UNIVERSITY OF AMSTERDAM

Appointment Scheduling in Healthcare

A. Kuiper

Appointment Scheduling in Healthcare

Alex Kuiper

Instituut voor Bedrijfs- en Industriële Statistiek

Publisher IBIS UvA, Amsterdam - www.ibisuva.nl
Printed by Gildeprint, Enschede - www.gildeprint.nl
Cover design Remco Wetzels - www.remcowetzels.nl
ISBN 978-94-6233-319-2

Appointment Scheduling in Healthcare

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op donderdag 30 juni 2016, te 12:00 uur

door

Alex Kuiper

geboren te Amsterdam

Promotiecommissie

Promotores

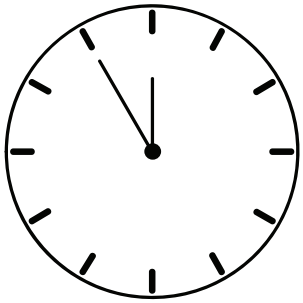
Prof. dr. M.R.H. Mandjes	Universiteit van Amsterdam
Prof. dr. J. de Mast	Universiteit van Amsterdam

Overige leden

Dr. T. Çayırlı	Özyeğin University
Prof. dr. R.J.M.M. Does	Universiteit van Amsterdam
Prof. dr. ir. E.W. Hans	Universiteit van Twente
Prof. dr. R. Núñez-Queija	Universiteit van Amsterdam
Prof. dr. M. Salomon	Universiteit van Amsterdam
Prof. dr. ir. M.F. van Assen	Tilburg University

Faculteit Economie en Bedrijfskunde

Voor mijn vader Cas Kuiper



Contents

1 Introduction **1**
- 1.1 Healthcare operations . . . 2
- 1.2 Dealing with variation in healthcare operations . . . 3
- 1.3 Literature review . . . 4
  - 1.3.1 Static versus dynamic appointment scheduling . . . 5
  - 1.3.2 Situational characteristics in appointment scheduling . . . 5
  - 1.3.3 Appointment rules with fixed block length . . . 6
  - 1.3.4 Appointment rules with variable block length . . . 7
  - 1.3.5 Appointment schedules generated by optimization . . . 7
  - 1.3.6 Challenges . . . 8
- 1.4 Motivation and objectives . . . 8
  - 1.4.1 Problem structuring . . . 8
  - 1.4.2 Quantifying the performance of schedules . . . 10
  - 1.4.3 Schedule selection . . . 11
- 1.5 Outline . . . 12
- 1.6 Scientific contribution . . . 13

2 Computational Approach for a Single Server **15**
- 2.1 Introduction . . . 15
- 2.2 Background and model . . . 16
- 2.3 The phase-type approach . . . 18
  - 2.3.1 Phase-type fit of the service-time distribution . . . 19
  - 2.3.2 Recursive procedure to derive sojourn-time distributions . . . 20
  - 2.3.3 Optimal schedules for sequential and simultaneous approach . 22
- 2.4 Optimal scheduling in a transient environment . . . 24

2.5 Optimal scheduling in a steady-state environment 27
2.5.1 Steady-state results in case scv $= 1$ 29
2.5.2 Steady-state results in case scv $\neq 1$ 31
2.5.3 Computational results in a steady-state environment 33
2.6 Discussion 34
2.6.1 Robustness of phase-type approach in steady state 34
2.6.2 Robustness of phase-type approach in transient environment . 37
2.6.3 Comparison with the approach by Lau and Lau 37
2.6.4 The effect of overtime on the schedule 38
2.6.5 Computational effort of the various numerical approaches . . . 40
2.6.6 Comparison of sequential and simultaneous approaches 41
2.7 Conclusion 44

3 Practical Principles to Appointment Scheduling 45
3.1 Introduction 45
3.2 Modeling approach 47
3.2.1 Framework, risk function 47
3.2.2 Phase-type distribution 48
3.2.3 Recursive approach, incorporating no-shows 49
3.3 Experiments and results 50
3.4 Conclusion and discussion 54

4 Computational Approach for Two Servers in Tandem 57
4.1 Introduction 57
4.2 Problem description 59
4.3 Methodology 60
4.3.1 Phase-type distribution 60
4.3.2 Computing expected idle and waiting times 62
4.3.3 Recursive procedure to compute the sojourn-time distribution . 63
4.4 Extensions 69
4.4.1 Heterogeneous service-time distributions 70
4.4.2 Models with blocking 71
4.5 Optimal schedules in a transient environment 73
4.5.1 Effect of coefficient of variation 73
4.5.2 Effect of mean 74
4.5.3 Effect of weight 75
4.5.4 Comparison with single-server system 75
4.6 Optimal schedules in steady state 77
4.6.1 Procedure 77
4.6.2 Computational results 78
4.7 Conclusion and discussion 80

5 Lag Order Approximation Method 83
5.1 Introduction 83
5.2 Problem statement 85
5.3 The lag order approximation method 86

5.3.1 Loss functions 86
5.3.2 Technical background of the lag order procedure 87
5.3.3 Technical background for generating LF values 88
5.3.4 Example with exponentially distributed service times 88
5.4 The lag order approximation method in practice 90
5.4.1 Approximating realistic service-time distributions 90
5.4.2 Application in a CT-scan process 91
5.5 Conclusion 93

6 Efficient Procedures for Appointment Scheduling in Healthcare 97
6.1 Introduction 97
6.2 Model and approach 98
6.2.1 Preliminaries 98
6.2.2 Objective function 99
6.2.3 Stationarity 102
6.2.4 Phase-type fit 103
6.3 Stationary schedules 104
6.3.1 Numerical determination of stationary schedules 104
6.3.2 Analytical derivation of stationary schedules 105
6.3.3 Heavy-traffic derivation of stationary schedules 109
6.3.4 Impact of phase-type approximation 110
6.3.5 Phase-type approximation and extreme distributions 111
6.3.6 Example 112
6.4 Transient schedules 113
6.4.1 Operation of the webtool 114
6.4.2 No-shows and walk-ins 116
6.4.3 Overtime 117
6.4.4 Discrete slots 118
6.5 Performance evaluation 118
6.6 Conclusion 121

7 Webtool for Appointment Scheduling 123
7.1 Filling in parameter values in the webtool 123
7.2 Interpreting the output of the webtool 125

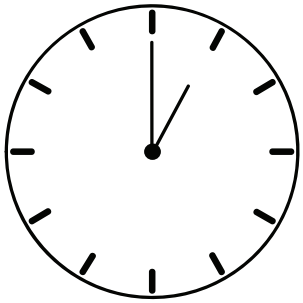
8 Summary 127
8.1 Appointment scheduling in healthcare 127
8.2 Motivation 128
8.3 Methodology and results 128
8.4 Practical implications 129

9 Bibliography 131

10 Samenvatting **137**
10.1 Roosteren van afspraken in de gezondheidszorg 137
10.2 Motivering 138
10.3 Methodologie en resultaten 138
10.4 Praktische implicaties 139

11 Curriculum Vitae **141**

12 Acknowledgments **143**



1. Introduction

Quality of care occupies center stage for healthcare providers. Competition, scarcity of resources and funding force providers to balance this ambition with efficiency and a productive utilization of staff, specialists and facilities. A method that helps providers to achieve a trade-off in this force-field is appointment scheduling. This dissertation deals with the design of optimal appointment schedules.

As an introductory example, consider a familiar situation: a dental practice. At the practice, clients arrive on their appointed arrival times. Since the treatment of the preceding client may be longer than anticipated, it may happen that a next client, upon arrival, needs to wait in the waiting room. Alternatively, the treatment of the preceding client may be shorter than scheduled, and as a consequence it may be the dentist who finds herself sitting idle for some time, waiting for the next client to arrive. Thus, variability and unpredictability in the treatment times result in waiting time for clients and idle time for the dentist.

Clients deem waiting time undesirable, as it negatively affects their perceived quality of service (Huang 1994, Anderson et al. 2007). Idle time has a negative impact on the utilization, which is measured as the percentage of available hours that the dentist effectively treats clients. A low utilization, as the result of much idle time, implies that the dentist sees fewer clients in the time available, and hence it degrades the dentist's effective capacity. In addition, a lower utilization implies that the fixed costs of the dentist and other resources (in terms of facilities, assistants and equipment) are spread over a smaller number of clients, and therefore leads to a higher per-client cost ('unit-cost').

The issue of appointment scheduling is not only relevant in the context of man-

aging the dentist's or other healthcare provider's available time, but also plays an important role when setting up procedures to optimally utilize MRI or CT scanners, operating rooms, and so forth. The commonality in these examples is that a scarce resource acts as a server handling patients (clients). Often the server uses an appointment schedule to regulate the demand to the resource's capacity.

This thesis is on appointment scheduling, with a focus on its application in healthcare. We structure the problem and establish a solid mathematical framework that facilitates the design of schedules. Within this framework, we propose different evaluation techniques to generate schedules that find a suitable balance between the interests of the patients and the provider. We assess the resulting schedules against methods that were earlier proposed and commonly used.

1.1 Healthcare operations

Most hospitals are not-for-profit in nature and exist to serve their communities. The rising expenditures for healthcare, however, have created general awareness that their performance should be evaluated in terms of the delivered care relative to the expenses incurred, see Porter (2010). This in turn has drawn attention to the performance of the operating and management practices involved. Two influential reports of the Institute of Medicine (2001, 2006) have urged the use of operational management methods and information technologies to improve the quality and efficiency in hospitals, and healthcare applications of operations management theory and techniques have become a thriving field.

In an evaluation of the needs of the U.S. healthcare system by Berwick et al. (2008), the authors conclude that there are three directions for improvement: the customer experience of the healthcare service, the quality of healthcare and the cost of healthcare. A healthcare provider is therefore confronted with opposing ambitions: on the one hand there is a need to control (or even reduce) costs, on the other hand, there is great pressure to improve service quality. The trade-off described in the introductory example, between waiting time for clients at a dentist and idle time for the resources, is a manifestation of this challenge. The challenge is more involved to the extent that there are more variability and unpredictability, for example caused by the randomness in the service-time durations and demand fluctuations. A good appointment schedule is one that finds an efficient balance between idle and waiting times.

Healthcare services can be classified into three categories as described by Gupta and Denton (2008): primary care, specialty care and elective surgery. Predictability and variability are relatively minor in primary care, whereas in specialty care and surgery they are a major complication for scheduling, and greatly depend on the sort of diagnosis and type of surgery.

In primary care *physicians* mostly divide their available clinic time into appointment slots of fixed given lengths between 10 to 30 minutes. The service time needed

is quite predictable. For example, it is known that new intakes require two time slots while follow-up visits require only one slot. In order to maintain continuity of care, patients are treated by a single, dedicated physician. Challenges in primary care are how to respond to walk-ins (unscheduled patients that require treatment) and unplanned understaffing (e.g., physician illness or emergency leaves) such that the impact on the patients' waiting times is limited. No-shows (patients who do not show up for their appointments) are not so prevalent in primary care due to the fact that appointments for primary care are often made a couple of days ahead.

Specialty care, also known as secondary care, delivers more specialized health services, such as medical imaging (MRI or CT scan), urgent care and other services in which the help of a specialist is needed. Providers are highly specialized physicians (*specialists*). Many specialty clinics require a referral from a primary care provider. The service time of a diagnosis or treatment depends greatly on the medical condition of the patient and the specific diagnosis. Although there is variability in service times, many facilities use an appointment schedule with fixed slot length. Since the specialist is a more expensive resource, a high utilization is desirable. The challenge for appointment scheduling is to realize high utilization, to reduce unit-cost and at the same time have sufficient leeway for emergency cases.

Elective surgery occurs either in an inpatient setting (admitted to a ward) or outpatient setting (poly-clinically). Surgery requires an operating room, one or more *surgeons*, supporting personnel and equipment, all of which are highly expensive resources. To utilize these resources efficiently there are two approaches focused on the usage of the operating room. The first approach is to reserve the operating room for a specific surgeon for an extended period of time, during which she may use the operating room; this is called *block-scheduling*. The second approach is called *open-scheduling*, in which surgeons make requests for specific time slots. The time slots needed for a specific type of surgery are usually based on historical averages by type and provider. Both approaches are viable, and can be used simultaneously. Open-scheduling has the advantage of more flexibility to deal with variation in demand and emergency patients, while block-scheduling is more efficient as the changeover times of personnel and equipment are reduced.

1.2 Dealing with variation in healthcare operations

Primary care, specialty care and elective surgery each have their own characteristics of variability and unpredictability, which make it challenging to synchronize supply (availability of healthcare providers and other resources) and demand (patients' requests for care). Variability and unpredictability are greatly reduced by *scheduling demand*. Scheduling appointments for patients at pre-set times is the most common choice for matching supply and demand, except for emergency care. The advantage of appointment scheduling is that it levels out patient demand over the available

time, thus making it easier to deploy resources efficiently without creating long waiting times. Emergency care is a form of *unscheduled demand* in which the healthcare service is started whenever an emergency patient shows up.

Designing an appointment schedule would be straightforward if patients showed up on time, service times were constant or perfectly predictable, and no-shows, walk-ins, cancelations and other disruptions did not occur. The challenge is to design schedules to handle such variability as well as possible, with as little waiting and idle times as possible. Appointment scheduling is therefore an instance of the more general problem of dealing with process variability; see Hopp and Spearman (2008). A sensible first step is to try to reduce variability and uncertainty. Some hospitals, for instance, bring down no-shows and last-minute cancelations by employing reminders and/or sanctions (Barron 1980, Johnson et al. 2007). A second step is to counterbalance variability by flexibility. Healthcare providers sometimes handle peak loads by stretching their working day or shrinking lunch time, or they may put unanticipated idle time to effective use by catching up on administrative work or other pending tasks.

After variability has been reduced or counterbalanced as far as possible, the *variability buffering law* of Hopp and Spearman (2008) predicts that the remaining variability will be absorbed by a combination of three buffers. In the first place, the provider may build up in advance an inventory of finished 'products' as a buffer to absorb peaks in workload. This is rarely an option, however, for the type of services that we consider, because products in our setting are treatments and diagnoses. Therefore, production cannot start until patient and provider come together. This leaves us with two other types of buffering:

- A queue of patients waiting to get served.
- An excess of unutilized capacity of the healthcare provider (which implies idle time).

As a consequence, a schedule's performance degradation due to variability is a combination of waiting time for patients and idle time for servers, which act as communicating vessels. On the one hand, it is efficient for a healthcare provider if there are some patients waiting, which act as a buffer of work on standby that prevents that the provider has idle time when treatments are shorter than anticipated or patients do not show up. On the other hand, it is convenient for the patient if the waiting room is empty and the provider is sitting idle ready to start treating a next patient.

1.3 Literature review

The appointment scheduling literature focuses on finding heuristics (rules of thumb) and proposing approaches to derive (close-to-)optimal appointment schedules, because the problem itself is analytically intractable as concluded by Robinson and

Chen (2003). However, various analytical approaches succeed, under specific conditions, in finding computational methods to evaluate and optimize appointment schedules. A clear trend is that early work, mid 20th century, is focused on empirical and simulation studies. In such studies current practices are observed and their performance is compared to alternatives, either in a simulation model or sometimes in a real clinical setting. Later the focus of attention has shifted to more computational studies. A plausible explanation for this shift are the technological advances over the last decades that have introduced a variety of new numerical techniques. These techniques are proposed by researchers to quantify the performance of and optimize appointment schedules.

Comprehensive literature reviews are given by Çayırlı and Veral (2003), Mondschein and Weintraub (2003) and Gupta and Denton (2008). We provide an overview of the field of appointment scheduling by the facets of the appointment scheduling problem that we are to consider. Our review is not exhaustive and therefore each chapter will contain a separate introduction containing references relevant to its contents.

1.3.1 Static versus dynamic appointment scheduling

Most healthcare services use an appointment book in which the patients' appointments are scheduled in advance (before the session starts). In such a setting the problem of designing the appointment schedule is merely a static problem, whereas in a dynamical setting requests for appointments are assigned during a session (e.g., Fries and Marathe 1981, Liao et al. 1993, Klassen and Rohleder 1996, Liu and Liu 1998a). Since 2000 there has been great interest for dynamical scheduling in practice, instigated by open-access and advanced-access policies. In these policies, patients are seen the same day or make an appointment in the near future, for example studied by Liu et al. (2010).

The potential benefits of open-access scheduling are that it may reduce no-shows and cancelations and improve access to healthcare providers (Murdock et al. 2002, Gallucci et al. 2005, Steinbauer et al. 2006). However, open-access scheduling makes it more difficult to control and predict patient demand, and therefore, capacity management may be much more difficult; see discussions in Green and Savin (2008), Robinson and Chen (2010) and Liu et al. (2010). Appointment scheduling in the static setting is still the more common and relevant problem to study.

1.3.2 Situational characteristics in appointment scheduling

Literature identifies situational characteristics (clinical environments) that should be taken into account in designing an appointment schedule. A first characteristic is the structure of the service process. Literature on appointment scheduling often considers a single-server process in which patients, typically 10 to 30, are being served

during a session. There are healthcare services that require a sequence of activities, such as an x-ray followed by a consult, such as studied in Rising et al. (1973) and Swisher et al. (2001).

A second situational characteristic are the stochastic properties of service times. In healthcare operations it has been observed that the service times vary greatly across type of care and specialty. Furthermore, studies find that the service times follow a uni-modal and rightly-skewed distribution (Bailey 1952, Welch 1964, Goldman et al. 1969, Brahimi and Worthington 1991). For simplicity, it is assumed in literature that the service times are independent and identically distributed, which is an assumption that is contradicted by various empirical studies (Bailey 1952, Rising et al. 1973, Babes and Sarma 1991) as health service providers tend to increase their speed when multiple patients are waiting.

Another important characteristic for the design of an appointment schedule is the no-show probability, which introduces uncertainty in the workload. In a comparison study of situational characteristics, Ho and Lau (1999) conclude that the no-show probability and the number of patients are the most influential factors determining the performance of an appointment schedule. Zacharias and Pinedo (2014) study how no-shows can be handled by overbooking slots. In addition, the walk-in probability is another situational characteristic that increases unpredictability in the workload and varies per type of care considered. The effects of both factors have extensively been studied (Fetter and Thompson 1966, Vissers and Wijngaard 1979, Vissers 1979, Çayırlı et al. 2006, Luo et al. 2012, Çayırlı et al. 2012).

Lastly, punctuality can also be a crucial situational characteristic (Çayırlı and Veral 2003). On the one hand, we have unpunctuality of patients, where literature mostly assumes that patients arrive early (rather than late). Furthermore, in Çayırlı et al. (2006) it is claimed that patients' tardiness is a less critical factor in the design of appointment schedules. In simulation studies their unpunctuality is modeled by an extra random variable, but in analytical or numerical studies patients are typically assumed to be punctual. On the other hand, also tardiness of providers may be considered.

1.3.3 Appointment rules with fixed block length

The literature of appointment scheduling originated under the paradigm where the available time (*session*) is divided into a number of intervals (*blocks*) spread evenly over the session. The pioneering works by Bailey (1952), Welch and Bailey (1952) and Welch (1964) introduce heuristics where the session is divided in blocks with lengths equal to the average service time. The first block starts with two or more scheduled patients and all subsequent blocks are assigned to only one patient. By simulation they found that starting a session with 2 patients works surprisingly well; the birth of the famous Bailey-Welch rule. White and Pike (1964), Soriano (1966), Fetter and Thompson (1966), and Rockart and Hofmann (1969) further study this kind

of heuristics. For example, Soriano (1966) introduces the heuristic to have the block length equal to twice the average service time and schedule two patients per block. Fries and Marathe (1981), Liao et al. (1993), Liu and Liu (1998b), Vanden Bosch et al. (1999), and Zacharias and Pinedo (2014) generalize this heuristic by determining the optimal number of patients to be scheduled for *each* block (still with constant block lengths).

1.3.4 Appointment rules with variable block length

As realized by Charnetski (1984), a clear shortcoming of heuristics with fixed interval lengths is that variability can only be taken into account by scheduling more (or fewer) patients per block. Charnetski proposes a heuristic based on setting block lengths equal to the average service time plus a multiple of the standard deviation in service times. Thus, the focus in scheduling is no longer to determine how many patients to schedule in a block, but rather to decide how long the blocks should be. In the same spirit, Ho and Lau (1992, 1999) study and compare sophisticated alternatives to traditional appointment rules in a comprehensive simulation study. Interestingly, they conclude that such rules with variable block lengths work better in specific situations, but that across a range of situational characteristics the Bailey-Welch rule (with 2 patients scheduled in the first block) has the most robust performance. Robinson and Chen (2003) propose a well performing heuristic in which two parameters, the block length of the first patient and the equally-sized block lengths for the remaining patients, are optimized (cf. an optimized version of the Bailey-Welch rule). Yang et al. (1998) propose an appointment rule that has been optimized over a given set of situational characteristics, which is further enriched in Çayırlı et al. (2012) by implementing no-shows and walk-ins.

1.3.5 Appointment schedules generated by optimization

The appointment rules discussed above have the main advantage of being simple, implying that no or just a few parameters are to be optimized. In its most general form, however, the scheduling problem has as many degrees of freedom as there are patients to be scheduled in a session, with the decision variables being the scheduled arrival epochs of all individual patients. The resulting objective function is typically hard to evaluate, but this problem can be solved by choosing tractable service-time distributions, such as the exponential (Healy 1992, Stein and Côté 1994, Hassin and Mendel 2008, Kaandorp and Koole 2007), phase-type (Wang 1997, Vanden Bosch et al. 1999, Vanden Bosch and Dietz 2000) or beta distribution (Lau and Lau 2000).

Also discretized versions of the service-times or schedules can be used and this facilitates a fast evaluation of schedules for optimization (e.g., Brahimi and Worthington 1991, Vanden Bosch et al. 1999, Vanden Bosch and Dietz 2000, Swisher et al. 2001, Kaandorp and Koole 2007, De Vuyst et al. 2011). Other, less obvious approaches are:

modeling the problem as a (two-stage) stochastic linear program (Denton and Gupta 2003); minimizing maximum loss (Mak et al. 2015); considering the problem *sequentially*, introduced by Weiss (1990) and studied in detail by Kemper et al. (2014).

The focus in many of these studies is more on the computational approach and less on a realistic characterization of the healthcare environment, or they fail to give operational, well-structured tools for practitioners.

1.3.6 Challenges

Literature has shown various trends, such as a tendency to focus on novel computational approaches without taking actual healthcare settings into consideration, or producing approaches that are very narrow in the situations in which they can be used.

In our view an approach to appointment scheduling should fulfill a number of requirements. In the first place it should be *fast*, in that it should be possible to generate optimal schedules in minimal computational time. If the underlying algorithm can be performed fast, then it facilitates sensitivity analysis: we can study and quantify the impact of a change in the weight in the objective function, or in the random service times. In the second place, the procedure should be *robust*: it should not overly depend on detailed distributional information. In many studies it is assumed that in principle the service-time distribution is fully known, whereas in practice maybe only the first two moments can be estimated; e.g., Çayırlı and Veral (2003), Mak et al. (2015). In the third place, the methodology should be *flexible* in the sense that it should be able to incorporate additional features such as no-shows, walk-ins, and restrictions on the possible values of the appointment durations (to cover the situation that slots are multiples of e.g. 5 minutes).

1.4 Motivation and objectives

The goal of this thesis is to develop an approach to the appointment scheduling problem that is capable of capturing specific and realistic characteristics of healthcare operations and that outperforms existing methods. We motivate our approach and clarify the various objectives to be solved by our procedure. Additionally, we sketch the model considered throughout this dissertation: the appointment scheduling problem is cast in a queueing-theoretic framework.

1.4.1 Problem structuring

The design of an appointment schedule consists of determining the individual scheduled arrival times for a certain number of patients. The chosen arrival times determine a trade-off in idle and waiting times for patients and healthcare providers, which arises due to variability in service times and uncertainty in the workload. The

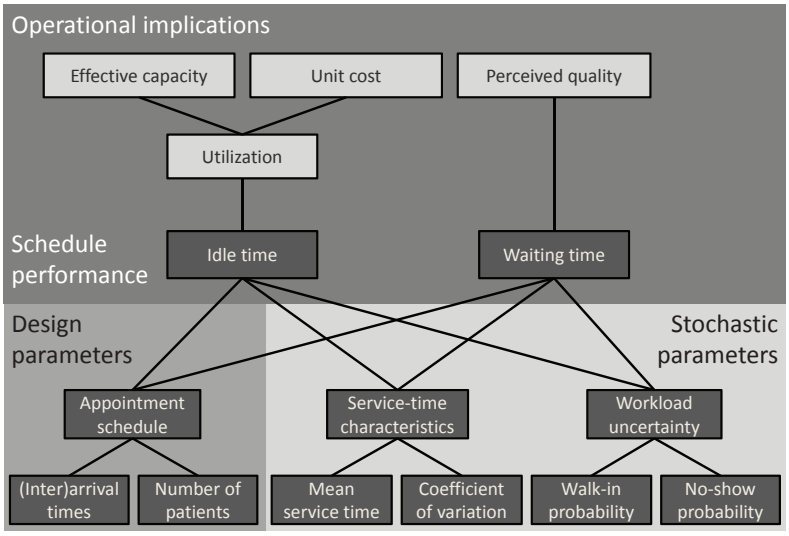


Figure 1.1: A model for the relation of stochastic and decision parameters relevant to the performance of the schedule (in terms of idle and waiting time) and their operational implications.

idle times affect the utilization of specialists and other resources. The waiting times have impact on the perceived quality experienced by the patients. The goal is therefore to choose the arrival times (*design parameters*) such that the schedule minimizes the expected idle and waiting times, as depicted in Figure 1.1.

In addition, the idle and waiting times are affected by situational characteristics, which we named *stochastic parameters* in Figure 1.1. First, the variability in the service times is split into the mean and the coefficient of variation (cv), which is the ratio of the standard deviation to the mean. Note that only two moments are used to describe the service times. As reported by Çayırlı and Veral (2003) and Mak et al. (2015) these moments greatly determine the performance of the schedule. In healthcare settings the cv typically lies in the interval of $[0.35, 0.85]$ (Bailey 1952, White and Pike 1964, Rising et al. 1973, O'Keefe 1985, Brahimi and Worthington 1991, Çayırlı et al. 2006). Second, the uncertainty in the workload is affected by both the no-show and walk-in rates, which typically range from 5% up to 30%, as observed by Fetter and Thompson (1966), Rockart and Hofmann (1969), O'Keefe (1985), Cox et al. (1985), but can be substantially higher in special cases as reported by Çayırlı et al. (2006). Fetter and Thompson (1966), Moore et al. (2001) and Çayırlı et al. (2006) indicate that the no-show and walk-in rates vary by the type of care considered, the time of day and the patient population.

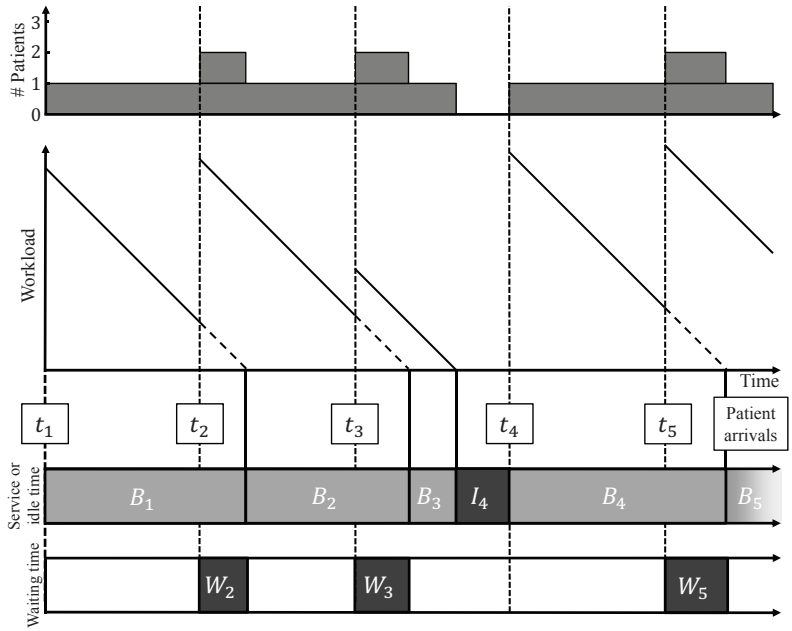


Figure 1.2: Various metrics related to an appointment schedule. On the top the total number of patients in the system, in the middle the workload of the server and on the bottom the relation to idle times and waiting times.

1.4.2 Quantifying the performance of schedules

As shown in Figure 1.1, the expected idle and waiting times quantify the performance of an appointment schedule. We mathematically model the problem in terms of a queueing system, which is illustrated in Figure 1.2. This figure has time at the horizontal axis, with the start of a session at the origin. We assume punctuality: all patients precisely arrive at their scheduled arrival epoch. An appointment schedule consists of arrival epochs t_i for $i = 1, \ldots, n$ where n is the number of patients to be seen in a single session. The service times $B_1, \ldots, B_n$ are the random variables in the model. Waiting time arises when the preceding patient is still being served when a new patient enters the clinic, which is the case for patients 2, 3 and 5 in Figure 1.2. The waiting time is zero when the system is empty on arrival of a new patient; this is the case for the 4-th patient in Figure 1.2. When this happens, the server is idle until the arrival of the next patient. For example, the idle time prior to the arrival of the 4-th patient is indicated as I_4 in the figure. The relations between number of patients in the system, workload, idle times and waiting times are graphically illustrated in Figure 1.2. The performance of an appointment schedule is evaluated in terms of the expected idle and waiting time that it implies.

The objectives of minimal idle and waiting times are a trade-off: minimizing one goes at the expense of the other. This trade-off is visualized in Figure 1.3. The graph

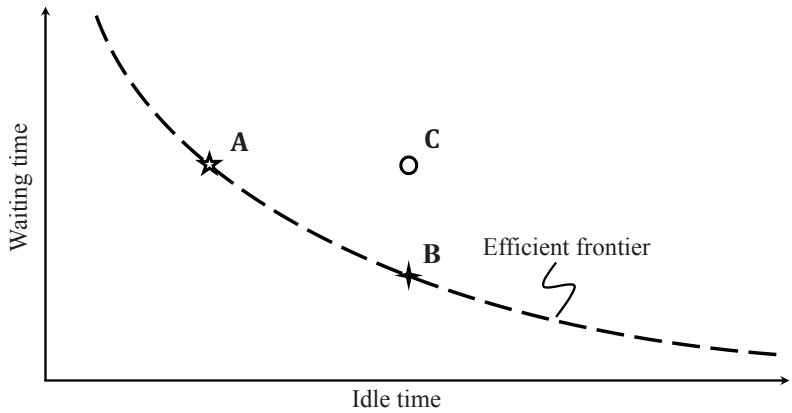


Figure 1.3: Three schedules (**A**, **B** and **C**) and the efficient frontier that set a trade-off in terms of idle and waiting times.

positions, for a hypothetical situation, three schedules in terms of their implied idle and waiting time. Note how, for example, schedule **A** has the smaller idle time, but **B** implies smaller waiting time. Given the variability in the service times, no-shows and walk-ins, there is a boundary that demarcates all possible combinations. This boundary is named the efficient frontier, and all possible schedules are above and to the right of this curve. The shape of the curve shows the trade-off: minimizing idle time inevitably results in long waiting times, and vice versa. Any position on the efficient frontier corresponds with a strategic decision in terms of the relative valuation of idle and waiting time.

Note that schedule **C** is not on the curve. That means that it is not an efficient schedule: there is a schedule which has less idle time time without compromising waiting time (or vice versa). The techniques developed in this thesis can be used to determine for a real schedule the distance to the efficient frontier. In case a schedule is on the curve, its position reveals the underlying strategic choice in terms of the trade-off. The other way around, given the desire to balance between idle and waiting times we provide techniques to generate efficient schedules.

1.4.3 Schedule selection

When being able to quantify the performance of appointment schedules, the next step is to systematically evaluate such schedules, with the objective to select one that has certain desirable (or even optimal) properties. A few considerations play a role here. In the first place, the candidate schedules should be extensively tested using a set of representative scenarios. In the second place, the schedules should be developed in such a way that the most relevant situational characteristics are modeled; think in this respect of the variability of the service times, and the no-show and the walk-in probabilities. The resulting model enables the assessment of the impact of those char-

acteristics. Ideally, one method uniformly outperforms all other approaches, across all scenarios. In practice however, one could see that the specific circumstances dictate which of the candidate techniques works best.

1.5 Outline

We now proceed by providing an outline of the thesis. Each of the Chapters 2 up to 5 covers a specific aspect of the appointment scheduling problem that we introduced. The findings are combined in Chapter 6, which presents a very general and practically applicable technique. This technique was implemented in a webtool that is presented in Chapter 7.

Chapter 2 deals with the problem in which the distribution of the individual patients is given and the planner is confronted with the task to determine the arrival epochs of the patients. We demonstrate how to generate schedules that have certain optimality properties. We express the performance of a schedule in terms of its associated utility, which incorporates both waiting times and idle times. In a first class of schedules (referred to as the *simultaneous approach*), the arrival epochs are chosen such that the sum of the utilities of all patients as well as the service provider are minimized. In a second class (*sequential approach*), the arrival epoch of a next patient is scheduled given the scheduled arrival epochs of all previous patients. Our approach is applied in several examples, that provide insight in the impact of the variability of the service times on the schedule; it also shows the impact of the utility function selected.

In Chapter 3, we compare the optimal schedules generated by the approach presented in Chapter 2 with a number of easily-evaluated heuristics. In our setup it is assumed throughout that a given fraction of the patients does not show up. Our results are particularly useful in situations in which there is significant variation in the service times, which is the case in various healthcare-related settings.

Chapter 4 deals with multi-node appointment-based service systems that arise in a broad variety of healthcare settings (for example an outpatient clinic or a dentist). Where most existing algorithms specifically consider the situation of the patient undergoing a *single* service, in many practical situations *multiple* services have to be *sequentially* performed. Modeling the service system as a tandem queue, the main objective of this chapter is to generate schedules that soundly balance the interests of patients (i.e., low waiting times) and staff (i.e., low idle times). Importantly, following up on prior work for the single-node queue, as given in Chapter 2, we advocate a phase-type based technique that can deal with *any* service-time distribution (which may, in addition, vary across patients). Relying on a novel recursive scheme to evaluate the sojourn-time distribution of patients in such tandem systems, we show how optimal schedules can be computed. Our technique is illustrated by extensive numerical experimentation, also leading to practical guidelines that apply to a broad

range of parameter settings.

In Chapter 5 we try an alternative approach to schedule patients in continuous time using the actual service-time distributions as opposed to approximations and hypothesized distributions (the approach in the preceding chapters). As the optimal schedule is notoriously hard to derive within reasonable computation times, we develop the lag order approximation method, that sets the patient's optimal appointment time based on only a part of his predecessors. We show that a lag order of two, i.e., taking two predecessors into account, results in nearly optimal schedules but the required computation time may be relatively long. We illustrate our approximation method with an appointment scheduling problem in a CT-scan process.

Chapter 6 presents our ultimate model, which combines various aspects of the lessons learned in the previous chapters. We first consider the situations with a relatively large number of patients having stochastically identical service times (stationary schedules). We give accurate closed-form approximations that either exploit the distributional form of specific phase-type distributions or explicit heavy-traffic results. We then focus on the situation with a limited number of patients, for which we develop an approach for generating optimal schedules including relevant phenomena such as no-shows, walk-ins and overtime. Finally, we present an easy-to-use webtool, which allows healthcare providers to generate appointment schedules that significantly outperform existing approaches. A detailed instruction on how to use the webtool is given in Chapter 7.

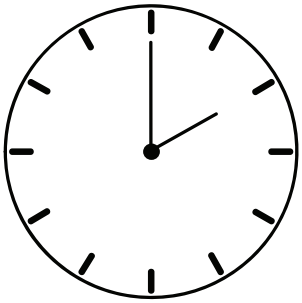
Finally, Chapter 8 provides a summary of the dissertation.

1.6 Scientific contribution

Chapter 2 has appeared as an article in *Queueing Systems* (Kuiper et al. 2015), which originated from my master thesis project under supervision of Dr. Benjamin Kemper and Prof. Michel Mandjes. The modeling and approach has been joint work, and most of the numerical results were attributed to me. I had the lead in the work of Chapter 3, which compares and relates our approach, extended with no-shows, to well-known heuristics. The content of Chapter 3 is published in *Quality and Reliability Engineering International* (Kuiper and Mandjes 2015b). Prof. Mandjes also came up with the idea to extend our work to tandem-type systems and together we explored this line of research, where I extended the numerical procedures to facilitate optimization of these systems, resulting in a publication in *Omega* (Kuiper and Mandjes 2015a).

Chapter 5 is based on an article that has been published in the *European Journal of Operational Research* (Vink et al. 2015). The article was initially written as a spin-off of the master thesis project of Wouter Vink M.Sc. supervised by Dr. Kemper and Dr. Sandjai Bhulai. Their submission was rejected, but with my help (providing additional numerical work and rewriting the paper) a revision has led to publication.

The content of Chapter 6 has been combined work with Prof. Mandjes, Ruben Brokkelkamp M.Sc. and Prof. Jeroen de Mast. As a result of our previous research Prof. Mandjes and I found interesting patterns in the steady-state solutions. This provided a basis for new research, which, with the help of Brokkelkamp M.Sc., were uncovered. Furthermore, Prof. Mandjes introduced heavy-traffic analysis and Prof. De Mast aided in structuring the problem in the concepts of operations management. His writings also inspired my thoughts for topics to be included in this introduction. Brokkelkamp M.Sc. also played a significant role in making the webtool work. At this stage the manuscript, called *Efficient procedures for appointment scheduling* (Kuiper et al. 2016), has been submitted for publication.



2. A Computational Approach to Appointment Scheduling for a Single Server

The goal of this chapter is to study the computational feasibility of scheduling algorithms that minimize an objective function that depends on idle and waiting times. The objective function under study requires knowledge of the first and second moments of the idle times and waiting times. Relying on a phase-type approximation, we analyze this problem in two environments. First, a transient environment, where a finite number of clients are scheduled, and second, in a steady-state environment, corresponding to the situation that the number of stochastically identical clients to be scheduled tends to infinity.

2.1 Introduction

An optimized schedule is such that the system's risk (the expectation of a loss function that involves both waiting times and idle times) is minimized, thus realizing an optimal trade off between the interest of the providers and the clients (patients). Here, a schedule is the vector of appointed arrival times. In queueing-theoretic terms: the planner is given the distributions of the service times, and then it is her task to determine the corresponding optimal (that is, disutility-minimizing) arrival epochs. In our work we limit ourselves to the situation in which the sequence of the arrivals is fixed.

Consider independent and identically distributed (i.i.d.) service times. In case the number of clients is relatively low, the optimal interarrival times may vary substantially (realize that the first client finds the system empty with certainty). If, to the contrary, there are many clients, the schedule will converge to a steady state: the optimal interarrival times of successive clients will be constant. By making a connection to appropriately chosen D/G/1 queues (with the service times following a phase-type distribution), we demonstrate how to determine the corresponding stationary schedules, both in the simultaneous and sequential approach.

Methodologically, our work is related to Wang (1997) (also using phase-type distributions), Lau and Lau (2000) (using beta distributions), Hassin and Mendel (2008) (using the exponential distribution) and De Vuyst et al. (2011) (using discretized versions of the service, idle, waiting and sojourn times). The phase-type and beta distributions are particularly attractive, as they allow a selection of the parameters in such a way that, e.g., the first moments match with those estimated from measurements. In this chapter we will assess the differences between these candidate approaches.

This chapter is organized as follows. In Section 2.2 we introduce our mathematical model, and define the risk functions considered. The approach is presented in Section 2.3. Section 2.4 demonstrates our approach for transient schedules, while Section 2.5 considers the stationary counterpart. In Section 2.6 we discuss the potential and limitations of our approach; in particular, we show that the error due to the phase-type fit is small. Section 2.7 concludes and suggests ideas for future work.

Various graphs illustrate a number of interesting effects. We quantify the following features: (i) the convergence of transient schedules to their stationary counterparts; (ii) the impact of the choice of the risk function on the schedule; (iii) the impact of the service times' variability on the schedule. Also the differences between the simultaneous and sequential approach are studied in greater detail. Evidently, replacing a non-phase-type distribution by its phase-type counterpart introduces an error; a simulation study shows that the impact of this error is negligible.

2.2 Background and model

The mathematical treatment of the appointment scheduling problem with one server dates back to at least the seminal works of Bailey (1952) and Welch and Bailey (1952). Since then, a sizeable number of papers has appeared in the operations research literature. The results in these papers tend to be rather case-specific, in terms of the service-time distribution under consideration as well as the risk function chosen. One often relies on simulations to overcome the inherent computational complexities. Such an approach has clear limitations: it evidently lacks general applicability, and, more importantly, it does not provide us with structural insights and generalizable solutions. Our aim is therefore: develop an approach that works for general service times, general risk functions, and that is numerically feasible.

A common way to reason about an appointment scheduling problem is to define for each arrival i a so-called *risk*. This risk is the expectation of a loss function that reflects the idle time (I_i) and the waiting time (W_i) of each client i. A natural choice is $\mathbb{E}g(I_i)+\mathbb{E}h(W_i)$, where it makes sense to choose non-decreasing loss functions $g(\cdot)$ and $h(\cdot)$ with $g(0) = h(0) = 0$. Observe that these risks clearly depend on the arrival epochs t_i and service times B_i; more precisely, the risk associated to the i-th client depends on the arrival epochs t_1 up to t_i and service times B_1 up to B_{i-1}. The optimal schedule corresponding to the simultaneous approach then follows from solving the minimization problem over the arrival epochs only

$$\min_{t_1,\ldots,t_n} \sum_{i=1}^{n} \left(\mathbb{E}g(I_i) + \mathbb{E}h(W_i)\right), \tag{2.1}$$

whereas its sequential counterpart minimizes $\mathbb{E}g(I_i)+\mathbb{E}h(W_i)$ over t_i, with $t_1, \ldots, t_{i-1}$ given.

In this chapter we focus on a quadratic and a linear loss function, but, importantly, the setup carries over to any loss function in the class defined above. For a quadratic loss the risk is defined by

$$R_i^{(q,\omega)}(t_1, \ldots, t_i) := \omega\mathbb{E}I_i^2 + (1-\omega)\mathbb{E}W_i^2, \quad i = 1, \ldots, n \quad \text{and} \quad \omega \in (0, 1).$$

Due to the well-known *Lindley recursion* (Lindley 1952),

$$I_i = \max\{t_i - t_{i-1} - W_{i-1} - B_{i-1}, 0\}, \tag{2.2}$$

and

$$W_i = \max\{W_{i-1} + B_{i-1} - t_i + t_{i-1}, 0\}. \tag{2.3}$$

Let $S_i := W_i + B_i$ denote the sojourn time of the i-th client, with distribution function $F_{S_i}(\cdot)$. In addition, define by $x_{i-1} := t_i - t_{i-1}$ the time between the $(i-1)$-st and i-th arrival. Then, with (2.2) and (2.3) in mind, we may write the system's risk (in relation to the i-th client) as

$$\begin{aligned} R_i^{(q,\omega)}(t_1, \ldots, t_{i-1}, t_{i-1} + x_{i-1}) &:= \omega\mathbb{E}I_i^2 + (1-\omega)\mathbb{E}W_i^2 \\ &= \omega\mathbb{E}\left(x_{i-1} - S_{i-1}\right)^2 \mathbb{1}_{x_{i-1} > S_{i-1}} \\ &\quad + (1-\omega)\mathbb{E}\left(S_{i-1} - x_{i-1}\right)^2 \mathbb{1}_{x_{i-1} < S_{i-1}}. \end{aligned} \tag{2.4}$$

This is a nonnegative convex function of x_{i-1}. Below we specialize to the case of equal weights, that is, $\omega = \frac{1}{2}$. In that case the risk related to the i-th client reduces to $\frac{1}{2}\mathbb{E}\left(S_{i-1} - x_{i-1}\right)^2$ (where we can leave out the factor $\frac{1}{2}$). For $\omega \neq \frac{1}{2}$ there is no such a simplification of the expressions. The computation time required to determine optimal schedules does not depend on the choice of ω, however: all cases can be evaluated in essentially the same amount of computation time. At the end of Section 2.5.3 we

assess the effect of the weights in steady state.

In the case of a linear loss function, the risk associated with the i-th client equals the sum of the expected waiting time and the expected idle time. Again, due to (2.2) and (2.3), we obtain, again with $\omega \in (0,1)$,

$$\begin{aligned} R_i^{(\ell,\omega)}(t_1,\ldots,t_{i-1},t_{i-1}+x_{i-1}) &:= \omega\mathbb{E}I_i + (1-\omega)\mathbb{E}W_i \\ &= \omega\mathbb{E}\left(x_{i-1}-S_{i-1}\right)\mathbb{1}_{x_{i-1}>S_{i-1}} \\ &\quad + (1-\omega)\mathbb{E}\left(S_{i-1}-x_{i-1}\right)\mathbb{1}_{x_{i-1}<S_{i-1}}, \end{aligned} \tag{2.5}$$

which is a nonnegative convex function of x_{i-1}. Again, we consider in this chapter only the case of equal weights, so that the risk related to the i-th client reduces to $\frac{1}{2}\mathbb{E}|S_{i-1}-x_{i-1}|$ (where again we can leave out the factor $\frac{1}{2}$).

