage assignment policy in carousel system. *Computers & Industrial Engineering*, 26(3):489–499, July 1994. *[4]*

[75] N. Hadidi. Queues with partial correlation. *SIAM Journal on Applied Mathematics*, 40(3):467–475, June 1981. *[142]*

[76] N. Hadidi. Further results on queues with partial correlation. *Operations Research*, 33(1):203–209, January–February 1985. *[142]*

[77] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, second edition, 1952. *[6]*

[78] C. M. Harris. Queues with state-dependent stochastic service rates. *Operations Research*, 15(1):117–130, January–February 1967. *[144]*

[79] C. M. Harris. Some results for bulk arrival queues with state-dependent service times. *Management Science: Theory Series*, 16(5):313–326, January 1970. *[144]*

[80] J. M. Harrison and S. I. Resnick. The recurrence classification of risk and storage processes. *Mathematics of Operations Research*, 3:57–66, 1977. *[18]*

[81] E. Hassini and R. G. Vickson. A two-carousel storage location problem. *Computers & Operations Research*, 30(4):527–539, April 2003. *[4, 13, and 45]*

[82] H. Heffes. A class of data traffic processes – covariance function characterization and related queueing results. *The Bell System Technical Journal*, 59(6):897–929, July–August 1980. *[140]*

[83] H. Heffes and D. M. Lucantoni. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications*, 4(6):856–868, September 1986. Special Issue on Congestion Control in High-Speed Packet Switched Networks. *[140]*

[84] A. Horváth and M. Telek. Approximating heavy tailed behavior with phase type distributions. In G. Latouche and P. G. Taylor, editors, *Advances in Algorithmic Methods for Stochastic Models, Proceedings of the Third International Conference on Matrix Analytic Methods*, pages 191–214, Leuven, Belgium, June 2000. Notable Publications. *[121]*

[85] H. Hwang and J.-W. Ha. Cycle time models for single/double carousel system. *International Journal of Production Economics*, 25:129–140, 1991. *[4, 11, and 12]*

[86] H. Hwang and J.-W. Ha. An optimal boundary for two class-based storage assignment policy in carousel system. *Computers & Industrial Engineering*, 27(1–4):87–90, September 1994. *[6]*

[87] H. Hwang, C.-S. Kim, and K.-H. Ko. Performance analysis of carousel systems with double shuttle. *Computers & Industrial Engineering*, 36(2):473–485, April 1999. *[11, 12, and 165]*

[88] H. Hwang, Y.-K. Song, and K.-H. Kim. The impacts of acceleration/deceleration on travel time models for carousel systems. *Computers & Industrial Engineering*, 46(2):253–265, April 2004. *[10]*

[89] O. C. Ibe. Analysis of polling systems with mixed service disciplines. *Communications in Statistics. Stochastic Models*, 6(4):667–689, 1990. *[167]*

[90] V. I. Istrăţescu. *Fixed Point Theory*, volume 7 of *Mathematics and its Applications*. D. Reidel Publishing Co., Dordrecht, 1981. *[33]*

[91] D. P. Jacobs, J. C. Peck, and J. S. Davis. A simple heuristic for maximizing service of carousel storage. *Computers & Operations Research*, 27(13):1351–1356, November 2000. *[4 and 7]*

[92] P. Jacquet. Subexponential tail distribution in LaPalice queues. *Performance Evaluation Review*, 20(1):60–69, June 1992. *[144]*

[93] M. A. Johnson and M. R. Taaffe. Matching moments to phase distributions: Mixtures of Erlang distributions of common order. *Communications in Statistics. Stochastic Models*, 5(4):711–743, 1989. *[121]*

[94] M. A. Johnson and M. R. Taaffe. Matching moments to phase distributions: Density function shapes. *Communications in Statistics. Stochastic Models*, 6(2):283–306, 1990. *[121]*

[95] M. A. Johnson and M. R. Taaffe. Matching moments to phase distributions: Nonlinear programming approaches. *Communications in Statistics – Stochastic Models*, 6(2):259–281, 1990. *[121]*

[96] V. Kalashnikov. Stability bounds for queueing models in terms of weighted metrics. In Y. Suhov, editor, *Analytic Methods in Applied Probability*, volume 207 of *American Mathematical Society Translations Ser. 2*, pages 77–90. American Mathematical Society, Providence, RI, 2002. *[18 and 120]*