2.3 The phase-type approach

As argued earlier in this chapter, the main problem when generating schedules of a realistic size concerns the fact that neither explicit expressions are available for the expected idle and waiting times (or the corresponding second moments), nor for the distributions of the sojourn times — these are needed to be able to evaluate the risk (which then needs to be optimized, either sequentially or simultaneously). This section proposes an approach to circumvent this problem, by replacing the service times by a phase-type counterpart (of relatively low dimension). For these approximate service times, we *can* evaluate the first and second moments of the client's sojourn time and therefore, through (2.4) and (2.5), client i's risk associated with I_i and W_i, as it will turn out.

The approach we propose in this chapter consists of three steps:

1: Based on the mean and variance of the service times (or, equivalently, the mean and the coefficient of variation), we fit a phase-type distribution.

2: With a recursive procedure we derive, for each client, the sojourn-time distribution (for the fitted phase-type distribution).

3: The phase-type based sojourn-time distribution enables us to evaluate the objective function. Relying on standard numerical packages, we can then solve the simultaneous optimization problem as stated in (2.1). In the sequential counterpart it suffices to compute the expected value (in case of a quadratic loss) or median (in case of a linear loss) of the clients' sojourn times.

In this section we provide further details on our approach; in Section 2.4 and Section 2.5 we demonstrate the resulting procedure in transient (relatively few clients) and steady-state (relatively many clients) settings.

2.3.1 Phase-type fit of the service-time distribution

In the first step of our approach we use phase-type distributions to fit the service-time distributions in the system under study. Phase-type distributions are mixtures and convolutions of exponential distributions and include (mixtures of) Erlang distributions and hyperexponential distributions. It is well known from the literature that they can approximate any positive distribution arbitrarily accurately, see e.g. Tijms (1986) and Asmussen et al. (1996).

The reason to use phase-type distributions is twofold. In the first place, due the enforced Markovianity, the resulting system often enables the computation of explicit expressions for various queueing-related metrics, such as the waiting times distribution (where 'explicit' means in terms of eigenvalues/eigenvectors of an associated eigensystem). In the second place, restricting ourselves to a phase-type distribution of a certain dimension, estimating this distribution from data can be done via a (semi-)parametric density estimation procedure.

In our study we use the idea presented in Tijms (1986) to match the first and second moment of the service-time distribution, or, equivalently, the mean and the *squared coefficient of variation* (scv). The scv of the random variable X is defined as its variance divided by the square of the mean. In line with Kemper and Mandjes (2012), we choose to match a mixture of two Erlang distributions in case the actual service-time distribution has an scv smaller than 1, and a hyperexponential distribution in case of an scv larger than (or equal to) 1. More precisely:

- In case scv < 1 we match the service-time distribution with a mixture of two Erlang distributions with the same scale parameter, denoted as $E_{K-1,K}(\mu; p)$. A sample from this distribution is obtained by sampling from an Erlang distribution with $K-1$ phases and mean $(K-1)/\mu$ with probability p, and from an Erlang distribution with K phases and mean K/μ with probability $1-p$. Its n-th moment is given by

  $$\mathbb{E}\left[E_{K-1,K}^n\right] = p\frac{(K+n-2)!}{(K-2)!}\frac{1}{\mu^n} + (1-p)\frac{(K+n-1)!}{(K-1)!}\frac{1}{\mu^n},$$

  with $p \in [0,1]$. The corresponding scv equals

  $$\frac{K-p^2}{(K-p)^2},$$

  which lies between $1/K$ and $1/(K-1)$ for $K \in \{2,3,\ldots\}$. We can thus uniquely identify an $E_{K-1,K}(\mu; p)$ distribution matching the first two moments of the target distribution, as long as scv < 1.

- In case scv ≥ 1 we match the service-time distribution with a specific example of the hyperexponential distribution, namely a mixture of two exponential

distributions, denoted by $H_2(\boldsymbol{\mu}; p)$, with $\boldsymbol{\mu} = (\mu_1, \mu_2)$. Its n-th moment is given by

$$\mathbb{E}\left[H_2^n\right] = p\frac{n!}{\mu_1^n} + (1-p)\frac{n!}{\mu_2^n}.$$

We impose the additional condition of *balanced means*, see Eq. (A.16) in Tijms (1986). That is, we require $\mu_1 = 2p\mu$ and $\mu_2 = 2(1-p)\mu$ for some $\mu > 0$. The corresponding scv equals $(2p(1-p))^{-1}$, which is larger than (or equal to) 1. It can be verified that

$$p = \frac{1}{2}\left(1 \pm \sqrt{\frac{\text{scv}-1}{\text{scv}+1}}\right).$$

It is readily checked that for the special case scv $= 1$, the fit results in an exponential distribution (with $p = \frac{1}{2}$).

2.3.2 Recursive procedure to derive sojourn-time distributions

We now present a procedure to compute the sojourn-time distribution of any specific client, in case the service times are of phase type. We concentrate on mixtures of Erlangs (i.e., $E_{K-1,K}(\mu; p)$) and hyperexponentials (i.e., $H_2(\boldsymbol{\mu}; p)$), as these are the ones to which we fitted our service-time distributions. The procedure works, however, for any phase-type distribution; see e.g. Wang (1997). We assume that the service times are i.i.d., but the procedure can be extended to independent, *non*-identically distributed phase-type service times, at the expense of rather involved notation.

A phase-type distribution is characterized by an $m \in \mathbb{N}$, an m-dimensional row vector $\boldsymbol{\alpha}$ with nonnegative entries adding up to 1, and $\boldsymbol{S} = (s_{ij})_{i,j=1}^m$ an $(m \times m)$-dimensional matrix such that $s_{ii} < 0$, $s_{ij} \geq 0$ and $\sum_{j=1}^m s_{ij} \leq 0$ for any $i \in \{1, \ldots, m\}$.

- ▷ In case scv < 1, we use an $E_{K-1,K}(\mu; p)$ distribution. Then $m = K$, and the vector $\boldsymbol{\alpha}$ such that $\alpha_1 = 1$ and $\alpha_i = 0$ for $i = 2, \ldots, K$. In addition $s_{ii} = -\mu$ for $i = 1, \ldots, K$ and $s_{i,i+1} = -s_{ii} = \mu$ for $i = 1, \ldots, K-2$, while $s_{K-1,K} = (1-p)\mu$; all other entries are 0.

- ▷ In case scv ≥ 1, we use a $H_2(\boldsymbol{\mu}; p)$ distribution (as explained in Section 2.3.1). Then $m = 2$, and $\alpha_1 = p = 1 - \alpha_2$. Also, $s_{ii} = -\mu_i$, for $i = 1, 2$, while the other two entries of $\boldsymbol{S}$ equal 0.

For more background on phase-type distributions, see Asmussen (2003).

Next, we briefly describe the algorithm, presented in Wang (1997), that determines the clients' sojourn-time distributions. To this end, we consider the bivariate process $\{N_i(t), K_i(t), t \geq 0\}$ for client $i = 1, \ldots, n$. Here $N_i(t)$ is the number of clients in front of the i-th arriving client, t time units after her arrival. Obviously, $N_i(t) \in \{0, \ldots, i-1\}$. The second component, $K_i(t) \in \{1, \ldots, m\}$, represents the phase of the client being served, t time units after the arrival of the i-th client, where

$N_i(t) = 0$ refers to the case that the last arriving client is being served. We also introduce the probabilities, for $t \geq 0$, $i = 1, \ldots, n$, $j = 0, \ldots, i-1$, and $k = 1, \ldots, m$,

$$p_{j,k}^{(i)}(t) = \mathbb{P}\left(N_i(t) = j, K_i(t) = k\right).$$

In addition, the following vector (of dimension mi) plays a crucial role

$$\boldsymbol{P}_i(t) := \left(p_{i-1,1}^{(i)}(t), \ldots, p_{i-1,m}^{(i)}(t), p_{i-2,1}^{(i)}(t), \ldots, p_{i-2,m}^{(i)}(t), \ldots, p_{0,1}^{(i)}(t), \ldots, p_{0,m}^{(i)}(t)\right).$$

The sojourn-time distribution of the i-th client can be computed from $\boldsymbol{P}_i(t)$ through the identity (with $\boldsymbol{e}_{mi}$ an all-one vector of dimension mi):

$$F_i(t) := \mathbb{P}(S_i \leq t) = 1 - \sum_{j=0}^{i-1} \sum_{k=1}^{m} p_{j,k}^{(i)}(t) = 1 - \boldsymbol{P}_i(t)\boldsymbol{e}_{mi}.$$

For the first client arriving at $t_1 = 0$, we have that $\boldsymbol{P}_1(t) = \boldsymbol{\alpha}\exp(\boldsymbol{S}t)$ (which is an m-dimensional vector). For the second client, arriving x_1 after the first client, we have

$$\boldsymbol{P}_2(t) = (\boldsymbol{P}_1(x_1), \boldsymbol{\alpha}F_1(x_1))\exp(\boldsymbol{S}_2 t), \quad t \geq 0,$$

which is an object of dimension $2m$; here, with $\boldsymbol{s} := -\boldsymbol{S}\boldsymbol{e}_m$ and $\boldsymbol{0}_{m,m}$ an $(m \times m)$-dimensional all-zero matrix,

$$\boldsymbol{S}_2 := \begin{pmatrix} \boldsymbol{S} & \boldsymbol{s}\boldsymbol{\alpha} \\ \boldsymbol{0}_{m,m} & \boldsymbol{S} \end{pmatrix}.$$

The sojourn-time distributions of the other clients can be found recursively in a similar manner. To this end, define the matrix $\boldsymbol{T}_i$ of dimension $(i-1)m \times m$ as

$$\boldsymbol{T}_i := (\boldsymbol{0}_{m,m}, \boldsymbol{0}_{m,m}, \ldots, \boldsymbol{0}_{m,m}, \boldsymbol{s}\boldsymbol{\alpha})^{\mathrm{T}},$$

so that

$$\boldsymbol{S}_i := \begin{pmatrix} \boldsymbol{S}_{i-1} & \boldsymbol{T}_i \\ \boldsymbol{0}_{m,(i-1)m} & \boldsymbol{S} \end{pmatrix}.$$

Then the vector $\boldsymbol{P}_i(t)$ (dimension mi) can be found from $\boldsymbol{P}_{i-1}(t)$ (dimension $m(i-1)$) by the recursion

$$\boldsymbol{P}_i(t) = (\boldsymbol{P}_{i-1}(x_{i-1}), \boldsymbol{\alpha}F_{i-1}(x_{i-1}))\exp(\boldsymbol{S}_i t), \quad t \geq 0.$$

Realize that in our examples the matrix $\boldsymbol{S}$ is upper triangular, and hence so are the matrices $\boldsymbol{S}_i$. As a consequence, the eigenvalues can be read off from the diagonal. This property facilitates easy computation of the matrix exponent $\exp(\boldsymbol{S}_i t)$. In case of the $E_{K-1,K}(\mu; p)$ all eigenvalues are μ.

2.3.3 Optimal schedules for sequential and simultaneous approach

Above we explained how to approximate any distribution on $[0, \infty)$ by a phase-type distribution of relatively low dimension (either a mixture of Erlang distributions or a hyperexponential distribution, depending on the value of the scv). We also showed how to compute the corresponding sojourn-time distributions. The next step is to use these findings to determine optimal schedules, for the sequential and simultaneous optimization approach, and for quadratic and linear loss functions, as in Kemper et al. (2014).

The sequential optimization approach

In the sequential optimization approach, we minimize for each arriving client i the corresponding risk. This means that we minimize the expected loss over t_i, for given values of $t_1(=0), \ldots, t_{i-1}$. In suggestive notation, the optimization program

$$\min_{t_i} R_i(t_i \mid t_{i-1}, \ldots, t_1) = \min_{t_i} \mathbb{E}g(I_i) + \mathbb{E}h(W_i).$$

As we argued earlier, to solve this sequential optimization problem, we only need to know the sojourn-time distribution of the previous arrival, S_{i-1}, given $t_1, \ldots, t_{i-1}$ (Kemper et al. 2014). We now show in greater detail how this works for the *weighted-linear* and the *weighted-quadratic* loss function.
Weighted-linear loss function. Let the risk for each arrival be a weighted expected linear loss over the idle time and waiting time, i.e.,

$$\min_{t_i} R_i^{(\ell,\omega)}(t_i|t_{i-1}, \ldots, t_1) = \min_{t_i} \omega\mathbb{E}I_i + (1-\omega)\mathbb{E}W_i, \quad i = 1, \ldots, n, \quad \omega \in (0,1).$$

Given (2.5) we may write for $i = 2, \ldots, n$ and again for $\omega \in (0,1)$

$$\min_{x_{i-1}} \omega\mathbb{E}\left(x_{i-1} - S_{i-1}\right)\mathbb{1}_{x_{i-1}>S_{i-1}} + (1-\omega)\mathbb{E}\left(S_{i-1} - x_{i-1}\right)\mathbb{1}_{x_{i-1}<S_{i-1}},$$

where the interarrival time x_{i-1} equals $t_i - t_{i-1}$.

Then the optimal interarrival time $x_{i-1}^\star$ can be found by solving the first-order equation

$$\omega F_{S_{i-1}}(x) - (1-\omega)\left(1 - F_{S_{i-1}}(x)\right) = F_{S_{i-1}}(x) - 1 + \omega = 0.$$

This leads to the optimal schedule

$$t_1^\star := 0 \quad \text{and} \quad t_i^\star := \sum_{j=1}^{i-1} F_{S_j}^{-1}(1-\omega), \quad i = 2, \ldots, n.$$

For $\omega = \frac{1}{2}$ we obtain that client i is scheduled to arrive after a time that equals the

sum of the *medians* of the sojourn times of all previous clients.

Weighted-quadratic loss function. Let the risk for each arrival be a weighted expected quadratic loss over the idle time and waiting time

$$\min_{t_i} R_i^{(q,\omega)}(t_i|t_{i-1},\ldots,t_1) = \min_{t_i} \omega \mathbb{E} I_i^2 + (1-\omega)\mathbb{E} W_i^2, \quad i = 1,\ldots,n, \quad \omega \in (0,1).$$

Given (2.4) we write for $i = 2,\ldots,n$ and again for $\omega \in (0,1)$, with $x_{i-1} = t_i - t_{i-1}$,

$$\min_{x_{i-1}} \omega \mathbb{E}\left(x_{i-1} - S_{i-1}\right)^2 \mathbb{1}_{x_{i-1} > S_{i-1}} + (1-\omega)\mathbb{E}\left(S_{i-1} - x_{i-1}\right)^2 \mathbb{1}_{x_{i-1} < S_{i-1}}.$$

As above, the optimal interarrival time $x_{i-1}^\star$ follows from the first-order equation, which now reads

$$\omega(x - \mathbb{E}S_{i-1}) - (1-2\omega)\int_x^\infty \mathbb{P}(S_{i-1} > s)\mathrm{d}s = 0.$$

For $\omega = \frac{1}{2}$ we obtain the optimal schedule

$$t_1^\star := 0 \quad \text{and} \quad t_i^\star := \sum_{j=1}^{i-1} \mathbb{E}S_j, \quad i = 2,\ldots,n.$$

This means that for $\omega = \frac{1}{2}$ we obtain that client i is scheduled to arrive after a time that equals the sum of the *means* of the sojourn times of all previous clients.

The simultaneous optimization approach

In case of a simultaneous optimization approach we set the optimal schedule that jointly minimizes

$$\min_{t_1,\ldots,t_n} R(t_1,\ldots,t_n) = \min_{t_1,\ldots,t_n} \sum_{i=1}^n \left(\mathbb{E}g(I_i) + \mathbb{E}h(W_i)\right).$$

It is known that this joint optimization in general has no tractable solution, as was the case in the sequential approach. Only in case of an exponential service-time distribution and a linear loss function it has a tractable solution, see Wang (1997). Therefore, we rely on numerical analysis software to find the optimal schedule.

In the next sections we present numerical examples that feature schedules generated by our approach. Section 2.4 concentrates on the situation of a relatively low number of clients, whereas Section 2.5 uses results for the steady-state of the D/G/1 queue (with phase-type service times) to analyze the situation of a relatively high number of clients. In Section 2.6 we discuss the potential and limitations of our approach; in particular, we show that the error due to the phase-type fit is small.

2.4 Optimal scheduling in a transient environment

If the number of clients in the schedule, n, is relatively high, and their service times are i.i.d., then one will obtain schedules with more or less constant interarrival times. This section presents results for optimal schedules in the opposite case, i.e., situations in which the number of clients is relatively low. Particularly at the beginning of the schedule (and in the simultaneous approach also at the end) it is expected that the optimal interarrival times will vary substantially. Our experiments show that, for the loss functions and the range of scv values that we consider, this 'transient effect' has significant impact up to, say, $n = 25$ clients.

Normalizing time such that the mean service time equals 1, we use four different values of the scv (scv $\in \{0.1225, 0.7186, 1.0000, 1.6036\}$). These can be considered typical for services and healthcare processes, see Kemper and Mandjes (2012). The latter three values are used in Wang (1997). We added scv $= 0.1225 = 0.35^2$ to be consistent with the healthcare literature, where it is reported that the cv ranges from 0.35 up to 0.85 (Çayırlı and Veral 2003).

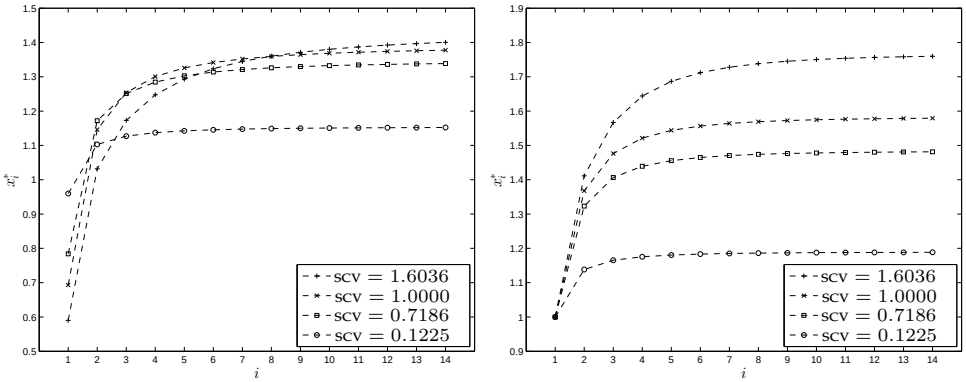
Based on our approach as proposed in Section 2.3, we first find the corresponding phase-type service-time distribution, then we derive for each arrival the sojourn-time distribution, and finally we compute the optimal schedule (for the sequential and simultaneous approach, with linear and quadratic loss functions).

- ▷ We model an scv $= 0.1225 < 1$ with an $E_{K-1,K}(\mu; p)$ distribution with parameters $K = 9$ (realize that scv $\in [\frac{1}{9}, \frac{1}{8}]$), $\mu = 8.3958$ and $p = 0.6042$.
- ▷ We model an scv $= 0.7186 < 1$ with an $E_{K-1,K}(\mu; p)$ distribution with parameters $K = 2$, $\mu = 1.6003$ and $p = 0.3997$. The resulting parameters are $\boldsymbol{\alpha} = (1, 0)$ and
$$\boldsymbol{S} = \begin{pmatrix} -1.6003 & 0.9606 \\ 0 & -1.6003 \end{pmatrix}.$$
- ▷ We model an scv $= 1$ with an exponential distribution with parameter $\mu = 1$.
- ▷ We model an scv $= 1.6036$ with a $H_2(\boldsymbol{\mu}; p)$ distribution under the condition of *balanced means* and with parameters chosen by the matching method explained above. The resulting parameters are $\boldsymbol{\alpha} = (p, 1-p) = (0.7407, 0.2593)$ and
$$\boldsymbol{S} = \begin{pmatrix} -1.4815 & 0 \\ 0 & -0.5185 \end{pmatrix}.$$

Note that for all cases the mean service time is given by

$$-\boldsymbol{\alpha}\boldsymbol{S}^{-1}\boldsymbol{e}_m = 1,$$

($\boldsymbol{e}_m$ is the all-ones vector) as desired.



(a) Linear loss, phase-type fit.

(b) Quadratic loss, phase-type fit.

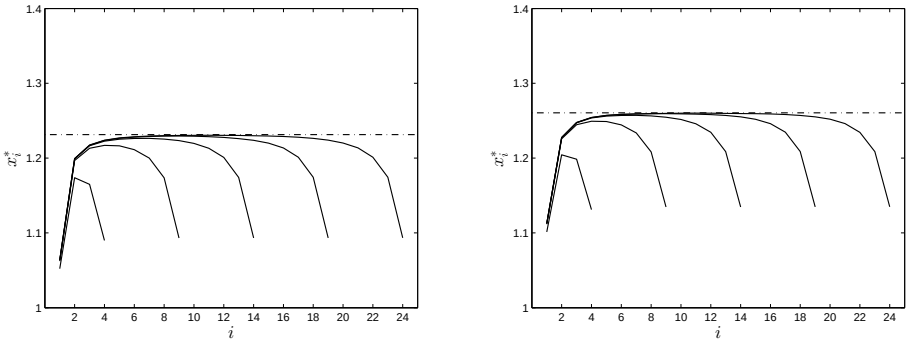
Figure 2.1: The optimal schedule in $x_i^\star$ s by sequential optimization for different scv s.

The sojourn-time distribution of each client is then found by performing the second step of our approach, as explained in Section 2.3. Next, based on these sojourn-time distributions we compute the optimal interarrival times $x_i^\star$ through both the sequential approach and the simultaneous approach. Both approaches are studied in case of an equally weighted ($\omega = \frac{1}{2}$) linear loss function and an equally weighted quadratic loss function. In our experiments for the simultaneous case, we study various schedule sizes ($n = 5, 10, \ldots, 25$ arrivals).

The results for the sequential approach are shown in Figure 2.1. From these figures we observe for linear loss that in case scv > 1 the interarrival times in the beginning of the scheme (up to the 5-th arrival) are smaller than the interarrival times for the cases scv $= 1$ or scv < 1. Later in the schedule (from arrival 10 onwards) the optimal interarrival times are approximately identical in size (that is, the curves for different scv values are close together), but increasing in the value of the scv. In case of a quadratic loss only the first and second interarrival time are close together, but from arrival 3 onwards the interarrival times differ substantially; again they are increasing in the scv, as expected. Overall, for any scv the quadratic loss yields larger optimal interarrival times than the linear loss.

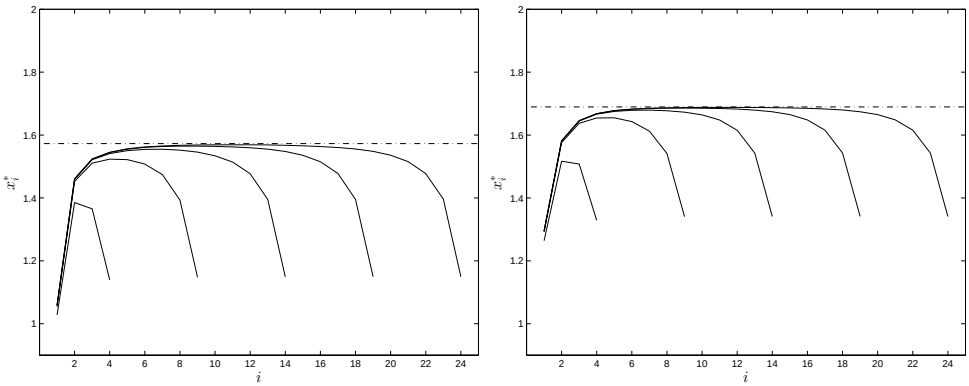
The five graphs in Figures 2.2, 2.3, 2.4, and 2.5 show the optimal schedules for $n = 5, 10, \ldots, 25$ arrivals under simultaneous optimization. From these figures we observe two interesting features. First, we observe that the interarrival times tend to increase in the value of the scv, and, again, for any of the scv values the quadratic loss leads to larger interarrival times than linear loss.

Second, the simultaneous approach leads to schemes for which the optimal interarrival times increase in the beginning and decrease towards the end of the scheme. The short interarrival times in the beginning of the schedule are essentially due to the fact that there the risk of waiting is relatively low. The short interarrival times at



(a) Linear loss, phase-type fit. (b) Quadratic loss, phase-type fit.

Figure 2.2: The optimal schedules in $x_i^\star$ s by simultaneous optimization for scv $= 0.1225 < 1$.

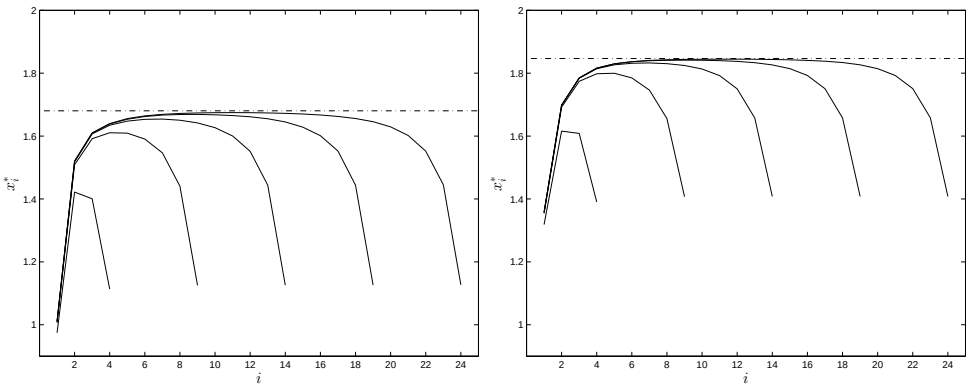


(a) Linear loss, phase-type fit. (b) Quadratic loss, phase-type fit.

Figure 2.3: The optimal schedules in $x_i^\star$ s by simultaneous optimization for scv $= 0.7186 < 1$.

the end can be explained from the fact that, despite a potentially substantial risk of high waiting times, there are few clients suffering from this (e.g., the last client having a large service time does not affect the waiting time of any subsequent clients). In the middle part the interarrival times are nearly constant indicating that the system is not affected by start- or end-of-session effects. The steady-state solution, the top horizontal line, is added in each case. In all settings the system seems to converge fast to the steady state. This justifies considering the steady-state solution in which all transient effects are neglected; see Section 2.5 for more results.

The pattern described above is the so-called *dome shape*, which was also found in the literature. The dome shape was found in Hassin and Mendel (2008) and Wang (1993), who minimized the expected waiting times and expected session-end time. Robinson and Chen (2003) and Vink et al. (2015) found it when minimizing the com-



(a) Linear loss. (b) Quadratic loss.

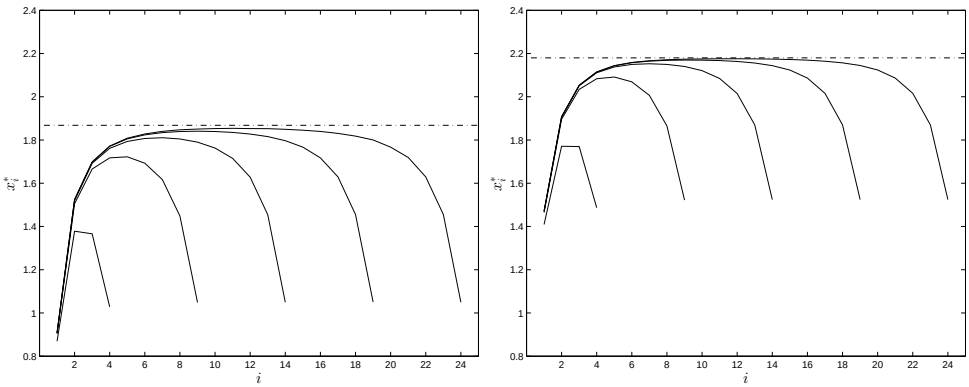
Figure 2.4: The optimal schedules in $x_i^\star$ s by simultaneous optimization for scv $= 1$.

bination of expected waiting and idle times. In Kaandorp and Koole (2007) also expected overtime is added to the latter minimization problem; for a more detailed discussion on the effect of overtime on the schedule, see Section 2.6.4. Furthermore, optimal interarrival times computed by the recursive beta distribution approximation, as advocated in Lau and Lau (2000), show a dome-shape pattern as well. In Section 2.6.3 we further compare this method with the phase-type approach. In addition, in case of a linear loss function the dome-shape pattern is also found when minimizing expected quadratic waiting and idle times, see Kemper et al. (2014).

2.5 Optimal scheduling in a steady-state environment

In the previous section we showed that our approach enables us to derive optimal interarrival times for different levels of scv, for both the sequential and simultaneous optimization, and for various risk functions and scheme sizes. Note that we chose the *equally weighted* linear and quadratic loss (that is, $\omega = \frac{1}{2}$), which we continue to do in this section, apart from Section 2.5.3 in which also the effect of ω on the optimal schedule will be studied. The primary goal of this section is to analyze the case of a large number of clients with i.i.d. service times. In this situation the schedules have constant interarrival times, and we show in detail how to determine these.

To study the steady-state interarrival time, given the value of the service-time distribution's scv, we need to derive the steady-state sojourn-time distribution of the corresponding D/G/1 queue, with the service times having either a mixture of Erlang distributions or hyperexponential distribution. We first show how we derive the steady-state sojourn-time distribution for various scv values; then we model the optimal interarrival time as a function of the scv for both sequential and simultaneous optimization using the loss functions mentioned above.



(a) Linear loss, phase-type fit. (b) Quadratic loss, phase-type fit.

Figure 2.5: The optimal schedules in $x_i^\star$ s by simultaneous optimization for scv $= 1.6036 > 1$.

We first point out that the optimality condition, that determines the optimal interarrival time $x^\star$, depends on the choice of the specific case (simultaneous vs. sequential, linear vs. quadratic). This optimality condition is a relation that involves both the distribution of the steady-state sojourn time S and the interarrival time $x^\star$. We first observe that (using the Lindley recursion, and the fact that W_i and I_i cannot be both positive)

$$\mathbb{E}g(I_i) + \mathbb{E}h(W_i) = \mathbb{E}\ell(S_{i-1} - x_{i-1}),$$

with $\ell(\cdot)$ defined through

$$\ell(x) := g(-x)\mathbf{1}_{\{x<0\}} + h(x)\mathbf{1}_{\{x\geq 0\}}.$$

In case of the sequential optimization approach, Kemper et al. (2014) proves that for any convex loss function the optimal interarrival time solves

$$\frac{\mathrm{d}\mathbb{E}\ell(S-x)}{\mathrm{d}x} = 0.$$

In the transient case we have to take the sojourn-time distributions of the individual clients, whereas in the steady-state case we have to take the stationary sojourn-time distribution. This representation leads to some appealing results: for equally weighted loss functions we obtain for linear loss the median of the sojourn time, i.e., $x^\star = F_S^{-1}\left(\frac{1}{2}\right)$, and for quadratic loss the mean of the sojourn time, i.e., $x^\star = \mathbb{E}S$.

In case of the simultaneous optimization approach, we are to evaluate, for large n,

$$\min_{x_1,\ldots,x_n} \sum_{i=1}^{n} \mathbb{E}\ell(S_i(x) - x_i) \approx n \cdot \min_x \mathbb{E}\ell(S(x) - x).$$

We write $S(x)$ rather than S to emphasize that the sojourn times depend on the interarrival time x. The optimal interarrival time then follows from the first-order condition

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}\ell(S(x)-x)=0.$$

For linear loss this yields the condition

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}|S(x)-x|=\frac{\mathrm{d}}{\mathrm{d}x}\left(\int_x^\infty (t-x)f_{S(x)}(t)\,\mathrm{d}t+\int_0^x (x-t)f_{S(x)}(t)\,\mathrm{d}t\right)=0, \tag{2.6}$$

whereas for quadratic loss we obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}\left(S(x)-x\right)^2=\frac{\mathrm{d}}{\mathrm{d}x}\left(\mathbb{E}S(x)^2-2x\mathbb{E}S(x)+x^2\right)=0. \tag{2.7}$$

The above formula suggests that the linear case requires knowledge of the distribution function of $S(x)$, but, interestingly, only $\mathbb{E}S(x)$ is needed. This can be seen as follows. Note that

$$\sum_{i=1}^n(\mathbb{E}I_i+\mathbb{E}W_i)=\sum_{i=1}^n((\mathbb{E}I_i+\mathbb{E}B_i)+(\mathbb{E}W_i+\mathbb{E}B_i))-2\sum_{i=1}^n\mathbb{E}B_i.$$

Now realize that, in addition to $W_i+B_i=S_i$, we also have that for the total length of the schedule

$$\sum_{i=1}^n(I_i+B_i)=t_n+S_n,$$

which can be recognized as the makespan. Realizing that the value of $\sum_{i=1}^n\mathbb{E}B_i$ does not affect the optimization, we conclude that minimizing the linear loss is equivalent to minimizing $\sum_{i=1}^n\mathbb{E}S_i+t_n+\mathbb{E}S_n$. Because $t_n\approx(n-1)x$, we are to minimize $\mathbb{E}S(x)+x$.

2.5.1 Steady-state results in case $\mathrm{scv}=1$

We illustrate the steady-state results in case of $\mathrm{scv}=1$, since it leads to pleasingly explicit results. Based on results of a G/M/1 queue (Tijms 1986), we have the following expression for the sojourn-time distribution in case $\mathrm{scv}=1$

$$\mathbb{P}(S\le x)=1-e^{-\mu(1-\sigma_x)x},\quad x\ge 0, \tag{2.8}$$

where $\sigma_x\in(0,1)$ solves $\sigma_x=e^{-(\mu-\mu\sigma_x)x}$.

Results for a sequential approach. In case the loss function is assumed linear, we solve

$x^\star = F_S^{-1}(1/2)$. We find

$$x = F_S^{-1}\left(\frac{1}{2}\right) = \frac{\log 2}{\mu(1-\sigma_x)} \quad \text{and} \quad F_S(x) = \frac{1}{2} = 1 - \sigma_x,$$

leading to an optimal schedule with interarrival times

$$x^\star = \frac{2\log 2}{\mu} \approx \frac{1.3862}{\mu}.$$

For the case of quadratic loss we solve

$$x = \mathbb{E}S = \frac{1}{\mu(1-\sigma_x)} \quad \text{and} \quad \log \sigma_x = -1,$$

and obtain

$$x^\star = \frac{e}{\mu(e-1)} \approx \frac{1.5820}{\mu}.$$

These limiting results are in line with those corresponding to the transient schemes in Figure 2.1a. For schedules of more than, say, 15 clients, the middle part of the schedule is close to the steady-state schedule. In this sequential setup quadratic loss leads to larger optimal interarrival times than linear loss.

Results for a simultaneous approach. To obtain the steady-state results in the simultaneous case and linear loss we solve the first order condition (2.6)

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}x}\left(\int_x^\infty (t-x) f_{S(x)}(t)\,\mathrm{d}t + \int_0^x (x-t) f_{S(x)}(t)\,\mathrm{d}t\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}x}\frac{1 - 2e^{-\mu(1-\sigma_x)x} - \mu(1-\sigma_x)x}{\mu(\sigma_x - 1)} = -\sigma_x' \frac{1 + \sigma_x(\log\sigma_x - 2)}{\mu(\sigma_x-1)^2\sigma_x} \\
&= \frac{1 + (\log\sigma_x - 2)\sigma_x}{\mu(\sigma_x-1)(1-\sigma_x+\sigma_x\log\sigma_x)} = 0,
\end{aligned}$$

where we used that

$$\sigma_x' = \frac{\mu\sigma_x(\sigma_x-1)}{1-\mu\sigma_x x} \quad \text{and} \quad x = \frac{\log\sigma_x}{\mu(\sigma_x-1)}.$$

This equation is solved for $\sigma_x \approx 0.32$, and we obtain

$$x^\star \approx \frac{1.6803}{\mu}.$$

The case of quadratic loss can be dealt with analogously; now the first-order condi-

tion (2.7) needs to be solved. We eventually obtain

$$x^\star \approx \frac{1.8466}{\mu}.$$

Again these limiting results agree well with the results of large transient schemes, see Figure 2.1b. We observe that, as in the sequential approach, in the simultaneous approach quadratic loss leads to larger optimal interarrival times than linear loss.

2.5.2 Steady-state results in case $\text{SCV} \neq 1$

As pointed out in Section 2.3, in the first step of our approach we fit a phase-type distribution to our service-time distribution. The special case of $\text{SCV} = 1$ (i.e., exponentially distributed service times) was dealt with in the previous subsection; now we focus on the cases in which $\text{SCV} \neq 1$.

Steady-state analysis for the $\mathrm{D}/E_{K-1,K}/1$ queue

As presented in Section 2.3.1, we use the $E_{K-1,K}(\mu, p)$ distribution to approximate service-time distributions with an SCV between $1/K$ and $1/(K-1)$, for $K \in \{2, 3, \ldots\}$. We analyze the resulting $\mathrm{D}/E_{K-1,K}/1$ queue through the sequence $(N_0, N_1, \ldots)$ with $N_0 = 0$ (the system starts empty), and N_i referring to the number of phases in the system just before the i-th arrival. These phases are exponentially distributed with mean $1/\mu$.

First observe that $(N_0, N_1, \ldots)$ follows a (discrete-time) Markov chain. The transition probabilities $p_{m,n} = \mathbb{P}(N_{i+1} = n \mid N_i = m)$ can easily be expressed in terms of the parameters K, p, μ, and the interarrival time x. Writing $\boldsymbol{P} = (p_{m,n})_{m,n=0}^{\infty}$ for the transition matrix, the steady-state distribution of N follows from:

$$\boldsymbol{a} = \boldsymbol{a}\boldsymbol{P}. \tag{2.9}$$

In addition the normalization constraint $a_0 + a_1 + a_2 + \ldots = 1$ needs to be imposed. Based on the limiting probabilities $\boldsymbol{a}$, we can derive the steady-state sojourn-time distribution and its moments, and hence we can deal with the various first-order conditions of Section 2.5. In order to solve (2.9), we truncate the state space to $\{0, \ldots, M\}$. Since the a_n decay roughly exponentially in n (with a decay rate that can be evaluated explicitly), it is not hard to select an appropriate value for M. Generally speaking, we saw in this $\text{SCV} < 1$ regime that the choice $M = 10 + K$ works well in nearly all situations.

Now with the sojourn-time distribution

$$\mathbb{P}(S \leq t) = \mathbb{P}(W + B \leq t) = \int_0^t F_W(t-u) f_B(u)\, \mathrm{d}u \tag{2.10}$$

and the vector $\boldsymbol{a}$ that solves (2.9), we may write

$$\begin{aligned}
\mathbb{P}\left(S \leq t\right) &= a_0 F_B(t) + \sum_{m=1}^{M} a_m \int_0^t \mu \frac{(\mu(t-u))^{m-1}}{(m-1)!} e^{-\mu(t-u)} f_B(u)\,\mathrm{d}u, \\
\mathbb{E}S &= \mathbb{E}W + \mathbb{E}B = \sum_{m=0}^{M} a_m \left(p\frac{m+K-1}{\mu} + (1-p)\frac{m+K}{\mu} \right), \\
\mathbb{E}S^2 &= \mathbb{E}W^2 + 2\mathbb{E}W\,\mathbb{E}B + \mathbb{E}B^2 \\
&= \sum_{m=1}^{M} a_m \frac{m(m+1)}{\mu^2} + 2\sum_{m=1}^{M} a_m \left(p\frac{m}{\mu}\frac{K-1}{\mu} + (1-p)\frac{m}{\mu}\frac{K}{\mu} \right) \\
&\quad + \left(p\frac{K(K-1)}{\mu^2} + (1-p)\frac{(K+1)K}{\mu^2} \right).
\end{aligned}$$

Steady-state results for the $\mathrm{D}/H_2/1$ queue

Mimicking the procedure described in Section 2.5.2, we now sketch a procedure to generate the steady-state sojourn-time distribution of a $\mathrm{D}/H_2/1$ system, so as to cover the case scv > 1. To do so, we analyze the queue through the sequence

$$((N_0, K_0), (N_1, K_1), \ldots) \quad \text{with} \quad (N_i, K_i) = (m, k)$$

meaning that the number of clients in the system just before the i-th arrival is m, and the phase of the client in service is k; if $k = 1$ the client in service is served with rate μ_1, and if $k = 2$ with rate μ_2. Evidently, $(N_i, K_i) \in \{0, 1, 2, \ldots\} \times \{1, 2\}$, and $(N_i, K_i) = (0, 0)$ corresponds to the empty system.

Again we truncate the state-space (in terms of the number of clients) to M, generate the transition probability matrix $\boldsymbol{P}$, and solve (2.9). Define $a_{m,k}$ as the steady-state probability of m clients in the system just before an arrival epoch, jointly with the phase of the client in service being k. We then evaluate (2.10), which leads to the following expressions. Writing $(H_{j,2})_j$ for a sequence of i.i.d. samples from a $H_2(\boldsymbol{\mu}; p)$ distribution, and assuming that $B^{(k)}$ has an exponential distribution with mean $1/\mu_k$ ($k = 1, 2$) we obtain for the distribution function:

$$P(S \leq t) = a_{0,0} F_B(t) + \sum_{m=0}^{M-1} \sum_{k=1}^{2} a_{m,k} \int_0^t \mathbb{P}\left(\sum_{j=1}^{m} H_{j,2} + B^{(k)} < t - u \right) f_B(u)\,\mathrm{d}u.$$

This expression can be evaluated further. With D a binomial distribution with parameters m and p, we have that

$$\sum_{j=1}^{m} H_{j,2} \stackrel{\mathrm{d}}{=} \sum_{i=1}^{D} B_i^{(1)} + \sum_{i=1}^{m-D} B_i^{(2)},$$

where $B_i^{(k)}$ are i.i.d. copies of $B^{(k)}$. For the corresponding first and second moment we obtain (with $\mathbb{E}H_2 = p/\mu_1 + (1-p)/\mu_2$)

$$\mathbb{E}S = \sum_{m=0}^{M-1} \left(a_{m,1} \left(m\mathbb{E}H_2 + \frac{1}{\mu_1} \right) + a_{m,2} \left(m\mathbb{E}H_2 + \frac{1}{\mu_2} \right) \right) + \mathbb{E}H_2,$$

$$\begin{aligned}\mathbb{E}S^2 = \sum_{m=0}^{M-1} \Bigg(& a_{m,1} \sum_{j=0}^{m} \binom{m}{j} \left(\frac{(j+1)(j+2)}{\mu_1^2} + 2\frac{(j+1)(m-j)}{\mu_1\mu_2} + \frac{(m-j)(m-j+1)}{\mu_2^2} \right) \\ & + a_{m,2} \sum_{j=0}^{m} \binom{m}{j} \left(\frac{j(j+1)}{\mu_1^2} + 2\frac{j(m-j+1)}{\mu_1\mu_2} + \frac{(m-j+1)(m-j+2)}{\mu_2^2} \right) \Bigg) \\ & + 2\sum_{m=0}^{M-1} \left(a_{m,1} \left(m\mathbb{E}H_2 + \frac{1}{\mu_1} \right) + a_{m,2} \left(m\mathbb{E}H_2 + \frac{1}{\mu_2} \right) \right) \mathbb{E}H_2 + \left(\frac{2p_1}{\mu_1^2} + \frac{2p_2}{\mu_2^2} \right).\end{aligned}$$

Evidently, by letting M grow large we get arbitrarily close to the true vector of stationary probabilities. We validated that the choice of $M = 25$ works well for the range of scv $\in (0, 3)$ when ω equals $\frac{1}{2}$. However, when ω is closer to 1 the truncation level M should be suitably increased.

2.5.3 Computational results in a steady-state environment

In this section we studied the optimal interarrival time as a function of the service-time distribution's scv $\in (0, 3)$. We did this for the sequential and the simultaneous optimization approach, in case of both an (equally weighted) linear loss function and an (equally weighted) quadratic loss function. From these results, depicted in Figure 2.6, we conclude that the steady-state optimal interarrival time is increasing in the scv for any of the four scenarios considered, as expected. In line with earlier findings, we observe that for each approach and for any scv $\in (0, 3)$ the quadratic loss function yields larger optimal interarrival times than the linear loss function. Furthermore, for any scv $\in (0, 3)$ the sequential approach yields smaller optimal interarrival times, for both quadratic and linear loss. Loosely speaking, this says that the sequential approach favors the service provider, since smaller interarrival times lead to smaller idle times.

Given an arbitrary scv, we now consider the effect of the weight parameter, ω. Since an increasing ω results in more weight assigned to the service provider's time, the interarrival times will decrease. Indeed, this is observed in Figure 2.7, where we plotted the dependence of the steady-state solutions resulting from the various optimization programs. We did these computations for scv $= 0.5625$, that is, the coefficient of variation (cv) equalling 0.75. This value is in $[0.35, 0.85]$, which, according to Çayırlı and Veral (2003) is the common range for the cv.

We observe that the solution curves of the simultaneous optimization and its sequential counterpart have a similar shape. When ω tends to 1, the occupation rate approaches 1, so that we need to increase the truncation level M to reliably compute

the steady-state solution. For this reason we set M to 50 when generating Figure 2.7.

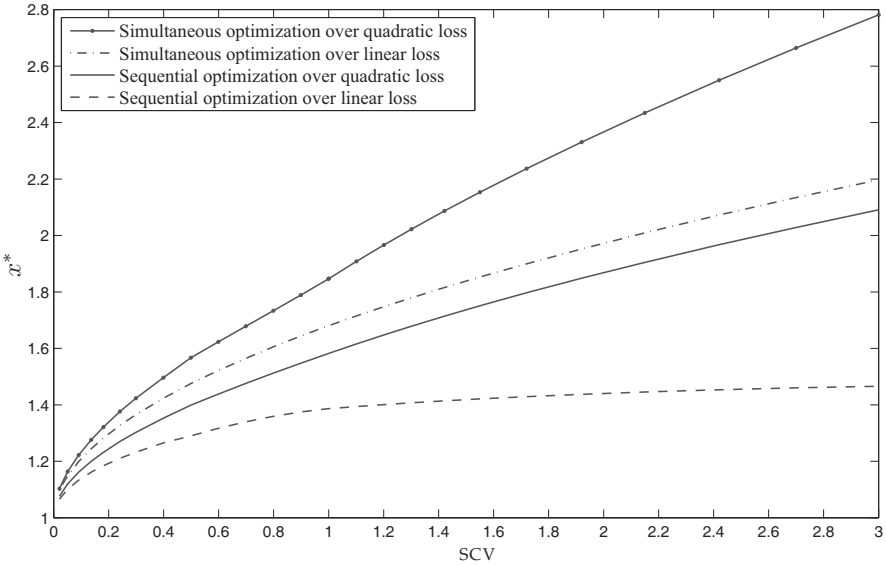


Figure 2.6: An overview of the optimal steady-state interarrival times $x^\star$ for four different optimization settings as a function of the SCV, where we take $M = 25$.

2.6 Discussion

In this section we systematically study different aspects of the schedules we developed. (i) In the first place we consider the *robustness* of our approach (both steady state and transient), so as to assess the effect of replacing generally distributed non-negative service times by their phase-type counterparts. (ii) Secondly, we compare our approach with the approach based on the characteristics of the beta distribution introduced by Lau and Lau (2000). (iii) Furthermore, we briefly discuss the effect of overtime in two transient settings. (iv) Also, we provide an account of the computational effort (in terms of computation time) for the various approaches. (v) We conclude this section with a comparison of the sequential and simultaneous optimization approach, in terms of the disutilities perceived by the individual agents.

2.6.1 Robustness of phase-type approach in steady state

To study the robustness of our approach for the optimal steady-state interarrival times as presented in the previous section, we apply our approach to a D/G/1 setting in which the service-time distribution is non-phase-type. We concentrate on the Weibull distribution and the lognormal distribution, as often seen in practice (Babes

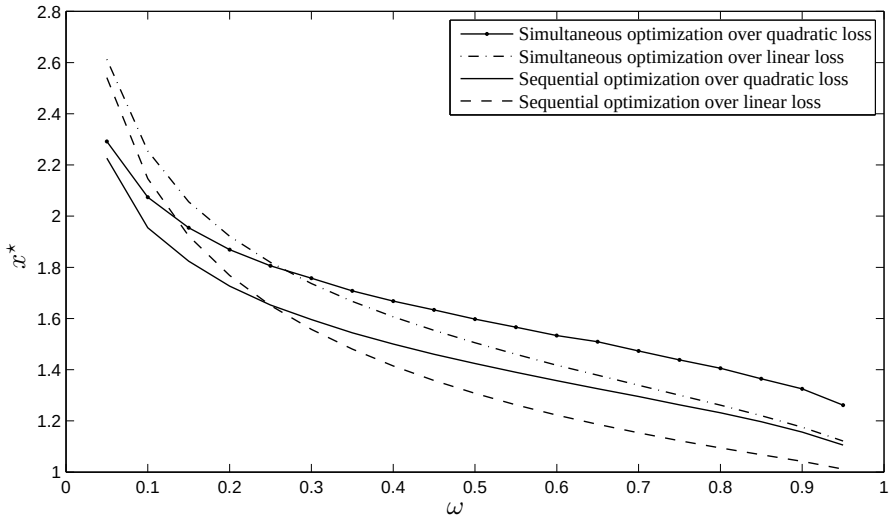


Figure 2.7: An overview of the optimal steady-state interarrival times $x^\star$ for four different optimization settings as a function of ω, where we take $M = 50$.

and Sarma 1991, Klassen and Rohleder 1996, Vink et al. 2015). In our study we assume again that the scv $= 0.5625$ (contained in the interval identified by Çayırlı and Veral (2003)).

Our study is set up as follows. We consider the following 2-parameter distributions:

▷ the Weibull distribution, with density

$$\frac{kx^{k-1}}{\lambda^k} e^{-(\frac{x}{\lambda})^k}$$

with parameters $k \approx 1.3476$, and $\lambda \approx 1.0902$, and

▷ the lognormal distribution, with density

$$\frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

with parameters $\mu = -\frac{1}{2}\log 1.5625$ and $\sigma = \sqrt{\log 1.5625}$,

which both lead to scv $= 0.5625$.

For all four scenarios (sequential or simultaneous approach, and quadratic or linear loss) we determined the optimal interarrival times by simulation, as follows. For a given steady-state interarrival time x, we simulate the queueing system using $100\,000$ clients (with a 'warm-up' corresponding to $1\,000$ clients), to estimate the value of the loss function for this specific x. In the loop around this routine, we identify the x

that minimizes the loss; this is done using MATLAB's minimization routine. We perform this optimization 100 times, and estimate the 'real' optimal interarrival time and risk, $\tilde{x}$ and $\tilde{R}$, by the average of the optimal interarrival times of the 100 individual experiments.

We compare these results with the optimal interarrival times resulting from our phase-type based technique with the SCV, i.e., 0.5625. In Table 2.1, we compare for the lognormal service-time distribution both the optimal interarrival time and the risk per client in steady state. The values resulting from the phase-type-based approach are denoted by $x^\star$ and $R^\star$. Finally, x^e and R^e refer to an approach where one assumes exponential service times instead (with mean 1 and SCV $= 1$). In a similar way, in Table 2.2 we compare for the Weibull service-time distribution the optimal interarrival time and the risk per client in steady state.

Setting	$\tilde{x}$	$\lvert\tilde{x} - x^\star\rvert$	$\lvert\tilde{x} - x^e\rvert$	$\tilde{R}$	$\lvert\tilde{R} - R^\star\rvert$	$\lvert\tilde{R} - R^e\rvert$
Sim. & quad.	1.6661	0.0631	0.1804	1.2866	0.0190	0.1075
Sim. & lin.	1.5085	0.0033	0.1718	0.8680	0.0002	0.0464
Seq. & quad.	1.4398	0.0156	0.1422	1.6147	0.0646	0.2937
Seq. & lin.	1.2749	0.0326	0.1114	1.0826	0.0724	0.1726

Table 2.1: The Monte Carlo optimal steady-state interarrival times and risk in case of lognormal service times compared with our approach and with an approach based on exponential service times.

Setting	$\tilde{x}$	$\lvert\tilde{x} - x^\star\rvert$	$\lvert\tilde{x} - x^e\rvert$	$\tilde{R}$	$\lvert\tilde{R} - R^\star\rvert$	$\lvert\tilde{R} - R^e\rvert$
Sim. & quad.	1.5946	0.0084	0.2519	1.0307	0.0005	0.2099
Sim. & lin.	1.5058	0.0007	0.1745	0.8260	0.0001	0.0488
Seq. & quad.	1.4223	0.0019	0.1597	1.2395	0.0051	0.2078
Seq. & lin.	1.3138	0.0063	0.0725	0.9542	0.0114	0.0878

Table 2.2: The Monte Carlo optimal steady-state interarrival times and risk in case of Weibull service times compared with our approach and with an approach based on exponential service times.

From Tables 2.1 and 2.2 we observe that the phase-type approximation has just a modest impact on the accuracy of the the optimal interarrival time and risk. It also shows that the naïve approach of assuming exponential distributed service times (thus completely ignoring SCV values different from 1) leads to large deviations from the optimal scheme.

2.6.2 Robustness of phase-type approach in transient environment

To study the robustness of the phase-type approach in a transient environment, we considered the same service-time distributions as used in Section 2.6.1: Weibull and lognormal. We took $n = 15$ clients to be scheduled resulting in 14 interarrival times. Again, we ran Monte Carlo simulation experiments to determine the optimal interarrival times and associated *total risk* (defined as the aggregate of the risks of all individual clients). In this case, however, the simulations were more involved than in the steady-state counterpart. For a given schedule $(x_1, \ldots, x_{14})$ we estimate the risk (by using $100\,000$ repetitions). Then we apply MATLAB's optimization procedure to identify the schedule that minimizes the risk. This optimization is performed 100 times. We estimate the 'real' optimal interarrival times and total risk ($\tilde{x}_1, \ldots, \tilde{x}_{14}$ and $\tilde{R}$) by the average of the 100 individual schedules.

In Table 2.3 we compare the simulation results with the phase-type approach and the assumption of exponential service times in case of optimization with a linear loss function, while in Table 2.4 we do the same in case of optimization with a quadratic loss function. Similar to Section 2.6.1, the values resulting from the approach based on the phase-type approximation method are denoted by $x_i^\star$ and $R^\star$, whereas x_i^e and R^e refer to the optimal arrival times and risk assuming exponential service times.

As in Section 2.6.1 we see that the phase-type approach results in a significant gain, in terms of the total risk, compared to the results obtained when assuming exponential service times. We did not include simulations related to the sequential optimization, since these are only affected by a start-of-session effect resulting in rapid convergence to steady state, as seen in Figure 2.1. Therefore these simulations are redundant.

2.6.3 Comparison with the approach by Lau and Lau

Instead of using phase-type distributions to compute optimal schedules, one can opt for using a recursive method based on the beta distribution, see Lau and Lau (2000). This approach involves four parameters, which are set by matching the first four moments of the service-time distributions. In Table 2.5 we compare for both methods the optimized schedules in terms of arrival times, expected waiting times and expected idle times per client. The clients' ($n = 20$) service times are i.i.d. with mean 1, variance 0.25, skewness 1, and kurtosis 4; the risk per client to be minimized is $R_i^{(\ell,10/11)}$, i.e., $\omega = \frac{10}{11}$. These settings are chosen such that they match the problem considered by Lau and Lau (2000). To compare the *total risk* found by Lau and Lau (2000), denoted by $\mathbb{E}C_s$, the risk per client R_i can be scaled arbitrarily, since it does not affect the optimal schedule (cf. Equation (2.5)). Noting that $R_1 = 0$ we have

$$\sum_{i=2}^{20} R_i^{(\ell,10/11)} = \frac{10}{11}\left(\sum_{i=2}^{20} \mathbb{E}I_i + \frac{1}{10}\sum_{i=2}^{20} \mathbb{E}W_i\right) = \frac{10}{11}\mathbb{E}C_s.$$

Setting	Lognormal service times			Weibullian service times		
i	$\tilde{x}_i$	$\lvert\tilde{x}_i - x_i^\star\rvert$	$\lvert\tilde{x}_i - x_i^e\rvert$	$\tilde{x}_i$	$\lvert\tilde{x}_i - x_i^\star\rvert$	$\lvert\tilde{x}_i - x_i^e\rvert$
1	1.0101	0.0546	0.0031	1.0739	0.0093	0.0634
2	1.3546	0.0543	0.1625	1.4188	0.0100	0.0983
3	1.4282	0.0322	0.1789	1.4652	0.0062	0.1418
4	1.4551	0.0232	0.1796	1.4808	0.0049	0.1539
5	1.4702	0.0158	0.1767	1.4879	0.0041	0.1590
6	1.4762	0.0137	0.1775	1.4918	0.0041	0.1620
7	1.4773	0.0130	0.1766	1.4916	0.0046	0.1622
8	1.4748	0.0128	0.1751	1.4898	0.0045	0.1602
9	1.4666	0.0147	0.1751	1.4834	0.0049	0.1583
10	1.4530	0.0189	0.1740	1.4739	0.0042	0.1531
11	1.4312	0.0231	0.1694	1.4579	0.0047	0.1427
12	1.3911	0.0321	0.1606	1.4283	0.0059	0.1234
13	1.3066	0.0461	0.1364	1.3606	0.0080	0.0823
14	1.0884	0.0535	0.0379	1.1525	0.0106	0.0262
Total risk	$\tilde{R}$	$\lvert\tilde{R} - R^\star\rvert$	$\lvert\tilde{R} - R^e\rvert$	$\tilde{R}$	$\lvert\tilde{R} - R^\star\rvert$	$\lvert\tilde{R} - R^e\rvert$
	5.6083	0.0093	0.2201	5.5264	0.0003	0.1637

Table 2.3: The Monte Carlo optimal times and risk in simultaneous optimization of linear risk in a transient environment with lognormal or Weibull service times compared with our approach and with an approach based on exponential service times.

We find that the phase-type fit approach based on the first two moments $\mu = 1$, scv $= 0.25$ gives nearly identical results, in terms of the optimal schedule and the corresponding waiting and idle times. Furthermore, in case of a linear loss function the phase-type fit approach uses explicit expressions for the expected idle and waiting times, so that it should perform at roughly the same speed as the method by Lau and Lau.

The major strength of the phase-type approach is that it requires only the first two moments of the service-time distribution. This tends to be sufficient to determine the optimal schedule (see the discussion in Çayırlı and Veral 2003, Section 2.5). In addition, estimating higher moments such as skewness and kurtosis requires a large sample size to obtain accurate estimates.

2.6.4 The effect of overtime on the schedule

In our optimization problem we only considered the minimization of risk in terms of waiting and idle time. This allowed us to study the difference between the sequential and simultaneous approach, and for both cases how they are affected by the scv. Another performance measure in healthcare, which could be modeled easily, is the overtime O. Overtime is defined as the actual session-end time set minus the sched-

Setting	Lognormal service times			Weibullian service times		
i	$\tilde{x}_i$	$\lvert\tilde{x}_i - x_i^\star\rvert$	$\lvert\tilde{x}_i - x_i^e\rvert$	$\tilde{x}_i$	$\lvert\tilde{x}_i - x_i^\star\rvert$	$\lvert\tilde{x}_i - x_i^e\rvert$
1	1.2672	0.0099	0.0897	1.2550	0.0044	0.1019
2	1.5221	0.0124	0.1753	1.5087	0.0041	0.1887
3	1.5955	0.0309	0.1878	1.5597	0.0056	0.2236
4	1.6244	0.0413	0.1895	1.5762	0.0073	0.2378
5	1.6388	0.0481	0.1878	1.5831	0.0078	0.2435
6	1.6442	0.0505	0.1875	1.5859	0.0078	0.2458
7	1.6461	0.0521	0.1865	1.5862	0.0082	0.2464
8	1.6452	0.0528	0.1851	1.5846	0.0078	0.2457
9	1.6397	0.0513	0.1847	1.5804	0.0082	0.2440
10	1.6281	0.0473	0.1850	1.5730	0.0080	0.2401
11	1.6084	0.0418	0.1834	1.5598	0.0069	0.2321
12	1.5711	0.0327	0.1788	1.5329	0.0057	0.2170
13	1.4937	0.0184	0.1636	1.4723	0.0036	0.1851
14	1.3033	0.0039	0.1047	1.2994	0.0020	0.1085
Total risk	$\tilde{R}$	$\lvert\tilde{R} - R^\star\rvert$	$\lvert\tilde{R} - R^e\rvert$	$\tilde{R}$	$\lvert\tilde{R} - R^\star\rvert$	$\lvert\tilde{R} - R^e\rvert$
	8.1236	0.0364	0.5801	6.7542	0.0012	0.9764

Table 2.4: The Monte Carlo optimal times and risk in simultaneous optimization of quadratic risk in a transient environment with lognormal or Weibull service times compared with our approach and with an approach based on exponential service times.

uled end time T, that is

$$O := \max\{\textsc{set} - T, 0\} = \max\left\{\sum_{i=1}^{n}(I_i + B_i) - T, 0\right\}.$$

To stress that O depends on the value of T, we add a subscript and write O_T. To study the effect of overtime we extend the simultaneous optimization approach with expected overtime. We focus on linear risk, in a schedule of $n = 15$ clients (cf. Equation (2.5)), i.e., we consider

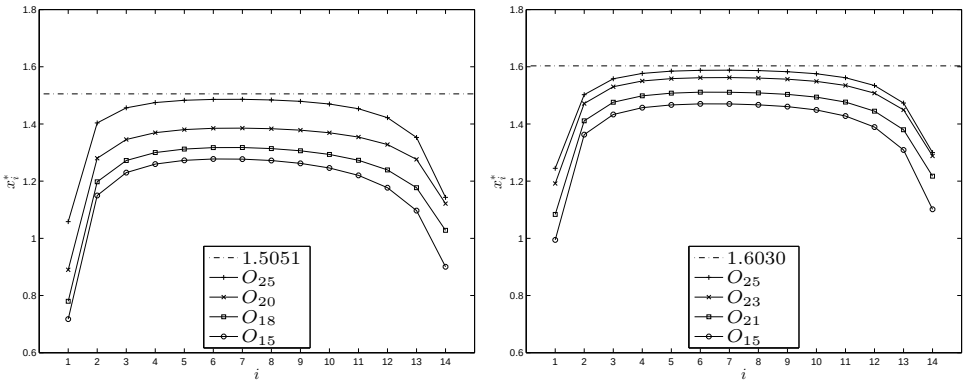
$$\min_{t_1,\ldots,t_n} \sum_{i=1}^{15} R_i + \beta\mathbb{E}O_T = \min_{t_1,\ldots,t_n} \sum_{i=1}^{15} (\omega\mathbb{E}I_i + (1-\omega)\mathbb{E}W_i) + \beta\mathbb{E}O_T.$$

Take $\omega = 0.5$ (equal weights) and $\beta/\omega = 1.5$, which models the situation in which overtime is valued roughly 50% higher than idle time (Çayırlı et al. 2012). In Figure 2.8 we see the influence of overtime on the schedule; here the service times are chosen by the phase-type approach, so as to generate a distribution with mean 1 and $\textsc{scv} = 0.5625$ as in Section 2.6.1. We consider for both linear as quadratic loss four cases, where $T \geq 15$ varies. A special case is O_{15} where all clients must be served with

Method	Beta distribution approach			Phase-type approach		
Client (i)	Arrival times	$\mathbb{E}W_i$	$\mathbb{E}I_i$	Arrival times	$\mathbb{E}W_i$	$\mathbb{E}I_i$
2	0.542	0.477	0.022	0.535	0.489	0.024
5	3.395	0.792	0.068	3.424	0.780	0.069
10	8.603	0.969	0.072	8.635	0.951	0.077
15	13.785	1.146	0.070	13.815	1.127	0.065
20	18.467	1.698	0.017	18.514	1.644	0.021
$\sum \mathbb{E}W_i$ or $\sum \mathbb{E}I_i$		19.514	1.139		19.165	1.160
Total risk	2.810			2.798		

Table 2.5: The optimized schedules for the beta distribution approach and phase-type fit approach. The schedules minimize the total risk $\sum R_i^{(\ell,10/11)}$ as defined in Equation (2.5).

their expected service time in order to avoid overtime a queue with load 1. Indeed, we see that the schedule gets tighter when the scheduled session-end time decreases. Including overtime has a similar effect as assigning a higher weight to the idle times in the risk function, viz. result in tighter schedules. When T tends to infinity we are in the case of our original models optimized in Section 2.4.



(a) Linear loss. (b) Quadratic loss.

Figure 2.8: The effect of overtime on the schedule with simultaneous optimization over linear and quadratic loss, $n = 15$, with the corresponding steady-state solutions.

2.6.5 Computational effort of the various numerical approaches

We now give a brief account of the computational effort required to evaluate the schedules, and further describe how our code has been set up. A general remark is that, for obvious reasons, determining steady-state schedules is substantially less expensive than determining transient schedules. In our numerical experiments, we generated transient schedules of up to 25 clients. All programming was done in

MATLAB, benefiting from its built-in function for determining roots, its minimization routine, and its numerical integration routine. As a result the code to be developed was relatively minimal. The most complicated cases (25 clients, SCV > 1) required a computation time of a few minutes, but usually the computations were considerably faster.

The structure of the code is as follows. Here x is the steady-state interarrival time, whereas $\boldsymbol{x} = (x_1, \ldots, x_{n-1})$ is the transient schedule.

1. Determine the phase-type fit (hyperexponential or Erlang mixture) for given mean and SCV.

2. The corresponding loss function is computed as follows.

   (i) Regarding the steady state, for a given x, the equilibrium probabilities are found through the embedded Markov chain, choosing the truncation level suitably. These probabilities yield the steady-state distribution of the waiting time for an arriving client, see Section 2.5. Then one computes the steady-state sojourn-time distribution by evaluating the convolution of the waiting time and service time.

   (ii) In the transient case one uses the recursive method outlined in Section 2.3 to evaluate the sojourn-time distribution for given $\boldsymbol{x}$.

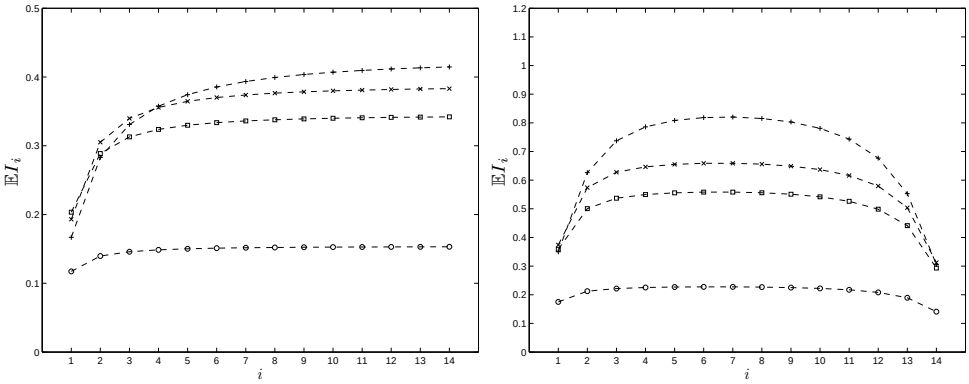
   We now evaluate the chosen loss function (sequential or simultaneous approach, and quadratic or linear loss).

3. Given the loss function, we perform the minimization. In the sequential approach this is implemented by solving the first order condition.

Obviously, the computational effort can be substantially reduced by tailoring the software more directly to our specific needs, e.g. by using 3rd generation programming environments (such as C++). Also, a significant reduction of the computational effort can be achieved by using optimal values of a previously calculated, 'nearby' scenario as starting values when determining a next schedule; this idea can be exploited for instance when generating optimal schedules for a range of SCV values.

2.6.6 Comparison of sequential and simultaneous approaches

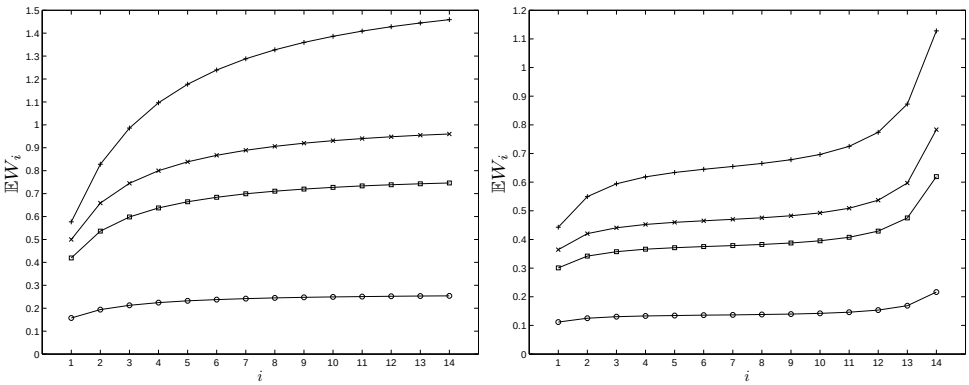
In this section we study the expected waiting time and idle time associated with each individual client, so as to compare the impact of the chosen approach (i.e., sequential vs. simultaneous). In Figures 2.9 and 2.10 we do so for linear loss, whereas Figures 2.11a and 2.11b relate to quadratic loss. The graphs in the figures are labeled as in Figure 2.1: that is, the plus signs refer to an SCV $= 1.6036$, the crosses refer to an SCV $= 1.0000$, the squares to an SCV $= 0.7186$, and the circles to an SCV $= 0.1225$. In all experiments we focus on $n = 15$ clients and hence 14 interarrival times, but other values of n show very similar behavior.



(a) Linear loss, sequential approach. (b) Linear loss, simultaneous approach.

Figure 2.9: The optimal idle times by sequential and simultaneous approach for the various SCV values, in case of a linear loss.

Figures 2.9a and 2.9b show the idle times for each arrival for the sequential (2.9a) and simultaneous (2.9b) optimization approach, with linear loss. From these results we observe that the mean idle times in the sequential approach are in general smaller than those in the simultaneous approach. Furthermore, the patterns of the mean idle times resonate the patterns of the optimal individual interarrival times — see Figure 2.1a for the sequential approach, and Figures 2.2a, 2.3a, 2.4a, and 2.5a for the simultaneous approach.

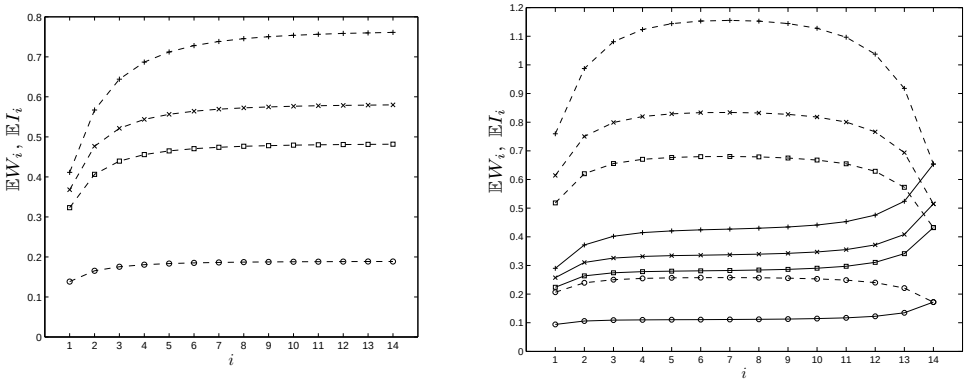


(a) Linear loss, sequential approach. (b) Linear loss, simultaneous approach.

Figure 2.10: The optimal waiting times by sequential and simultaneous approach for the various SCV values, in case of a linear loss.

Next, Figures 2.10a and 2.10b show the mean waiting times for both approaches, with linear loss. From these results we observe that the individual waiting times are larger

in case of the sequential approach. This means that, together with the results of the idle times, we conclude that the sequential approach favors the service provider. Furthermore, we observe that the individual waiting times are more variable for the simultaneous approach than for the sequential approach; this salient feature illustrates the difference in 'fairness' between both schemes.



(a) Quadratic loss, sequential approach; the curves corresponding to the mean idle and waiting times lie on top of each other.

(b) Quadratic loss, simultaneous approach; top curves are mean idle times, bottom curves are mean waiting times.

Figure 2.11: The optimal waiting times by sequential and simultaneous approach for the various scv values, in case of a quadratic loss.

Finally, we discuss the mean idle and waiting times for quadratic loss, as shown in Figures 2.11a and 2.11b. From the sequential results of Figure 2.11a, we observe that for each arrival the mean idle time equals the mean waiting time. This follows from the risk function presented in (2.4) with $\omega = \frac{1}{2}$, and its corresponding first order condition. The optimal interarrival time follows from $\mathbb{E}(S_{i-1} - x_{i-1}) = 0$ for clients $i = 2, \ldots, n$, entailing that $x^\star$ is chosen so that $\mathbb{E}I_i = \mathbb{E}W_i$.

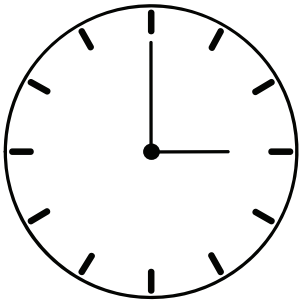
From the simultaneous results of Figure 2.11b, we again conclude that the mean idle times are larger than for the sequential approach; at the same time, the mean waiting times are smaller. From this observation it is seen that also for quadratic loss the sequential approach favors the server. Also, we see that the *dome shape* is reflected in the pattern of the mean idles times; cf. Figures 2.2b, 2.3b, 2.4b, and 2.5b; and the mean waiting times of each individual arrival are more variable than for the sequential approach, in line with what we observed for linear loss. The fact that the mean idle time equals the mean waiting time for the final arrival essentially follows from the fact that the final arrival is 'sequentially' scheduled, since no subsequent client is to be scheduled.

2.7 Conclusion

This chapter presents an approach for generating optimal appointment schedules. In our procedure we replace service-time distributions by their phase-type counterparts, and then (either sequentially or simultaneously) optimize a utility function. The procedures are backed by a series of numerical experiments, that also shed light on the impact of the utility function and the service-times' variability (expressed in terms of the squared coefficient of variation, scv) on the optimal interarrival times.

These numerical studies provide evidence for the feasibility of the procedure. At the same time we assessed its robustness; in particular it was shown that approximating non-phase-type distributions (Weibull, lognormal) by phase-type distributions, based on a two-moment fit, hardly affects the optimality of the produced schedule.

There are various directions for future research. (i) In the first place the setup can be made more realistic, that is, more in line with specific conditions in healthcare settings. For instance, ideally schedules should be flexible enough to be able to deal with walk-ins. This requires the possibility to adapt the schedule *on the fly*. (ii) In the second place one could think of situations with multiple servers, in which it also needs to be determined to which server each client is assigned. In addition, it would be interesting to study settings where clients have to undergo multiple (rather than just one) services. (iii) In the numerical examples we considered the situation of all clients having the same service-time distribution. It is readily checked, however, that the modeling framework does not require such a uniformity: all computations can be performed for heterogeneous service times as well.



3. Comparing Optimal Schedules to Practical Principles in Appointment Scheduling

The main contribution of this chapter is the generation of appointment schedules that incorporate random service times and no-shows. Therefore, we extend the approach presented in Chapter 2. More precisely, we adapt the phase-type approach presented in Chapter 2 to incorporate no-shows. In addition, we provide a comparison of such schedules with those resulting from straightforward heuristics (with environments in which the parameters match those observed in practice). We consider the problem in a transient environment, where a finite number of clients (patients) are scheduled.

3.1 Introduction

When appointment scheduling research took off, computational power was limited, and one therefore primarily focused on heuristics. Perhaps the most classical example is the 'equidistant' schedule in which the block lengths (slot sizes) equal the clients' average service time. However, as is known from the pioneering work of Welch and Bailey (Welch and Bailey 1952), such a scheme performs poorly in many cases; to remedy this, they propose to overbook the first slot with an additional client. It was shown by Ho and Lau (1992) that this rule, often referred to as the Bailey-Welch rule, is fairly robust over a broad range of situations. It has been proven by Kemper et al. (2014), however, that equidistant schedules ultimately lead to long waiting times

when the number of clients grows large (with the mean waiting time of the n-th client roughly behaving as $\sqrt{n}$). In that regime the Bailey-Welch rule may lead to schedules that are highly unattractive to clients.

In many studies one relies on extensive, and often case-specific, simulations (see e.g., Bailey 1952, Welch and Bailey 1952, Ho and Lau 1992, 1999). A more generic approach is to assume a specific service-time distribution that allows explicit expressions for the waiting-time and idle-time distributions, so as to analytically generate schedules. The easiest distribution to work with is the exponential distribution, as studied by Hassin and Mendel (2008), but this choice, corresponding with $\text{scv} = 1$, typically overestimates the variability. One has therefore looked into methods in which the service-time distribution is fitted by a distribution which provides more freedom but that still allows a (semi-)analytic solution. More specifically, a fit with the beta distribution was advocated by Lau and Lau (2000), whereas a phase-type distribution was proposed by e.g. Wang (1997) and Kuiper et al. (2015), see also Chapter 2. In the latter reference the validity of the phase-type approach, in which the first two moments of the service-time distribution are fit, has been thoroughly examined for typical service-time distributions observed in healthcare. A final option is to rely on discrete-time versions of the continuous schedules, thus facilitating a very fast evaluation of any schedule; see e.g. Brahimi and Worthington (1991) and De Vuyst et al. (2011).

A fundamentally different approach was followed by Zacharias and Pinedo (2014). In that setup service times are deterministic and equal to the block length (both normalized to 1) and the only stochastic component in the model is the no-show probability. No-shows are indeed a prevalent problem in many healthcare scheduling practices and no-show percentages are typically 5-30%, as reported by Çayırlı and Veral (2003). Moreover, in an assessment by Ho and Lau (1992) of environmental factors that affect appointment schedules, it was found that no-shows and the service-time variability have a profound impact on the performance of the appointment schedule, which motivates why a proper design should take both stochastic characteristics into account.

The remainder of the chapter is organized as follows. In Section 3.2 we introduce the concept of a *risk function* that balances the interests of the service provider (doctor) and the clients, and then we extend the phase-type approach given in Chapter 2. The framework thus obtained enables us to evaluate an optimal schedule, i.e., the schedule that minimizes the risk function. Then, in Section 3.3 we apply commonly used scheduling heuristics and numerically compare them with the optimal schedule. We conclude this chapter with a discussion of the results in Section 3.4.

3.2 Modeling approach

In this section we outline the stochastic model and the method used for evaluation and optimization of appointment schedules. The main focus is on extending the framework given in Chapter 2 to a setup that incorporates clients' no-shows. This extension is non-trivial as there are some subtleties to be dealt with. First we describe the risk function that represents the expected loss per client in terms of mean idle times and mean waiting times. Then we describe the phase-type approach, and point out how the recursive method should be adapted to deal with no-shows. We assume clients and the physician, specialist or surgeon (also referred to as *provider*) to be punctual.

3.2.1 Framework, risk function

In mathematical terms, the appointment scheduling problem aims at determining suitable epochs t_1 up to t_n at which the n clients are scheduled to arrive. We denote by $\mathcal{V} := (t_1, \ldots, t_n)$ the resulting schedule. In this chapter, the service times B_1 up to B_n are assumed independent and identically distributed (but this assumption can be alleviated). We write I_i for the server's (random) idle time prior to the i-th arrival, and W_i for the (random) waiting time of the i-th client.

The *risk* associated with client i, defined as weighted sum of the expected idle and the expected waiting time, is given by

$$R_i^{(\omega)}(t_1, \ldots, t_i) = \omega\, \mathbb{E}I_i + (1-\omega)\, \mathbb{E}W_i,$$

where the $\omega \in (0,1)$ is a weight factor that reflects the importance of the provider's (idle) time versus the clients' (waiting) time. Note that the random variables I_i and W_i are affected by the arrival epochs $t_1, \ldots, t_i$ of the preceding clients. The *aggregate risk* is given by

$$R^{(\omega)}(t_1, \ldots, t_n) = \sum_{i=1}^{n} R_i^{(\omega)} = \sum_{i=1}^{n} (\omega\, \mathbb{E}I_i + (1-\omega)\, \mathbb{E}W_i). \tag{3.1}$$

Since we consider *expected* idle and waiting times, we do not have to compute explicit idle and waiting-time distributions to evaluate Eqn. (3.1). Instead, we rely on the definition of the *sojourn time* as the sum of waiting and service time:

$$S_i = W_i + B_i, \tag{3.2}$$

and also on the fact that the total duration of a session (the makespan) equals the sum of idle and service times:

$$t_i + S_i = \sum_{j=1}^{i} (I_j + B_j)\,. \tag{3.3}$$

If we take the expected value in Eqns. (3.2) and (3.3), we end up with a formula for the expected waiting time and a recursion for the expected idle time of the i-th client in terms of his expected sojourn time:

$$\mathbb{E}W_i = \mathbb{E}S_i - \mathbb{E}B;$$
$$\mathbb{E}I_i = t_i + \mathbb{E}S_i - i\,\mathbb{E}B - \sum_{j=1}^{i-1} \mathbb{E}I_j.$$

We thus conclude that the sojourn-time distribution (and in particular its mean) enables a recursive algorithm to find the mean waiting times and the mean idle times, with which we can evaluate our objective function.

Next, we propose to approximate the service-time distributions by phase-type distributions. It is well known that phase-type distributions, which are mixtures and convolutions of exponential distributions, can be used to approximate any positive distribution with arbitrary precision, see e.g. Tijms (1986) and Asmussen et al. (1996).

3.2.2 Phase-type distribution

We approximate the service-time distribution of B by a phase-type counterpart based on the mean and the scv, in the way proposed by Tijms (1986). The candidate distributions that we rely on are mixtures of Erlang distributions $E_{K-1,K}(\mu;p)$ and the hyperexponential distribution $H_2(\boldsymbol{\mu};p)$. These phase-type distribution are characterized by an $m \in \mathbb{N}$, an m-dimensional (row) vector $\boldsymbol{\alpha}$ with nonnegative entries adding up to 1, and an $(m \times m)$-dimensional matrix $\boldsymbol{S} = (s_{ij})_{i,j=1}^m$ such that $s_{ii} < 0$, $s_{ij} \geq 0$ and $\sum_{j=1}^m s_{ij} \leq 0$ for any $i \in \{1,\ldots,m\}$. For the two specific phase-type distributions mentioned above the representations in terms of m, $\boldsymbol{\alpha}$, and $\boldsymbol{S}$ are:

- In case scv < 1, we use an $E_{K-1,K}(\mu;p)$ distribution. In this case $m = K$, and the vector $\boldsymbol{\alpha}$ is such that $\alpha_1 = 1$ and $\alpha_i = 0$ for $i = 2,\ldots,K$. In addition, $s_{ii} = -\mu$ for $i = 1,\ldots,K$ and $s_{i,i+1} = -s_{ii} = \mu$ for $i = 1,\ldots,K-2$, while $s_{K-1,K} = (1-p)\mu$; all other entries of $\boldsymbol{S}$ are 0.
- In case scv ≥ 1, we use a $H_2(\boldsymbol{\mu};p)$ distribution. Then $m = 2$, and $\alpha_1 = p = 1 - \alpha_2$. Also, $s_{ii} = -\mu_i$, for $i = 1,2$, while the other two entries of $\boldsymbol{S}$ equal 0.
- If scv $= 1$ then the exponential distribution Exp(μ) is used.

Observe that the first case (scv < 1) is particularly relevant in healthcare as it contains the typical cv values in the range of 0.35 and 0.85.

By $B =_{\mathrm{d}} \mathbb{P}\mathrm{h}(\boldsymbol{\alpha},\boldsymbol{S})$ we denote that B has a phase-type distribution. An attractive property of phase-type distributions is that the moments have explicit forms (see e.g., Asmussen 2003). For the mean we have

$$\mathbb{E}B = -\boldsymbol{\alpha}\boldsymbol{S}^{-1}\boldsymbol{e}_m, \tag{3.4}$$

with $\boldsymbol{e}_m$ being an m-dimensional column vector consisting of ones. This can be evaluated fast for the phase-type distributions, since $\boldsymbol{S}$ is an upper diagonal matrix in case of a mixture of Erlang distributions or a diagonal matrix for the hyperexponential distribution.

Our approach accommodates that each scheduled arrival has a $q \in (0,1)$ of being a *no-show*. The phase-type distribution is adapted reflecting that each client requires no service with probability q and a service time B with probability $(1-q)$. As a consequence, the vector $\boldsymbol{\alpha}$ is multiplied by $(1-q)$, that is, $B =_{\mathrm{d}} \mathbb{P}\mathrm{h}((1-q)\boldsymbol{\alpha}, \boldsymbol{S})$.

3.2.3 Recursive approach, incorporating no-shows

The key idea is to use the recursive procedure proposed by Wang (1997) to compute each client's sojourn-time distribution. These are of phase-type, and hence the objective is to identify the $\boldsymbol{\alpha}$ and $\boldsymbol{S}$ in its representation $\mathbb{P}\mathrm{h}(\boldsymbol{\alpha}, \boldsymbol{S})$. At each moment in time we keep track of the number of clients in the system together with the phase of the client in service; the current state of the system is given by these two variables. Notice that the i-th client's sojourn time is only affected by his $i-1$ predecessors. Since typically the number of clients to be scheduled is relatively small, the dimensionality of the problem stays manageable and our techniques are effective for realistic numbers of clients. For large numbers of clients, one could neglect some of the dependence between the clients by introducing a so-called *lag order*, as proposed in Vink et al. (2015), see also Chapter 5. If the lag order is k, then this means that only clients $i-k$ up to $i-1$ can affect the sojourn time of the i-th client.

To outline the procedure under no-shows we define the following bivariate process $\{N_i(t), K_i(t), t \geq 0\}$ for client $i = 1, \ldots, n$, where $N_i(t) \in \{0, \ldots, i-1\}$ represents the number of clients in front of the i-th arriving client, t time units after his arrival. The second component, $K_i(t) \in \{1, \ldots, m\}$, represents the phase of the client in service at t time units after the arrival. We introduce the corresponding probabilities, for $t \geq 0$, $i = 1, \ldots, n$, $j = 0, \ldots, i-1$, and $k = 1, \ldots, m$:

$$p_{j,k}^{(i)}(t) = \mathbb{P}\left(N_i(t) = j, K_i(t) = k\right).$$

In addition, the vector $\boldsymbol{P}_i(t)$ (of dimension mi) is given by

$$\left(p_{i-1,1}^{(i)}(t), \ldots, p_{i-1,m}^{(i)}(t), p_{i-2,1}^{(i)}(t), \ldots, p_{i-2,m}^{(i)}(t), \ldots, p_{0,1}^{(i)}(t), \ldots, p_{0,m}^{(i)}(t)\right).$$

The sojourn-time distribution of the i-th client can be computed from $\boldsymbol{P}_i(t)$ through the following identity, with $\boldsymbol{e}_{mi}$ an all-ones vector of dimension mi:

$$F_i(t) := \mathbb{P}(S_i \leq t) = 1 - \sum_{j=0}^{i-1} \sum_{k=1}^{m} p_{j,k}^{(i)}(t) = 1 - \boldsymbol{P}_i(t)\boldsymbol{e}_{mi}.$$

The sojourn time of the first client, arriving at $t_1 = 0$, is determined by his service-time distribution:

$$\boldsymbol{P}_1(t) = (1-q)\boldsymbol{\alpha}\exp(\boldsymbol{S}t), \quad \text{for} \quad t \geq 0,$$

which is an m-dimensional object. The second client, arriving $x_1 := t_2 - t_1$ time units after the first client, either shows up with probability $(1-q)$ (thus increasing the number of clients by one), or not. For any $t \geq 0$, with $\mathbf{0}_m$ denoting an all-zeros vector of dimension m, this leads to

$$\boldsymbol{P}_2(t) = ((1-q)(\boldsymbol{P}_1(x_1), \boldsymbol{\alpha}F_1(x_1)) + q(\mathbf{0}_m, \boldsymbol{P}_1(x_1)))\exp(\boldsymbol{S}_2 t),$$

which is an object of dimension $2m$; here,

$$\boldsymbol{S}_2 := \begin{pmatrix} \boldsymbol{S} & \boldsymbol{s}\boldsymbol{\alpha} \\ \mathbf{0}_{m,m} & \boldsymbol{S} \end{pmatrix},$$

with $\boldsymbol{s} := -\boldsymbol{S}\boldsymbol{e}_m$ and $\mathbf{0}_{m,m}$ denoting an $(m \times m)$-dimensional all-zeros matrix. For the other clients the vector $\boldsymbol{P}_i(t)$ (dimension mi) can be found from $\boldsymbol{P}_{i-1}(t)$ (dimension $m(i-1)$) by the recursion, for $t \geq 0$,

$$\boldsymbol{P}_i(t) = ((1-q)(\boldsymbol{P}_{i-1}(x_{i-1}), \boldsymbol{\alpha}F_{i-1}(x_{i-1})), q(\mathbf{0}_m, \boldsymbol{P}_{i-1}(x_{i-1})))\exp(\boldsymbol{S}_i t).$$

Here $x_{i-1} := t_i - t_{i-1}$ (the *interarrival time*) and the matrix $\boldsymbol{S}_i$ is defined recursively by

$$\boldsymbol{S}_i := \begin{pmatrix} \boldsymbol{S}_{i-1} & \boldsymbol{T}_i \\ \mathbf{0}_{m,(i-1)m} & \boldsymbol{S} \end{pmatrix},$$

with $\boldsymbol{T}_i$ a matrix of dimension $(i-1)m \times m$ defined by

$$\boldsymbol{T}_i := (\mathbf{0}_{m,m}, \ldots, \mathbf{0}_{m,m}, \boldsymbol{s}\boldsymbol{\alpha})^{\mathrm{T}}.$$

Above we have outlined the procedure for evaluating the aggregate risk of *any* schedule $\mathcal{V}$. Using this recursive procedure, we can use standard numerical tools to *optimize* over all possible schedules, so as to find the *optimal* schedule. The optimal interarrival times $(x_1^\star, \ldots, x_{n-1}^\star)$ that minimize the risk function (for a given weight ω) give the optimal schedule $\mathcal{V}^\star = (t_1^\star, \ldots, t_n^\star)$ by $t_i^\star = \sum_{j=1}^{i-1} x_j^\star$ for $i = 2, \ldots, n$. We will use this procedure in Section 3.3 to evaluate commonly used scheduling heuristics and compare those with the optimal schedule $\mathcal{V}^\star$.