[97] V. Kalashnikov and R. Norberg. Power tailed ruin probabilities in the presence of risky investments. *Stochastic Processes and their Applications*, 98(2):211–228, April 2002. *[144]*

[98] M. Keane, A. G. Konheim, and I. Meilijson. The organ pipe permutation. *SIAM Journal on Computing*, 13(3):531–540, 1984. *[6]*

[99] F. P. Kelly. *Reversibility and Stochastic Networks*. Wiley, Chichester, 1979. *[97]*

[100] D. G. Kendall. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics*, 24(3):338–354, September 1953. *[16]*

[101] B. Kim. Maximizing service of carousel storage. *Computers & Operations Research*, 32(4):767–772, April 2005. *[7]*

[102] L. Kleinrock. Time-shared systems: A theoretical treatment. *Journal of the Association for Computing Machinery*, 14:242–261, 1967. *[167]*

[103] L. Kleinrock. *Queueing Systems, Vol. 2: Computer Applications*. Wiley, New York, 1976. *[69]*

[104] L. Kleinrock, R. Muntz, and E. Rodemich. The processor-sharing queueing model for ime-shared systems with bulk arrivals. *Networks*, 1:1–13, 1971. *[167]*

[105] E. Koenigsberg. Analysis of the efficiency of carousel and tote-stacker performance. In J. White, editor, *Proceedings of the 7th International Conference on Automation in Warehousing*, pages 173–183, San Francisco, California, October 1986. Springer. *[12, 13, and 45]*

[106] D. Korshunov. On distribution tail of the maximum of a random walk. *Stochastic Processes and their Applications*, 72(1):97–103, December 1997. *[26]*

[107] A. Lang and J. L. Arthur. Parameter approximation for phase-type distributions. In S. Chakravarty and A. S. Alfa, editors, *Matrix-Analytic Methods in Stochastic Models*, volume 183 of *Lecture Notes in Pure and Applied Mathematics*, pages 151–206. Marcel Dekker, New York, 1996. *[122]*

[108] G. M. Laslett, D. B. Pollard, and R. L. Tweedie. Techniques for establishing ergodic and recurrence properties of continuous-valued Markov chains. *Naval Research Logistics Quarterly*, 25(3):455–472, 1978. *[144]*

[109] E. L. Lehmann. Ordered families of distributions. *The Annals of Mathematical Statistics*, 26(3):399–419, September 1955. *[92]*

[110] A. J. Lemoine. Waiting time and workload in queues with periodic Poisson input. *Journal of Applied Probability*, 26(2):390–397, June 1989. *[141]*

[111] C.-L. Li and G. Wan. Improved algorithm for maximizing service of carousel storage. *Computers & Operations Research*, 32(8):2147–2150, August 2005. *[7]*

[112] S. Li and J. Garrido. On ruin for the Erlang(n) risk process. *Insurance: Mathematics & Economics*, 34(3):391–408, June 2004. *[141]*

[113] S. Li and J. Garrido. On a general class of renewal risk process: Analysis of the Gerber-Shiu function. *Advances in Applied Probability*, 37(3):836–856, 2005. *[141]*

[114] W. K. Lim, J. J. Bartholdi, and L. K. Platzman. Storage schemes for carousel conveyors under real time control. Material Handling Research Center Technical Report MHRC-TR-85-10, Georgia Institute of Technology, 1985. *[6]*

[115] D. V. Lindley. The theory of queues with a single server. *Proceedings Cambridge Philosophical Society*, 48:277–289, 1952. *[16 and 143]*

[116] N. Litvak. *Collecting n Items Randomly Located on a Circle*. Ph.D. Thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, January 2001. Available at http://alexandria.tue.nl/extra2/200210141.pdf. *[4 and 7]*

[117] N. Litvak. Optimal picking of large orders in carousel systems. *Operations Research Letters*, 34(2):219–227, March 2006. *[8]*

[118] N. Litvak and I. J.-B. F. Adan. The travel time in carousel systems under the nearest item heuristic. *Journal of Applied Probability*, 38(1):45–54, March 2001. *[7]*

[119] N. Litvak and I. J.-B. F. Adan. On a class of order pick strategies in paternosters. *Operations Research Letters*, 30(6):377–386, December 2002. *[8]*

[120] N. Litvak, I. J.-B. F. Adan, J. Wessels, and W. H. M. Zijm. Order picking in carousel systems under the nearest item heuristic. *Probability in the Engineering and Informational Sciences*, 15(2):135–164, April 2001. *[7]*