3.3 Experiments and results

The primary objective of this section is to examine how frequently used scheduling heuristics perform relative to each other, and relative to optimal schedules (i.e., schedules that minimize $R^{(\omega)}(t_1, \ldots, t_n)$ for some $\omega \in [0, 1]$). We do so by evaluat-

ing the so-called *efficient frontier*, consisting of all combinations of the averaged (over all clients) mean waiting times and aggregated mean idle times when varying the weight ω. In our computations we rely on the phase-type approach, augmented to incorporate no-shows, as has been described in Section 3.2.

We consider five heuristics, each of them based on the average service time $\mathbb{E}B$.

$\mathcal{A}$ An equidistant schedule: $t_i = (i-1)\,\mathbb{E}B$ for all i. This is the simplest rule.

$\mathcal{B}$ The Bailey-Welch rule with 2 clients in the first time slot: $t_1 = t_2 = 0$; $t_i = (i-2)\,\mathbb{E}B$ for $i > 2$. It was shown by Ho and Lau (1999) that this heuristic is very robust.

$\mathcal{C}$ Adaptation of the Bailey-Welch rule with 3 clients in the first time slot: $t_1 = t_2 = t_3 = 0$; $t_i = (i-3)\,\mathbb{E}B$ for $i > 3$.

$\mathcal{D}$ Adaptation of the Bailey-Welch rule with 4 clients in the first time slot: $t_1 = t_2 = t_3 = t_4 = 0$; $t_i = (i-4)\,\mathbb{E}B$ for $i > 4$. Rule $\mathcal{C}$ and $\mathcal{D}$ are adaptions of the original Bailey-Welch rule.

$\mathcal{A}^2$ Block appointment rule with two clients arriving for a double slot: $t_i = t_{i+1} = 2(i-1)\,\mathbb{E}B$ for $i = 1, 3, 5, \ldots$. This rule is also known as the *two-at-a-time scheduling rule* studied by Soriano (1966).

We also consider variants of these heuristics adapted to deal with no-shows. Thus $\mathcal{A}_q$ is rule $\mathcal{A}$ but with block length equal to $(1-q)\mathbb{E}B$, and analogously for $\mathcal{B}_q, \mathcal{C}_q, \mathcal{D}_q$ and $\mathcal{A}_q^2$.

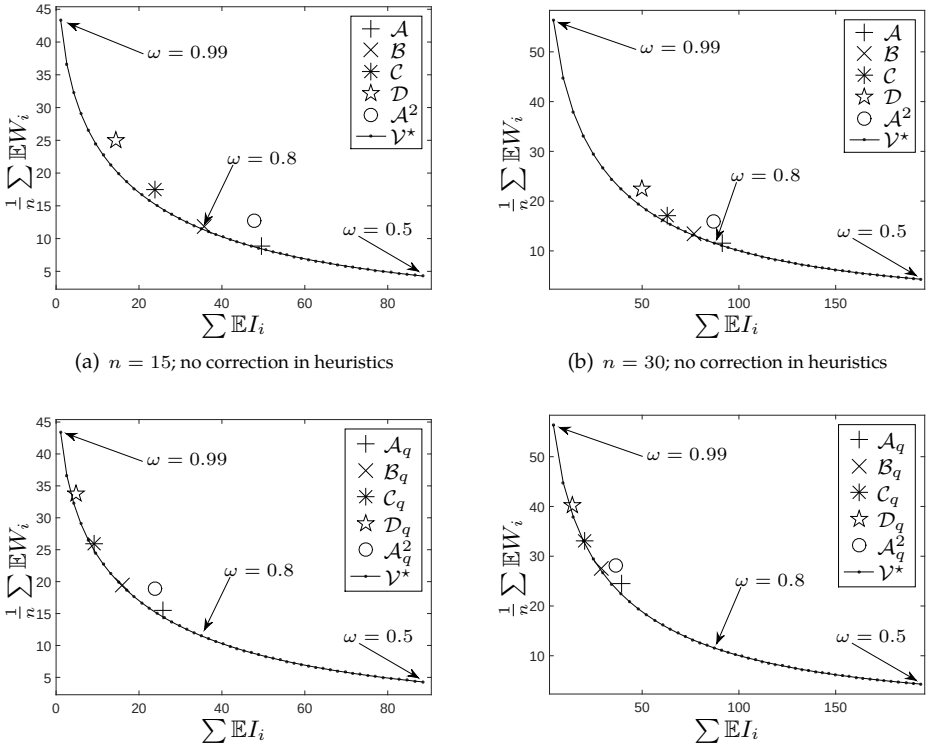
The ten scheduling rules are evaluated in a range of scenarios. The scenarios vary in terms of their scv, q and n. In each scenario we compute the expected idle and waiting times. And in addition we compute in each scenario the optimal schedule $\mathcal{V}^\star$ for $\omega \in (0.5, 0.99)$ (in steps of 0.01). The value $\omega = 0.5$ corresponds to equally weighted idle and waiting time. In typical healthcare settings ω is larger. For $\omega = 0.99$ the provider's idle time is valued 99 times more important than the clients' waiting time. Choosing $\omega = 1$ corresponds to the trivial schedule in which all clients arrive at time zero, such that the risk function has the value (in the case the schedules are not corrected for no-shows)

$$\frac{1}{n}\sum_{i=1}^{n-1} i\,\mathbb{E}B = \frac{(n-1)}{2}\,\mathbb{E}B,$$

as the (expected) idle times are zero. Computing the optimal schedules $\mathcal{V}^\star$ for each ω results in what we have called the efficient frontier. They are obtained by optimizing the risk function, and therefore no schedule can outperform them.

First we study the impact of the number of clients on the schedules. To this end, we choose scv $= 0.4225$ and $q = 0.175$, and set the number of clients first to 15 and then to 30, see Fig. 3.1. It is first observed that implementing the no-show correction

in the scheduling rules has a substantial effect. Furthermore, comparing Fig. 3.1(a) with Fig. 3.1(b) shows, for the situation without no-show correction, that the heuristics converge to each other as the number of clients grows. For large n the scheduling rules have very similar performance as they correspond to equal slot sizes in a stationary queue with load smaller than 1 (apart from the beginning of the session). Figs. 3.1(c) and 3.1(d) show the performance of the heuristics with no-show correction. In these scenarios the queues have load 1 and there is no convergence to steady state. Consequently the various heuristics do not converge in performance (even for large n).



(a) $n = 15$; no correction in heuristics

(b) $n = 30$; no correction in heuristics

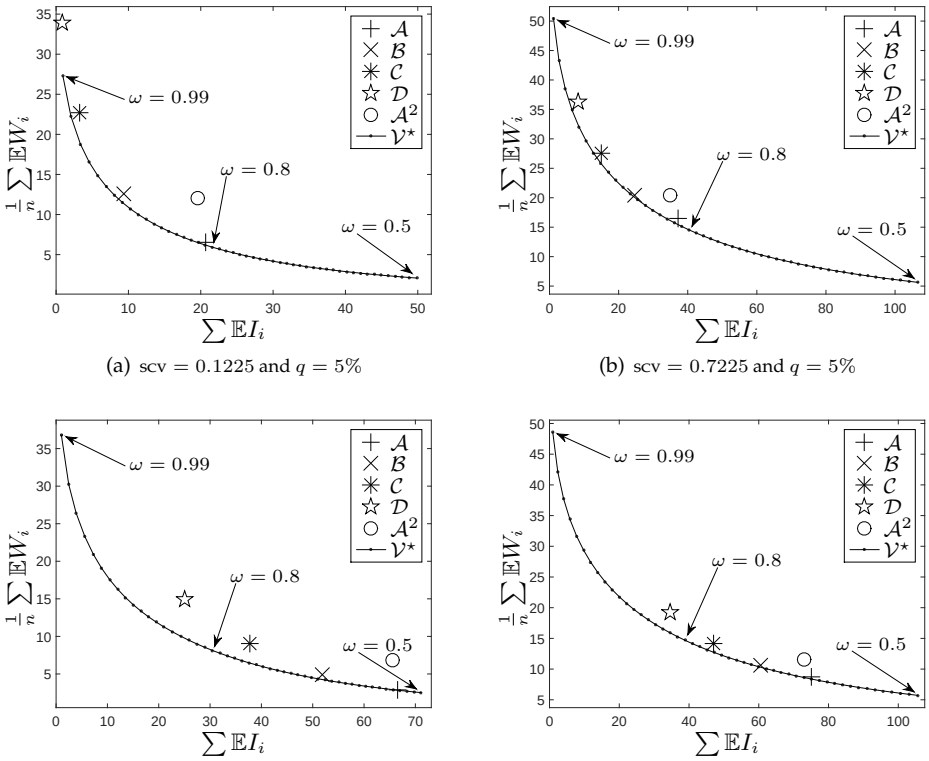
(c) $n = 15$; no-show correction in heuristics

(d) $n = 30$; no-show correction in heuristics

Figure 3.1: Various appointment schedules in the setting of scv $= 0.4225$ and no-show probability $q = 17.5\%$, where the number of clients n is varied horizontally, and where there is a correction for the no-show probability q in the bottom graphs.

In a second series of experiments, we let the number of clients be $n = 15$, and consider *four settings* that match the boundaries of typical healthcare situations that have been reported by Çayırlı and Veral (2003). We take (i) the cv (scv) equal to 0.35 (0.1225) and 0.85 (0.7225), and (ii) we let the no-show probability q have the values 0.05 and 0.30. Figures 3.2 and 3.3 present the performance of the heuristics with and

without no-show correction.



(a) scv $= 0.1225$ and $q = 5\%$

(b) scv $= 0.7225$ and $q = 5\%$

(c) scv $= 0.1225$ and $q = 30\%$

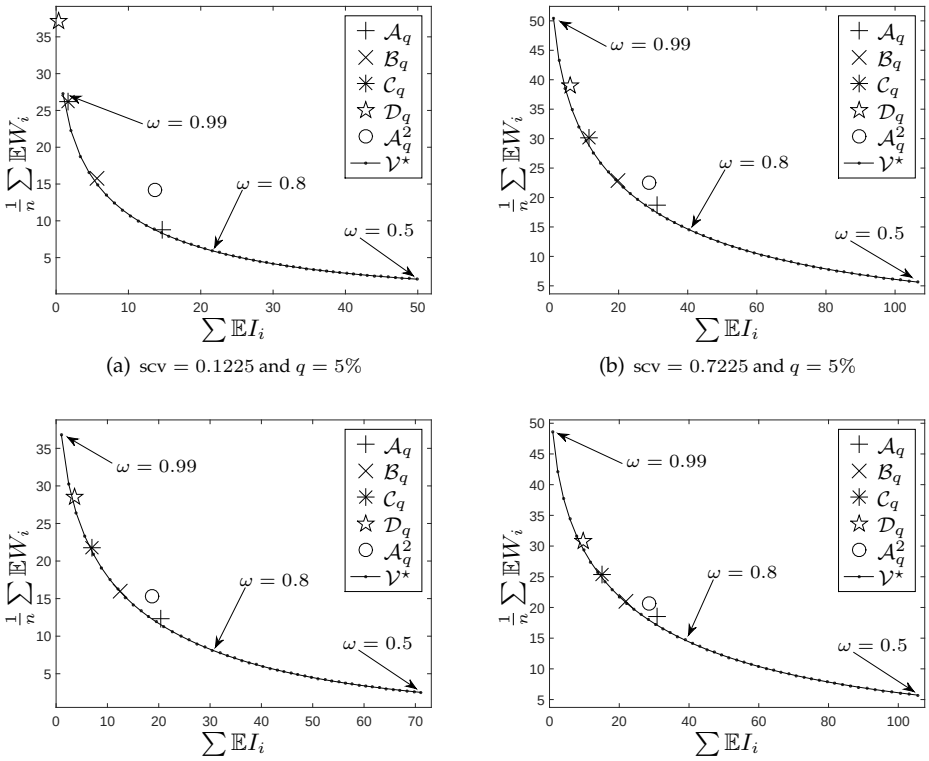
(d) scv $= 0.7225$ and $q = 30\%$

Figure 3.2: Various appointment scheduling heuristics based on the average service time $\mathbb{E}B$ in four healthcare settings: scv increases from left to right and no-show probability increases from top to bottom.

Comparing the uncorrected and corrected scheduling rules, it is observed that the corrected rules lead to lower expected idle times (and hence higher expected waiting times). Furthermore, the corrected scheduling rules cover only a small range of the trade-off parameter ω. Zooming in on Fig. 3.2(b) and Fig. 3.2(c) (equivalently, for the corrected versions, Fig. 3.3(b) and Fig. 3.3(c)) one finds that from the two environmental factors an increase in the service-time variability has a more significant impact on the schedule than an increase in the no-show probability.

It is remarkable that many heuristics lie close to the efficient frontier, which could be indicative for a flat risk function Eqn. (3.1). The exception is $\mathcal{A}^2$, which is clearly not an efficient rule. In this rule if both clients show up at the same time one of the clients has an expected waiting time that is at least equal to the average service time.

Finally, the figures show that the relative position of the scheduling rules on the efficient frontier is preserved across scenarios, except for $\mathcal{A}^2$ and $\mathcal{A}_q^2$. Therefore these



(a) scv $= 0.1225$ and $q = 5\%$

(b) scv $= 0.7225$ and $q = 5\%$

(c) scv $= 0.1225$ and $q = 30\%$

(d) scv $= 0.7225$ and $q = 30\%$

Figure 3.3: Various appointment scheduling heuristics based on the *no-show corrected* average service time $(1-q)\mathbb{E}B$ in four healthcare settings: scv increases from left to right and no-show probability increases from top to bottom.

rules implicate a choice in the tradeoff of idle and waiting times and thus also in ω. The expected idle time is lower for rules with additional clients at the beginning of the session, and therefore the order from low to high expected idle times is $\mathcal{D}, \mathcal{C}, \mathcal{B}, \mathcal{A}$. Obviously sorting by waiting time gives the reverse order.

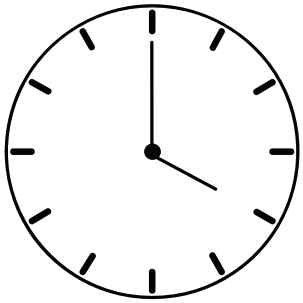
3.4 Conclusion and discussion

In this chapter we have considered appointment schedules in a setting with service-time variability. We have pointed out how to adapt the approach proposed in Chapter 2 to incorporate the possibility of clients not showing up. Despite the fact that service-time variability and no-shows are highly relevant in healthcare, we are among the first to systematically assess both effects in a computational study.

In our numerical study, we compared a number of heuristic schedules by eval-

uating for each of them the sum of the expected idle times and the average of the expected waiting times. From the two stochastic components we find that the service-time variability (expressed in terms of the scv) has a more significant impact than the no-show probability.

There are several directions for further research. In the first place, one could consider alternative risk functions. In this chapter we only considered the sum of the expected idle and waiting times, which has become the standard risk function in the literature. Given the fact that for both the provider and clients a modest leeway is hardly negatively perceived, one could argue that a quadratic loss function may be more appropriate than a linear one. The choice of the risk function, however, may have a substantial impact on the optimal schedule, as demonstrated in Chapter 2.



4. A Computational Approach to Appointment Scheduling for Two Servers in Tandem

In this chapter we extend the single-server framework as described in Chapter 2 to multi-stage systems. A practical limitation of the former framework is that in many healthcare settings patients do not necessarily undergo just one service. Instead, clients (patients) may sequentially be served at multiple service stations (or: *nodes*). There are numerous examples of this, such as a patient who first has an x-ray made and then sees a doctor, or a patient who first has an intake and is then examined by a doctor. In those contexts, representing the system as a single queue is obviously not appropriate: one should rather consider a (two-node) *tandem* network (sometimes referred to as an $D/G/1 \rightarrow G/1$ queue), where the individual queues correspond to the two service stages.

4.1 Introduction

Despite the relevance of multi-stage systems, the vast majority of all papers focuses on single nodes; see e.g. some remarks on this in Section 2.1 of Çayırlı and Veral (2003). Notable exceptions that do cover multi-node situations are the case study (backed by Monte Carlo simulation) presented by Rising et al. (1973) and the visual simulation-based approach by Swisher et al. (2001). An elementary queueing model, designed for a specific multi-stage application (i.e., an ear, nose & throat outpatient

clinic), has been developed by Cox et al. (1985). While there is a variety of situations in which single-stage systems are a sufficiently accurate representation of the real system, one would ideally like to have appointment scheduling algorithms that can deal with more complex structures as well, such as the ones presented by Côté and Stein (2007).

To set up appointment schedules one needs to be able to evaluate the transient distribution of the underlying queueing model; this transient distribution facilitates the computation of an objective function, which is then optimized over the arrival epochs. Single queues have a reasonable level of tractability, (multi-node) queueing networks on the contrary are known to allow such an explicit transient analysis only in specific cases (e.g. Jackson networks, relying heavily on various restrictive exponentiality assumptions). In light of this, this chapter, based on Kuiper and Mandjes (2015a), is among the first contributions to appointment scheduling in a multi-stage context. Importantly, our framework does not impose any restrictive assumptions on the service-time distributions.

The approach proposed in this chapter uses the transient distribution of the tandem queue to set up schedules. In systems in which the number of clients to be scheduled is relatively large and in which (per node) the clients' service times stem from the same distribution, however, we can work with the corresponding *stationary* distributions. The second main novelty of this chapter lies in the way we evaluate such steady-state distributions; it is noted that the approach we present here is significantly more efficient than the one we developed in Chapter 2.

The primary application area of multi-stage appointment scheduling lies in healthcare, but there is potential use in several other areas as well. In industrial applications, where jobs pass through multiple stations (e.g. machines) in a flow line, the cost function can be expressed in terms of holding cost and the (opportunity) cost of station idleness (e.g., Dallery and Gershwin 1992).

The structure of the chapter is as follows. In Section 4.2 we state the scheduling problem for the two-node tandem in terms of idle and waiting times, which will be addressed in the rest of this chapter. Section 4.3 explains in detail how one can exploit phase-type characterizations of the service-time distributions to compute idle and waiting times; various extensions of the 'base model' (viz. heterogeneous service-time distributions, the situation in which the second node may block the first node) are dealt with in Section 4.4. Then, in Section 4.5, we use the developed methodology to numerically compute optimal schedules and study the effect of various parameters on the optimal schedule.

We also see that, for the special case that all service times are identically distributed, schedules in this transient setting rapidly approach steady-state, and hence one could approximate the transient schedule by its stationary counterpart (which has the evident advantage of being easier to evaluate). In Section 4.6 we demonstrate an efficient technique to compute the optimal steady-state schedule, and we use this procedure to evaluate such schedules (thus showing the impact of the various model

parameters). The chapter is concluded by a brief discussion in Section 4.7.

4.2 Problem description

As argued in the introduction, appointment schedules are intended to properly balance the *disutilities* experienced by both the server (i.e., the provider) and the clients. More concretely, the schedules should be such that the server's idle time is kept sufficiently low, while at the same time controlling the clients' waiting times. In this section, we first recapitulate how a mathematical framework can be set up in the single-server setting, and then extend this to the two-node tandem.

A central role is played by the notion of a *risk function*, measuring the system's (aggregate) disutility, which captures the effect of idle times and waiting times in a single expression. A common choice (see e.g., Wang 1993, Hassin and Mendel 2008, Kuiper et al. 2015) is the (potentially weighted) sum of the mean idle times and the mean waiting times, i.e.,

$$R(t_1, \ldots, t_n) = \sum_{i=1}^{n} (\omega \mathbb{E}[I_i] + (1-\omega)\mathbb{E}[W_i]) \quad \text{with} \quad \omega \in (0,1); \tag{4.1}$$

here the t_i s (for $i \in \{1, \ldots, n\}$) denote the arrival epochs of the n clients, W_i is the waiting time of the i-th client, and I_i is the idle time prior to the arrival of the i-th client. Observe that shifting the value of ω amounts to trading off the interests of the server and the clients.

The idea is to balance idle and waiting times by optimizing the risk function (4.1) over all arrivals epochs $0 \leq t_1 \leq \ldots \leq t_n$:

$$\min_{t_1,\ldots,t_n} R(t_1, \ldots, t_n) = \min_{t_1,\ldots,t_n} \sum_{i=1}^{n} (\omega \mathbb{E}[I_i] + (1-\omega)\mathbb{E}[W_i]) \,. \tag{4.2}$$

As pointed out in Kemper et al. (2014), this optimization problem, cast in terms of mean idle and waiting times, can also be expressed in terms of the clients' sojourn times. More precisely, it turns out that, as a direct consequence of the Lindley recursion,

$$\min_{t_1,\ldots,t_n} \sum_{i=1}^{n} (\omega \mathbb{E}[I_i] + (1-\omega)\mathbb{E}[W_i]) = \min_{x_1,\ldots,x_{n-1}} \sum_{i=1}^{n-1} \mathbb{E}[\ell(S_i - x_i)], \tag{4.3}$$

where $\ell(\cdot)$, evaluated in $S_i - x_i$ denotes the so-called *loss function*, defined for any $x \in \mathbb{R}$ by

$$\ell(x) := -\omega x \mathbf{1}_{\{x<0\}} + (1-\omega) x \mathbf{1}_{\{x>0\}},$$

the variable S_i is the sojourn time of the i-th client (i.e., waiting time W_i plus service time B_i), and the non-negative numbers $x_{i-1} := t_i - t_{i-1}$ (with $t_1 = 0$) correspond to the *interarrival times*.

In the tandem setting clients are sequentially served by two servers. It is assumed throughout that the service times of client i at node r, for $i \in \{1, \ldots, n\}$ and $r \in \{1, 2\}$, are independent non-negative random variables $B_{r,i}$. As before, appointment schedules are sequences of epochs $t_1, \ldots, t_n$ at which the n clients are supposed to arrive at the first node. However, now both servers generate their own risk, so that the problem we are faced with is to find $t_1, \ldots, t_n$ that minimize a risk function that incorporates idle times and waiting times at *both* nodes. In self-evident notation, we are therefore to evaluate

$$\min_{t_1,\ldots,t_n} \sum_{i=1}^{n} \left\{ w\left(\omega_1 \mathbb{E}[I_{1,i}] + (1-\omega_1)\mathbb{E}[W_{1,i}]\right) + (1-w)\left(\omega_2 \mathbb{E}[I_{2,i}] + (1-\omega_2)\mathbb{E}[W_{2,i}]\right) \right\}, \tag{4.4}$$

with $\omega_1, \omega_2, w \in (0, 1)$. At each node, we balance the loss incurred by idle and waiting times, as before, reflected by ω_1 and ω_2 respectively, whereas w weighs the disutilities corresponding to both nodes. Observe that setting w equal to 1 in (4.4) reduces to the familiar D/G/1 queue, i.e., the single-node appointment scheduling problem, see Chapter 2 and 3.

4.3 Methodology

The goal of this chapter is to devise techniques to solve the optimization problem in (4.4) for general service times $B_{r,i}$. In this generality this is problematic, as evaluating the objective function essentially requires us to determine the sojourn-time distributions of the clients in our tandem system of the D/G/1 $\rightarrow$ G/1 type, for which no closed-form solution is available. We remedy this by relying on the approach proposed and validated in Chapter 2: as we explain in Section 4.3.1, we approximate the service times by their *phase-type* counterparts, for which computations turn out to be feasible.

The second subsection points out how the mean waiting time and idle times, to be used in (4.4), can be computed from the mean sojourn times; it is thus sufficient to be able to determine the clients' sojourn-time distributions. Where Chapter 2 focused on the single D/G/1 queue, we demonstrate how to extend this procedure to its tandem counterpart. This tandem case turns out to be substantially more involved; we present in Section 4.3.3 in detail the (recursive) method that yields the sojourn-time distribution of each client.

4.3.1 Phase-type distribution

In our study we use the idea, advocated in Tijms (1986), to match the first and second moment of the service-time distribution by a so-called phase-type distribution. Observe that it is equivalent to fitting the mean and the *squared coefficient of variation* scv; the scv of a random variable is defined as its variance divided by the square of the

mean. In line with the previous chapters, we choose to use a mixture of two Erlang distributions ($E_{K-1,K}(\mu;p)$) in case the service-time distribution has an scv smaller than 1; an exponential distribution in case scv equals one; and a hyperexponential distribution ($H_2(\boldsymbol{\mu};p)$) in case of an scv larger than 1.

Next, we point out how to express the mixture of Erlang distributions, the exponential distribution, and hyperexponential distribution as a phase-type distribution. A phase-type distribution is characterized by a 'dimension' $m \in \mathbb{N}$, an m-dimensional row vector $\boldsymbol{\alpha}$ with nonnegative entries adding up to 1, and an $(m \times m)$-dimensional matrix $\boldsymbol{S} = (s_{ij})_{i,j=1}^m$ such that $s_{ii} < 0$, $s_{ij} \geq 0$, and $\sum_{j=1}^m s_{ij} \leq 0$ for any $i \in \{1, \ldots, m\}$. If B has a phase-type distribution with representation $(\boldsymbol{\alpha}, \boldsymbol{S})$ — which we denote by $B =_{\mathrm{d}} \mathbb{P}\mathrm{h}(\boldsymbol{\alpha}, \boldsymbol{S})$ — then its first moment equals

$$\mathbb{E}[B] = -\boldsymbol{\alpha}\boldsymbol{S}^{-1}\mathbf{1}_m, \tag{4.5}$$

where $\mathbf{1}_m$ is an all-one column vector of dimension m; higher moments can be given in closed form as well, as can be found in e.g. (Asmussen 2003, Section III.4).

As indicated above, the following three types of phase-type distributions cover all values of the mean and scv.

- ▷ In case scv < 1, we use an $E_{K-1,K}(\mu;p)$ distribution, which corresponds to an Erlang distribution of $K-1$ phases and mean $(K-1)/\mu$ with probability p, and an Erlang distribution with K phases and mean K/μ with probability $1-p$. Then $m = K$, and the vector $\boldsymbol{\alpha}$ is such that $\alpha_1 = 1$ and $\alpha_i = 0$ for $i = 2, \ldots, K$. In addition $s_{ii} = -\mu$ for $i = 1, \ldots, K$ and $s_{i,i+1} = -s_{ii} = \mu$ for $i = 1, \ldots, K-2$, while $s_{K-1,K} = (1-p)\mu$; all other entries are 0. The corresponding scv equals

  $$\frac{K - p^2}{(K-p)^2},$$

  which lies between $1/(K-1)$ and $1/K$ for $K \in \{2, 3, \ldots\}$. We can thus uniquely identify an $E_{K-1,K}(\mu;p)$ distribution matching the first two moments of the target distribution, as long as scv < 1.

- ▷ In case scv $= 1$, we use an *Exp*(μ) distribution. Then $m = 1$, $\alpha_1 = 1$ and $\boldsymbol{S} = s_{11} = -\mu$.

- ▷ In case scv > 1, we use a $H_2(\boldsymbol{\mu};p)$ distribution: we sample from *Exp*(μ_1) distribution with probability p, and from an *Exp*(μ_2) distribution with probability $1-p$. Then $m = 2$, and $\alpha_1 = p = 1 - \alpha_2$. Also, $s_{ii} = -\mu_i$, for $i = 1, 2$, while the other two entries of $\boldsymbol{S}$ equal 0. Notice that we have three parameters that we can freely choose to make sure that the first two moments match; to reduce the number of degrees of freedom by 1, we impose the additional condition of *balanced means*, i.e., $\mu_1 = 2p\mu$ and $\mu_2 = 2(1-p)\mu$ for some $\mu > 0$. The corre-

sponding scv then equals

$$\frac{1}{2p(1-p)} - 1,$$

which is larger than or equal to 1 (where we remark that it is obviously equal to 1 only if $p = 1/2$, corresponding to the exponential distribution).

4.3.2 Computing expected idle and waiting times

Section 4.3.3 presents an algorithm to compute the clients' sojourn-time distributions in both queues (where it is recalled that the sojourn time is the sum of the waiting time and the service time). Above we pointed out for the single-server queue that, with S_i denoting the sojourn time of the i-th client, our objective function can be expressed in terms of the loss function $\ell(\cdot)$, evaluated in $S_i - x_i$. This suggests that we need to know the full distribution of the sojourn times to be able to evaluate the objective function. Perhaps counter-intuitively, this is *not* the case, as we explain in this section: as it turns out, one only needs to know the *mean* sojourn times.

We show that the expected idle and waiting times at both nodes can be expressed in terms of the expected sojourn times. Let us first consider the first node. Realize that for all $i \in \{1, \ldots, n\}$, in self-evident notation,

$$\mathbb{E}[S_{1,i}] = \mathbb{E}[W_{1,i}] + \mathbb{E}[B_{1,i}]. \tag{4.6}$$

$\mathbb{E}[B_{1,i}]$ being known, we have found $\mathbb{E}[W_{1,i}]$ (in terms of the expected sojourn times, that is).

Now notice that the time the i-th client leaves the system can be expressed in two ways. In the first place it is the sum of the idle and service times of the first i clients, but in the second place also the arrival epoch of the i-th client plus his sojourn time. As a consequence, we have, for any client i,

$$\sum_{j=1}^{i} \left(\mathbb{E}[B_{1,j}] + \mathbb{E}[I_{1,j}]\right) = t_i + \mathbb{E}[S_{1,i}] = \sum_{j=1}^{i-1} x_j + \mathbb{E}[S_{1,i}]. \tag{4.7}$$

Hence we can recursively compute the expected idle time at the first server, prior to the arrival of the i-th client through

$$\mathbb{E}[I_{1,i}] = \mathbb{E}[S_{1,i}] - \mathbb{E}[B_{1,i}] + \sum_{j=1}^{i-1} \left(x_j - \mathbb{E}[B_{1,j}] - \mathbb{E}[I_{1,j}]\right).$$

A similar procedure works for the second node. Instead of looking at the first node only, we now consider S_i, i.e., the client-specific sojourn time when traversing both nodes:

$$\mathbb{E}[S_i] = \mathbb{E}[W_{1,i}] + \mathbb{E}[B_{1,i}] + \mathbb{E}[W_{2,i}] + \mathbb{E}[B_{2,i}]. \tag{4.8}$$

Noting that $\mathbb{E}[W_{1,i}]$ follows from Eqn. (4.6), we are left to compute $\mathbb{E}[W_{2,i}]$. We now have, similar to (4.7),

$$\sum_{j=1}^{i} (\mathbb{E}[B_{2,j}] + \mathbb{E}[I_{2,j}]) = \sum_{j=1}^{i-1} x_j + \mathbb{E}[S_i] \tag{4.9}$$

(notice that $\mathbb{E}[I_{2,1}] > 0$ whereas $\mathbb{E}[I_{1,1}] = 0$). From the above we conclude that by knowing the clients' expected sojourn time at the first server and in the total system, we are able to compute all expected idle and waiting times by the above formulas. In the next subsection we show how we can recursively generate the sojourn-time distributions.

4.3.3 Recursive procedure to compute the sojourn-time distribution

In this subsection we describe an algorithm that determines the sojourn-time distributions, assuming the service times at both nodes have phase-type distributions. For the first node, the procedure relies on the principles developed in Wang (1997); the derivation of the sojourn-time distribution for the *entire* system (i.e., for each client i the time spent at the first node plus the time spent at the second node), however, is novel and more involved. More specifically, there are various ways to represent the jobs flowing through the tandem network, each having its own probabilistic description (and associated state space); the one we have chosen to work with in this chapter keeps the dimensionality relatively low. It is noted that, when setting up such a description, there are various additional subtleties to be dealt with; see the way we introduce the 'idle states' $\dagger$ (single node), and $\dagger_1$ and $\dagger_2$ (tandem case) below.

In the sequel we assume that, for each server, the service times are independent and identically distributed, and that there is independence between these two sequences of random variables. Let the service time at the first node follow a phase-type distribution in $\mathbb{P}\mathrm{h}(\boldsymbol{\alpha}^{(1)}, \boldsymbol{S}^{(1)})$, whereas for the service time at the second node we have the representation $\mathbb{P}\mathrm{h}(\boldsymbol{\alpha}^{(2)}, \boldsymbol{S}^{(2)})$; the dimensions of both phase-type distributions are m_1 and m_2, respectively.

It is noted, however, that the procedure we developed extends to independent, *non*-identically distributed service times, albeit at the expense of rather 'heavy' notation. This explains why we restrict ourselves to the case of (per node) identically distributed service times in this section; presenting the procedure directly in full generality obscures the reasoning behind it (but we point out how to deal with the 'heterogeneous case' in the next section).

Recursive procedure for the first node

To compute the sojourn-time distribution at the first server, we aim to derive the phase-type representation of the sojourn-time distribution of each client i at this node, that is, $S_{1,i} =_{\mathrm{d}} \mathbb{P}\mathrm{h}(\boldsymbol{\alpha}_i^{(1)}, \boldsymbol{S}_i^{(1)})$, where the subscript '1' is added to denote that

for the moment we are only considering the first server. We first define the following bivariate process:

$$\{N_i(t), K_i(t), t \geq 0\} \tag{4.10}$$

for client $i = 1, \ldots, n$. Here $N_i(t)$ is the number of the first i clients that are present in the system, t time units after the arrival of the i-th client; obviously $N_i(t) \in \{1, \ldots, i\}$. The second component, $K_i(t) \in \{1, \ldots, m_1\}$, represents the phase of the client in service t time units after the arrival of the i-th client. Observe that one state needs to be added to the state space $\{1, \ldots, i\} \times \{1, \ldots, m_1\}$, corresponding to the situation that at time t all i clients have left. We associate the symbol '†' with this state.

In the sequel the probabilities

$$p_{j,k}^{(i)}(t) := \mathbb{P}\left(N_i(t) = j, K_i(t) = k\right)$$

play a crucial role, with $t \geq 0$, $i = 1, \ldots, n$, $j = 1, \ldots, i$, and $k = 1, \ldots, m_1$. It is evident that

$$\mathbb{P}(S_{1,i} \leq t) = \mathbb{P}\left((N_i(t), K_i(t)) = \dagger\right) = 1 - \sum_{j=1}^{i} \sum_{k=1}^{m_1} p_{j,k}^{(i)}(t). \tag{4.11}$$

In addition, we introduce the vector $\boldsymbol{P}_i(t)$ (of dimension $m_1 i$), defined by

$$\left(p_{i,1}^{(i)}(t), \ldots, p_{i,m_1}^{(i)}(t), p_{i-1,1}^{(i)}(t), \ldots, p_{i-1,m_1}^{(i)}(t), \ldots, p_{1,1}^{(i)}(t), \ldots, p_{1,m_1}^{(i)}(t)\right).$$

The sojourn-time distribution of the i-th client can be computed from $\boldsymbol{P}_i(t)$, as, by virtue of Eqn. (4.11),

$$F_{1,i}(t) := \mathbb{P}(S_{1,i} \leq t) = 1 - \boldsymbol{P}_i(t)\mathbf{1}_{m_1 i};$$

here $\mathbf{1}_{m_1 i}$ represents an all-one column vector of dimension $m_1 i$. The question we now focus on, is how $\boldsymbol{P}_i(t)$ can be computed, for $t \geq 0$, and $i \in \{1, \ldots, n\}$. In the sequel, $\mathbf{0}_{m,n}$ denotes an $(m \times n)$ all-zero matrix.

▷ Considering the first client, to arrive at $t_1 = 0$, it is a standard result that $\boldsymbol{P}_1(t) = \boldsymbol{\alpha}^{(1)} \exp(\boldsymbol{S}^{(1)} t)$; conclude that, as a consequence,

$$(\boldsymbol{\alpha}_1^{(1)}, \boldsymbol{S}_1^{(1)}) = (\boldsymbol{\alpha}^{(1)}, \boldsymbol{S}^{(1)}),$$

thus defining the phase-type representation of $S_{1,1}$.

▷ Concerning the second client, arriving x_1 after the first client, realize that there are two scenarios: he can find still some work in the system upon his arrival, and he can find the system empty. It can be argued that it thus follows that the initial distribution

of the phase-type distribution, associated with the sojourn time of client 2, reads

$$\boldsymbol{\alpha}_2^{(1)} = (\boldsymbol{P}_1(x_1), \boldsymbol{\alpha}^{(1)} F_{1,1}(x_1)),$$

a (row) vector of dimension $2m_1$. It then follows (with the same arguments as the ones used in Kuiper et al. (2015), Wang (1997)) that

$$\boldsymbol{P}_2(t) = (\boldsymbol{P}_1(x_1), \boldsymbol{\alpha}^{(1)} F_{1,1}(x_1)) \exp(\boldsymbol{S}_2^{(1)} t) \tag{4.12}$$

(being an object of dimension $2m_1$ as well); here, with $\boldsymbol{s}^{(1)} := -\boldsymbol{S}^{(1)} \mathbf{1}_{m_1}$, and

$$\boldsymbol{S}_2^{(1)} := \begin{pmatrix} \boldsymbol{S}^{(1)} & \boldsymbol{s}^{(1)} \boldsymbol{\alpha}^{(1)} \\ \mathbf{0}_{m_1 \times m_1} & \boldsymbol{S}^{(1)} \end{pmatrix}.$$

We have thus identified $(\boldsymbol{\alpha}_2^{(1)}, \boldsymbol{S}_2^{(1)})$, i.e., the phase-type representation of $S_{1,2}$.
$\triangleright$ The sojourn-time distributions of the other clients can be found recursively in a similar manner. To this end, we introduce the matrix $\boldsymbol{T}_i$ of dimension $(i-1)m_1 \times m_1$ through

$$\boldsymbol{T}_i := \left(\mathbf{0}_{m_1 \times m_1}, \ldots, \mathbf{0}_{m_1 \times m_1}, \boldsymbol{s}^{(1)} \boldsymbol{\alpha}^{(1)} \right)^{\mathrm{T}};$$

in addition, we introduce

$$\boldsymbol{S}_i^{(1)} := \begin{pmatrix} \boldsymbol{S}_{i-1}^{(1)} & \boldsymbol{T}_i \\ \mathbf{0}_{m_1 \times (i-1)m_1} & \boldsymbol{S}^{(1)} \end{pmatrix}.$$

Then the vector $\boldsymbol{P}_i(t)$ (of dimension $m_1 i$) can be found from $\boldsymbol{P}_{i-1}(t)$ (of dimension $m_1(i-1)$) by the recursion

$$\boldsymbol{P}_i(t) = \left(\boldsymbol{P}_{i-1}(x_{i-1}), \boldsymbol{\alpha}^{(1)} F_{1,i-1}(x_{i-1}) \right) \exp(\boldsymbol{S}_i^{(1)} t), \quad t \geq 0. \tag{4.13}$$

This provides us with $(\boldsymbol{\alpha}_i^{(1)}, \boldsymbol{S}_i^{(1)})$.

Realize that for the specific phase-type distributions we are working with, the matrix $\boldsymbol{S}^{(1)}$ is upper triangular (in the hyperexponential case in fact even diagonal), and hence so are the matrices $\boldsymbol{S}_i^{(1)}$, for $i \in \{1, \ldots, n\}$. As a consequence, the eigenvalues can be read off from the diagonal. This property facilitates easy computation of the matrix exponent $\exp(\boldsymbol{S}_i^{(1)} t)$; in case of the $E_{K-1,K}(\mu; p)$ distribution all eigenvalues are μ; and, in case of the $H_2(\boldsymbol{\mu}; p)$ all eigenvalues are entries of the vector $\boldsymbol{\mu}$.

Recursive procedure for the two-node tandem

Where we above determined the sojourn-time distribution at the first queue, this subsection describes the extension of an algorithm that facilitates the computation of the distribution of the *total sojourn time*. More specifically, for each client we de-

termine the phase-type distribution of the time he spends in the system, denoted by $S_i =_{\mathrm{d}} \mathbb{P}\mathrm{h}(\boldsymbol{\alpha}_i, \boldsymbol{S}_i)$. Such a sojourn time S_i covers the waiting times and service times at both nodes, and can be used to evaluate our objective function, by using the approach presented in Section 4.3.2.

To this end, we define the *tandem counterpart* of (4.10): for client $i = 1, \ldots, n$, we record the process,

$$\{L_{1,i}(t), L_{2,i}(t), t \geq 0\},$$

with, for $r = 1, 2$, $L_{r,i}(t) := (N_{r,i}(t), K_{r,i}(t))$. Here $N_{r,i}(t)$ is the number of first i clients present at the r-th server (i.e., clients who are waiting plus potentially a client who is in service), and $K_{r,i}(t)$ represents the phase of the client in service on the r-th server (for $r = 1, 2$), t time units after the arrival (at the first node) of the i-th client. Again we have to augment the state space; we do so by adding states '$\dagger_1$' ('$\dagger_2$', respectively), representing the situation that no clients are present at node 1 (node 2).

We will study the probabilities, for $j_1, j_2 \in \mathscr{J}_i$, where

$$\mathscr{J}_i := \{j_1 \in \{1, \ldots, i-1\}, j_2 \in \{1, \ldots, i-1\} : j_1 + j_2 \in \{1, \ldots, i\}\},$$

and $k_r \in \{1, \ldots, m_r\}$,

$$p^{(i)}_{j_1 k_1, j_2 k_2}(t) := \mathbb{P}\left(L_{1,i}(t) = (j_1, k_1), L_{2,i}(t) = (j_2, k_2)\right),$$

as well as, for $j_r \in \{1, \ldots, i\}$, $k_r \in \{1, \ldots, m_r\}$,

$$\begin{aligned} p^{(i)}_{\dagger_1, j_2 k_2}(t) &:= \mathbb{P}\left(L_{1,i}(t) = \dagger_1, L_{2,i}(t) = (j_2, k_2)\right), \\ p^{(i)}_{j_1 k_1, \dagger_2}(t) &:= \mathbb{P}\left(L_{1,i}(t) = (j_1, k_1), L_{2,i}(t) = \dagger_2\right). \end{aligned}$$

If the number of clients in both queues is positive (say j_1 and j_2), the client in service at the r-th node can be in m_r states. This explains why we, in this situation, work with the vector (of dimension $m_1 m_2$)

$$\boldsymbol{p}^{(i)}_{[j_1, j_2]}(t) = \left((p^{(i)}_{j_1 1, j_2 1}(t), \ldots, p^{(i)}_{j_1 1, j_2 m_2}(t)), \ldots, (p^{(i)}_{j_1 m_1, j_2 1}(t), \ldots, p^{(i)}_{j_1 m_1, j_2 m_2}(t))\right).$$

In addition, we have the vector of dimension m_1 covering the cases that the second queue is empty, and the first is not:

$$\boldsymbol{p}^{(i)}_{[j_1, \dagger_2]}(t) = \left(p^{(i)}_{j_1 1, \dagger_2}(t), \ldots, p^{(i)}_{j_1 m_1, \dagger_2}(t)\right),$$

and a vector of dimension m_2 for the cases that the first queue is empty, and the second is not:

$$\boldsymbol{p}^{(i)}_{[\dagger_1, j_2]}(t) = \left(p^{(i)}_{\dagger_1, j_2 1}(t), \ldots, p^{(i)}_{\dagger_1, j_2 m_2}(t)\right).$$

Let $\bar{\boldsymbol{p}}^{(i)}_{[j]}(t)$ correspond to all situations in which j clients are present, t time units after

the arrival of the i-th client; by concatenating the vectors defined above, we obtain the following vector of dimension $m_1 + m_2 + (j-1)m_1m_2$:

$$\bar{\boldsymbol{p}}^{(i)}_{[j]}(t) := \left(\boldsymbol{p}^{(i)}_{[j,\dagger_2]}(t), \boldsymbol{p}^{(i)}_{[j-1,1]}(t), \ldots, \boldsymbol{p}^{(i)}_{[1,j-1]}(t), \boldsymbol{p}^{(i)}_{[\dagger_1,j]}(t)\right).$$

Finally, we define $\boldsymbol{P}^{(i)}(t)$ corresponding to all possible system states t time units after arrival of the i-th client:

$$\boldsymbol{P}^{(i)}(t) = \left(\bar{\boldsymbol{p}}^{(i)}_{[i]}(t), \ldots, \bar{\boldsymbol{p}}^{(i)}_{[1]}(t)\right);$$

the dimension of this vector is

$$m[i] := \sum_{j=1}^{i} \left((m_1+m_2) + (j-1)m_1m_2\right) = i(m_1+m_2) + \frac{1}{2}i(i-1)m_1m_2.$$

In order to compute the sojourn-time distribution, the option of both queues being empty does not need to be incorporated in the vector $\boldsymbol{P}^{(i)}(t)$, since we have

$$\begin{aligned} F_i(t) := \mathbb{P}(S_i \leq t) &= 1 - \sum_{j_1,j_2 \in \mathscr{J}_i} \sum_{k_1=1}^{m_1} \sum_{k_2=1}^{m_2} p^{(i)}_{j_1k_1,j_2k_2}(t) \\ &\quad - \sum_{j_1=1}^{i} \sum_{k_1=1}^{m_1} p^{(i)}_{j_1k_1,\dagger_2}(t) - \sum_{j_2=1}^{i} \sum_{k_2=1}^{m_2} p^{(i)}_{\dagger_1,j_2k_2}(t) \\ &= 1 - \boldsymbol{P}^{(i)}(t)\mathbf{1}_{m[i]}. \end{aligned}$$

The goal is now to construct a (recursive) algorithm to identify $\boldsymbol{P}^{(i)}(t)$.

▷ For the first client, to arrive at $t_1 = 0$, we have

$$\boldsymbol{P}^{(1)}(t) = (\boldsymbol{\alpha}^{(1)}, \mathbf{0}_{m_2}) \exp\left(\begin{pmatrix} \boldsymbol{S}^{(1)} & \boldsymbol{s}^{(1)}\boldsymbol{\alpha}^{(2)} \\ \mathbf{0}_{m_2\times m_1} & \boldsymbol{S}^{(2)} \end{pmatrix} t\right).$$

which is an (m_1+m_2)-dimensional object. As a consequence, we have for the phase-type description of the random variable S_1 that

$$\boldsymbol{\alpha}_1 = (\boldsymbol{\alpha}^{(1)}, \mathbf{0}_{m_2}), \quad \boldsymbol{S}_1 = \begin{pmatrix} \boldsymbol{S}^{(1)} & \boldsymbol{s}^{(1)}\boldsymbol{\alpha}^{(2)} \\ \mathbf{0}_{m_2\times m_1} & \boldsymbol{S}^{(2)} \end{pmatrix}.$$

▷ Concerning the second client, arriving x_1 time units after the first client, standard arguments yield that, using standard Kronecker notation,

$$\boldsymbol{\alpha}_2 = \left(\boldsymbol{p}^{(1)}_{[1,\dagger_2]}(x_1), \boldsymbol{\alpha}^{(1)} \otimes \boldsymbol{p}^{(1)}_{[\dagger_1,1]}(x_1), \mathbf{0}_{m_2}, \boldsymbol{\alpha}^{(1)}F_1(x_1), \mathbf{0}_{m_2}\right),$$

where the dimensions of these five vectors are m_1, m_1m_2, m_2, m_1 and m_2, so that the whole vector has dimension $m[2] = 2(m_1 + m_2) + m_1m_2$, as desired. Now we wish to identify the matrix $\boldsymbol{S}_2$ (of dimension $m[2] \times m[2]$) corresponding to the phase-type representation of the distribution of S_2:

$$\boldsymbol{P}^{(2)}(t) = \boldsymbol{\alpha}_2 \exp\left(\boldsymbol{S}_2 t\right), \quad t \geq 0.$$

To this end, we first define the following two matrices, for ease sometimes leaving out the dimensions of the $\mathbf{0}$-matrices,

$$\begin{aligned}
\boldsymbol{U}_2 &:= \begin{pmatrix} \boldsymbol{S}^{(1)} & -\boldsymbol{S}^{(1)}\mathbf{1}_{m_1}\boldsymbol{A}^{(0)} & \mathbf{0}_{m_1\times m_2} \\ \mathbf{0}_{m_1m_2\times m_1} & \boldsymbol{S}^{(1)} \oplus \boldsymbol{S}^{(2)} & -\boldsymbol{S}^{(1)}\mathbf{1}_{m_1} \otimes I_{m_2} \\ \mathbf{0}_{m_2\times m_1} & \mathbf{0}_{m_2\times m_1m_2} & \boldsymbol{S}^{(2)} \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{S}^{(1)} & \boldsymbol{s}^{(1)}\boldsymbol{A} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}^{(1)} \oplus \boldsymbol{S}^{(2)} & \boldsymbol{s}^{(1)} \otimes I_{m_2} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{S}^{(2)} \end{pmatrix},
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{V}_2 &:= \begin{pmatrix} \mathbf{0}_{m_1\times m_1} & \mathbf{0}_{m_1\times m_2} \\ -\boldsymbol{I}_{m_1} \otimes \boldsymbol{S}^{(2)}\mathbf{1}_{m_2} & \mathbf{0}_{m_1m_2\times m_2} \\ \mathbf{0}_{m_2\times m_1} & -\boldsymbol{S}^{(2)}\mathbf{1}_{m_2}\boldsymbol{\alpha}^{(2)} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \boldsymbol{I}_{m_1} \otimes \boldsymbol{s}^{(2)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{s}^{(2)}\boldsymbol{\alpha}^{(2)} \end{pmatrix},
\end{aligned}$$

where $\boldsymbol{A}^{(0)} := \boldsymbol{\alpha}^{(1)} \otimes \boldsymbol{\alpha}^{(2)}$, and $\boldsymbol{s}^{(r)} := -\boldsymbol{S}^{(r)}\mathbf{1}_{m_r}$. It is concluded that $\boldsymbol{U}_2$ is a square matrix with $m_1 + m_2 + m_1m_2$ rows and columns, whereas $\boldsymbol{U}_2$ is of dimension $(m_1 + m_2 + m_1m_2) \times (m_1 + m_2)$. We can now construct the $(m[2] \times m[2])$-dimensional matrix $\boldsymbol{S}_2$ by

$$\boldsymbol{S}_2 = \begin{pmatrix} \boldsymbol{U}_2 & \boldsymbol{V}_2 \\ \mathbf{0} & \boldsymbol{S}_1 \end{pmatrix}.$$

▷ For the other clients, the same iterative procedure can be followed. We first define the following two 'start matrices', relating to which server starts serving a new client:

$$\boldsymbol{A}^{(1)} := \boldsymbol{\alpha}^{(1)} \otimes \boldsymbol{I}_{m_2} \quad \text{and} \quad \boldsymbol{A}^{(2)} := \boldsymbol{I}_{m_1} \otimes \boldsymbol{\alpha}^{(2)}.$$

In addition, we introduce the following vector of dimension $m_1 + m_2 + jm_1m_2$, for $j \in \{1, \ldots, i\}$,

$$\check{\boldsymbol{p}}^{(i)}_{[j]}(t) := \left(\boldsymbol{p}^{(i)}_{[j,\dagger_2]}(t), \boldsymbol{p}^{(i)}_{[j-1,1]}(t), \ldots, \boldsymbol{p}^{(i)}_{[1,j-1]}(t), \boldsymbol{\alpha}^{(1)} \otimes \boldsymbol{p}^{(i)}_{[\dagger_1,j]}(t), \mathbf{0}_{m_2}\right).$$

Regarding the start distribution corresponding to the phase-type description of the sojourn time S_i, it follows that

$$\boldsymbol{\alpha}_i = \left(\check{\boldsymbol{p}}^{(i-1)}_{[i-1]}(x_{i-1}), \ldots, \check{\boldsymbol{p}}^{(i-1)}_{[1]}(x_{i-1}), \boldsymbol{\alpha}^{(1)} F_{i-1}(x_{i-1}), \mathbf{0}_{m_2}\right),$$

which can be verified to be of dimension $m[i]$.

Regarding the matrix $\boldsymbol{S}_i$ in $\boldsymbol{P}^{(i)}(t) = \boldsymbol{\alpha}_i \exp\left(\boldsymbol{S}_i t\right)$, this has the form

$$\boldsymbol{S}_i = \begin{pmatrix} \boldsymbol{U}_i & \boldsymbol{V}_i \\ \mathbf{0} & \boldsymbol{S}_{i-1} \end{pmatrix}.$$

Here

$$\boldsymbol{U}_i := \begin{pmatrix} \boldsymbol{S}^{(1)} & \boldsymbol{s}^{(1)}\boldsymbol{A}^{(0)} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}^{(1)} \oplus \boldsymbol{S}^{(2)} & \boldsymbol{s}^{(1)} \otimes \boldsymbol{A}^{(1)} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & & \ddots & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{S}^{(1)} \oplus \boldsymbol{S}^{(2)} & \boldsymbol{s}^{(1)} \otimes \boldsymbol{A}^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{S}^{(1)} \oplus \boldsymbol{S}^{(2)} & \boldsymbol{s}^{(1)} \otimes I_{m_2} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{S}^{(2)} \end{pmatrix},$$

of dimension $((i-1)m_1m_2 + m_1 + m_2) \times ((i-1)m_1m_2 + m_1 + m_2)$, and

$$\boldsymbol{V}_i = \left(\begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{I}_{m_1} \otimes \boldsymbol{s}^{(2)} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{s}^{(2)} \otimes \boldsymbol{A}^{(2)} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{s}^{(2)} \otimes \boldsymbol{A}^{(2)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{s}^{(2)}\boldsymbol{\alpha}^{(2)} \end{pmatrix}, \mathbf{0}\right)$$

of dimension $((i-1)m_1m_2 + m_1 + m_2) \times m[i-1]$. We conclude that $\boldsymbol{S}_i$ indeed has dimension $m[i] \times m[i]$.

It can be verified that the analysis simplifies greatly in the case in which both service times have exponentially distributions, $B_{r,i} =_{\mathrm{d}} \mathbb{P}\mathrm{h}(1, \mu_r)$ for $r = 1, 2$. In that situation, one needs to record only the number of clients present at both nodes (as a consequence of the fact that a service time corresponds to just a single exponential phase).

4.4 Extensions

In the previous section we have considered our 'base model'; in this section we point out how a number of variants can be dealt with. In the first subsection we consider the situation of heterogeneous service times, whereas the second subsection concentrates

on models in which the second node may *block* the first node.

4.4.1 Heterogeneous service-time distributions

The model analyzed in the previous section considers the situation in which at each station r (for $r = 1, 2$) the n service times, say $B_{r,1}$ up to $B_{r,n}$, are i.i.d. samples, distributed as a random variable B_r; importantly, the distributions of B_1 and B_2 do not necessarily coincide. We already indicated in Section 4.3.3 that our procedure extends to the situation in which the service-time distributions (at each of the nodes) are *client-specific*: i.e., each of the $B_{r,i}$ has an own distribution. As this extension is notationally involved, we restrict ourselves to explaining the main ideas behind it. We let job $B_{r,i}$ corresponds to a phase-type representation $(\boldsymbol{\alpha}^{(r,i)}, \boldsymbol{S}^{(r,i)})$, for $r = 1, 2$ and $i = 1, \ldots, n$, with 'dimension' $m_{r,i} \in \mathbb{N}$.

We start our exposition by pointing out how the recursive procedure for the first node needs to be adapted. As it turns out, essentially all steps carry over after minor modifications. The vector $\boldsymbol{P}_i(t)$, defined as in Section 4.3.3, has now dimension $\bar{m}_i := m_{1,1} + \ldots + m_{1,i}$. Regarding the first client, we obviously have

$$\boldsymbol{P}_1(t) = \boldsymbol{\alpha}^{(1,1)} \exp(\boldsymbol{S}^{(1,1)} t).$$

For the second client, (4.12) still applies, but with $\boldsymbol{\alpha}^{(1)}$ replaced by $\boldsymbol{\alpha}^{(1,2)}$ and $\boldsymbol{S}_2^{(1)}$ now being the $\bar{m}_2 \times \bar{m}_2$ matrix given by

$$\boldsymbol{S}_2^{(1)} := \begin{pmatrix} \boldsymbol{S}^{(1,1)} & \boldsymbol{s}^{(1,1)} \boldsymbol{\alpha}^{(1,1)} \\ \boldsymbol{0}_{m_{1,2} \times m_{1,1}} & \boldsymbol{S}^{(1,2)} \end{pmatrix},$$

where $\boldsymbol{s}^{(1,i)} := -\boldsymbol{S}^{(1,i)} \boldsymbol{1}_{m_{1,i}}$. This idea carries over to any client i, in the sense that the recursive relation (4.13) remains valid, but with $\boldsymbol{\alpha}^{(1)}$ replaced by $\boldsymbol{\alpha}^{(1,i)}$ and $\boldsymbol{S}_i^{(1)}$ now a matrix of size $\bar{m}_i \times \bar{m}_i$ defined as

$$\boldsymbol{S}_i^{(1)} := \begin{pmatrix} \boldsymbol{S}_{i-1}^{(1)} & \boldsymbol{T}_i \\ \boldsymbol{0}_{m_{1,i} \times \bar{m}_{i-1}} & \boldsymbol{S}^{(1,i)} \end{pmatrix},$$

where $\boldsymbol{T}_i$ is a matrix of size $\bar{m}_{i-1} \times m_{1,i}$:

$$\boldsymbol{T}_i := \left(\boldsymbol{0}_{m_{1,1} \times m_{1,i}}, \ldots, \boldsymbol{0}_{m_{1,i-2} \times m_{1,i}}, \boldsymbol{s}^{(1,i-1)} \boldsymbol{\alpha}^{(1,i)} \right)^{\mathrm{T}}.$$

In the same way we can extend the procedure for the two-node tandem (as developed in Section 4.3.3) to the situation of heterogeneous service times, but this becomes notationally rather involved. We therefore do not provide all details here, but restrict ourselves to a couple of general remarks.

In the first place, observe that a full description of our system now consists, at

the moment that i clients have entered the system, of the number j_1 present at the first node and the number j_2 at the second node (where obviously $j_1 + j_2 = j \in \{0, 1, \ldots, i\}$), together with the phases of the clients in service. It is seen that client $\ell := i - j + 1$ is in service at node 2 (if $j_2 > 0$), since $i - j$ clients already left the system. This also means that clients $i - j + 1$ up to $i - j_1$ are present at the second node. It thus follows that the client in service there has a service-time distribution $B_{2,\ell}$ (represented by a phase-type distribution of dimension $m_{2,\ell}$). Likewise, clients $i - j_1 + 1$ up to i are present at node 1, with client $k := i - j_1 + 1$ in service (as long as $j_1 > 0$), with service-time distribution $B_{1,k}$ (represented by a phase-type distribution of dimension $m_{1,k}$).

The above extension allows us to study the effect of all sorts of correlations. If client i tends to take relatively long at both nodes (relative to the other clients), one could put this information into the random variables $B_{1,i}$ and $B_{2,i}$ (for instance by giving them larger means than the other clients).

4.4.2 Models with blocking

The general setup we have considered in Section 4.3.3 is a model in which there is an *infinite* buffer (i.e., waiting room) after stage 1, and thus clients waiting for service at the second node do not prevent the first node from processing work. Models in which there is such a blocking effect (Dallery and Gershwin 1992), however, are relevant in specific cases. They turn out to be relatively easy to model, and simpler than the base model. In this subsection we show how to adapt the base model to incorporate two common types of blocking; these adaptations to the model still follow the recursive methods outlined in Section 4.3.3. For ease we consider the case that for a given r the $B_{r,i}$ are distributed as a random variable B_r for all $i \in \{1, \ldots, n\}$, but the situation of heterogeneity among the $B_{r,i}$ can be dealt with as described in Section 4.4.1.

- ▷ In a first type of blocking, so called *blocking-before-service*, the first server can only start a new job when the second server, is empty. Such a system can obviously be modeled by a single-node system, in which the phase-type representation of the per client service-time distribution B_i is derived by taking the convolution of the individual service times at both nodes (for client i represented by $B_{1,i} =_{\mathrm{d}} \mathbb{P}\mathrm{h}(\boldsymbol{\alpha}^{(1)}, \boldsymbol{S}^{(1)})$ and $B_{2,i} =_{\mathrm{d}} \mathbb{P}\mathrm{h}(\boldsymbol{\alpha}^{(2)}, \boldsymbol{S}^{(2)})$), that is,

$$B_i =_{\mathrm{d}} \mathbb{P}\mathrm{h}\left((\boldsymbol{\alpha}^{(1)}, \mathbf{0}_{m_2}), \begin{pmatrix} \boldsymbol{S}^{(1)} & \boldsymbol{s}^{(1)}\boldsymbol{\alpha}^{(2)} \\ \mathbf{0} & \boldsymbol{S}^{(2)} \end{pmatrix} \right)$$

  In addition, the server-specific costs cannot be differentiated between servers as the servers are considered as a single system. Therefore, it is natural to compute the objective function as in the single-node case; see Eqn. (4.1).

- ▷ Another type of blocking is called *blocking-after-service*, a client can only move

to the second node when this node is idle. It means that the client stays at the first node when he has been served at this first node, but cannot move on to the second node (as a consequence of the fact that there is a client being served there). With 'blocking' we refer to the situation that the next client-in-line cannot commence service at node 1, although the service time of the client in service has finished. As such, when 1 or 2 clients have entered the system, no blocking can occur; only for $i \geq 3$ one can be confronted with blocking. (For $i = 2$ the second server can still be serving the first client while the second client already finished his service at the first node, but this case does not require an adaptation of the algorithm presented in Section 4.3.3, since the second client is *de facto* waiting at the second node.)

let $\bar{\boldsymbol{p}}^{(i,b)}_{[j]}(t)$ correspond to all situations in which j clients are present in our model with blocking, t time units after the arrival of the i-th client; we obtain the following vector of dimension $m_1 + m_2 + (\min\{j, 2\} - 1)m_1 m_2$:

$$\bar{\boldsymbol{p}}^{(i,b)}_{[j]}(t) := \begin{cases} \left(\boldsymbol{p}^{(i)}_{[j,\dagger_2]}(t), \boldsymbol{p}^{(i)}_{[j-1,1]}(t), \boldsymbol{p}^{(i)}_{[j-2,2]}(t)\right) & \text{if } j \geq 3, \\ \left(\boldsymbol{p}^{(i)}_{[j,\dagger_2]}(t), \boldsymbol{p}^{(i)}_{[j-1,1]}(t), \boldsymbol{p}^{(i)}_{[\dagger_1,j]}(t)\right) & \text{if } j \leq 2. \end{cases}$$

Finally, let $\boldsymbol{P}^{(i,b)}(t)$ correspond to the probability vector related all possible system states t time units after arrival of the i-th client:

$$\boldsymbol{P}^{(i,b)}(t) = \left(\bar{\boldsymbol{p}}^{(i,b)}_{[i]}(t), \ldots, \bar{\boldsymbol{p}}^{(i,b)}_{[1]}(t)\right);$$

the dimension of this vector is less than $m[i]$. The transitions given by the matrix $\boldsymbol{S}^b_i$ can be found by

$$\boldsymbol{S}^b_i = \begin{pmatrix} \boldsymbol{U}^b_i & \boldsymbol{V}^b_i \\ \boldsymbol{0} & \boldsymbol{S}^b_{i-1} \end{pmatrix},$$

where the matrices $\boldsymbol{U}^b_i$ and $\boldsymbol{V}^b_i$ can be constructed as in in Section 4.3.3, but have just $(\min\{i, 2\} - 1)$ diagonal elements (instead of $i-1$), due to the fact that when $j_2 = i-2$ the first node is blocked, and only the second node is busy. The matrix $\boldsymbol{S}^b_i$ can then be used in $\boldsymbol{P}^{(i,b)}(t) = \boldsymbol{\alpha}^b_i \exp(\boldsymbol{S}^b_i t)$ (where the initial probabilities $\boldsymbol{\alpha}^b_i$ are adapted accordingly).

The above setup enables us to compute the sojourn-time distributions. We can only use Eqns. (4.8) and (4.9) to compute the expected idle and waiting times, since the sojourn time at the first server is affected by the performance of the second server; due to the blocking effect the first server has to be analyzed separately. In the case of equal weights (i.e., $w = 0.5$ and $\omega_1 = \omega_2$) this issue is trivially resolved. In other situations, one could opt for explicitly keeping track of the epoch that the first server finishes its service. Importantly, where the above setup corresponds with the situation of no waiting room between the

nodes, one can easily generalize the procedure to the case of $b \in \mathbb{N}$ positions in the waiting room.

4.5 Optimal schedules in a transient environment

In this section we present a numerical assessment related to the *transient* case, i.e., we determine, for various model instances, the optimal arrival times for n clients. The methodology outlined in the previous section enables us to compute the aggregate risk of a given schedule, and this risk is then to be minimized over the arrival epochs of the n clients (where the first client arrives at $t_1 = 0$), so as to obtain the optimal schedule. This minimization can be done relying on standard numerical packages.

Indeed, the phase-type representation, as obtained by the recursive method presented in the previous section, allows us to evaluate the sojourn-time distributions of the individual clients, and hence also the associated risk. Being able to compute optimal schedules, the impact of various parameters can be assessed. More specifically, in this section we perform such sensitivity analysis with respect to (i) both servers' scv s; (ii) both servers' means; (iii) the weight parameter w. In all experiments we assume that clients are homogeneous, in that their service times at node 1 (node 2, respectively) are identically distributed.

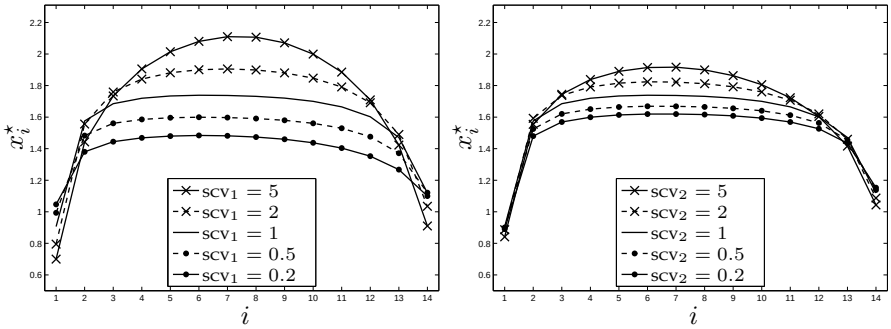
In general, a schedule consisting of n clients can be written as a vector of n arrival epochs $(t_1, \ldots, t_n)$, or, equivalently, $n-1$ interarrival times $\boldsymbol{x} = (x_1, \ldots, x_{n-1})$. In the sequel we represent the optimal schedule by the vector of interarrival times $\boldsymbol{x}^\star = (x_1^\star, \ldots, x_{n-1}^\star)$.

4.5.1 Effect of coefficient of variation

First we examine the effect of the variability of the service times. In healthcare applications the typical range for the scv is $0.35 - 0.85$, see Çayırlı and Veral (2003). In our experiments, however, we do not restrict the scv to this range; realize that the method can be used in other areas as well, such as the planning of jobs in a manufacturing environment, in which potentially other scv values apply.

In Figure 4.1 we consider optimal schedule for various values of the scv s, while keeping the mean service times fixed. In Figure 4.1(a) we plot the optimal interarrival times when varying the scv of the first server, whereas in Figure 4.1(b) the scv of the second server is varied. It is seen that the schedule has a so-called dome shape, as described in Chapter 2: the optimal interarrival times are relatively short at the beginning (as there is still little uncertainty in the system) and the end (as there are few later clients suffering from long service times) of the schedule.

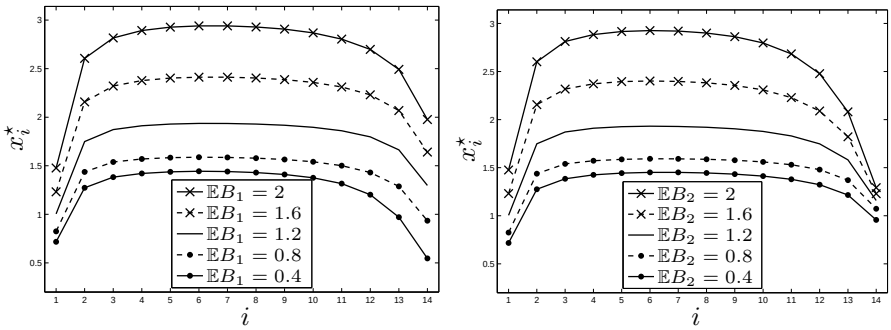
From the graphs we observe that the variability at the first server has a more pronounced impact. An explanation lies in the very nature of the tandem queue: variations in the service times at the first server are propagated to the second server. As



(a) scv_1 varies, while $\text{scv}_2 = 1$.

(b) scv_2 varies, while $\text{scv}_1 = 1$.

Figure 4.1: The other parameters are kept such that $\mathbb{E}B_1 = \mathbb{E}B_2 = 1$ and $w = 0.5$.



(a) $\mathbb{E}B_1$ varies, while $\mathbb{E}B_2 = 1$.

(b) $\mathbb{E}B_2$ varies, while $\mathbb{E}B_1 = 1$.

Figure 4.2: The other parameters are kept such that $\text{scv}_1 = \text{scv}_2 = 1$ and $w = 0.5$.

a consequence, fluctuations in the service times at the first server affect the schedule more than additional variations at the second server.

4.5.2 Effect of mean

In Figure 4.2, we systematically assess the effect of the mean service times on the schedule. Figure 4.2(a) shows how the optimal schedule is affected by the mean service time at the first node, whereas Figure 4.2(b) visualizes the effect of the mean service time at the second node. It is observed that these mean service times have less impact than the scv s. More precisely, nearly until the last client, the computed optimal schedules behave virtually identically; only at the very end of the session we see a (mild) discrepancy. In addition, it is seen that in Figure 4.2(a) the optimal interarrival times for the last client have clearly distinct values, whereas in Figure 4.2(b)

these are considerably closer together.

4.5.3 Effect of weight

We now assess the effects of the weight parameter w, by varying w from 0 to 1 (in steps of 0.2). We set the mean service times and coefficients of variation equal to 1 (at both servers). In the case $w = 1$ we are optimizing over the first server only, i.e., we are in the setting of the well-known D/M/1 queue (with non-homogeneous arrival times), see also Chapter 2. The other extreme situation, $w = 0$, is equivalent to only optimizing over the second server. The resulting schedules are presented in Figure 4.3. From this graph we observe that the interarrival times essentially decrease in w. The reason for the phenomenon is that, in order to control the sojourn times in node 2 relatively long interarrival times are needed (compared to the sojourn times in node 1); this is an immediate consequence of the fact that node 2 is facing a non-deterministic arrival process (as opposed to node 1). As a result, giving node 1 more weight (i.e., increasing w) leads to 'more predictability in the objective function', and hence shorter optimal interarrival times. Similar graphs are obtained when choosing other values for the mean service times and coefficients of variation.

4.5.4 Comparison with single-server system

Finally we study the difference between the two-node tandem with a corresponding single-node system. We consider the situation of a tandem network (with both mean service times equal to 1), and a single-server queue (with mean service time equal to 1). We set all scv values equal to a half (which is a common setting in healthcare). As observed in Figure 4.4, the second node is fed by a non-deterministic arrival process, thus explaining that the optimal interarrival times in the two-node case are higher than those for the single node.

To be able to compare the per-client loss of the tandem network with the loss in the single-node setting, we consider the average of the two expected waiting times, and for the idle times we did the same. We see for both systems that the mean waiting times are increasing functions (turning from concave to convex somewhere in the middle). For the mean idle times we observe the familiar dome-shape pattern, cf. Figure 4.4(a). Obviously the mean per-client loss in the two-node tandem is substantially higher than in the single-node system, as a result of the extra variation the second node is facing.

To further explore the effect of the tandem structure on the schedule, in relation with the corresponding single-node system, we also consider the corresponding optimal *steady-state* schedule. In Section 4.6 we point out how this schedule can be efficiently evaluated. It is stressed that transient solutions converge relatively rapidly to their steady-state counterparts, as is pictorially illustrated in Figure 4.4(a); there we additionally plotted the optimal interarrival times in steady state, leading to the hori-

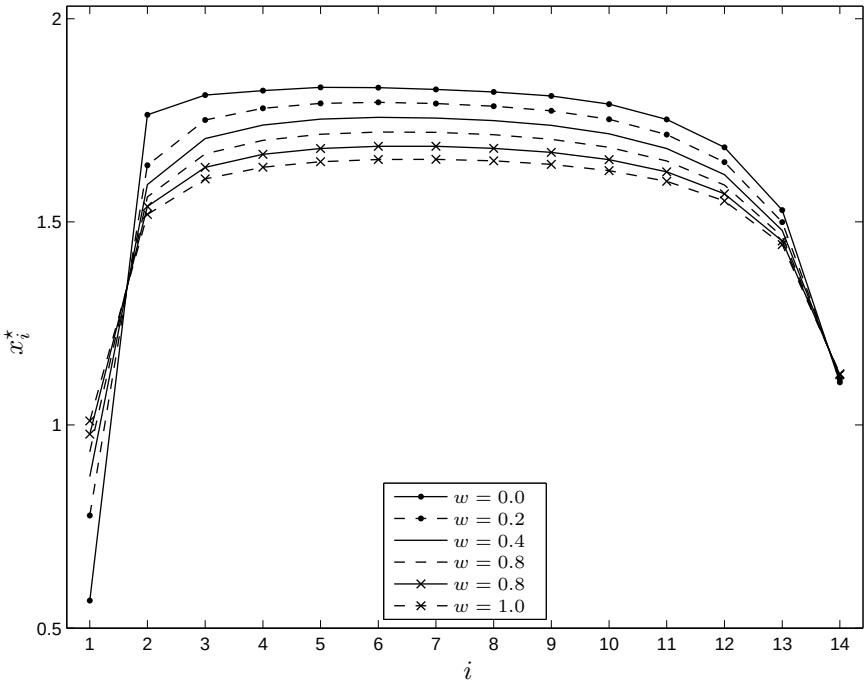
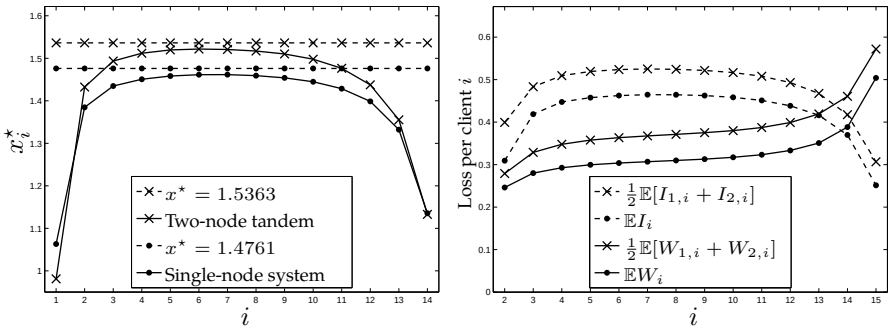


Figure 4.3: The optimal schedule is computed for different weights $w \in [0, 1]$. The other parameters are kept such that $\mathbb{E}B_1 = \mathbb{E}B_2 = 1$ and $\text{scv}_1 = \text{scv}_2 = 1$.



(a) Comparing the optimal solution $\boldsymbol{x}^\star$. (b) Comparing the losses corresponding to $\boldsymbol{x}^\star$.

Figure 4.4: The parameters are set such that every server operates with mean 1 and squared coefficient of variation of 0.5.

zontal lines at 1.5363 and 1.4761 for the two-node tandem and the single-node system respectively. Another motivation for using steady-state schedules is that they are easier to compute and conceptually simpler than transient schedules, as they consist of just a single value.

4.6 Optimal schedules in steady state

In this section we consider the situation that the number of clients grows large, assuming that at both nodes the service-time distribution is identical across the clients. As a consequence, the optimal interarrival time tends to a constant, the *steady-state optimal interarrival time*, which we explain how to evaluate. As before, we restrict ourselves to the situation of phase-type service times at both nodes. The second part of this section presents a series of experiments.

The evaluation of the steady-state optimal interarrival time relies directly on the transition matrix computed, so as to compute the invariant distribution of the embedded discrete-time Markov chain. This idea reduces the computational effort drastically, in that it is not necessary to compute specific integrals and summations in the way proposed in Section 2.5.2 in this thesis. To the best of our knowledge, this new approach to compute the stationary distribution of a D/G/1 or D/G/1 $\rightarrow$ G/1 queue has not been pointed out before.

4.6.1 Procedure

The optimal interarrival time in steady state is particularly important, because, as indicated by the experiments reported in Chapter 2, schedules for a finite number of clients converge rapidly to their steady-state counterparts. In addition, as we will show, the steady-state solution can be determined with relatively low computational effort. It is further remarked that it can be used as an upper bound for transient schedules, which is illustrated by the dashed lines in Figure 4.4(a). Our approach originates naturally from the phase-type framework featuring in Section 4.3.3. The method borrows elements from the one presented for the single node (see Chapter 2), but is significantly more efficient.

Denote with b^u the largest of the two mean service times. In steady state, the optimal interarrival time, to be denoted by $x^\star$, should evidently be larger than b^u; we denote $\rho := b^u/x^\star$. It is further noted that, in contrast with the transient setting, the number of clients can attain any positive integer. However, when $\rho < 1$ holds, the stationary probability of having more than, say, M clients in the system decays essentially geometrically in M. This fact justifies truncating the state space such that, at both queues, we do not allow more than M clients to be simultaneously present.

In our stationary setting, the optimization problem can be rephrased as

$$\min_x w\left(\omega_1 \mathbb{E}[I_1(x)] + (1-\omega_1)\mathbb{E}[W_1(x)]\right) + (1-w)\left(\omega_2 \mathbb{E}[I_2(x)] + (1-\omega_2)\mathbb{E}[W_2(x)]\right), \tag{4.14}$$

where $W_r(x)$ ($I_r(x)$, respectively) is the steady-state waiting time (idle time, respectively) at node r ($r = 1, 2$) given the interarrival times are x. The mean idle and waiting times can be derived from the steady-state sojourn-time distribution, as pointed out in Section 4.3.2.

Now consider the number of clients in both queues, as well as the phase of the client in service (if any), *just before* arrival epochs (at the first node, that is), i.e., the epochs $nx-$, for $n \in \mathbb{N}$). This process evidently constitutes a discrete-time Markov chain. Since the number of clients is truncated at M we can work with the matrix $\boldsymbol{S}_M$ that we identified in Section 4.3.3. The transition matrix of the embedded discrete-time process follows from the matrix exponent $\boldsymbol{Q}_M = \exp(x\, S_M(x))$, where a minor correction needs to be applied, in order to take care of the arrival that takes place immediately after the embedded epochs. In more detail, let $\boldsymbol{\pi}_{M-1}$ be the stationary probabilities, with the state space truncated at $M-1$. Starting with this vector $\boldsymbol{\pi}_{M-1}$ of dimension $m[M-1]+1$ (including the state of *no* clients in the system), we first perform a shift by one client (c.f. the transient setting), due to the arrival at nx, resulting in a vector of dimension $m[M]$. This vector can be multiplied by the transition matrix of the embedded discrete-time Markov chain $\boldsymbol{Q}_M$, and, because $\boldsymbol{\pi}_{M-1}$ was the stationary distribution, this should equal $\boldsymbol{\pi}_{M-1}$ again. Written in a compact way, we are therefore to solve

$$\boldsymbol{\pi}_{M-1} = t(\boldsymbol{\pi}_{M-1})\boldsymbol{Q}_M, \tag{4.15}$$

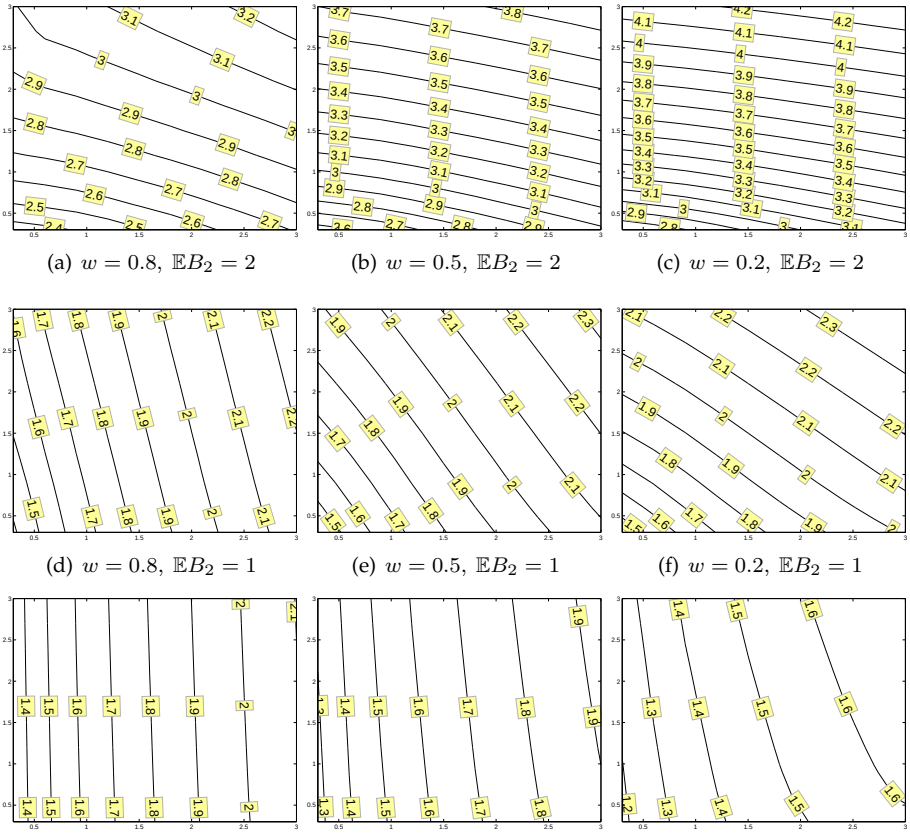
where the function $t(\cdot)$ corresponds to the shift operation applied to the vector $\boldsymbol{\pi}_{M-1}$, as described above. Using the normalizing equation $\boldsymbol{\pi}_{M-1} \cdot \mathbf{1}_{m[M-1]+1} = 1$ and Eqn. (4.15), we find the equilibrium distribution. Having found this vector, the objective function can be evaluated, by the phase-type representation for the steady-state sojourn-time distribution, given by $\mathbb{P}\mathrm{h}(t(\boldsymbol{\pi}_{M-1}), \boldsymbol{S}_M)$. Having a procedure to evaluate the objective function for given x, we can then optimize it over $x^\star > b^u$.

Along the same lines one can derive the steady-state sojourn-time distribution for the first server only, which is computationally less involved. Combining both steady-state sojourn-time distributions, we can find all expected idle and waiting times by the relations derived in Section 4.3.2.

4.6.2 Computational results

In this subsection we evaluate the effect of (i) both scv_1 and scv_2, (ii) the heterogeneity in the mean service times, i.e., $\mathbb{E}B_1$ and $\mathbb{E}B_2$, and (iii) the weight w on the steady-state optimal arrival time. To this end, we have considered 9 scenarios: all combinations of 3 different values of the weight w and three different values of $\mathbb{E}B_2$ (fixing, without loss of generality, $\mathbb{E}B_1$ at 1). For all these scenarios we vary the scv

values corresponding to the two service times. In Figure 4.5 the resulting graphs are given. The computational time per data point to compute the steady-state optimal interarrival time is less than 1 minute, which is considerably less, roughly ten times, than computing the corresponding transient schedule for $n = 25$ clients.



(a) $w = 0.8,\ \mathbb{E}B_2 = 2$ (b) $w = 0.5,\ \mathbb{E}B_2 = 2$ (c) $w = 0.2,\ \mathbb{E}B_2 = 2$

(d) $w = 0.8,\ \mathbb{E}B_2 = 1$ (e) $w = 0.5,\ \mathbb{E}B_2 = 1$ (f) $w = 0.2,\ \mathbb{E}B_2 = 1$

(g) $w = 0.8,\ \mathbb{E}B_2 = 0.5$ (h) $w = 0.5,\ \mathbb{E}B_2 = 0.5$ (i) $w = 0.2,\ \mathbb{E}B_2 = 0.5$

Figure 4.5: The optimal interarrival times $x^\star$ as function of the scv s for different scenarios. The scv_1 is varying along the horizontal axis, while scv_2 is varying along the vertical axis. The mean of the first server is set by $\mathbb{E}B_1 = 1$, while $\mathbb{E}B_2$ and w vary.

Perhaps the most striking observation from Figure 4.5 is that, when moving from the top/right graph to the bottom/left graph, the level curves per figure change from nearly flat (gradient is 'orientated in the scv_2 direction') to almost vertical (gradient is orientated in the scv_1 direction). Evidently, if w is close to 1 and service times in the first queue are substantially bigger than those in the second queue, then the impact of scv_2 is small. Likewise, for w small and service times in the first queue being small relative to those in the second queue, then the impact of scv_1 is small.

It is further observed that the level curves are nearly linear in scv_1 and scv_2. The

exceptions to this rule are Figures 4.5(a) and 4.5(i); in Figure 4.5(a) both the weight of node 1 and the service time at node 2 are relatively high, whereas in Figure 4.5(i) the weight of node 2 and the service time at node 1 are high. In addition, in some of the scenarios the distance between the contour lines is nearly constant.

An evident global conclusion is that both scv_1 and scv_2 have a significant impact on the schedule. In the case that both nodes are equally important ($w = \frac{1}{2}$) and have similar means, that is, Figure 4.5(e), we see that the first server's scv has a more profound impact than the second server. This finding is in line with the observations made in Section 4.5, where we argued that this is a consequence of the fact that the variability of the first queue propagates to the second queue.

Moreover, our procedure makes it possible to verify whether it is justified to make use of steady-state schedules rather than their transient counterparts. In specific situations already after three clients the optimal interarrival times are hardly distinguishable from those obtained when evaluating the steady-state schedule; see e.g. the examples in Section 4.5 on transient schedules. This fact can be used by managers: the steady-state schedules depicted in Figure 4.5 can then serve as some sort of 'cookbook' to determine the optimal interarrival times in the specific situation they encounter. A pragmatic view is that one could use the equidistant schedule as resulting from a steady-state analysis, which is in particular cases already close to optimal, and that one further improves it by slightly modifying the schedule at the start and end of the schedule (so as to obtain a dome-shape pattern, similar to the ones found in the section on transient schedules).

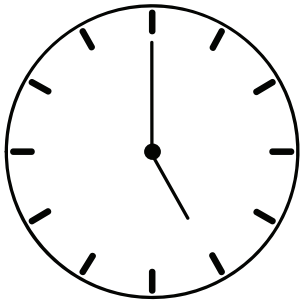
4.7 Conclusion and discussion

In this chapter we have considered the problem of finding appointment schedules that balance the clients' mean waiting times and the server's idle times. We have extended the approach that was developed in Chapter 2 for the single-node queue to its tandem counterpart. A key step in our procedure is that we approximate the service times by appropriately chosen phase-type random variables. Importantly, phase-type distributions allow for (relatively) easy calculations; in particular, the sojourn-time distributions of the individual clients can be determined recursively. Furthermore, we show how to efficiently compute the steady-state sojourn-time distribution. Having the sojourn-time distribution at our disposal, optimization techniques can be used to determine optimal schedules. We note that it was shown in Chapter 2 that replacing service-time distributions (Weibull and lognormal) by their phase-type counterparts (of low dimension) hardly affects the optimal schedule.

The experiments in Sections 4.5 and 4.6 (for schedules in a transient environment and in stationarity, respectively) give insight into the behavior of the optimal schedules under a broad variety of parameter settings (corresponding to the weights between both servers, and the mean and scv of each server).

There are several directions for further research. (i) In the first place, one could study the optimal schedules in alternative multi-node settings, such as the fork-join queue. In addition, it would be interesting to systematically assess the impact of the risk function on the the optimal schedule; in this chapter all results are based on a specific risk function (the linear one). One could also investigate the *sequential* approach (as proposed in Kemper et al. (2014) for the single node), which assigns optimal arrival times sequentially to individual users (i.e., the schedule gradually fills). (ii) In the second place, the optimal schedule can be studied in situations in which additional features play a role. One such feature are 'urgent clients', whose arrivals correspond to a Poisson process or more realistically by the process identified by Alexopoulos et al. (2008). Also, one could consider different client types, where each type has a distinctive service-time distribution. And finally, one could incorporate the occurrence of no-shows. These could be easily incorporate in the phase-type representation $(\boldsymbol{\alpha}, \boldsymbol{S})$ by adapting the initial probability vector $\boldsymbol{\alpha}$ (see Chapter 3).

In the heterogeneous scenario, where client types have distinctive distributions, one could also study the issue of optimizing the sequence in which clients are scheduled. We have observed that the 'variation' at the first server (expressed in terms of scv_1) propagates to the second server. As this variation results in higher expected idle and waiting times, it is anticipated that scv_1 has a crucial impact on the objective function. This suggests the heuristic, for situations in which the values of the scv_2 are roughly equal, to schedule clients in ascending order of their scv_1. Unfortunately, even for the single-node system there are almost no rigorous results for such properties yet, notable exceptions being Kemper et al. (2014) (focusing on the sequential approach mentioned above), Rohleder and Klassen (2000) (focusing on simulation studies where clients with low variance are scheduled first) and Mak et al. (2015) (focusing on a 'distribution-free' setup, minimizing the worst-case expected waiting and overtime over all probability distributions with given moments).



5. An Alternative Approach to Appointment Scheduling

In this chapter we propose an alternative method to study the appointment scheduling problem in continuous time. Instead of using phase-type distributions, as used in the preceding chapters, we develop an approximation method that is generic in terms of the service time distribution. The method is numerically tractable for large problem instances while offering good performance.