[121] N. Litvak and W. R. Van Zwet. On the minimal travel time needed to collect n items on a circle. *The Annals of Applied Probability*, 14(2):881–902, May 2004. *[8]*

[122] R. M. Loynes. On a property of the random walks describing simple queues and dams. *Journal of the Royal Statistical Society. Series B (Methodological)*, 27(1):125–129, 1965. *[163]*

[123] D. M. Lucantoni. New results on the single-server queue with a batch Markovian arrival process. *Stochastic Models*, 7:1–46, 1964. *[141]*

[124] C. J. Malmborg and K. Bhaskaran. A revised proof of optimality for the cube-per-order index rule for stored item location. *Applied Mathematical Modelling*, 14(2):87–95, February 1990. *[6]*

[125] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, March 1947. *[92]*

[126] M. Masujima. *Applied Mathematical Methods in Theoretical Physics*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2005. *[49 and 107]*

[127] L. F. McGinnis, M. H. Han, and J. A. White. Analysis of rotary rack operations. In J. White, editor, *Proceedings of the 7th International Conference on Automation in Warehousing*, pages 165–171, San Francisco, California, October 1986. Springer. *[10]*

[128] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Communications and Control Engineering Series. Springer-Verlag London Ltd., London, 1993. *[46]*

[129] S. G. Mikhlin. *Integral Equations and their Applications to Certain Problems in Mechanics, Mathematical Physics and Technology*, volume 4 of *International Series of Monographs on Pure and Applied Mathematics*. Pergamon Press, London, 1957. Translation of the 2nd Russian edition by A. H. Armstrong. *[107 and 113]*

[130] T. Mikosch. Regular variation, subexponentiality and their applications in probability theory. Technical Report 1999-013, Eurandom, Eindhoven, The Netherlands, 1999. Available at http://www.math.ku.dk/∼mikosch/preprint. *[34]*

[131] C. R. Mitchell, A. S. Paulson, and A. A. Beswick. The effect of correlated exponential service times on single server tandem queues. *Naval Research Logistics Quarterly*, 24(1):95–112, 1977. *[142]*

[132] P. M. Morse. Stochastic properties of waiting lines. *Journal of the Operations Research Society of America*, 3(3):255–261, August 1955. *[27]*

[133] V. I. Mudrov. Queueing with "impatient" customers and variable service times, linearly dependent on the queueing time of the customer. *Problems of Cybernetics*, 5:375–378, 1964. English translation. *[144]*

[134] R. B. Nelsen. *Introduction to Copulas*. Springer Series in Statistics. Springer, New York, second edition, 2006. *[164]*

[135] B. Noble. *Methods Based on the Wiener-Hopf Technique for the Solution of Partial Differential Equations*, volume 7 of *International Series of Monographs on Pure and Applied Mathematics*. Pergamon Press, New York, 1958. *[18, 19, 21, 106, and 107]*

[136] R. Norberg. Ruin problems with assets and liabilities of diffusion type. *Stochastic Processes and their Applications*, 81(2):255–269, June 1999. *[144]*

[137] T. Osogami. *Analysis of Multi-server Systems via Dimensionality Reduction of Markov Chains*. Ph.D. Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, June 2005. Available at http://www.cs.cmu.edu/∼osogami/. *[121]*

[138] T. Ozawa. Analysis of a single server model with two queues having different service disciplines. *Electronics and Communications in Japan. Part III: Fundamental Electronic Science*, 73(3):18–27, 1990. *[167]*

[139] A. G. Pakes. The serial correlation coefficients of waiting times in the stationary GI/M/1 queue. *The Annals of Mathematical Statistics*, 42(5):1727–1734, October 1971. *[27 and 101]*

[140] B. C. Park, J. Y. Park, and R. D. Foley. Carousel system performance. *Journal of Applied Probability*, 40(3):602–612, September 2003. *[13, 23, 45, 46, 47, 48, 51, 55, and 60]*

[141] M. J. M. Posner. Single-server queues with service times depending on waiting time. *Operations Research*, 21(2):610–616, March–April 1973. *[144]*

[142] N. U. Prabhu. On the ruin problem of collective rist theory. *Annals of Mathematical Statistics*, 32:757–764, 1961. *[18]*

[143] A. Riska, V. Diev, and E. Smirni. An EM-based technique for approximating long-tailed data sets with PH distributions. *Performance Evaluation*, 55(1–2):147–164, January 2004. *[121]*