5.1 Introduction

The basic setting of the appointment scheduling problem, as described in the Introduction (Chapter 1), belongs to the so-called –*static*– class of appointment scheduling approaches, in which a finite number of appointments are scheduled prior to the beginning of the actual service, see Çayırlı and Veral (2003). The origin of such an approach dates back to the work of Bailey (1952) and Welch and Bailey (1952), and generated substantial interest over the last decades.

Suppose that the service provider is given n clients with random service times that are to be scheduled in a session. Further, suppose that the service-time distribution is known as well as the loss functions of the client (in terms of waiting time) and the server (idle time and overtime). The goal is to minimize a (weighted)sum of these two loss functions. Explicit calculation of the optimal appointment schedule is problematic when there are many clients, since it requires the evaluation of high-dimensional integrals (Denton and Gupta 2003).

Most of the contributions on appointment scheduling are based on exponential service times, such as in Wang (1999), Kaandorp and Koole (2007), Hassin and Mendel (2008) and Turkcan et al. (2011); or a phase-type distribution for the service times, such as in Wang (1997), Vanden Bosch et al. (1999), Vanden Bosch and Dietz (2000) and the previous chapters. Also, it is common to assume that the service times are independent and identically distributed.

Simulation approaches are used to evaluate the performance of scheduling heuristics; see, for example, Ho and Lau (1992), Robinson and Chen (2003), and references mentioned in the overview of Günal and Pidd (2010). We note, however, that the evaluation of heuristics with the help of simulation studies can be a time consuming effort or is often limited to specific service settings, including service-time distributions and cost ratios (Yang et al. 1998). To the best of our knowledge, the number of studies that use simulation in order to trace an optimal schedule are modest, but for an example see Zhu et al. (2012).

An alternative approach to deal with the high-dimensional optimization problem is to impose restrictions, such as equally-spaced interarrival times, see for example Hassin and Mendel (2008).

We also mention the sequential approach of Kemper et al. (2014), that enables the service provider to sequentially optimize the client's appointment time. The sequential approach clearly reduces the dimensions of the optimization problem. It is shown to be generic and flexible allowing nonidentical service-time distributions and loss functions, and accommodating phenomena such as no-shows and walk-in clients. However, the computation gets involved for larger schedules with service-time distributions other than the exponential.

Given the importance and relevance of the problem, and the fact that there is, to the best of our knowledge, no fully satisfactory solution available, we decided to explore an alternative approach. Our approach is able to deal with larger schedules, and general service-time distributions, such as the lognormal (Klassen and Rohleder 1996) or the Weibull (Babes and Sarma 1991) as often seen in practice. The key idea is that we reduce the computational complexity by a lag order approximation: in determining a client's optimal arrival time (minimizing the combined loss function), we only take into account the effects of the two immediately preceding clients. We refer to this approach as the *lag order approximation method*, with the lag order being two.

The organization of this chapter is as follows. In Section 5.2 we formulate the problem in mathematical terms. The lag order approximation method is then presented in Section 5.3. The performance of the lag order approximation method is evaluated in Section 5.4 by studying some numerical examples and a real-life example from a radiology department. The results show that our method needs significantly less computational effort, and is able to derive appointment schedules that are close to optimal. Finally, we conclude and discuss directions for further research in Section 5.5.

5.2 Problem statement

Consider a service system in which clients $i = 1, \ldots, n$ are scheduled to arrive at times $t_1, \ldots, t_n \in \mathbb{R}^+$. Each client has a service-time duration, which is denoted by the random variable B_i for client i. The service system has a single server and if upon arrival client i finds the server idle, service starts immediately. If the server is busy, then client i waits for his turn until all clients that are scheduled before client i have finished their service. We assume that both the clients and the server are punctual, and we do not allow for no-shows and walk-in clients.

The vector $(t_1, \ldots, t_n)$ is called an appointment schedule for this service system. For a given schedule, we denote by I_i the time that the server has been idle when service for client n starts. We denote by W_i the waiting time of client i. Note that the sojourn time S_i of client i is defined by $S_i = W_i + B_i$. In most settings the total duration of the session (that is, the available time, T, in which clients can be scheduled) is finite. However, it can happen that after the planned session-end time there are still clients that need to be served. We therefore define the system's lateness O as the overtime that the server has to make in order to finish all services. It is useful to define the *interarrival times* by

$$x_i = t_{i+1} - t_i, \quad i = 1, \ldots, n-1.$$

Due to the Lindley recursion (Lindley 1952) the idleness I_i can be written as

$$I_i = \max\{x_{i-1} - S_{i-1}, 0\}, \quad i = 2, \ldots, n, \tag{5.1}$$

and the waiting time W_i as

$$W_i = \max\{S_{i-1} - x_{i-1}, 0\}, \quad i = 2, \ldots, n. \tag{5.2}$$

From Eqns. (5.1) and (5.2) it follows that $W_i + I_i = |S_{i-1} - x_{i-1}|$, for $i \geq 2$. The overtime can be expressed as

$$O = \max\{t_n + S_n - T, 0\}. \tag{5.3}$$

Since $t_1 = 0$, both $W_1 = 0$ and $I_1 = 0$. The objective of the appointment scheduling problem is to find a schedule $(t_2, \ldots, t_n)$, or equivalently $(x_1, \ldots, x_{n-1})$, such that a loss function LF, which depends on the I_i, W_i, and O, is minimized. Throughout this chapter, we assume that LF has the form

$$LF(x_1, \ldots, x_{n-1}) = \sum_{i=2}^{n} (\mathbb{E}g(I_i) + \mathbb{E}h(W_i)) + \mathbb{E}v(O), \tag{5.4}$$

with $g(\cdot), h(\cdot)$, and $v(\cdot)$ nondecreasing continuous and convex functions.

5.3 The lag order approximation method

In this section we present the lag order approximation method in its general form. The optimal schedule is the one that minimizes (5.4):

$$\left(x_1^\star, \ldots, x_{n-1}^\star\right) = \underset{x_1,\ldots,x_{n-1}}{\arg\min}\ LF(x_1, \ldots, x_{n-1}). \tag{5.5}$$

The waiting times $W_1, \ldots, W_i$ of successive clients are related because of Eqn. (5.2), as

$$\begin{aligned} W_i &= \max\left\{W_{i-1} + B_{i-1} - x_{i-1}, 0\right\}; \\ &= \max\left\{\max\left\{W_{i-2} + B_{i-2} - x_{i-2}, 0\right\} + B_{i-1} - x_{i-1}, 0\right\}; \\ &\ \ \vdots \\ &= \max\left\{\max\left\{\ldots \max\left\{B_1 + W_1 - x_1, 0\right\} + B_2 - x_2, 0\right\} + \ldots, 0\right\}. \end{aligned}$$

In a similar way the idle time I_i is related with $W_1, \ldots, W_{i-1}$. The idea of the approximation is to ignore (that is, set to zero) terms in this recursion beyond K terms. Let $k = \min\{K, i-1\}$ then $\tilde{W}$ is the recursion expanded in the first k steps, and W_{i-k-1} is set to zero (thus stopping further development of the recursion). Similarly, we express the approximated idle times and overtime as $\tilde{I}_i$ and $\tilde{O}$. The optimization with respect to this partial information in (5.4) is called the lag order approximation method of order K, which minimizes

$$\min_{x_1,\ldots,x_{n-1}} LF_K(x_1, \ldots, x_{n-1}) = \sum_{i=2}^{n} \left(\mathbb{E}g(\tilde{I}_i) + \mathbb{E}h(\tilde{W}_i)\right) + \mathbb{E}v(\tilde{O}). \tag{5.6}$$

Note that $K = n-1$ corresponds to the original optimization problem of (5.4). The advantage of this approach is that by limiting the dependence on predecessors we are able to use convolution formulas to compute the $\tilde{W}_i$ and $\tilde{I}_i$, and the schedule's overtime $\tilde{O}$.

5.3.1 Loss functions

In this subsection, we present two loss functions that are commonly used in the literature. The loss function includes the expected waiting times, the expected idle times, and the expected overtime with different weighting factors.

One often chooses general polynomial functions and sets $g(x) = \omega x^\lambda$, $h(x) = (1-\omega)x^\lambda$ and $v(x) = \beta x$, where $\omega,\ \beta \geq 0$, and $\lambda > 0$. However, setting $\omega = \frac{1}{2}$ and taking $\lambda = 1, 2$ gives us two valuable insights, which we will refer to as the absolute value loss function and the quadratic loss function. Note that in these cases the idle and waiting times are weighted equally, similar to Fries and Marathe (1981)

and Chapter 2. In this chapter we adopt the convention to take the same weighting factors for the idle and waiting times, but remark that the method does not require it.

Absolute value loss function. The absolute value loss function can be obtained by taking $g(x) = h(x) = \frac{1}{2}x$ and $v(x) = \beta x$, with $\omega, \beta \in \mathbb{R}^+$. We know from Section 5.2 that the expectation of the loss function reduces to

$$LF(x_1, \ldots, x_{n-1}) = \frac{1}{2}\sum_{i=1}^{n-1} \mathbb{E}|S_i - x_i| + \beta \mathbb{E}O.$$

This loss function penalizes deviations from the schedule (either caused by waiting or by idling) linearly. It has been used (with $\beta = 0$) by, e.g., Wang (1997), Vanden Bosch et al. (1999), Kaandorp and Koole (2007) and Kuiper et al. (2015).

Quadratic loss function. The quadratic loss function penalizes the deviation from the schedule quadratically instead of linearly. We have $g(x) = h(x) = \frac{1}{2}x^2$ and $v(x) = \beta x^2$. Since $W_i^2 + I_i^2 = (S_{i-1} - x_{i-1})^2$ for $i > 1$, the expected losses reduce to

$$LF(x_1, \ldots, x_{n-1}) = \frac{1}{2}\sum_{i=1}^{n-1} \mathbb{E}\big(S_i - x_i\big)^2 + \beta \mathbb{E}O^2.$$

This loss function has been used (with $\beta = 0$) by, e.g., Schild and Fredman (1961) and Kemper et al. (2014).

5.3.2 Technical background of the lag order procedure

To compute the solution $x_1, \ldots, x_{n-1}$ through the lag order approximation method of order K, we use the following derivations. In case of lag order 0 the sojourn-time distribution for each client is equal to his service time, i.e., $S_i = B_i$. Obviously, the interdependence between interarrivals is removed in this way. For lag order I, the sojourn-time distribution is linked through the waiting time of the previous client $i > 1$ (for $i = 1$ we use the lag order 0: $S_1 = B_1$)

$$S_i = B_i + W_i \approx B_i + \tilde{W}_i(x_{i-1}) = B_i + \max\{B_{i-1} - x_{i-1}, 0\}. \tag{5.7}$$

In lag order II the interdependence increases to the waiting times of the two clients before client $i > 2$ (for $i = 2$ we use (5.7))

$$\begin{aligned} S_i &= B_i + W_i \approx B_i + \tilde{W}_i(x_{i-2}, x_{i-1}) \\ &= B_i + \max\{B_{i-1} + \max\{B_{i-2} - x_{i-2}, 0\} - x_{i-1}, 0\}. \end{aligned} \tag{5.8}$$

Since we have now derived approximations of the client specific sojourn times we can implement those in (5.1), (5.2) and (5.3) to compute the waiting and idle times, and the overtime respectively. The algorithms to find optimal lag order K solutions are written in `MATLAB R2012b`. This program optimizes (5.6) with either the linear loss function or the quadratic loss function, for any service-time distribution, exploiting `MATLAB R2012b`'s built-in minimization routines.

We outline the algorithm for lag order II (the procedure can be easily adapted to incorporate other lag orders). The program contains three stages.

1. The lag ordered loss functions are implemented in a `for loop`, using the lag order 0 for the second client; lag order I for the third client (as in (5.7)); and the lag order II for the remaining clients, c.f. (5.8).

2. The sum of all losses is computed using `MATLAB`'s adaptive Simpson quadrature routines (tolerance is 10^{-5}), over the different lag ordered loss functions.

3. Finally, we jointly optimize the aggregate over $n-1$ interarrival times with tolerance level of 10^{-4} and an all-ones vector (with dimension $n-1$) as start vector.

5.3.3 Technical background for generating LF values

In order to simulate the different distributions we apply a similar approach. Instead of computing integrals we use a Monte Carlo simulation study. The procedure is as follows: in a set of $n \times 10^5$ random drawings from a particular distribution one minimizes the loss function over the interarrivals $x_1, \ldots x_{n-1}$. We repeat this 100 times (so in total $n \times 10^7$ random drawings are needed) to get a sample mean, $\bar{LF}$, and a sample variance, s^2_{LF}, of system's losses. The Central Limit Theorem dictates that with 95% confidence the average loss is containted in the interval given by $\bar{LF} \pm z_{0.05} \frac{s_{LF}}{10}$, which then can be compared with the lag order method.

In this section we first explore the approach for exponential service times. Then, in Section 5.4, we study the lognormal and Weibull service-time distributions. For all results generated in the various studies in this chapter, the simulation study stops when the estimated values of the loss function have a confidence interval with a width smaller than 1‰ of the estimates.

5.3.4 Example with exponentially distributed service times

We illustrate the results of our lag order approximation method for a system with $n = 11$ clients in Figure 5.1. We assume that the service times of the clients are independent and exponentially distributed (i.i.d.) with parameter $\mu_i = 1$ for $i = 1, \ldots, n$. We optimize the loss function with respect to the quadratic loss function. Figure 5.1 displays the optimal slot sizes $x_1^\star, \ldots, x_{10}^\star$ (interarrival times) for various lag orders. The figure supports the conclusion that optimal scheduling according to

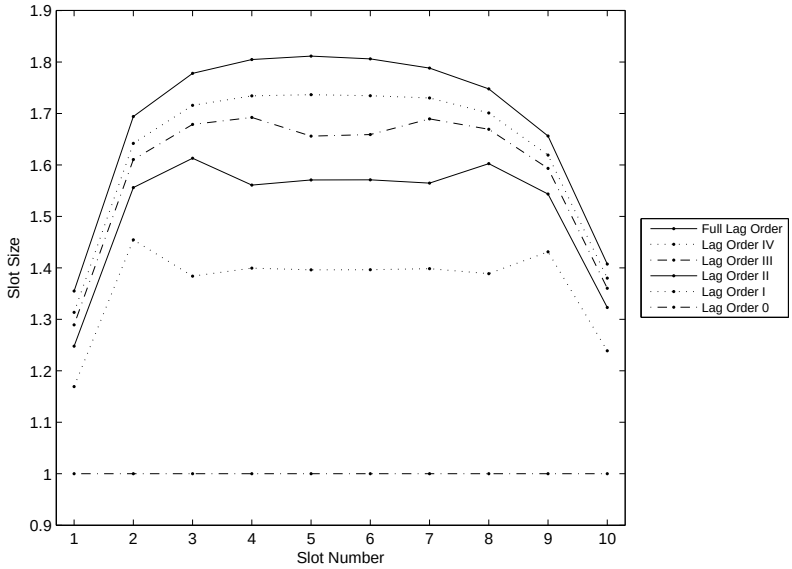


Figure 5.1: Optimal slot sizes for the lag order approximation method with quadratic loss. $n = 11$, i.i.d. exponential($\mu = 1$) service times.

lag order 0 corresponds to setting each interarrival time equal to the average service time, which behaves as a $D/M/1$ queue with load 1.

When the lag order increases the approach results in larger slot sizes; for example, for the first slot the interarrival time increases from about 1.18 based on a lag order I to a slot size of about 1.32 based on a lag order IV. Furthermore, for lower lag orders, between I and III, the slot sizes alternate in the beginning of the schedule and towards the end of the schedule. For higher lag orders, from order IV and onwards, the slots in the middle of the schedule are larger than the slots in the beginning and the end of the schedule.

The full lag order corresponds to a lag order of $n-1$. These results are obtained by the procedure proposed in Wang (1993), where we extended the algorithm to handle quadratic loss functions. The results show that as the lag order increases, the appointment schedule converges to the optimal schedule. Figure 5.1 also shows that the lag order method results in shorter interarrival times than optimal schedules, thus yielding longer waiting times and shorter idle times on average. As seen in the figure, the slot sizes increase, then decrease, and then increase and decrease again, hence exhibiting a non-unimodal pattern. However, as the lag order increases, this effect diminishes. In fact, the optimal appointment schedule (the full lag order) does not show this behavior, but instead follows a dome shape pattern that is also found and described in Chapter 2.

Table 5.1 summarizes the performance of the lag order method in practice by displaying the value of the loss function for the first two lag orders including the full lag

	Linear loss			Quadratic loss		
Method	Value LF	Δ Opt	Time (s)	Value LF	Δ Opt	Time (s)
Lag order 0	22.220	111.1%	1.2E-1	47.627	160.1%	4.7E-2
Lag order I	12.720	20.8%	1.1E1	22.918	25.2%	2.6E1
Lag order II	11.109	5.5%	3.9E2	19.476	6.4%	1.2E3
Simulation	10.526	–	5.0E2	18.311	–	5.9E3

Table 5.1: Optimization results of the lag order approximation method compared with simulation. $n = 11$, i.i.d. exponential($\mu = 1$) service times.

order obtained by extensive simulation. We see that increasing the lag order reduces the expected total loss of the system, which approaches the loss in the full lag order.

5.4 The lag order approximation method in practice

In this section we apply the lag order approximation method to realistic appointment scheduling problems. In the previous section, we were able to compute the optimal schedule for exponentially distributed service times. However, empirical evidence shows that this setting is generally unrealistic (see, e.g., Çayırlı and Veral 2003). Commonly seen service-time distributions in realistic settings are the Weibull and the lognormal distribution. We study these distributions in Section 5.4.1. Then, in Section 5.4.2 we apply our approach to a real-life example of a radiology process in a hospital.

5.4.1 Approximating realistic service-time distributions

According to Çayırlı and Veral (2003) service-time distributions have a coefficient of variation (cv) typically in the range of 0.35 to 0.85. Given this range, we illustrate our method by using the Weibull and lognormal distribution with coefficients of variation: 0.35, 0.5, and 0.85.

In Tables 5.2 and 5.3 we show the results of the various lag orders to this problem. The results are compared to an optimal schedule derived through simulation only, as described in the background paragraph of Section 5.3.

For each approach in our study the tables report the values of the loss function, the difference between the LF values of the approach and that of the simulated optimal value, and the computer's CPU time of each method, where for example 4E3 (s) means 4×10^3 seconds.

From the tables we conclude that the application of the lag order approximation method results in a small loss of quality of the appointment schedule. Across cases our approach with lag order 0 generates schedules that are at least about 63% from the optimal LF value, and hence the lag order 0 is not useful in designing optimal appointment schedules. In case of a linear loss function (and either lognormal or

Weibull	Linear loss			Quadratic loss		
Method	Value LF	Δ Opt	Time (s)	Value LF	Δ Opt	Time (s)
			cv $= 0.35$			
Lag order 0	5.488	63.3%	9.3E-2	5.168	193.6%	3.1E-2
Lag order I	3.659	8.9%	1.2E1	2.167	23.1%	7.7
Lag order II	3.435	2.2%	6.4E2	1.855	5.4%	3.5E2
Simulation	3.360	–	5.1E2	1.760	–	7.7E3
			cv $= 0.5$			
Lag order 0	8.808	77.0%	1.1E-1	10.951	188.3%	3.1E-2
Lag order I	5.534	11.2%	9.3	4.689	23.4%	1.7E1
Lag order II	5.115	2.8%	4.5E2	4.008	5.5%	6.3E2
Simulation	4.977	–	5.0E2	3.799	–	7.7E3
			cv $= 0.85$			
Lag order 0	18.172	104.8%	2.2E-1	33.746	169.4%	6.2E-2
Lag order I	10.489	18.2%	2.3E1	15.597	24.5%	4.4E1
Lag order II	9.268	4.5%	1.8E3	13.282	6.0%	3.3E3
Simulation	8.871	–	6.7E2	12.526	–	9.2E3

Table 5.2: Optimization results of the lag order approximation method compared with the optimal values generated by simulation. $n = 11$, i.i.d. Weibull service times.

Weibull service-time distribution), a lag order I approximation generates schedules that are within 25% from the optimal LF value, and that are more than 10 (and even 100 times in case of a quadratic loss function) faster to compute, c.f. Table 5.1.

Furthermore, from the tables we see that lag order II generates schedules that are reasonably close to the optimal LF value (around 2% to 7%). However, the computation time may be too substantial when using the linear loss function. By taking a quadratic loss function computation times are further reduced: about 3 times faster when cv $= 0.85$, up to about 20 times faster in case of cv $= 0.35$. We note that the program code is rather straightforward and that further enhancements to reduce time are possible, see Section 5.3.

In Figures 5.2 and 5.3 we compare the schedules derived by the lag order method with the simulated optimal schedule in both a Weibull and lognormal setting with coefficient of variation equal to 0.5. We see that the typical shape of the lag order method does not depend on the actual distribution nor on the loss function.

In the following section we apply our approximation approach to a realistic example, where the client's loss function is quadratic. The service times follow a Weibull distribution.

5.4.2 Application in a CT-scan process

Let us consider a real-life scheduling problem in a CT-scan process, with the following parameters: $n = 20$, $T = 300$ (min). We choose the quadratic loss function with

Lognormal	Linear loss			Quadratic loss		
Method	Value LF	Δ Opt	Time (s)	Value LF	Δ Opt	Time (s)
			cv = 0.35			
Lag order 0	6.537	84.3%	1.2E-1	5.519	173.6%	4.7E-2
Lag order I	4.054	14.3%	1.5E1	2.507	24.3%	1.2E1
Lag order II	3.682	3.8%	6.0E2	2.134	5.8%	4.6E2
Simulation	3.546	–	4.6E2	2.017	–	7.6E3
			cv = 0.5			
Lag order 0	9.843	91.5%	1.2E-1	11.560	162.7%	3.1E-2
Lag order I	6.020	17.1%	1.1E1	5.503	25.0%	1.5E1
Lag order II	5.379	4.7%	6.5E2	4.683	6.4%	3.7E2
Simulation	5.139	–	4.8E2	4.401	–	9.2E3
			cv = 0.85			
Lag order 0	17.395	98.1%	1.1E-1	35.064	134.8%	4.7E-2
Lag order I	10.726	22.1%	2.0E1	18.750	25.6%	2.6E1
Lag order II	9.372	6.7%	6.7E2	16.037	7.4%	1.2E3
Simulation	8.783	–	5.9E2	14.932	–	7.8E3

Table 5.3: Optimization results of the lag order approximation method compared with the optimal values generated by simulation. $n = 11$, i.i.d. lognormal service times.

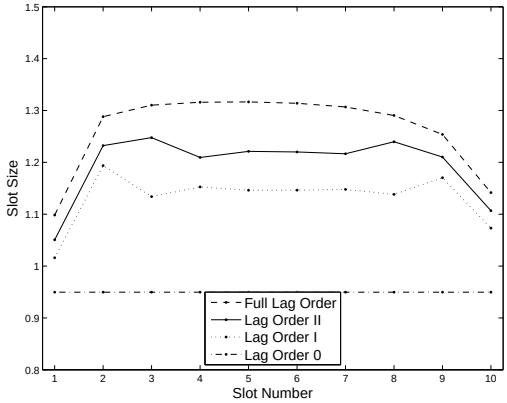
weights $\omega = 0.75$ and $\beta = 1.5$. Thus, we have

$$LF = \sum_{i=1}^{20} \left(\frac{3}{4}\mathbb{E}I_i^2 + \frac{1}{4}\mathbb{E}W_i^2 \right) + \frac{6}{4}\mathbb{E}O.$$

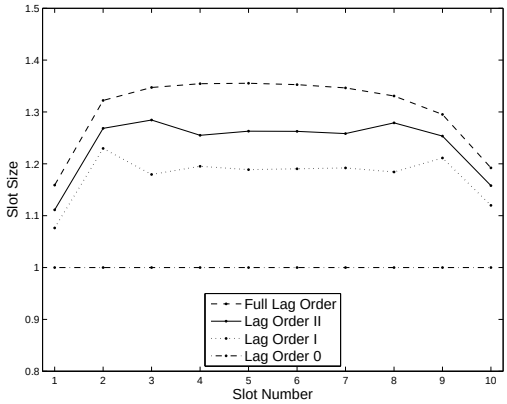
We obtained service-time data from the Deventer Hospital (described in De Mast et al. 2011). The best fit to these data is a lognormal distribution with location parameter $\mu = 2.4$ and scale parameter $\sigma = 0.58$. The coefficient of variation equals 0.63.

We compare several approximation approaches in Table 5.4. The table includes the hospital's current schedule and the simulated optimal schedule. For each approach in our study the table reports the values of the loss function, the difference between the LF values of the approach and that of the simulated optimal value, and the computer's CPU time.

We observe that, given the loss function, the current schedule differs about 20% compared to the simulated optimal value of the LF. Also, we see that the lag order I and lag order II differ about 17% and 7.7% from the optimal schedule. The advantage of the lag order method is that it reduces computation time considerably. Finally, the table also offers, for comparison, the performance of an equidistant schedule, where all slot sizes are equal (Hassin and Mendel 2008). The performance of such schedule is good, but computation of the optimal length of the (identical) slot sizes costs 2.5 times more CPU time than our lag order II approximation.



(a) Linear loss.



(b) Quadratic loss.

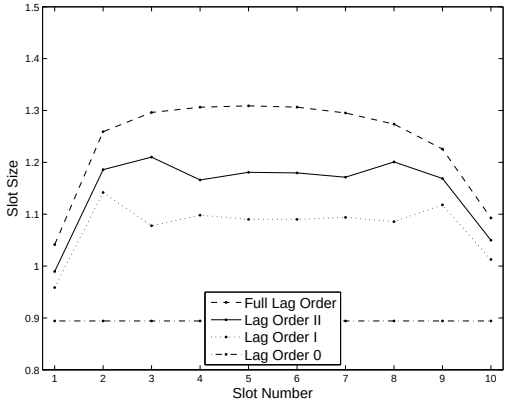
Figure 5.2: Optimal slot sizes for the lag order approximation method and simulation. $n = 11$, i.i.d. Weibull distributed service times, with mean 1 and CV $= 0.5$.

Method	Value LF	Δ Opt	Time (s)
Hospital's schedule	2 000	20%	–
Lag order I	1 940	17%	4.3
Lag order II	1 788	7.7%	2E2
Equidistant simulation	1 670	<1%	5E2
Simulation	1 660	–	12E3

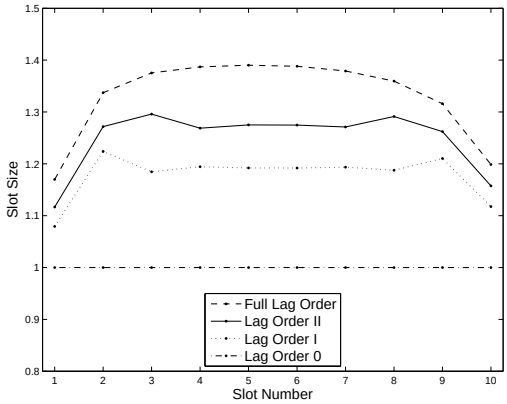
Table 5.4: The hospital's current schedule and lag orders I and II compared with simulation. Weighted-quadratic ($\omega = 0.75$) loss function with overtime ($\beta = 1.5$), $n = 20$, i.i.d. lognormal (cv $= 0.63$) service times.

5.5 Conclusion

In this chapter we study the problem of designing a suitable appointment schedule for a session with n clients. The clients are punctual but have random service times.



(a) Linear loss.



(b) Quadratic loss.

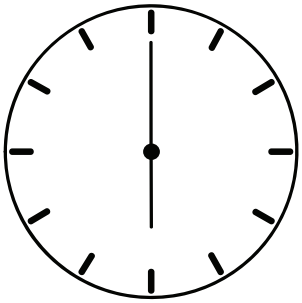
Figure 5.3: Optimal slot sizes for the lag order approximation method and simulation. $n = 11$, i.i.d. lognormal distributed service times, with mean 1 and CV $= 0.5$.

We do not allow for no-shows and walk-in clients. We develop a lag order approximation method that minimizes a loss function of the waiting time of the clients, the idle time of the server, and the overtime of the schedule. The approximation method with a lag order I yields near-optimal schedules (about 20% from the performance of the optimal schedule) in substantially smaller computation times. The lag order II approximation yields schedules that are about 5% from the optimal LF value, and that can be up to 20 times faster than simulation in case of a quadratic loss function and when the coefficient of variation of the service times is relatively small (say, cv $= 0.35$).

There are a few interesting directions for further research. A logical next step would be to extend the lag order approximation method to allow for no-shows and walk-in clients. Including both presents opportunities, since no-shows create gaps in

the schedule which in turn can be potentially filled by walk-ins.

In addition, we studied the approach to schedule patients in a CT-scan process. One could also study a setting in which there are several patient groups, see for example Creemers et al. (2012). The case of nonidentical service time distributions can easily be incorporated in the lag order approach presented in this chapter.



6. Efficient Procedures for Appointment Scheduling in Healthcare

This chapter presents a general approach for setting up appointment schedules, combining the insights obtained in the previous chapters. It has been implemented in an easy-to-use webtool, which is described in Chapter 7, that generates schedules essentially instantaneously.

The underlying computational machinery relies on the phase-type techniques developed in Chapter 2. The base model has been extended so as to cover various add-ons and relevant situational characteristics. For example, it incorporates no-shows, but does so in a more pragmatic way than in Chapter 3. As a result the approach presented in this chapter meets the requirements stated in the introduction of the thesis: it is at the same time flexible, robust, and fast. Additionally, we offer novel analytical results for the situation in which there are relatively many patients whose service times are independent and stochastically identical.

6.1 Introduction

The key difficulty in scheduling is to deal effectively with uncertainty and variability. Early work yielded computationally simple heuristics, such as the widely used rules of Bailey (1952) and Welch and Bailey (1952). These rules schedule appointments in equidistant time slots (blocks), with lengths equal to the mean service time, and

overbook the first slot by one or more patients. Such approaches do not deal effectively with variability in service times and may result in excessive waiting times when variability is substantial (as is typical for specialty care and surgery; see Gupta and Denton (2008) and Kemper et al. (2014)). More recent studies typically incorporate variability in service times, and sometimes random no-shows, cancelations, walk-ins and emergencies (Gupta and Denton 2008, Hassin and Mendel 2008, Çayırlı et al. 2012, Zacharias and Pinedo 2014). Tardiness of patients is claimed to be a less critical factor in the design of appointment schedules (Çayırlı et al. 2006). Moreover, the performance of scheduling approaches depends on characteristics such as the number of patients to be scheduled in a session and the relative costs of waiting times and idle time (Ho and Lau 1992).

Our approach aims to strike a proper balance between waiting times for patients and idle times for healthcare providers; in the next section we introduce and motivate an objective function weighting these. The approach is based on techniques originating in queueing theory, where we rely on the idea of approximating the distribution of the service times by its phase-type counterpart. We distinguish between two situations: one in which there is a relatively large number of stochastically identical patients (the stationary scheduling problem), and one in which there are relatively few patients to be scheduled (the transient scheduling problem). In the former case, the challenge is to select the value of the interarrival time in the corresponding queueing system that minimizes the objective function. With a heavy-traffic argument, we explain why schedules based on the first two moments work so well. For the latter case, we devise a technique to evaluate the objective function, which can then be optimized using standard numerical techniques. By combining stored, precalculated schedules for specific parameter values with interpolations and fitting techniques, we have developed a tool that accurately and efficiently determines optimal schedules across a wide range of parameter values (covering those that are common in healthcare). Extensive numerical experiments show that in the test cases considered in the literature our method significantly outperforms existing appointment scheduling rules.

6.2 Model and approach

In this section we first sketch the model considered in this chapter: the appointment scheduling problem is cast in a queueing-theoretic framework. The final part of the section describes how the service times are approximated by an appropriately chosen phase-type counterpart.

6.2.1 Preliminaries

We model the situation as a single-server queueing model. Patients $i = 1, \ldots, n$ arrive at or before their scheduled arrival time t_i, with $t_1 = 0$, where n is the number of patients to be seen in a single session. We consider the situation in which the

patients have appointments with a specific healthcare provider (specialist, doctor), who therefore acts as a single server. We assume that the service times $B_1, \ldots, B_n$ are i.i.d. random variables. We define by W_i the net *waiting time* of the i-th patient, that is, the time in between his scheduled arrival and the moment he receives service, where we set $W_1 = 0$. Define I_i as the server *idle time* prior to the i-th patient's arrival, with $I_1 = 0$. It is a standard result that, by virtue of the Lindley recursion, with $x_i = t_{i+1} - t_i$ (the interarrival time), the I_i can be determined recursively:

$$I_i = \max\{x_{i-1} - W_{i-1} - B_{i-1}, 0\};$$

likewise,

$$W_i = \max\{W_{i-1} + B_{i-1} - x_{i-1}, 0\}. \tag{6.1}$$

Evidently, we cannot have that both W_i and I_i are strictly positive. This observation leads to the following identities, where $S_i = W_i + B_i$ denotes the *sojourn time* of the i-th patient:

$$I_i + W_i = |\, S_{i-1} - x_{i-1} \,| \quad \text{and} \quad W_i^2 + I_i^2 = (S_{i-1} - x_{i-1})^2.$$

The *makespan*, defined as the epoch that patient n has been fully served, can be written in two alternative ways, noting that $\sum_{i=1}^{n-1} x_i = t_n$,

$$\sum_{i=1}^{n} B_i + \sum_{i=2}^{n} I_i = \sum_{i=1}^{n-1} x_i + S_n. \tag{6.2}$$

In healthcare the makespan is also referred to as the *session end time*.

6.2.2 Objective function

In our approach the schedules are generated so as to optimize a specific objective function. The objective functions considered in this chapter aim at striking a proper balance between idle times (being unfavorable to the server) and waiting times (being unfavorable to the patients). The interarrival times should be chosen so as to minimize the objective function. We specifically consider the so-called weighted-linear objective function, where without loss of generality we assume that the two weights add up to 1: with the weights ω and $1 - \omega$ in the interval $(0, 1)$, we define

$$\mathscr{F}^{(\ell)}[x_1, \ldots, x_{n-1}] = \omega \sum_{i=1}^{n} \mathbb{E} I_i + (1 - \omega) \sum_{i=1}^{n} \mathbb{E} W_i. \tag{6.3}$$

For given weight ω, the optimal schedule is the sequence $\bar{x}_1, \ldots, \bar{x}_{n-1}$ that minimizes $\mathscr{F}^{(\ell)}[x_1, \ldots, x_{n-1}]$. Define $\bar{I}(\omega) = \sum_{i=1}^{n} \mathbb{E} I_i$ and $\bar{W}(\omega) = \sum_{i=1}^{n} \mathbb{E} W_i$ as the mean total idle and waiting time of the optimal schedule $\bar{x}_1, \ldots, \bar{x}_{n-1}$ for the weight ω. Generally, when ω approaches 1 (i.e., the situation in which the value of the objective

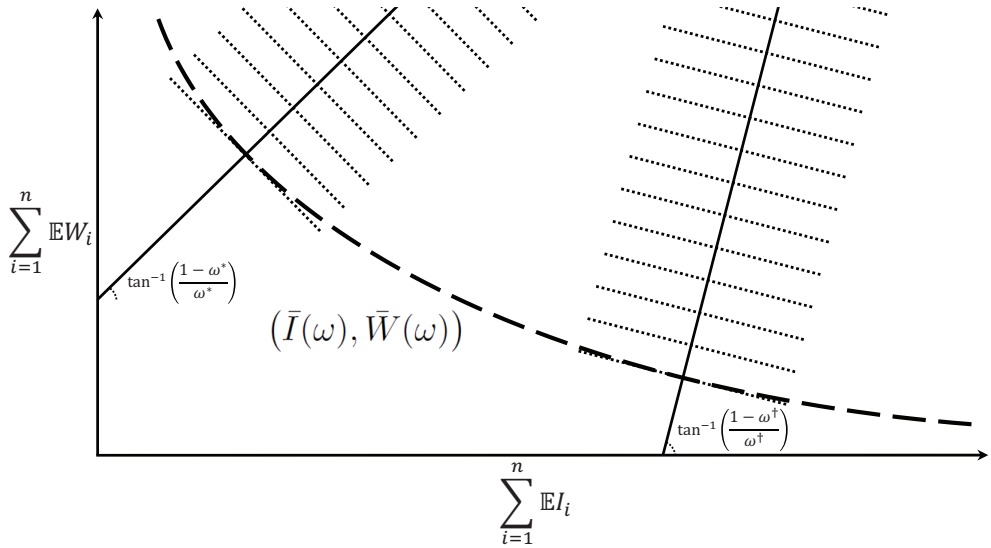


Figure 6.1: Efficient frontier showing optimal schedules given ω (dashed curve) and iso-performance lines for two particular values of ω (dotted lines).

function is essentially determined by the idle times only), $\bar{W}(\omega)$ explodes; cf. the *utilization law* of Hopp and Spearman (2008). On the other hand, when ω approaches 0 the contribution of the mean total idle time experienced by the server, i.e., $\bar{I}(\omega)$, increases sharply.

The dashed curve in Figure 6.1, consisting of combinations $\left(\bar{I}(\omega), \bar{W}(\omega)\right)$ for $\omega \in (0, 1)$, is named the efficient frontier in the management literature. All feasible schedules correspond to combinations on or above this curve. The efficient frontier conceptualizes that some differences between the performance of schedules are due to the trade-off between $\bar{I}(\omega)$ and $\bar{W}(\omega)$ and other differences are due to suboptimality. The first is expressed by ω and corresponds to a position on the curve. The latter are represented by the iso-performance lines (the dotted lines in Figure 6.1), whose angle is determined by the ω of choice. The given objective function thus breaks down the performance of schedules into a trade-off component (which ultimately is a strategic decision) and an optimality component (which is a matter of superiority of one schedule compared to another).

Interestingly, the optimization problem (6.3) can be rewritten in terms of only (expected) waiting times:

$$\underset{x_1,\ldots,x_{n-1}}{\arg\min}\ \mathscr{F}^{(\ell)}[x_1, \ldots, x_{n-1}] = \underset{x_1,\ldots,x_{n-1}}{\arg\min} \left(\mathbb{E}W_n + \sum_{i=1}^{n-1}(1-\omega)\mathbb{E}W_i + \omega \sum_{i=1}^{n-1} x_i \right), \quad (6.4)$$

as can be seen as follows. From Eqn. (6.2), by comparing the 'makespan' up to patient

i with that up to patient $i-1$, we obtain

$$B_i + I_i = \left(\sum_{j=1}^{i-1} x_j + S_i\right) - \left(\sum_{j=1}^{i-2} x_j + S_{i-1}\right) = x_{i-1} + S_i - S_{i-1}.$$

This directly leads to

$$I_i = x_{i-1} + W_i - W_{i-1} - B_{i-1}. \tag{6.5}$$

Relation (6.4) now follows by taking expectations in (6.5) and by noting that a number of terms (in a telescopic series) vanish. The $\mathbb{E}B_i$ terms can be dropped from the resulting expression for the objective function, as these constants do not affect the optimal interarrival times $\bar{x}_1, \ldots, \bar{x}_{n-1}$.

Since we aim to minimize the objective function in (6.4) it is important to know whether the function is convex in its arguments $x_1, \ldots, x_{n-1}$. Convexity of the optimization problem has attracted substantial attention in the past. Partial results, corresponding to the situation of exponentially distributed service times, were reported in Pegden and Rosenshine (1990), Kaandorp and Koole (2007) and Hassin and Mendel (2008). Note that this exponentiality assumption, which implies a coefficient of variation equal to 1, overestimates the variability that one typically encounters in healthcare settings, and as a consequence the resulting schedules tend to be overly conservative. For a specific objective function (different from ours), convexity has been proven in Wang (1993).

For the sake of completeness we provide a straightforward proof of the objective function $\mathscr{F}^{(\ell)}[x_1, \ldots, x_{n-1}]$ being convex in $\boldsymbol{x} \equiv (x_1, \ldots, x_{n-1})$. Define by $W_i(\boldsymbol{x})$ the waiting time of the i-th patient if the vector of interarrival times is given by $\boldsymbol{x}$. For a given i, define $Z_j = \sum_{k=i-j}^{i-1} B_k$ and $y_j(\boldsymbol{x}) = \sum_{k=i-j}^{i-1} x_k$, the following distributional equality can be obtained after repeated iteration of (6.1):

$$W_i(\boldsymbol{x}) \stackrel{\mathrm{d}}{=} \max_{j \in \{0,1,\ldots,i-1\}} (Z_j - y_j(\boldsymbol{x}))$$

(following the convention that empty sums are defined as 0). Now observe that, with $\lambda \in [0,1]$ given and $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}_+^{n-1}$,

$$\begin{aligned}
\mathbb{E}W_i(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) &= \mathbb{E}\left(\max_{j \in \{0,\ldots,i-1\}} (Z_j - (y_j(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2)))\right) \\
&= \mathbb{E}\left(\max_{j \in \{0,\ldots,i-1\}} \lambda(Z_j - y_j(\boldsymbol{x}_1)) + (1-\lambda)(Z_j - y_j(\boldsymbol{x}_2))\right) \\
&\leqslant \mathbb{E}\left(\max_{j \in \{0,\ldots,i-1\}} \lambda(Z_j - y_j(\boldsymbol{x}_1))\right) \\
&\quad + \mathbb{E}\left(\max_{j \in \{0,\ldots,i-1\}} (1-\lambda)(Z_j - y_j(\boldsymbol{x}_2))\right),
\end{aligned}$$

which equals $\lambda\,\mathbb{E}W_i(\boldsymbol{x}_1) + (1-\lambda)\mathbb{E}W_i(\boldsymbol{x}_2)$, so that $\mathbb{E}W_i(\boldsymbol{x})$ is convex. Because of (6.4), the objective function is a linear combination of expected waiting times (with positive weights), minus a function that is linear in $\boldsymbol{x}$, and therefore convex as well. As a consequence, there is a unique minimum on $\mathbb{R}_+^{n-1}$.

Throughout this chapter we primarily focus on the weighted-linear objective function, but most of our material carries over to alternative objective functions, e.g. the weighted-quadratic one:

$$\mathscr{F}^{(\mathrm{q})}[x_1,\ldots,x_{n-1}] = \omega\sum_{i=1}^{n}\mathbb{E}I_i^2 + (1-\omega)\sum_{i=1}^{n}\mathbb{E}W_i^2,$$

where ω is assumed to be in $(0,1)$. The 'mixed' objective functions $\mathscr{F}^{(\ell\mathrm{q})}$ (weighted-linear-quadratic) and $\mathscr{F}^{(\mathrm{q}\ell)}$ (weighted-quadratic-linear) are defined in the obvious way.

6.2.3 Stationarity

In this chapter, special attention will be paid to the situation that the B_i are governed by a single distribution, while n is large. Under the assumption that the patients arrive equidistantly with interarrival time x, the distribution of the waiting time is uniquely defined through the distributional fixed point equation, cf. Eqn. (6.1),

$$W = \max\{W + B - x, 0\}.$$

The resulting queueing system is of the D/G/1 type, which does not allow explicit solutions in general. In the cases of exponential and Erlang(2) service times, however, the stationary waiting-time distribution can be given in (semi-)closed-form (see Section 6.3).

We now point out that the first moment $\mathbb{E}I$ can easily be found. Dividing (6.2) by n, taking expectations, and considering the limit when $n\to\infty$, we conclude that

$$\mathbb{E}I = x - \mathbb{E}B. \tag{6.6}$$

In the stationary setting the weighted-linear objective function equals $\varphi^{(\ell)}[x] = \omega\,\mathbb{E}I + (1-\omega)\mathbb{E}W$, which is now a function of just the (constant) interarrival time x. The goal is to find the minimizer $\bar{x}$. It is easily seen that such a minimizer uniquely exists (and is larger than $\mathbb{E}B$), due to the fact that the objective function is convex. To this end, observe that $\mathbb{E}I$ is linear in x, whereas the proof of $\mathbb{E}W$ being convex in x can be done in essentially the same way as in the case of a finite number of patients.

The stationary version of the weighted-quadratic objective function reads $\varphi^{(\mathrm{q})}[x] = \omega\,\mathbb{E}I^2 + (1-\omega)\,\mathbb{E}W^2$. The 'mixed' stationary objective functions $\varphi^{(\ell\mathrm{q})}$ and $\varphi^{(\mathrm{q}\ell)}$ are defined in a self-evident manner.

6.2.4 Phase-type fit

Unfortunately, no analytical procedures are available to determine, for generally distributed service times, the distributions of the idle times I_i and the waiting times W_i. We remedy this by replacing the actual service times by their so-called phase-type counterparts. The rationale behind this approach is the well-known fact that phase-type distributions are capable of approximating any positive distribution with arbitrary accuracy; see e.g. Asmussen et al. (1996). The resulting queueing system allows (semi-)explicit computation of the objective function, as pointed out in Chapter 2. In a discrete-time setting, an entirely different approach that facilitates fast computations has been presented by De Vuyst et al. (2014).

We have chosen to characterize the service-time distributions by fitting a phase-type distribution with the correct first two moments; the values of these moments can be chosen in line with for instance the findings of Çayırlı and Veral (2003). This choice is motivated by the fact that it is cumbersome to estimate higher moments, where it is in addition anticipated that those higher moments have just a modest impact on the performance of an appointment schedule (a claim that we later corroborate in Section 6.3.4).

In line with the literature on scheduling, we represent the first two moments by (i) the mean, and (ii) the *squared coefficient of variation* (ϱ), a unitless quantity that is defined as the ratio of the variance and the square of the mean. We follow the standard procedure, advocated in e.g. Tijms (1986), to approximate the service time by a mixture of two Erlang random variables if it has a ϱ smaller than 1, and by a hyperexponential random variable if it has a ϱ larger than 1. In more detail, the approximation is constructed as follows.

- In case ϱ is smaller than 1 the service-time distribution is approximated by a mixture of Erlang distributions: it is an Erlang distribution with $K-1$ phases and mean $(K-1)/\mu$ with probability $p \in [0,1)$, and an Erlang distribution with K phases and mean K/μ with probability $1-p$. It can be verified that the ϱ of this distribution lies in the interval $(1/K, 1/(K-1)]$, for $K \in \{2,3,\ldots\}$. As a result, we can identify unique K, μ, and p such that our mixture of Erlangs has the desired mean and ϱ.

- In the other situation, in which ϱ is larger than 1, the service time is approximated by a hyperexponential random variable, which is constructed as an exponential random variable with mean μ_1^{-1} with probability p, and an exponential random variable with mean μ_2^{-1} with probability $1-p$. By imposing *balanced means*: $\mu_1 = 2p\mu$ and $\mu_2 = 2(1-p)\mu$ for some $\mu > 0$ one reduces the number of free parameters from three to two, so that for each mean and ϱ a unique hyperexponential distribution can be determined.

In Chapter 2 it is explained how to evaluate the objective functions in the case that the service times are of phase-type. As mentioned earlier, we comment in Section 6.3.4

on the error due to approximating the service times by their phase-type counterpart.

6.3 Stationary schedules

In the model we introduced in Section 6.2, we assume the number of patients n to be given. If n is relatively large and the service times are independent and identically distributed, we anticipate that in an optimal schedule the interarrival times of patients scheduled in the middle of a session are about equal, say $\bar{x}$. In this situation, the optimal schedule could be approximated by a schedule in which *all* interarrival times are set to $\bar{x}$, the so-called *stationary schedule*. The main objective of this section is to devise an efficient procedure to identify $\bar{x}$ when the service times stem from the phase-type distributions introduced in Section 6.2.

In our approach we renormalize time so that the mean service time equals 1. Thus, the only parameters the optimal interarrival time $\bar{x}$ depends on are (i) the weight ω and (ii) the ϱ of the service-time distribution. The main conclusion of the present section is the empirical finding that surprisingly simple functional forms can be used. In particular, we advocate the use of interarrival times of the form

$$\bar{x} \equiv \bar{x}(\omega, \varrho) = 1 + A(\omega)\varrho^{B(\omega)}, \tag{6.7}$$

for functions $A(\cdot)$ and $B(\cdot)$ that we can accurately determine.

We provide three (complementary) approaches to determine $A(\cdot)$ and $B(\cdot)$. The first of these is numerical, the second analytical, and the third based on a heavy-traffic approximation.

6.3.1 Numerical determination of stationary schedules

Our first approach is to empirically identify the functions $A(\cdot)$ and $B(\cdot)$ featuring in (6.7). To this end, we have implemented the following least-squares approach, for each $\omega \in \Omega := \{0.1, \ldots, 0.9\}$.

- For $\varrho \in \mathscr{R} = \{0.2, 0.3, \ldots, 3.0\}$ we numerically determine $\bar{x}(\omega, \varrho)$, as follows. For any given value of x, we can evaluate the objective function (requiring the computation of first and/or second moments of the idle and waiting times) relying on the machinery for phase-type service times presented in Chapter 2 using the computational approach presented in Section 4.6.

- We observe from (6.7) that $\log(\bar{x}(\omega, \varrho) - 1) = \log A(\omega) + B(\omega) \log \varrho$. This means that we can use the method of least squares to determine, for any ω, $A(\omega)$ and $B(\omega)$, based on the 29 data points determined in the first step.

The resulting curves, say $A_{\text{num}}(\cdot)$ and $B_{\text{num}}(\cdot)$, are depicted in Figs. 6.2–6.3. The fit is excellent: in the case of a linear objective function the coefficient of determination

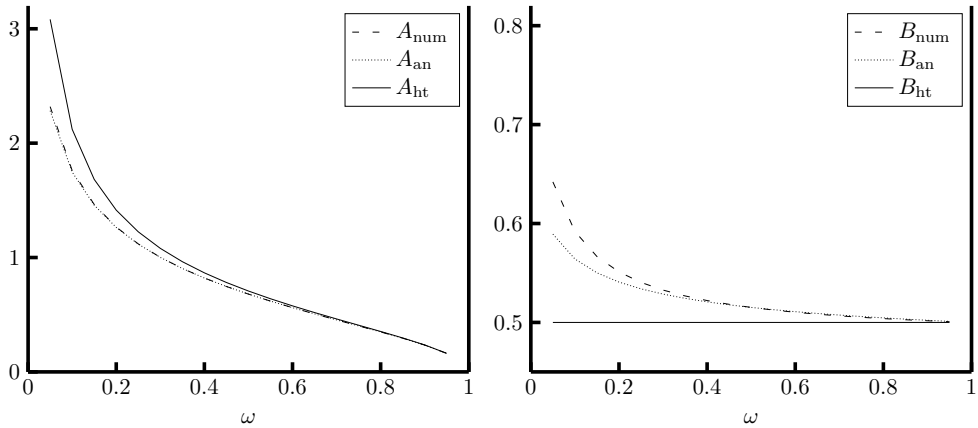


Figure 6.2: The curves $A(\omega)$ (left panel) and $B(\omega)$ (right panel) for the weighted-linear objective function. The dashed curves use the numerical approach described in Section 6.3.1, the dotted curves the analytical approach in Section 6.3.2, and the solid curves correspond to the heavy-traffic approach in Section 6.3.3.

is at least $R^2 = 0.9998$ for any $\omega \in \Omega$, and for a quadratic objective function at least $R^2 = 0.9988$. The mixed linear-quadratic objective functions give similar results.

6.3.2 Analytical derivation of stationary schedules

In this section we determine approximative analytical expressions for $A_{\text{an}}(\cdot)$ and $B_{\text{an}}(\cdot)$, based on results for the stationary distributions of the D/M/1 and D/E$_2$/1 systems. More specifically, $A_{\text{an}}(\cdot)$ is determined using D/M/1 results, and $B_{\text{an}}(\cdot)$ using the expression for $A_{\text{an}}(\cdot)$ in combination with D/E$_2$/1 results.

Recall that the case $\varrho = 1$ corresponds to a system of the type D/M/1, for which one can explicitly derive various quantities pertaining to the stationary queue. In particular, we have (near-)closed-form expressions for the distributions of the stationary waiting times and idle times for a given interarrival time x. This allows us to determine, for any weight ω, the optimal interarrival time $\bar{x}(\omega, 1)$, and we thus find the function $A_{\text{an}}(\cdot)$ from

$$\bar{x}(\omega, 1) = 1 + A_{\text{an}}(\omega) 1^{B_{\text{an}}(\omega)} = 1 + A_{\text{an}}(\omega),$$

and hence $A_{\text{an}}(\omega) = \bar{x}(\omega, 1) - 1$.

We now explain how to evaluate $\bar{x}(\omega, 1)$. We start by identifying the interarrival time $\bar{x}^{(\ell)}(\omega, 1)$ corresponding to the case of the weighted-linear objective function. As pointed out in Kemper et al. (2014), the distribution of the stationary waiting-time $W \equiv W(x)$ is given by

$$\mathbb{P}(W > y) = \sigma_x e^{-(1-\sigma_x)y},$$

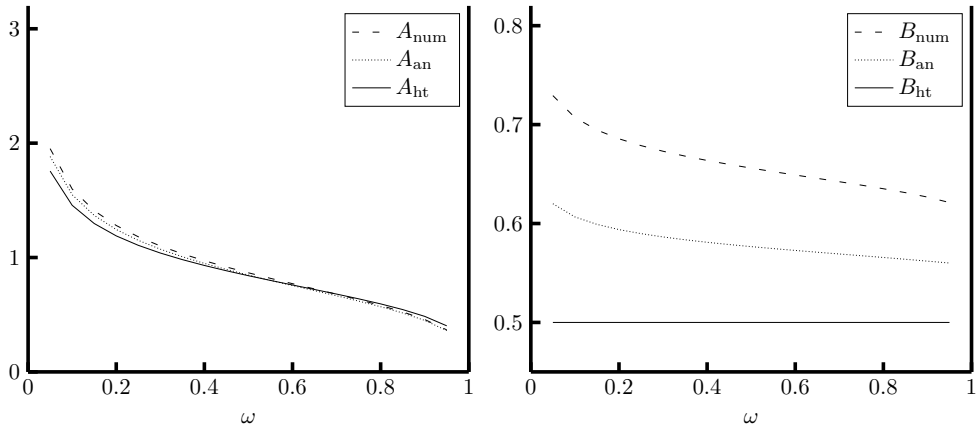


Figure 6.3: The curves $A(\omega)$ (left panel) and $B(\omega)$ (right panel) for the weighted-quadratic objective function. The dashed curves use the numerical approach described in Section 6.3.1, the dotted curves the analytical approach in Section 6.3.2, and the solid curves correspond to the heavy-traffic approach in Section 6.3.3.

where σ_x is the unique solution in $[0,1]$ of $e^{-(1-\sigma)x} = \sigma$ or, equivalently, $x \equiv x_\sigma \in [1,\infty)$ satisfies $x = -\log\sigma/(1-\sigma)$. It thus follows that $\mathbb{E}W = \sigma_x/(1-\sigma_x)$. Also, the distribution of the sojourn time S follows from

$$\mathbb{P}(S > y) = \mathbb{P}(W + B > y) = \int_0^y f_B(z)\mathbb{P}(W > y - z)\mathrm{d}z + \mathbb{P}(B > y) = e^{-(1-\sigma_x)y},$$

with $f_B(z) = e^{-z}$ denoting the service-time density. In other words, S has an exponential distribution with mean $1/(1-\sigma_x)$. Recalling that, due to (6.6), $\mathbb{E}I = x - \mathbb{E}B = x - 1$, the following objective function needs to be minimized with respect to $x \geqslant 1$:

$$\varphi^{(\ell)}[x] = \omega\,\mathbb{E}I + (1-\omega)\mathbb{E}W = \omega(x-1) + \frac{(1-\omega)\sigma_x}{1-\sigma_x} = \frac{1-\omega}{1-\sigma_x} + \omega x - 1,$$

and $\bar{x}^{(\ell)}(\omega, 1)$ thus solves

$$(1-\omega)\frac{\sigma_x'}{(1-\sigma_x)^2} + \omega = 0. \tag{6.8}$$

From the definition of σ_x, it directly follows that

$$\sigma_x' = \frac{1-\sigma_x}{\sigma_x x - 1}\sigma_x = \frac{(1-\sigma_x)^2\sigma_x}{\sigma_x - 1 - \sigma_x\log\sigma_x}.$$

From (6.8), after straightforward algebra, we find that

$$\bar{x}^{(\ell)}(\omega, 1) = -\log \sigma^{(\ell)}(\omega)/(1 - \sigma^{(\ell)}(\omega)),$$

where $\sigma^{(\ell)}(\omega)$ is the unique solution in $[0, 1]$ of

$$\log \sigma + \frac{1}{\sigma} = \frac{1}{\omega}.$$

Let $\mathscr{W}(\cdot) : [e^{-1}, \infty) \mapsto [-1, \infty)$ denote one of the two real branches of the Lambert W-function, i.e., $\mathscr{W}(x)$ is the largest solution for z in the equation $ze^z = x$; see e.g. Corless et al. (1996). It follows that $\sigma^{(\ell)}(\omega) \in [0, 1]$ can be written as

$$-\frac{1}{\sigma^{(\ell)}(\omega)} = \mathscr{W}\left(-e^{-1/\omega}\right), \quad \text{or} \quad \sigma^{(\ell)}(\omega) = -\frac{1}{\mathscr{W}\left(-e^{-1/\omega}\right)}.$$

We eventually obtain

$$A_{\text{an}}^{(\ell)}(\omega) = \bar{x}^{(\ell)}(\omega, 1) - 1 = -\frac{\log \sigma^{(\ell)}(\omega)}{1 - \sigma^{(\ell)}(\omega)} - 1.$$

A similar procedure can be followed for the other objective functions we consider. For completeness, we also show how $A_{\text{an}}(\omega)$ can be found in the quadratic case; the mixed linear-quadratic objective functions can be handled analogously. First observe that $\mathbb{E}W^2 = 2\sigma_x/(1 - \sigma_x)^2$. Also,

$$\mathbb{E}I^2 = \mathbb{E}(S - x)^2 - \mathbb{E}W^2 = \mathbb{E}(W + B - x)^2 - \mathbb{E}W^2, \tag{6.9}$$

with W and B in the first term of the rightmost expression being independent. It is an elementary exercise to verify that

$$\mathbb{E}(S - x)^2 = \frac{2}{(1 - \sigma - x)^2} - \frac{2x}{1 - \sigma_x} + x^2,$$

so that we eventually obtain

$$\mathbb{E}I^2 = x^2 - 2\left(\frac{x - 1}{1 - \sigma_x}\right).$$

We are to minimize, over $x \geqslant 1$,

$$\varphi^{(\text{q})}[x] = \omega\left(x^2 - 2\left(\frac{x - 1}{1 - \sigma_x}\right)\right) + (1 - \omega)\frac{2\sigma_x}{(1 - \sigma_x)^2},$$

or, equivalently over $\sigma \in [0,1]$ the function

$$\omega \left(\frac{(\log \sigma)^2 + 2\log \sigma + 2(1-\sigma)}{(1-\sigma)^2} \right) + (1-\omega)\frac{2\sigma}{(1-\sigma)^2}. \tag{6.10}$$

With $\sigma^{(\mathrm{q})}(\omega)$ the $\sigma \in [0,1]$ that minimizes (6.10), it thus follows that

$$A_{\mathrm{an}}^{(\mathrm{q})}(\omega) = \bar{x}^{(\mathrm{q})}(\omega, 1) - 1 = -\frac{\log \sigma^{(\mathrm{q})}(\omega)}{1 - \sigma^{(\mathrm{q})}(\omega)} - 1.$$

We are thus left with determining the function $B_{\mathrm{an}}(\cdot)$. Note that the case $\varrho = \frac{1}{2}$ corresponds to the D/E$_2$/1 queue for which analytic expressions are available too (see, e.g., Goddard 1963, p. 109). Once we have identified $\bar{x}(\omega, \frac{1}{2})$, we can find $B_{\mathrm{an}}(\omega)$ by solving

$$1 + A_{\mathrm{an}}(\omega) \left(\frac{1}{2}\right)^{B_{\mathrm{an}}(\omega)} = \bar{x}\left(\omega, \frac{1}{2}\right).$$

Using the $A_{\mathrm{an}}(\omega)$ derived above, it now follows that

$$B_{\mathrm{an}}(\omega) = \frac{\log A_{\mathrm{an}}(\omega) - \log(\bar{x}(\omega, \frac{1}{2}) - 1)}{\log 2}.$$

Since other cases are analogous to that of a weighted-linear objective function, we only demonstrate how $\bar{x}^{(\ell)}(\omega, \frac{1}{2})$ is determined. To this end, for a given $x > 1$, we consider the equation

$$f(\tau) := \left(1 - \frac{\tau}{2}\right)^2 = e^{-\tau x} =: g(\tau).$$

Essentially from (i) $f(\infty) = \infty$ and $g(\infty) = 0$, (ii) $f(2) = 0$ and $g(2) > 0$, (iii) $f(0) = g(0) = 1$, and (iv) $f'(0) = -1 > -x = g'(0)$, it follows that the above equation has two positive roots, one of which lies between 0 and 2 while the other is larger than 2. Call these roots $\tau_1 \equiv \tau_{1,x}$ and $\tau_2 \equiv \tau_{2,x}$. Then $\mathbb{P}(W > y) = c_1 e^{-\tau_1 y} + c_2 e^{-\tau_2 y}$, where $c_1 \equiv c_{1,x}$ and $c_2 \equiv c_{2,x}$ solve

$$\frac{c_1}{1 - \tau_1/2} + \frac{c_2}{1 - \tau_2/2} = 1 \quad \text{and} \quad \frac{c_1}{(1-\tau_1/2)^2} + \frac{c_2}{(1-\tau_2/2)^2} = 1,$$

i.e.,

$$c_1 = \frac{\tau_2}{\tau_2 - \tau_1}\left(1 - \frac{\tau_1}{2}\right)^2 = \frac{\tau_2 e^{-\tau_1/x}}{\tau_2 - \tau_1} \quad \text{and} \quad c_2 = \frac{\tau_1}{\tau_1 - \tau_2}\left(1 - \frac{\tau_2}{2}\right)^2 = \frac{\tau_1 e^{-\tau_2/x}}{\tau_1 - \tau_2}.$$

Realizing that again $\mathbb{E}I = x - 1$, we are to minimize

$$\varphi^{(\ell)}[x] = \omega\, \mathbb{E}I + (1-\omega)\mathbb{E}W = \omega(x-1) + \frac{(1-\omega)c_{1,x}}{\tau_{1,x}} + \frac{(1-\omega)c_{2,x}}{\tau_{2,x}}$$

in order to obtain $\bar{x}^{(\ell)}(\omega, \frac{1}{2})$. (Semi-)closed-form expressions cannot be derived now, but a straightforward numerical routine performs the minimization.

The resulting curves $A_{\mathrm{an}}(\cdot)$ and $B_{\mathrm{an}}(\cdot)$ are depicted in Figs. 6.2–6.3. Regarding the $A_{\mathrm{an}}(\cdot)$ curve, the fit is still remarkably good; it is also seen that the curve is at most 1% off the curve determined by the approach of Section 6.3.1. The $B_{\mathrm{an}}(\cdot)$ curve in the linear case is still rather precise, in the quadratic case the performance is slightly worse (but due to the fine scales chosen in the picture, the difference may seem more substantial than it actually is). The overall performance remains rather good, as reflected in the R^2, which measures the discrepancy between the $\bar{x}(\omega, \frac{1}{2})$ as predicted by the analytical approach presented in this subsection and the corresponding true values. In the case of a weighted-linear objective function, for all $\omega \in \Omega$ the R^2 is at least 0.9914. For the weighted-quadratic objective function the fit somewhat degrades: the R^2 ranges from 0.91 for $\omega = 0.1$ to 0.96 for $\omega = 0.9$.

6.3.3 Heavy-traffic derivation of stationary schedules

In this subsection we provide a theoretical justification for the use of schedules based on the form (6.7). More specifically, we show that this relation is exact in the heavy-traffic regime, i.e., the regime in which x is just slightly larger than the mean service time, which we normalized to 1.

It is well-known that $(x-1)W(x)$ for $x \downarrow 1$ converges to an exponential random variable with mean $\frac{1}{2}\,\mathbb{V}\mathrm{ar}\,B$ (Asmussen 2003, Section X.7). As a consequence, one can approximate $W(x)$ by an exponential distribution with mean μ_x^{-1}, with $\mu_x = 2(x-1)/\varrho$. Observe that due to our normalization $\mathbb{E}B = 1$ we have $\mathbb{V}\mathrm{ar}\,B = \varrho$. It is anticipated that for ω close to 1, the optimal interarrival times are relatively short as the system is relatively indifferent with respect to waiting times, and therefore one could expect that in this regime heavy-traffic approximations lead to accurate predictions.

Thus, the weighted-linear objective function reads

$$\varphi_{\mathrm{ht}}^{(\ell)}[x] = \omega(x-1) + (1-\omega)\frac{\varrho}{2(x-1)}.$$

The corresponding minimization allows an explicit solution:

$$\bar{x}_{\mathrm{ht}}^{(\ell)}(\omega, \varrho) = 1 + A_{\mathrm{ht}}^{(\ell)}(\omega)\varrho^{B_{\mathrm{ht}}^{(\ell)}(\omega)}, \quad \text{with } A_{\mathrm{ht}}^{(\ell)}(\omega) = \sqrt{\frac{1-\omega}{2\omega}}, \quad B_{\mathrm{ht}}^{(\ell)}(\omega) = \frac{1}{2}.$$

The weighted-quadratic objective function requires more care. Recall (6.9); the objective function equals

$$\omega\,\mathbb{E}I^2 + (1-\omega)\mathbb{E}W^2 = \omega\,\mathbb{E}(W+B-x)^2 + (1-2\omega)\mathbb{E}W^2.$$

Using standard properties of the exponential distribution, we have the heavy-traffic

approximation

$$\mathbb{E}W^2 = \int_0^\infty \mu_x e^{-\mu_x z}\, z^2 \mathrm{d}z = \frac{2}{\mu_x^2} = \frac{\varrho^2}{2(x-1)^2}.$$

Likewise, with $f_B(\cdot)$ the density of B,

$$\mathbb{E}(W+B-x)^2 = \int_0^\infty \int_0^\infty f_B(y)\, \mu_x e^{-\mu_x z}(y+z-x)^2 \mathrm{d}z\, \mathrm{d}y,$$

which by an elementary computation turns out to equal $\mathbb{E}W^2 + (x-1)^2$. We thus obtain the objective function

$$\varphi_{\mathrm{ht}}^{(\mathrm{q})}[x] = \omega(x-1)^2 + (1-\omega)\frac{\varrho^2}{2(x-1)^2},$$

being minimized by

$$\bar{x}_{\mathrm{ht}}^{(\mathrm{q})}(\omega,\varrho) = 1 + A_{\mathrm{ht}}^{(\mathrm{q})}(\omega)\varrho^{B_{\mathrm{ht}}^{(\mathrm{q})}(\omega)}, \quad \text{with } A_{\mathrm{ht}}^{(\mathrm{q})}(\omega) = \sqrt[4]{\frac{1-\omega}{2\omega}}, \quad B_{\mathrm{ht}}^{(\mathrm{q})}(\omega) = \frac{1}{2}.$$

Observe that in both the weighted-linear case and the weighted-quadratic case, we find that in heavy traffic the optimal interarrival times have the shape $1 + A(\omega)\sqrt{\varrho}$, but interestingly, for the mixed objective functions we obtain different structures:

$$\bar{x}_{\mathrm{ht}}^{(\ell\mathrm{q})}(\omega,\varrho) = 1 + \sqrt[3]{\frac{1-\omega}{\omega}}\varrho^{2/3} \quad \text{and} \quad \bar{x}_{\mathrm{ht}}^{(\mathrm{q}\ell)}(\omega,\varrho) = 1 + \sqrt[3]{\frac{1-\omega}{4\omega}}\varrho^{1/3}.$$

The resulting curves for the weighted-linear case and the weighted-quadratic cases, say $A_{\mathrm{ht}}(\cdot)$ and $B_{\mathrm{ht}}(\cdot)$, can be found in Figs. 6.2–6.3. The overall fit is remarkably good, and for ω approaching 1 even excellent.

6.3.4 Impact of phase-type approximation

In this subsection we assess the impact of replacing the service-time distribution by its phase-type counterpart as we proposed in Section 6.2.4. We present experiments with Weibull and lognormal service times in line with earlier healthcare-related studies (see e.g. Çayırlı and Veral 2003). Importantly, the Weibull and lognormal distributions do *not* belong to the class of phase-type distributions. Both distributions are characterized by two parameters, which are chosen such that the mean equals 1, whereas ϱ is varied in the experiment. It is stressed that there are no explicit results for D/G/1 systems with Weibull or lognormal service times, and therefore we have chosen to determine the optimal interarrival time by simulation. We show that the resulting stationary schedule virtually coincides with the one obtained using the phase-type service times with the same first two moments.

More specifically, we have used the following procedure. In each simulation run

we sampled the service times $B_1, \ldots, B_M$ of $M = 100\,000$ patients. We used this (single) run to estimate the distribution of the waiting time $W(x)$ (for all $x \geqslant 1$ simultaneously), by simulating for x on a fine grid the queue with interarrival times x and service times $B_1, \ldots, B_M$ and by interpolation for x between neighboring grid points. The estimates of the waiting-time distributions are further improved by repeating this experiment $1\,000$ times. Then we minimize the objective function, which is the weighted-linear objective function $\varphi^{(\ell)}$ in the output presented below, but the other objective functions give comparable results.

The experiments have revealed that there is almost a perfect fit, in terms of the maximum difference between the optimal interarrival time corresponding to the actual service-time distribution and the one based on the phase-time fit, when varying ϱ between 0.1 and 1.5. For Weibull service times this maximum difference is 2.7% for $\omega = 0.1$, but this sharply drops when ω increases, and is just 0.067% for $\omega = 0.9$. For lognormal service times we see the same pattern, with the maximum difference decreasing from 3.0% for $\omega = 0.1$ to 0.093% for $\omega = 0.9$.

6.3.5 Phase-type approximation and extreme distributions

In this subsection we present some experiments that have been inspired by a recent paper by Mak et al. (2015) that studies the setup in which there is *limited information* on the shape of the service-time distribution available. A specific example of this concerns the case in which only the mean $\mathbb{E}B$ and the variance $\mathbb{V}\mathrm{ar}\, B$ are given (for instance due to the fact that there are only few historical data available). The idea is then that, for a given value of the interarrival time x one finds the distribution B on $[0, \infty)$ with mean 1 (still on the normalized time scale used throughout this section) and second moment $\varrho + 1$ (such that $\mathbb{V}\mathrm{ar}\, B = \varrho$) that *maximizes* the objective function, so as to identify the *worst-case service-time distribution* for a given value of x. This yields a function, say, $\xi(x)$, which can be minimized over $x \geqslant 1$.

We use the following service-time distributions that are in some sense *extreme* (while still having mean 1 and squared coefficient of variation ϱ): B equals $\varrho + 1$ with probability $1/(\varrho + 1)$ and 0 otherwise. Our objective here is to investigate the impact of this choice of B on the resulting schedule.

With $Z_n := \sum_{i=1}^{n} B_i$, observe that $Z_n/(\varrho + 1)$ is binomially distributed with parameters n and $1/(\varrho + 1)$. It is well known (Asmussen 2003, Prop. VIII.4.5) that

$$\mathbb{E}W = \sum_{n=1}^{\infty} \frac{1}{n} \mathbb{E}(Z_n - nx)^+ = \sum_{n=1}^{\infty} \frac{1}{n} \sum_{\ell = \lceil nx/(\varrho+1) \rceil}^{n} (\ell(\varrho + 1) - nx)\, \mathbb{P}\left(\frac{Z_n}{\varrho + 1} = \ell \right);$$

here x^+ is defined as $\max\{0, x\}$. The weighted-linear objective function therefore

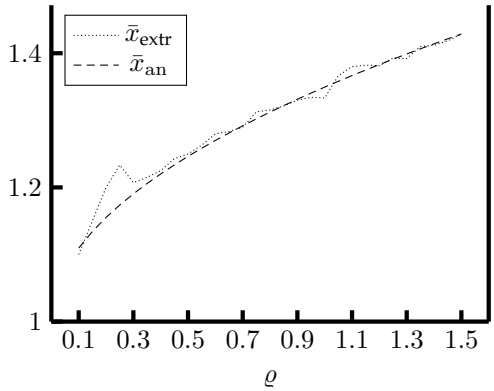


Figure 6.4: Values for $\bar{x}_{\text{an}}$ and $\bar{x}_{\text{extr}}$ as a function of ϱ, $\omega = 0.8$. The dashed curve corresponds to the phase-type service times, and the dotted to the values obtained using the extreme distribution.

reads

$$\varphi^{(\ell)}[x] = \omega(x-1)+(1-\omega)\sum_{n=1}^{\infty}\frac{1}{n}\sum_{\ell=\lceil nx/(\varrho+1)\rceil}^{n}(\ell(\varrho+1)-nx)\binom{n}{\ell}\left(\frac{1}{\varrho+1}\right)^{\ell}\left(\frac{\varrho}{\varrho+1}\right)^{n-\ell},$$

which can be rewritten as

$$\begin{aligned}\omega(x-1) \quad &+ \quad (1-\omega)\sum_{n=1}^{\infty}\sum_{\ell=\lceil nx/(\varrho+1)\rceil}^{n}\binom{n-1}{\ell-1}\left(\frac{1}{\varrho+1}\right)^{\ell-1}\left(\frac{\varrho}{\varrho+1}\right)^{n-\ell}\\ &- \quad x(1-\omega)\sum_{n=1}^{\infty}\sum_{\ell=\lceil nx/(\varrho+1)\rceil}^{n}\binom{n}{\ell}\left(\frac{1}{\varrho+1}\right)^{\ell}\left(\frac{\varrho}{\varrho+1}\right)^{n-\ell}.\end{aligned}$$

As an example we show in Fig. 6.4 for the weight $\omega = 0.8$, that the optimal interarrival times obtained using the extreme distribution are close to what would be obtained when using its phase-type counterpart. While for decreasing ω the fit (slightly) degrades, additional numerical experiments show that even in this extreme case the phase-type fit is typically not far off.

6.3.6 Example

We end this section by an example that demonstrates how our findings can be used in a practical situation. Suppose we use the weighted-linear objective function, and we consider a situation with parameter values suggested in Çayırlı and Veral (2003): $\varrho = 0.5$ and $\omega = 0.8$. We assume that the mean service time is 10 minutes.

The numerical method pointed out in Section 6.3.1 can be applied as follows.

From Fig. 6.2, we see that $A_{\mathrm{num}}(0.8) = 0.349$ and $B_{\mathrm{num}}(0.8) = 0.504$, leading to $\bar{x}^{(\ell)}_{\mathrm{num}}(0.8, 0.5) \approx 1 + 0.349 \cdot 0.8^{0.504} = 1.312$, and hence the optimal interarrival time is $10 \cdot 1.312 = 13.1$ minutes. The analytical method of Section 6.3.2 gives the same value for $\bar{x}_{\mathrm{an}}$ (up to three digits). The approach based on a heavy-traffic regime yields nearly the same result: $\bar{x}^{(\ell)}_{\mathrm{ht}}(0.8, 0.5) \approx 1 + \sqrt{1/8} \cdot 0.8^{0.5} = 1.316$.

6.4 Transient schedules

In the previous section we considered situations in which the number of patients is relatively large, so that a stationary schedule is justified. We depart from this setting now: in this section we analyze a setup with a relatively small number of patients (say $n \leq 35$). In principle all that needs to be done is to evaluate the preferred objective function (i.e., $\mathscr{F}^{(\ell)}$, $\mathscr{F}^{(\mathrm{q})}$, $\mathscr{F}^{(\ell\mathrm{q})}$, or $\mathscr{F}^{(\mathrm{q}\ell)}$, with a certain weight ω), and to minimize this function over the interarrival times $x_1, \ldots, x_{n-1}$.

Now that we have approximated the service-time distribution by means of the phase-type fit advocated in Section 6.2.4, we can exploit its structure in an iterative procedure to compute each patient's individual sojourn-time distribution, as proposed by Wang (1997). We outline the principles of such procedure for the case of i.i.d. service times in Chapter 2. This method can be seen as an extension to the work of Hassin and Mendel (2008), which only covers the case of exponentially distributed service times (i.e., $\varrho = 1$). Using this methodology we can, for each sequence of $\boldsymbol{x} = (x_1, \ldots, x_{n-1})$, compute the patients' sojourn times, which are used to compute the aggregate objective function $\mathscr{F}^{(\ell)}[\boldsymbol{x}]$ as given in Eqn. (6.3). This objective function is minimized over the interarrival times $\boldsymbol{x}$, which is an $(n-1)$–dimensional optimization. Such minimizations can be performed using standard software (such as Matlab), but are complicated by potentially long computation times (up to 15 minutes for a single problem instance). The numerical minimization is typically slow when

- n is larger than, say, 10, in which case the vector over which the optimization is performed is of (relatively) high dimensionality;
- ϱ is relatively small. Consider for instance the case that $\varrho \in (0.2, 0.25]$ and $n = 25$. Then potentially five exponential phases enter the system with each arrival, such that the state space of the sojourn time of the n-th patient is 125-dimensional.

The impact of the two complications above can be mitigated somewhat by an appropriate truncation of the state-space. For instance, it is implausible that the 25-th patient sees patients who arrived during the first part of the schedule; returning to the example of $\varrho \in (0.2, 0.25]$, the dimensionality of the phase-type distribution describing the sojourn time of this patient can be chosen substantially lower than 125 without numerical impact. It is not straightforward, however, to determine in advance

how far the dimensionality can safely be truncated: for instance when the weight ω is close to 1, waiting times are hardly penalized, and as a consequence the load in the optimizing schedule is relatively high, and we cannot truncate the dimensionality much. To overcome such complications we have developed a tool (which is available at `http://www.appointmentscheduling.info`) that generates schedules instantaneously.

The tool exploits a set of precalculated schedules for a grid of values of ϱ, ω, and n, and for each of the four objective functions, and uses interpolation techniques to find a suitable schedule from there. Extensive validation revealed that the error is negligible (well below 1%, and typically on the order of 0.05% or less). The approach offers various additional options, which are discussed in Sections 6.4.1–6.4.4. On the application of the webtool we refer to Chapter 7.

The curve of successive interarrival times in optimal schedules typically has a *dome* (concave) shape, as was observed earlier in Wang (1997), Robinson and Chen (2003), Kaandorp and Koole (2007) and Hassin and Mendel (2008), with shorter interarrival times at the beginning and end of the schedule. This shape may be understood intuitively by considering that early in the schedule, variability in the patients' completion times is limited, which allows relatively short interarrival times. At the end of the schedule, an overrun of an appointment affects only few patients, which also warrants shorter interarrival times.

In addition, we found in all our experiments that the curve is bounded from above by the optimal interarrival time in the stationary schedule. For given n, the entire curve increases in ϱ and decreases in ω. We have included a number of illustrative examples in Fig. 6.5.

6.4.1 Operation of the webtool

We explain the use of the webtool for transient schedules in practice. First, the user enters appropriate values for $\mathbb{E}B$ and ϱ, preferably based on the estimated first two moments of measured service times. Typical values for ϱ are between 0.1225 and 0.7225 (Çayırlı and Veral 2003). Produced schedules should be scaled back from normalized time ($\mathbb{E}B = 1$) to real time by multiplying them by $\mathbb{E}B$.

Second, an objective function should be selected. In the appointment scheduling literature one usually chooses the linear cost function (for both idle times and waiting times). We offer the extra option of quadratic objective functions (or mixed linear/quadratic). For instance, $\mathscr{F}^{(\ell\mathrm{q})}$ (and its steady-state counterpart $\varphi^{(\ell\mathrm{q})}$) can be used in order to penalize excessive waiting times more than excessive idle times.

Finally, the user enters two out of the triple n (the number of patients to be scheduled), ω (the desired weight) and T (the targeted session end time). As explained before, ω is ultimately a strategic choice reflecting the hospital's value proposition. Robinson and Chen (2011) studied mean queue lengths and utilizations in practice to determine how the trade-off is implicitly made in reality. They find implied values up to $\omega = 0.98$, corresponding to situations where utilization is maximized by

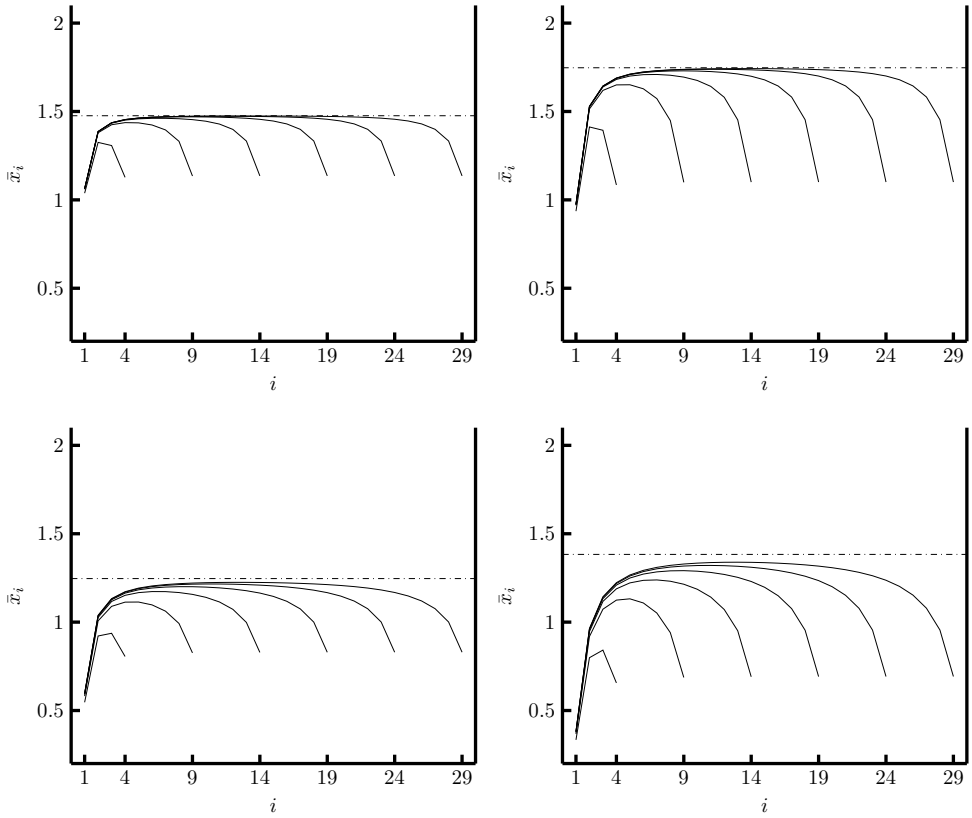


Figure 6.5: Optimal appointment schedules, as a function of the patient number, where in each figure the individual curves correspond to $n = 5, 10, 15, 20, 25,$ and 30, and the horizontal (dash-dotted) line to the stationary schedule. The top panels correspond to $\omega = 0.5$, and the bottom panels to $\omega = 0.8$; the left panels correspond to $\varrho = 0.5$, and the right panels to $\varrho = 1.2$.

allowing very long waiting times for patients. Shorter waiting times are obtained by choosing lower values for ω.

Depending on which two parameters are entered, the webtool produces the following results:

- If n and ω are provided, the webtool generates the resulting optimal schedule as well as its expected makespan T.
- Entering n and T (where, evidently, $T > n\,\mathbb{E}B$), the tool determines the implied value of ω and returns the optimal schedule that has expected makespan T.
- The third option is to select T and ω to find out how many patients can optimally be scheduled such that the expected makespan remains below T, given the weight ω.

6.4.2 No-shows and walk-ins

We now point out how no-shows and walk-ins can be dealt with. First consider the situations of no-shows only. We assume a per patient probability $q \in [0, 1)$ of a no-show. We apply the developed machinery, but with adapted service times. For the situation of i.i.d. unit-mean service times, the expected service time becomes $1 - q$, and the corresponding squared coefficient of variation

$$\frac{(1-q)\,\mathbb{E}B^2 - (1-q)^2}{(1-q)^2} = \frac{(1-q)\,\varrho + (1-q)q}{(1-q)^2} = \frac{\varrho + q}{1-q}, \tag{6.11}$$

with ϱ, as before, the squared coefficient of variation of B. Fig. 6.6 presents, as an illustrative example, the optimal schedule for various values of q. For q relatively high, it is optimal to apply *overbooking*: the optimal interarrival times are *shorter* than the mean service time 1.

Although ϱ is typically lower than 1 in the healthcare context, the adapted squared coefficient of variation in the situation with no-shows, $(\varrho + q)/(1 - q)$, *can* be larger than 1. This means that also the hyperexponential distribution is important when modeling healthcare applications using the phase-type approach; see Section 6.3.1.

Additionally, we can incorporate the option that an unscheduled patient walks in during a session. Suppose that with probability w an unscheduled patient arrives at the arrival epoch of a scheduled patient. One thus finds that the expected service time is $1 - q + w$, whereas it can be verified that the squared coefficient of variation equals

$$\frac{(1-q+w)\,\mathbb{E}B^2 - (1-q+w)^2}{(1-q+w)^2} = \frac{(1-q+w)\varrho + q(1-q) + w(1-w)}{(1-q+w)^2}.$$

The case of zero walk-ins ($w = 0$) is indeed equivalent with the situation of only no-shows, cf. Eqn. (6.11).

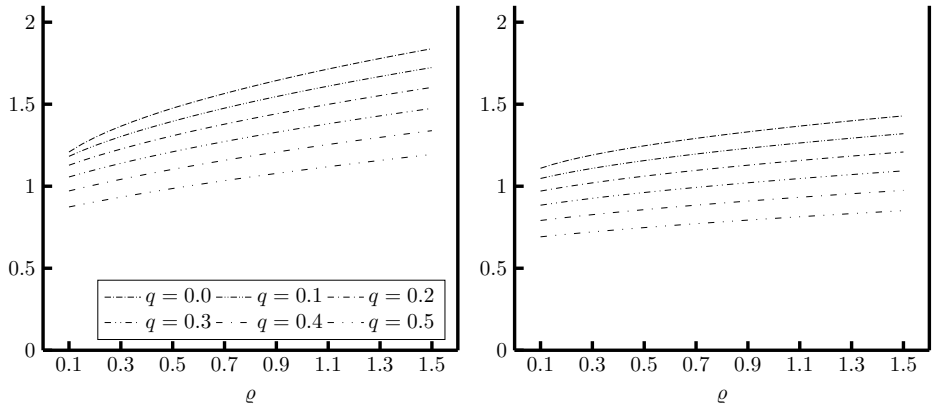


Figure 6.6: Both panels give values for $\bar{x}$ as a function of ϱ, for $q = 0.5$ (bottom), 0.4, 0.3, 0.2, 0.1, and 0 (top). Left panel corresponds to $\omega = 0.5$ and right panel to $\omega = 0.8$.

This adjustment procedure that incorporates no-shows and walk-ins into the service times can be generally applied. For example, the two-server tandem setting of Chapter 4 lends itself to this approach by modifying *each* server's ϱ value.

6.4.3 Overtime

The makespan, or equivalently the session end time as defined in (6.2), measures how long the service provider should be available. In the weighted-linear case, for instance, an objective function that penalizes overtime is

$$\omega \sum_{i=1}^{n} \mathbb{E}I_i + (1-\omega) \sum_{i=1}^{n} \mathbb{E}W_i + \bar{\omega} \left(\sum_{i=1}^{n} \mathbb{E}B_i + \sum_{i=1}^{n} \mathbb{E}I_i \right), \tag{6.12}$$

for some scalar $\bar{\omega} > 0$. Noting that the the expected service times are quantities that are exogenous, in that they are not affected by the choice of the schedule, it is seen that minimizing the above objective function is equivalent to minimizing

$$\frac{\omega + \bar{\omega}}{1 + \bar{\omega}} \sum_{i=1}^{n} \mathbb{E}I_i + \frac{1-\omega}{1+\bar{\omega}} \sum_{i=1}^{n} \mathbb{E}W_i.$$

This objective function is precisely of the form that was introduced in Section 6.2, but with adapted weights. As a consequence, the techniques we developed can be used in the setting that incorporates overtime in the objective function.

$\varrho = 0.5$

$\omega = 0.5$								$\omega = 0.8$						
Patient	Continuous		Discrete		Rounded			Patient	Continuous		Discrete		Rounded	
i	$\bar{x}_i$	$\bar{t}_{i+1}$	x_i	t_{i+1}	x_i	t_{i+1}		i	$\bar{x}_i$	$\bar{t}_{i+1}$	x_i	t_{i+1}	x_i	t_{i+1}
1	15.93	15.93	15	15	15	15		1	8.82	8.82	10	10	10	10
2	20.76	36.69	20	35	20	35		2	15.32	24.14	15	25	15	25
3	21.48	58.17	20	55	25	60		3	16.64	40.79	15	40	15	40
4	21.73	79.90	25	80	20	80		4	17.13	57.91	20	60	20	60
5	21.81	101.71	20	100	20	100		5	17.31	75.22	15	75	15	75
6	21.82	123.54	25	125	25	125		6	17.33	92.55	20	95	20	95
7	21.77	145.31	20	145	20	145		7	17.24	109.78	15	110	15	110
8	21.65	166.96	20	165	20	165		8	17.02	126.81	20	130	15	125
9	21.42	188.38	25	190	25	190		9	16.66	143.46	15	145	20	145
10	20.97	209.35	20	210	20	210		10	16.05	159.51	15	160	15	160
11	19.99	229.34	20	230	20	230		11	14.96	174.47	15	175	15	175
12	17.03	246.37	15	245	15	245		12	12.42	186.89	15	190	10	185
$\mathbb{E}$[Makespan]	268.92		268.51		268.55			$\mathbb{E}$[Makespan]	222.30		223.74		222.42	
Cost	66.57		67.04		67.04			Cost	52.46		52.77		52.79	

Table 6.1: Both tables give values for the interarrival and arrival times, for the optimal continuous schedule, the optimal discrete schedule, and the rounded schedule. Left table corresponds to $\omega = 0.5$ and right table to $\omega = 0.8$. In the tables, ϱ is assumed 0.5 and $\Delta = 5$.