[144] T. Rolski. Approximation of periodic queues. *Advances in Applied Probability*, 19(3):691–707, September 1987. *[141]*

[145] T. Rolski. Relationships between characteristics in periodic Poisson queues. *Queueing Systems. Theory and Applications*, 4(1):17–26, March 1989. *[141]*

[146] M. Rosenshine. Queues with state-dependent service times. *Transportation Research*, 1(2):97–104, August 1967. *[144]*

[147] S. M. Ross. *Stochastic Processes*. Wiley, New York, second edition, 1996. *[160]*

[148] B. Rouwenhorst, J. P. Van den Berg, G. J. Van Houtum, and W. H. M. Zijm. Performance analysis of a carousel system. In R. J. Graven, L. F. McGinnis, D. J. Medeiros, R. E. Ward, and M. R. Wilhelm, editors, *Progress in Material Handling Research: 1996*, pages 495–511. The Material Handling Institute, Charlotte, NC, 1996. *[8 and 9]*

[149] R. Schassberger. *Warteschlangen*. Springer-Verlag, Wien, 1973. *[58, 88, 121, 134, and 146]*

[150] H. L. Seal. Risk theory and the single server queue. *Mitteilungen der Vereinigung schweizerischer Versicherungsmathematiker*, 72(Heft 2):171–178, 1972. *[18]*

[151] K. Sigman. Appendix: A primer on heavy-tailed distributions. *Queueing Systems. Theory and Applications*, 33(1–3), March 1999. Queues with heavy-tailed distributions. *[34]*

[152] S. R. Smits, M. Wagner, and T. G. De Kok. Determination of an order-up-to policy in the stochastic economic lot scheduling model. *International Journal of Production Economics*, 90(3):377–389, August 2004. *[141]*

[153] D. Spee. Automatic order picking system with horizontal racks. In R. J. Graven, L. F. McGinnis, D. J. Medeiros, R. E. Ward, and M. R. Wilhelm, editors, *Progress in Material Handling Research: 1996*, pages 545–550. The Material Handling Institute, Charlotte, NC, 1996. *[10]*

[154] K. Sriram and W. Whitt. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE Journal on Selected Areas in Communications*, 4:833–846, 1986. *[140]*

[155] H. I. Stern. Parts location and optimal picking rules for a carousel conveyor automatic storage and retrieval system. In J. White, editor, *Proceedings of the 7th International Conference on Automation in Warehousing*, pages 185–193, San Francisco, California, October 1986. Springer. *[4, 6, 8, 9, and 13]*

[156] S. Sugawara and M. Takahashi. On some queues occurring in an integrated iron and steel works. *Journal of Operations Research Society of Japan*, 8(1):16–23, September 1965. *[144]*

[157] R. Szekli. *Stochastic Ordering and Dependence in Applied Probability*. Springer, New York, 1995. *[92 and 97]*

[158] L. Takács. *Introduction to the Theory of Queues*. Oxford University Press, Oxford, 1962. *[69]*

[159] Q. Tang and G. Tsitsiashvili. Precise estimates for the ruin probability in finite horizon in a discrete-time model with heavy-tailed insurance and financial risks. *Stochastic Processes and their Applications*, 108(2):299–325, December 2003. *[144]*

[160] H. C. Tijms. *Stochastic Models: An Algorithmic Approach*. John Wiley & Sons, Chichester, 1994. *[121]*

[161] H. C. Tijms. *A First Course in Stochastic Models*. John Wiley & Sons, Chichester, 2003. *[65]*

[162] E. C. Titchmarsh. *Theory of Functions*. Oxford University Press, London, second edition, 1968. *[137, 148, and 155]*

[163] F. G. Tricomi. *Integral Equations*. Dover Publications Inc., New York, fifth printing edition, March 1985. *[107]*

[164] J. P. Van den Berg. Multiple order pick sequencing in a carousel system: A solvable case of the rural postman problem. *Journal of the Operational Research Society*, 47(12):1504–1515, December 1996. *[4, 9, and 10]*

[165] J. P. Van den Berg. A literature survey on planning and control of warehousing systems. *IIE Transactions*, 31(8):751–762, August 1999. *[9]*

[166] M. Van Vuuren and E. M. M. Winands. Iterative approximation of k-limited polling systems. Technical Report 2006-06, Eindhoven University of Technology, May 2006. Available at http://www.win.tue.nl/math/bs/spor/. *[167]*

[167] W. Vervaat. On a stochastic difference equation and a representation of non-negative infinitely divisible random variables. *Advances in Applied Probability*, 11(4):750–783, December 1979. *[144]*

[168] R. G. Vickson and A. Fujimoto. Optimal storage locations in a carousel storage and retrieval system. *Location Science*, 4(4):237–245, 1996. *[6]*

[169] M. Vlasiou. A non-increasing Lindley-type equation. Technical Report 2005-015, Eurandom, Eindhoven, The Netherlands, 2005. Available at http://www.eurandom.nl. *[23, 24, 25, and 104]*

[170] M. Vlasiou and I. J.-B. F. Adan. An alternating service problem. *Probability in the Engineering and Informational Sciences*, 19(4):409–426, October 2005. *[23, 70, and 151]*

[171] M. Vlasiou and I. J.-B. F. Adan. Exact solution to a Lindley-type equation on a bounded support. *Operations Research Letters*, 2006. To appear. Available at http://www.eurandom.nl. *[23, 24, 49, and 119]*

[172] M. Vlasiou, I. J.-B. F. Adan, O. J. Boxma, and J. Wessels. Throughput analysis of two carousels. Technical Report 2003-037, Eurandom, Eindhoven, The Netherlands, 2003. Available at http://www.eurandom.nl. *[48]*

[173] M. Vlasiou, I. J.-B. F. Adan, and J. Wessels. A Lindley-type equation arising from a carousel problem. *Journal of Applied Probability*, 41(4):1171–1181, December 2004. *[4, 23, and 48]*

[174] M. Vlasiou and B. Zwart. Time-dependent behaviour of an alternating service queue. Technical Report 2005-061, Eurandom, Eindhoven, The Netherlands, 2005. Available at http://www.eurandom.nl. *[23, 25, and 71]*

[175] M. Wagner and S. R. Smits. A local search algorithm for the optimization of the stochastic economic lot scheduling problem. *International Journal of Production Economics*, 90(3):391–402, August 2004. *[141]*

[176] Y.-W. Wan and R. W. Wolff. Picking clumpy orders on a carousel. *Probability in the Engineering and Informational Sciences*, 18(1):1–11, January 2004. *[8]*

[177] D. J. Weiss. Computer controlled carousels. In *Proceedings of the 3rd International Conference on Automation in Warehousing*, pages 413–418, Stratford-upon-Avon, UK, June 1980. IFS(Publications). *[4]*

[178] U.-P. Wen. The order picking problem in the carousel systems. In J. White, editor, *Proceedings of the 7th International Conference on Automation in Warehousing*, pages 195–199, San Francisco, California, October 1986. Springer. *[8]*

[179] U.-P. Wen and D.-T. Chang. Picking rules for a carousel conveyor in an automated warehouse. *OMEGA: The International Journal of Management Science*, 16(2):145–151, 1988. *[8]*

[180] U.-P. Wen, J. T. Lin, and D.-T. Chang. Order picking for a two-carousel-single-server system in an automated warehouse. In J. A. White, editor, *Proceedings of the 10th International Conference on Automation in Warehousing*, pages 87–93, Dallas, Texas, October 1989. IFS(Publications). *[12]*

[181] W. Whitt. Queues with service times and interarrival times depending linearly and randomly upon waiting times. *Queueing Systems. Theory and Applications*, 6(1):335–351, December 1990. *[144 and 145]*

[182] D.-H. Yeh. A note on "A simple heuristic for maximizing service of carousel storage". *Computers & Operations Research*, 29(11):1605–1608, September 2002. *[4 and 7]*

[183] B. Zwart. *Queueing Systems with Heavy Tails*. Ph. D. Thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2001. Available at http://alexandria.tue.nl/extra2/200112999.pdf. *[34]*

About the author

Maria Vlasiou was born in Drama, Greece, on April 17, 1980. She studied Mathematics at the Aristotle University of Thessaloniki, where she graduated with honours from the Faculty of Exact Sciences in July 2002. From September 2001 until March 2002 she followed in parallel the Masters programme of the Faculty of Mathematics at the University of Bielefeld, Germany.

Maria started her Ph. D. research programme in September 2002 at EURANDOM, Eindhoven. Her research programme was supported by EURANDOM and by a scholarship for post-graduate studies abroad from the Legacy of L. Athanasoula awarded by the Aristotle University of Thessaloniki. She defends her thesis on September 25, 2006 at the Eindhoven University of Technology. Maria plans to continue as a Research Engineer II at the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Institute of Technology, Atlanta.