6.4.4 Discrete slots

The appointment schedules developed in this chapter are based on optimization routines *in continuous time*. In practice, however, slot lengths are typically discrete multiples of a resolution Δ (for instance 5 min.). An idea is to round the arrival epochs $t_1, \ldots, t_n$ to multiples of Δ. Such a procedure is computationally considerably more efficient than solving the corresponding integer-programming problem. As illustrated in Table 6.1, rounding typically leads to near-optimal solutions: the scheduled arrival epochs of the rounded schedule and those of the optimal discrete solution differ only for one (for $\omega = 0.5$) or two patients (for $\omega = 0.8$). Moreover, the difference in terms of the objective function is marginal.

The webtool has the option to impose any preferred resolution Δ on the schedule, and provides the optimal schedule in continuous time as well as the schedule in which the arrival epochs are rounded to multiples of Δ.

6.5 Performance evaluation

We compare the performance of our approach to three alternatives: Bailey's rule (2BEG), an optimized version of the individual-block/fixed-interval rule (IBFI$^+$) and the universal rule of Çayırlı et al. (2012) (DOME). Bailey's rule 2BEG (Bailey 1952) schedules two patients at the start of the session, and sets subsequent arrival times at intervals equal to the mean service time (after a correction for no-shows and walk-ins). The main merit of this rule is its simplicity. It does not take variability in service times into account, however, which may limit its performance in general. The optimized individual-block/fixed-interval rule IBFI$^+$ finds the best equidistant

schedule, and thus reduces the appointment scheduling problem to solving a one-dimensional minimization. The evaluation of the objective function is based on our phase-type approximations. This rule comes in a number of variants such as in Bailey (1952) and Hassin and Mendel (2008).

Probably the most refined approach is DOME, the 'universal appointment rule' of Çayırlı et al. (2012). This rule produces dome-shaped schedules, where the arrival times are set at

$$t_i := \max\left\{0, k(i-1) - \sqrt{\varrho\, i} \cdot \frac{n+i}{n-1}\right\}. \tag{6.13}$$

The schedule is adjusted for situational characteristics by the scalar k. It is a value that depends on ϱ, ω, q, w and n, and which optimizes a linear-weighted objective function. The relationship between k and the parameters ϱ, ω, q, w and n is approximated by a nonlinear regression equation, based on extensive simulations, and achieving an R^2 of 95.29%.

Before embarking on a numerical comparison, we discuss the most important differences between the DOME rule and ours.

- *Objective function.* Both approaches have the same set of situational parameters ϱ, ω, q, w and n. DOME optimizes a weighted-linear objective function of expected idle time, waiting time and overtime, and the cost ratio $\bar{\omega}/\omega$ of overtime to regular time is fixed at 1.5. Our approach offers more flexibility by offering a choice of four objective functions, and the cost of overtime can be chosen freely and is incorporated as explained in Section 6.4.3.

- *Optimization approach.* The DOME rule approximates the optimal schedule for a situation by the best fitting schedule adhering to (6.13), which in turn is approximated by the nonlinear regression equation for k. This fitted approximation has an R^2 of 95.29% when service times are lognormal, but its precision is unknown otherwise. Our approach approximates service times by their phase-type counterparts, and returns the corresponding optimal schedule (exactly on the grid points, or by interpolation otherwise).

The numerical performance comparison is based on 27 situations that are claimed to be representative for outpatient clinics (Çayırlı et al. 2012). These 27 situations correspond to the squared coefficient of variation of the service times (ϱ) equalling 0.16, 0.36 and 0.64; no-show probabilities (q) 0.05, 0.2, and 0.4, and walk-in probabilities (w) 0, 0.2, and 0.4. The number of patients to be scheduled (n) equals 10 or 20. The weight parameter $\omega/(1-\omega)$ has 3 levels: 2, 5, and 10 (and hence ω equals $2/3$, $5/6$, and $10/11$). To enable the comparison with DOME we fix the relative cost of overtime at $\bar{\omega}/\omega = 1.5$. As a result, $\omega^\star := \omega/(1+\bar{\omega})$, as used in (6.12), has the values $5/6$, $25/27$, and $26/27$. In addition, in our numerical assessment we assumed the service times to be lognormal, to comply with the setup of Çayırlı et al. (2012).

The first step in the numerical evaluation is that we computed for each of the four approaches the schedule $t_1, \ldots, t_n$. Then we simulated for each of these four

				$n = 10$						$n = 20$					
Environment				%	Rule	%	Rule	%	Rule	%	Rule	%	Rule	%	Rule
#	ϱ	q	w	$\omega^\star = 5/6$		$\omega^\star = 25/27$		$\omega^\star = 25/26$		$\omega^\star = 5/6$		$\omega^\star = 25/27$		$\omega^\star = 25/26$	
1	0.16	0.05	0.00	2.81	IBFI$^+$	5.67	2BEG	7.67	2BEG	7.25	DOME	4.96	2BEG	2.70	2BEG
2	0.16	0.05	0.20	4.03	IBFI$^+$	5.12	2BEG	12.19	2BEG	2.63	DOME	2.84	2BEG	5.68	2BEG
3	0.16	0.05	0.40	3.97	IBFI$^+$	4.33	2BEG	12.10	2BEG	3.83	DOME	2.86	2BEG	4.64	2BEG
4	0.16	0.20	0.00	1.31	DOME	0.65	2BEG	12.51	2BEG	0.87	DOME	0.20	2BEG	5.06	2BEG
5	0.16	0.20	0.20	2.26	DOME	1.86	DOME	5.04	DOME	2.18	DOME	0.13	DOME	0.68	DOME
6	0.16	0.20	0.40	3.06	IBFI$^+$	2.80	DOME	4.93	DOME	1.23	DOME	0.40	DOME	0.63	DOME
7	0.16	0.40	0.00	-0.91	DOME	2.52	DOME	10.88	DOME	4.27	DOME	-0.25	DOME	1.52	DOME
8	0.16	0.40	0.20	0.41	DOME	-0.67	DOME	1.71	DOME	6.21	DOME	0.10	2BEG	-0.18	DOME
9	0.16	0.40	0.40	1.74	DOME	-0.19	DOME	0.26	DOME	2.87	DOME	0.09	2BEG	-0.07	DOME
10	0.36	0.05	0.00	4.34	IBFI$^+$	4.78	2BEG	17.27	2BEG	2.15	DOME	1.31	2BEG	7.11	2BEG
11	0.36	0.05	0.20	4.51	IBFI$^+$	5.28	2BEG	18.20	DOME	1.92	DOME	1.18	2BEG	8.16	2BEG
12	0.36	0.05	0.40	4.22	IBFI$^+$	4.88	2BEG	18.46	2BEG	3.41	DOME	1.32	2BEG	7.42	2BEG
13	0.36	0.20	0.00	2.25	DOME	4.97	2BEG	14.78	DOME	1.30	DOME	0.35	2BEG	6.96	DOME
14	0.36	0.20	0.20	2.41	DOME	3.52	DOME	8.97	DOME	1.36	DOME	0.60	DOME	3.36	DOME
15	0.36	0.20	0.40	3.85	IBFI$^+$	4.70	2BEG	9.56	DOME	1.26	DOME	0.60	2BEG	3.19	DOME
16	0.36	0.40	0.00	-0.31	DOME	2.68	DOME	10.21	DOME	2.91	DOME	-0.29	DOME	2.64	DOME
17	0.36	0.40	0.20	0.56	DOME	0.91	DOME	4.13	DOME	3.02	DOME	0.18	DOME	0.38	DOME
18	0.36	0.40	0.40	2.33	2BEG	2.17	DOME	5.16	DOME	1.30	DOME	0.17	DOME	0.49	DOME
19	0.64	0.05	0.00	4.06	2BEG	8.40	2BEG	20.18	DOME	1.40	DOME	1.37	2BEG	12.32	2BEG
20	0.64	0.05	0.20	4.04	2BEG	8.36	2BEG	17.16	DOME	1.77	DOME	1.19	2BEG	11.27	2BEG
21	0.64	0.05	0.40	4.80	IBFI$^+$	7.44	2BEG	18.11	DOME	3.86	DOME	1.11	2BEG	10.87	2BEG
22	0.64	0.20	0.00	1.78	2BEG	5.66	DOME	13.59	DOME	1.22	DOME	1.36	2BEG	7.85	DOME
23	0.64	0.20	0.20	1.90	2BEG	4.75	DOME	10.16	DOME	1.04	DOME	1.44	2BEG	6.39	DOME
24	0.64	0.20	0.40	2.77	2BEG	6.88	DOME	13.13	DOME	1.70	DOME	0.87	2BEG	6.90	DOME
25	0.64	0.40	0.00	-0.36	2BEG	4.14	DOME	9.92	DOME	1.99	DOME	0.18	DOME	3.93	DOME
26	0.64	0.40	0.20	0.07	2BEG	3.44	DOME	7.73	DOME	1.27	DOME	0.21	DOME	2.68	DOME
27	0.64	0.40	0.40	1.00	2BEG	5.35	DOME	9.26	DOME	0.96	DOME	0.82	DOME	2.98	DOME
Average				2.33		4.09		10.86		2.41		0.94		4.65	

Table 6.2: The table presents the performance of our tool relative to three existing methods (2BEG, IBFI$^+$, DOME). The percentages are the gain of our method compared to the best competitor (which is identified in the columns headed 'Rule').

schedules 10 000 clinical sessions, and evaluated the weighted-linear cost function (6.12). Table 6.2 presents the performance of our tool relative to 2BEG, IBFI$^+$ and DOME. The percentages are the gain of our method compared to the best competitor; this best performing alternative is also identified. Our appointment schedule outperforms 2BEG, IBFI$^+$ and DOME in 153 of the 162 cases; in the other 9 cases the relative difference is below 1%. Note that all cases are based on lognormal service times, which are the situations for which the DOME rule has been optimized.

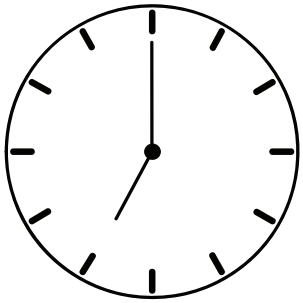
Of the three competitors, DOME performs substantially better than the other two rules, but averaged over all 162 instances about 5.5% worse than ours; IBFI$^+$ is over 6% worse, and 2BEG over 9%. Note that in Table 5 of Çayırlı et al. (2012), it was shown that DOME outperforms the *non-optimized* individual-block/fixed-interval rule 'IBFI', whereas in Table 6.2 we have included the results of the *optimized* version IBFI$^+$; for completeness, we mention that on average IBFI performs 19% worse than our tool (and in none of the instances better).

6.6 Conclusion

Appointment scheduling directly impacts the perceived quality, cost-efficiency and capacity of a substantial part of healthcare services. Our account establishes a framework for reasoning about the performance of appointment schedules, and structures the problem of designing a schedule. We frame the problem as one of achieving appropriate buffers to absorb variability and uncertainty in the arrivals and service times of patients. We further show that the appropriateness of schedules pertains to two issues. First, the achieved balance between idle and waiting time, reflected in the implied ω, should be consistent with the hospital's value proposition. And second, for given ω, the achieved expected idle and waiting times should be as near as possible to the efficient frontier. Our framework identifies a number of situational characteristics besides ω that should be taken into account in the design of a schedule: the number n of patients to be scheduled per session, the mean service time $\mathbb{E}B$, the squared coefficient of variation ϱ of service times, and the probabilities q of a no-show and w of a walk-in.

For the actual generation of schedules we propose approaches that are fast, robust and flexible. When the number n of patients increases, schedules based on a steady-state approximation become more and more appropriate (*stationary schedules*). For smaller numbers of patients (*transient schedules*), we offer a webtool that produces near-optimal schedules instantaneously. The scheduled interarrival times follow the familiar dome-shaped pattern and are based on an approximation of the service-time distribution by its phase-type counterpart. The tool offers a number of customizations such as four objective functions and imposing a resolution on the produced schedules. The tool implements three functionalities: it produces an efficient schedule given n and ω, or the implied ω given n and the expected session end time T, or the number of patients that can be scheduled for a given value of ω in a session with expected end time T. Extensive numerical simulations demonstrate that this tool outperforms competing approaches in the literature almost uniformly.

We believe that the framework set forth in this chapter establishes a solid basis for further refinements and additions. In Sections 6.4.2-6.4.4, we demonstrated how the core procedure for computing suitable schedules lends itself easily to the incorporation of situational specifics such as no-shows, walk-ins, overtime and discrete time slots. Further research could enrich the approach by incorporating additional relevant phenomena. Obvious candidates include the situation of multiple interchangeable healthcare providers, heterogeneous patient populations, and processes consisting of more than one stage. The integration of such additions into a single framework and webtool is a strong trump for finding adoption for the approach in practice.



7. Webtool for Appointment Scheduling

In this chapter we explain in detail how to use the webtool that is found online via the url: `http://www.appointmentscheduling.info`. The webtool facilitates practitioners to generate appointment schedules, and is based on the theory and concepts developed in Chapter 6. The tool implements three functions:

- Given the number of patients in a session and the relative weight ω of waiting versus idle time, the tool produces a (near) optimal schedule.
- Given the session (number of patients and duration), the tool calculates the implied relative weight ω of waiting versus idle time.
- Given the duration of a session and weight ω, the tool determines how many patients can be scheduled.

7.1 Filling in parameter values in the webtool

Figure 7.1 presents the interface of the webtool. We divide the situational characteristics that determine a clinical environment in two groups, addressing characteristics of the patients (workload) and specifications for schedule characteristics. The user enters her settings in the second column. In the third column we provide the range of values that are acceptable for the tool to work (the range is chosen such that it

meets the healthcare environments found in practice). The rightmost column indicates whether parameter values are optional or required.

Parameter	Value	Range	Necessity
Patient Characteristics			
Mean	15	$(0, \infty)$	Optional
SCV	0.5	$[0.1, 1.5]$	Required
q		$[0, \frac{3-2\cdot\text{SCV}}{5})$	Optional
w		$[0, 1]$	Optional
Schedule Characteristics			
n	13	$[2, 35]$	2 out of 3
ω	0.8	$[0.05, 0.99]$	2 out of 3
T		$(n \cdot \text{Mean}, \infty)$	2 out of 3
Δ	5	$[0, \infty)$	Optional
Objective Function			
$\min_{t_1,\ldots,t_n} \omega \sum_{i=1}^{n} \mathbb{E}I_i^{k_1} + (1-\omega) \sum_{i=1}^{n} \mathbb{E}W_i^{k_2}$			
$k_1 = 1$ & $k_2 = 1$	$k_1 = 1$ & $k_2 = 2$	$k_1 = 2$ & $k_2 = 1$	$k_1 = 2$ & $k_2 = 2$
Compute appointment schedule			

Figure 7.1: A print screen of the user interface of the webtool.

First the user enters the average service time a patient requires. If this field is left blank, then the webtool sets the average service times equal to 1. The second parameter is the scv, which expresses how much variance (relative to the squared mean) there is in the service times. A value close to zero means almost deterministic service times, whereas a high value means very unpredictable service times. The q and w are the no-show and walk-in probabilities. Leaving these fields empty corresponds to zero probability of no-shows or walk-ins.

The user specifies two out of the three parameters that characterize the schedule: the number of patients to be scheduled (n), the weight value (ω) and the (targeted)

expected session end time (T). As explained in Section 6.4.1, the operation of the webtool depends on which two of these parameters are entered:

1. If n and ω are provided, the webtool generates the resulting optimal schedule as well as its expected session end time T.

2. Entering n and T (where, evidently, $T > n\,\mathbb{E}B$), the tool determines the implied value of ω and returns the optimal schedule that has an expected session end time T.

3. The third option is to select T and ω to find out how many patients can optimally be scheduled such that the expected session end time remains below T, given the weight ω.

In addition, the user selects the type of objective function that is minimized by selecting k_1 and k_2. The parameters k_1 (contribution of idle time) and k_2 (waiting time) are equal to 1 or 2 corresponding to a linear or quadratic contribution to the objective function. The default setting is that both are linear. Finally, there is the option to impose a resolution on the schedule by setting the value of Δ. The schedule will be presented in multiples of Δ, where common choices are 5, 10 or 15 (minutes). This field can also be left blank resulting in the optimal schedule in continuous time.

7.2 Interpreting the output of the webtool

Irrespective of the chosen functionality (options 1, 2 and 3 above) the tool presents the generated schedule as in Figure 7.2. Subsequent patients are presented in the rows, and for each the optimal interarrival time (x_i) to the next patient is given as well as the scheduled arrival times (t_i). Since the generation of an appointment schedule is almost instantaneous, a user can easily modify her input and experiment with various settings.

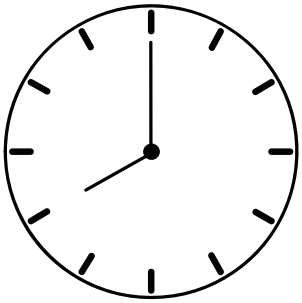
The output in Figure 7.2 corresponds to the parameter values in Figure 7.1: 13 patients are scheduled; the provider's (idle) time is weighted $\omega/(1-\omega) = 0.8/0.2 = 4$ times the patients' (waiting) time; the mean service time is 15 minutes and the scv is 0.5; and the schedule ignores no-shows and walk-ins (q and w are equal to 0). This example illustrates functionality 1, where n and ω are provided.

The resulting expected session end time (T) is 222 minutes. If process variability were nil then the patients could be scheduled 15 minutes apart and the total session duration would be 195 minutes. Therefore the excess (unutilized) capacity that the optimal schedule incorporates as a buffer to absorb variability is $222 - 195 = 27$ minutes. The provider's appointment book is based on the schedule that consists of the arrival times ($t_1, \ldots, t_{13}$) that have been rounded to multiples of 5 minutes.

Appointment Schedule		
Patient (i)	Interarrival time (x_i)	Arrival time (t_i)
1	10	0
2	15	10
3	15	25
4	20	40
5	15	60
6	20	75
7	15	95
8	15	110
9	20	125
10	15	145
11	15	160
12	10	175
13		185
Expected makespan (T)		222

Return to homepage

Figure 7.2: An example of the output that the webtool is presenting after filling in the default values as given in Figure 7.1.



8. *Summary* Appointment Scheduling in Healthcare

This dissertation is about structuring the problem of appointment scheduling, quantifying the performance of an appointment schedule and generating optimal appointment schedules. The resulting schedules are intended to be used in healthcare applications. The research presented in the dissertation comprises the results of the papers by Kuiper et al. (2015), Kuiper and Mandjes (2015a,b), Vink et al. (2015) and Kuiper et al. (2016).

8.1 Appointment scheduling in healthcare

There is great pressure to improve healthcare processes, on the one hand to control costs and on the other hand to guarantee good service. Appointment scheduling is generally applied in healthcare, and has to strike a satisfactory balance in these opposing ambitions. Generating appropriate appointment schedules is nontrivial due to the variability of various quantities, such as the patients' service times. An additional difficulty is that there is typically a broad variety of offered services, each of them involving specific situational characteristics.

The research on appointment scheduling started in the 1950s with pioneering works describing simple appointment rules. A clear shortcoming of such rules is that they do not deal effectively with the underlying variability. As computing power has increased significantly over the past decades, more recent studies propose to design schedules by optimizing an objective function.

8.2 Motivation

In the modern appointment scheduling literature various approaches have been proposed. Some of them are simulation-based, but these have the obvious shortcoming that the resulting guidelines tend to be case-specific. Other approaches cannot handle relevant characteristics of healthcare processes.

The goal of this dissertation is to develop an appointment scheduling methodology that covers realistic healthcare settings, and that outperforms existing methods. The objective is to develop a framework for appointment scheduling that is *fast*, *robust* and *flexible*. It should lead to a procedure that can produce schedules at low computational cost while capable of incorporating the most relevant situational characteristics.

8.3 Methodology and results

The main idea behind our approach is to cast appointment scheduling in a queueing-theoretic framework. Due to the randomness in the service-time durations, the healthcare provider can find herself idle sometimes, whereas at other moments the patients may have to wait before they can be served. An *objective function* should reflect these conflicting interests. A frequently used objective function is (for $k_1, k_2 \in \{1, 2\}$ and $\omega \in (0, 1)$):

$$\mathscr{F}^{(k_1,k_2)}[t_1, \ldots, t_n] = \omega \sum_{i=1}^{n} \mathbb{E} I_i^{k_1} + (1 - \omega) \sum_{i=1}^{n} \mathbb{E} W_i^{k_2},$$

where W_i the waiting time is of the i-th patient, and I_i the corresponding idle time of the service provider. This function has to be minimized over the patients' appointed arrival times $(t_1, \ldots, t_n)$. The evaluation of this objective function is in general not tractable. To overcome this complication we study various simplification approaches. In the first place, we approximate the service times by their phase-type counterparts, thus facilitating a fast and efficient recursive procedure for the evaluation of the distribution of the waiting times and idle times, which can be used to determine $\mathbb{E} I_i^{k_1}$ and $\mathbb{E} W_i^{k_2}$ for all n patients.

We also consider the *steady-state version* of our appointment scheduling problem, corresponding with the situation in which a large number of patients are scheduled. Steady-state schedules often provide useful direction for setting up schedules for smaller numbers of patients, as they help in understanding the impact of various parameters on the resulting schedule. This motivates why we include various methods to analyze and minimize the objective function in steady state. For this limiting setting we provide insightful analytical results relying on a *heavy-traffic* approximation.

Besides the phase-type approach, we have developed an alternative approach that uses the actual service-time distribution. The main idea behind this so-called *lag order*

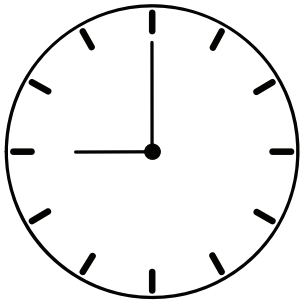
approximation is that it approximates the waiting-time and idle-time distributions by only taking into account a short history of two preceding patients. We show that this approach leads to accurate results, but tends to be prohibitively slow (and substantially slower than the phase-type approach).

For the most part this dissertation focuses on the situation of a single service provider, although we do analyze a two-node tandem system as well. This, technically considerably more demanding, system corresponds to the situation in which a patient is offered service sequentially by two providers; the schedule corresponds to the arrival epochs at the first provider. We show in detail how various healthcare-specific features, such as walk-ins and no-shows, can be incorporated in this approach.

8.4 Practical implications

The techniques developed in this dissertation yield appointment schedules, which are the optimal interarrival times of successive patients (with respect to a given objective function). These interarrival times turn out to have a so-called *dome shape*: shorter interarrival times early in the session (thus building up a buffer of workload in the form of waiting patients); longer interarrival times in the middle of the session, which are approximately equal to those in the steady-state schedule; and again shorter interarrival times at the end. As a consequence of the more condensed appointments at the end, expected waiting times increase for the last patients in the session, but since the number of subsequent patients is smaller, a schedule overrun affects fewer and fewer patients and therefore carries less and less weight compared to the opposing ambition to reduce expected idle time.

Our general conclusion is that the preferred technique is the phase-type approach, which outperforms competing approaches in the literature almost uniformly. We provide a webtool that implements this approach, which can be used directly by practitioners in healthcare. Importantly, it incorporates various relevant situational characteristics, such as service-time variability, walk-ins and no-shows. It meets the imposed requirements of being fast, robust and flexible, as it generates optimal schedules instantaneously for a broad range of relevant settings and achieves excellent performance.



9. Bibliography

Alexopoulos, C., D. Goldsman, J. Fontanesi, D. Kopald, J. Wilson. 2008. Modeling patient arrivals in community clinics. *Omega* **36**(1) 33–43.

Anderson, R., F. Camacho, R. Balkrishnan. 2007. Willing to wait?: The influence of patient wait time on satisfaction with primary care. *BMC Health Services Research* **7**(1) 7–31.

Asmussen, S. 2003. *Applied Probability and Queues*. Stochastic Modelling and Applied Probability, Springer-Verlag, New York, NY, USA.

Asmussen, S., O. Nerman, M. Olssen. 1996. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics* **23**(4) 419–441.

Babes, M., G. Sarma. 1991. Out-patient queues at the lbn-rochd health center. *Operational Research Society* **42**(10) 845–855.

Bailey, N.T.J. 1952. A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society. Series B (Methodological)* **14**(2) 185–199.

Barron, W.M. 1980. Failed appointments. Who misses them, why they are missed, and what can be done. *Primary care* **7**(4) 563–574.

Berwick, D.M., T.W. Nolan, J. Whittington. 2008. The triple aim: care, health, and cost. *Health Affairs* **27**(3) 759–769.

Brahimi, M., D. J. Worthington. 1991. Queueing models for outpatient appointment systems - a case study. *Journal of The Operational Research Society* **42** 733–746.

Çayırlı, T., E. Veral, H. Rosen. 2006. Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science* **9**(1) 47–58.

Çayırlı, T., K.K. Yang, S.A. Quek. 2012. A universal appointment rule in the presence of no-shows and walk-ins. *Production and Operations Management* **21**(4) 682–697.

Çayırlı, T., E. Veral. 2003. Outpatient scheduling in health care: A review of literature. *Production and Operations Management* **12**(4) 519–549.

Charnetski, J.R. 1984. Scheduling operating room surgical procedures with early and late completion penalty costs. *Journal of Operations Management* **5**(1) 91–102.

Corless, R.M., G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey, D.E. Knuth. 1996. On the Lambert W-function. *Advances in Computational Mathematics* **5**(1) 329–359.

Côté, M., W. Stein. 2007. A stochastic model for a visit to the doctor's office. *Mathematical and Computer Modelling* **45**(3) 309–323.

Cox, T., J. Birchall, H. Wong. 1985. Optimising the queuing system for an ear, nose and throat outpatient clinic. *Journal of Applied Statistics* **12**(2) 113–126.

Creemers, S., J. Belën, M. Lambrecht. 2012. The optimal allocation of server time slots over different classes of patients. *European Journal of Operational Research* **219**(3) 508–521.

Dallery, Y., S. Gershwin. 1992. Manufacturing flow line systems: A review of models and analytical results. *Queueing Systems* **12**(1-2) 3–94.

De Mast, J., B.P.H. Kemper, R.J.M.M. Does, M.R.H. Mandjes, H.W.J. Van der Bijl. 2011. Process improvement in healthcare: Overall resource efficiency. *Quality and Reliability Engineering International* **27**(8) 1095–1106.

De Vuyst, S., H. Bruneel, D. Fiems. 2011. Fast evaluation of appointment schedules for outpatients in health care. *Proc. ASMTA 2011* 113–131.

De Vuyst, S., H. Bruneel, D. Fiems. 2014. Computationally efficient evaluation of appointment schedules in health care. *European Journal of Operational Research* **237**(3) 1142–1154.

Denton, B., D. Gupta. 2003. A sequential bounding approach for optimal appointment scheduling. *IIE Transactions* **35**(11) 1003–1016.

Fetter, R.B., J.D. Thompson. 1966. Patients' waiting time and doctors' idle time in the outpatient setting. *Health Services Research* **1**(1) 66–90.

Fries, B.E., V. P. Marathe. 1981. Determination of optimal variable-sized multiple-block appointment systems. *Operations Research* **29**(2) 324–345.

Gallucci, G., W. Swartz, F. Hackerman. 2005. Brief reports: Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center.

Goddard, L.S. 1963. *Mathematical Techniques of Operational Research*. International Series of Monographs on Pure and Applied Mathematics, Pergamon, Oxford, United Kingdom.

Goldman, J., H.A. Knappenberger, E.W. Moore Jr. 1969. An evaluation of operating room scheduling policies. *Hospital Management* **107**(4) 40–51.

Green, L.V., S. Savin. 2008. Reducing delays for medical appointments: A queueing approach. *Operations Research* **56**(6) 1526–1538.

Günal, M.M., M. Pidd. 2010. Discrete event simulation for performance modelling in health care: a review of the literature. *Journal of simulation* **4**(1) 42–51.

Gupta, D., B. Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions* **40**(9) 800–819.

Hassin, R., S. Mendel. 2008. Scheduling arrivals to queues: a single-server model with no-shows. *Management Science* **54**(3) 565–572.

Healy, K. J. 1992. Scheduling arrivals to a stochastic service mechanism. *Queueing Systems* **12**(3) 257–272.

Ho, C. J., H. S. Lau. 1999. Evaluating the impact of operating conditions on the performance of appointment scheduling rules in service systems. *European Journal of Operational Research* **112**(3) 542–553.

Ho, C.J., H.S. Lau. 1992. Minimizing total cost in scheduling outpatient appointments. *Management Science* **38**(12) 1750–1764.

Hopp, W.J., M.L. Spearman. 2008. *Factory Physics*. 3rd ed. McGraw-Hill, Boston, MA, USA.

Huang, X. 1994. Patient attitude towards waiting in an outpatient clinic and its applications. *Health Services Management Research* **7**(1) 2–8.

Institute of Medicine. 2001. *Crossing the Quality Chasm: A New Health System for the 21st Century*. National Academy Press.

Institute of Medicine. 2006. *Hospital-Based Emergency Care: At the Breaking Point*. National Academy Press.

Johnson, B.J., J.W. Mold, J.M. Pontious. 2007. Reduction and management of no-shows by family medicine residency practice exemplars. *The Annals of Family Medicine* **5**(6) 534–539.

Kaandorp, G.C., G. Koole. 2007. Optimal outpatient appointment scheduling. *Health Care Management Science* **10**(3) 217–229.

Kemper, B., C.A.J. Klaassen, M. Mandjes. 2014. Optimized appointment scheduling. *European Journal of Operational Research* **239**(1) 243–255.

Kemper, B., M. Mandjes. 2012. Mean sojourn times in two-queue fork-join systems: bounds and approximations. *OR Spectrum* **34**(3) 723–742.

Klassen, K., T. Rohleder. 1996. Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management* **14**(2) 83–101.

Kuiper, A., B. Kemper, M. Mandjes. 2015. A computational approach to optimized appointment scheduling. *Queueing Systems* **79**(1) 5–36.

Kuiper, A., M. Mandjes. 2015a. Appointment scheduling in tandem-type service systems. *Omega* **57**(B) 145–156.

Kuiper, A., M. Mandjes. 2015b. Practical principles in appointment scheduling. *Quality and Reliability Engineering International* **31**(7) 1127–1135.

Kuiper, A., M. Mandjes, K.R. Brokkelkamp, J. De Mast. 2016. Efficient procedures to appointment scheduling. Submitted for publication.

Lau, H.S., A.H.L. Lau. 2000. A fast procedure for computing the total system cost of an appointment schedule for medical and kindred facilities. *IIE Transactions* **32**(9) 833–839.

Liao, C.Y., D.C. Pegden, C. Rosenshine. 1993. Planning timely arrivals to a stochastic production or service system. *IIE Transactions* **25**(5) 63–73.

Lindley, D.V. 1952. The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society* **48**(2) 277–289.

Liu, L., X. Liu. 1998a. Block appointment systems for outpatient clinics with multiple doctors. *The Journal of the Operational Research Society* **49**(12) 1254–1259.

Liu, L., Xi. Liu. 1998b. Dynamic and static job allocation for multi-server systems. *IIE Transactions* **30**(9) 845–854.

Liu, N., S. Ziya, V.G. Kulkarni. 2010. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing & Service Operations Management* **12**(2) 347–364.

Luo, J., V. Kulkarni, S. Ziya. 2012. Appointment scheduling under patient no-shows and service interruptions. *Manufacturing and Service Operations Management* **14**(4) 670–684.

Mak, H.Y., Y. Rong, J. Zhang. 2015. Appointment scheduling with limited distributional information. *Management Science* **61**(2) 316–334.

Mondschein, S. V., G. Y. Weintraub. 2003. Appointment policies in service operations: a critical analysis of the economic framework. *Production and Operations Management* **12**(2) 266–286.

Moore, C.G., P. Wilson-Witherspoon, J.C. Probst. 2001. Time and money: effects of no-shows at a family practice residency clinic. *Family Medicine-Kansas City-* **33**(7) 522–527.

Murdock, A., C. Rodgers, H. Lindsay, T.C.K. Tham. 2002. Why do patients not keep their appointments? prospective study in a gastroenterology outpatient clinic. *Journal of the Royal Society of Medicine* **95**(6) 284–286.

O'Keefe, R.M. 1985. Investigating outpatient departments: implementable policies and qualitative approaches. *Journal of the Operational Research Society* 705–712.

Pegden, C.D., M. Rosenshine. 1990. Scheduling arrivals to queues. *Computers & Operations Research* **17**(4) 343–348.

Porter, M.E. 2010. What is value in health care? *New England Journal of Medicine* **363**(26) 2477–2481.

Rising, E. J., R. Baron, B. Averill. 1973. A systems analysis of a university-health-service outpatient clinic. *Operations Research* **21**(5) 1030–1047.

Robinson, L.W., R.R. Chen. 2003. Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Transactions* **35**(3) 295–307.

Robinson, L.W., R.R. Chen. 2010. A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing & Service Operations Management* **12**(2) 330–346.

Robinson, L.W., R.R. Chen. 2011. Estimating the implied value of the customer's waiting time. *Manufacturing & Service Operations Management* **13**(1) 53–57.

Rockart, J.F., P.B. Hofmann. 1969. Physician and patient behavior under different scheduling systems in a hospital outpatient department. *Medical Care* **7**(6) 463–470.

Rohleder, T., K. Klassen. 2000. Using client-variance information to improve dynamic appointment scheduling performance. *Omega* **28**(3) 293–302.

Schild, A., I.J. Fredman. 1961. On scheduling tasks with deadlines and non-linear loss functions. *Management Science* **7**(3) 280–285.

Soriano, A. 1966. Comparison of two scheduling systems. *Operations Research* **14**(3) 388–397.

Stein, W. E., M. J. Côté. 1994. Scheduling arrivals to a queue. *Comput. Oper. Res.* **21**(6) 607–614.

Steinbauer, J.R., K. Korell, J. Erdin, S.J. Spann. 2006. Implementing open-access scheduling in an academic practice. *Family practice management* **13**(3) 59–64.

Swisher, J., S. Jacobson, J. Jun, O. Balci. 2001. Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. *Computers & Operations Research* **28**(2) 105–125.

Tijms, H.C. 1986. *Stochastic Modelling and Analysis — a Computational Approach*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, Chichester, UK.

Turkcan, A., B. Zeng, K. Muthuraman, M. Lawley. 2011. Sequential clinical scheduling with service criteria. *European Journal of Operational Research* **214**(3) 780–795.

Vanden Bosch, P. M., D. C. Dietz. 2000. Minimizing expected waiting in a medical appointment system. *IIE Transactions* **32**(9) 841–848.

Vanden Bosch, P. M., D. C. Dietz, J. R. Simeoni. 1999. Scheduling customer arrivals to a stochastic service system. *Naval Research Logistics (NRL)* **46**(5) 549–559.

Vink, W., A. Kuiper, B. Kemper, S. Bhulai. 2015. Optimal appointment scheduling in continuous time: The lag order approximation method. *European Journal of Operational Research* **240**(1) 213–219.

Vissers, J. 1979. Selecting a suitable appointment system in an outpatient setting. *Medical Care* **17**(12) 1207–1220.

Vissers, J., J. Wijngaard. 1979. The outpatient appointment system: Design of a simulation study. *European Journal of Operational Research* **3**(6) 459–463.

Wang, P.P. 1993. Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics* **40**(3) 345–360.

Wang, P.P. 1997. Optimally scheduling n customer arrival times for a single-server system. *Computers & Operations Research* **24**(8) 703–716.

Wang, P.P. 1999. Sequencing and scheduling n customers for a stochastic server. *European Journal of Operational Research* **119**(3) 729–738.

Weiss, E. 1990. Models for determining estimated start times and case orderings in hospital operating rooms. *IIE Transactions* **22**(2) 143–150.

Welch, J.D. 1964. Appointment systems in hospital outpatient departments. *OR* **15**(3) 224–232.

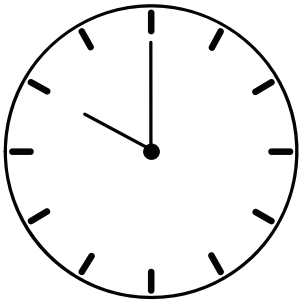
Welch, J.D., N.T.J. Bailey. 1952. Appointment systems in hospital outpatient departments. *The Lancet* **259**(6718) 1105–1108.

White, M. J. B., M. C. Pike. 1964. Appointment systems in out-patients' clinics and the effect of patients' unpunctuality. *Medical Care* **2**(3) 133–145.

Yang, K.K., M.L. Lau, S.A. Quek. 1998. A new appointment rule for a single-server, multiple-customer service system. *Naval Research Logistics (NRL)* **45**(3) 313–326.

Zacharias, C., M. Pinedo. 2014. Appointment scheduling with no-shows and overbooking. *Production and Operations Management* **23**(5) 788–801.

Zhu, Z., B.H. Heng, K.L. Teow. 2012. Analysis of factors causing long patient waiting time and clinic overtime in outpatient clinics. *Journal of Medical Systems* **36**(2) 707–713.



10. *Samenvatting* Roosteren van afspraken in de gezondheidszorg

Het structureren van de problematiek omtrent het roosteren van afspraken, het kwantificeren van de prestatie van een afsprakenrooster en het genereren van optimale roosters staan in dit proefschrift centraal. De met onze methodologie verkregen afsprakenroosters zijn geschikt voor toepassingen in de gezondheidszorg. Het onderzoek in dit proefschrift beschrijft de resultaten van verschillende onderzoeksartikelen: Kuiper et al. (2015), Kuiper and Mandjes (2015a,b), Vink et al. (2015) en Kuiper et al. (2016).

10.1 Roosteren van afspraken in de gezondheidszorg

De druk om processen in de gezondheidszorg te verbeteren neemt steeds verder toe. Enerzijds is het doel hiervan de kosten te reduceren, anderzijds beoogt men een degelijk serviceniveau te garanderen. De roostering van afspraken kan beschouwd worden als een techniek om een balans te vinden tussen deze tegenstrijdige ambities. Door de verschillende onzekere factoren die een rol spelen, bijvoorbeeld de duur van de behandeling van de patiënten (de bedieningstijd), is het maken van een goed afsprakenrooster niet triviaal.

Het onderzoek op het gebied van afsprakenroostering is gestart omstreeks 1950 met de analyse van eenvoudige roosteringsregels. Een duidelijke tekortkoming van zulke regels is dat inherente onzekerheden niet adequaat worden meegenomen. Met behulp van de recentelijk sterk toegenomen rekenkracht is de ambitie om optimale

roosters te verkrijgen, die berusten op doelfuncties die de onzekerheden op een juiste wijze verdisconteren.

10.2 Motivering

Gedurende de afgelopen decennia zijn verscheidene methodes geïntroduceerd voor het inroosteren van afspraken. Sommige daarvan zijn gebaseerd op simulaties, die echter de intrinsieke tekortkoming hebben dat de gegenereerde regels geen universele geldigheid hebben. Daarentegen kunnen meer generieke aanpakken in de regel niet goed omgaan met de grilligheid van de omgevingskenmerken die een rol spelen in de gezondheidszorg.

Het doel van dit proefschrift is om een methode te ontwikkelen die de meest realistische situaties afdekt. We beogen tot een raamwerk van het roosteren van afspraken te komen dat *snel*, *robuust* en *flexibel* is. Dit moet leiden tot een efficiënte procedure die tevens, met weinig extra rekenkracht, ook de relevante omgevingskenmerken in de gezondheidszorg meeneemt.

10.3 Methodologie en resultaten

Het probleem van het roosteren van afspraken wordt in dit proefschrift met behulp van wachtrijtheorie geanalyseerd. Vanwege de fluctuaties in de behandeltijden kan het zijn dat patiënten eerst moeten wachten alvorens ze behandeld kunnen worden. Als er geen patiënten in de wachtkamer zijn, kan het ook zo zijn dat juist de gezondheidszorgverlener (bijvoorbeeld een arts) moet wachten. Met behulp van een *doelfunctie* kunnen beide belangen vertegenwoordigd worden. Een veelgebruikte doelfunctie is (met $k_1, k_2 \in \{1, 2\}$ en $\omega \in (0, 1)$):

$$\mathscr{F}^{(k_1,k_2)}[t_1, \ldots, t_n] = \omega \sum_{i=1}^{n} \mathbb{E} I_i^{k_1} + (1 - \omega) \sum_{i=1}^{n} \mathbb{E} W_i^{k_2},$$

waarbij W_i de wachttijd is van de i-de patiënt en I_i de daarmee samenhangende wachttijd van de zorgverlener. Voor een optimaal rooster moet deze functie geminimaliseerd worden over de gewenste aankomstmomenten van de patiënten, die gegeven worden door $(t_1, \ldots, t_n)$. De doelfunctie kan in het algemeen niet expliciet bepaald worden. Daarom bestuderen we in dit proefschrift verschillende vereenvoudigingen die leiden tot een evalueerbare doelfunctie. In de eerste plaats benaderen we de verdeling van de behandeltijden met een fase-type verdeling met dezelfde verwachting en variantie. Deze vereenvoudiging maakt een snelle en efficiënte recursieve methode mogelijk, waarmee $\mathbb{E} I_i^{k_1}$ en $\mathbb{E} W_i^{k_2}$ bepaald kunnen worden.

Bovendien bekijken we ook de zogenaamde *stationaire setting* van het roosteringsprobleem. Deze setting komt overeen met de situatie waarin een zeer groot aantal patiënten geroosterd worden. Stationaire roosters geven waardevolle inzichten voor het roosteren van een klein aantal patiënten, omdat ze helpen met het kwantificeren van de impact van verschillende parameters op het rooster. We zijn tevens in staat

om in deze stationaire setting, met behulp van een *heavy-traffic* benadering, een aantal inzichtelijke limietresultaten te verkrijgen.

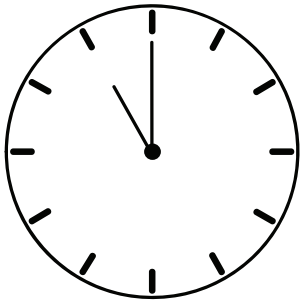
Naast de fase-type aanpak hebben we ook een andere methode ontwikkeld. Deze alternatieve methode benadert de wachttijden door steeds alleen de effecten van de twee voorgaande patiënten mee te nemen, waarbij de verdelingen van de behandeltijden onveranderd blijven. Deze zogenaamde *lag-order* benadering leidt tot nauwkeurige resultaten, maar is wel aanmerkelijk langzamer dan de fase-type aanpak.

Het grootste deel van dit proefschrift is gewijd aan de situatie met een enkele dienst- of zorgverlener (in wachtrij-terminologie: een wachtrij-netwerk met slechts één station). Daarnaast analyseren we echter ook netwerken met twee stations in een zogenaamde tandemopstelling. In dit complexere systeem doorloopt de patiënt twee stadia met elk een eigen dienstverlener; het rooster geeft de gewenste aankomstmomenten bij de eerste dienstverlener. Tot slot laten we in detail zien hoe verschillende relevante aspecten in de gezondheidszorg, zoals onaangekondigde spoedpatiënten en no-shows, in de procedures meegenomen kunnen worden.

10.4 Praktische implicaties

De technieken die in dit proefschrift gepresenteerd worden, leveren optimale afsprakenroosters (met betrekking tot een gegeven doelfunctie). Deze optimale roosters bestaan uit de gewenste aankomstmomenten voor de patiënten. In de hiermee samenhangende tussenaankomsttijden herkennen we een *koepelvorm*: kortere tussenaankomsttijden aan het begin van de sessie (die als doel hebben een werkbuffer met wachtende patiënten op te bouwen); langere tussenaankomsttijden in het midden van de sessie, die te benaderen zijn met stationaire tussenaankomsttijden; en aan het einde weer kortere tussenaankomsttijden. In zo´n rooster lopen de verwachte wachttijden voor de patiënten aan het eind van het rooster behoorlijk op. Echter, omdat op het eind van de sessie een steeds kleiner aantal patiënten hier last van heeft, weegt mogelijke uitloop van het rooster minder zwaar dan het tegengestelde belang om de tijd van de zorgverlener goed te gebruiken.

De conclusie van uitgebreide numerieke experimenten is dat de fase-type benadering in bijna alle situaties beter presteert dan concurrerende methodes, zoals die beschreven worden in de literatuur. De fase-type aanpak is geïmplementeerd in een webtool en kan door praktijkbeoefenaars in de gezondheidszorg gebruikt worden. Relevante omgevingskenmerken zijn verwerkt in de tool, zowel de variabiliteit in behandeltijden, als de kans op spoedpatiënten en no-shows. De aanpak voldoet hiermee aan de eerder gestelde eisen, want het genereert nagenoeg *real-time* een optimaal rooster voor iedere setting.



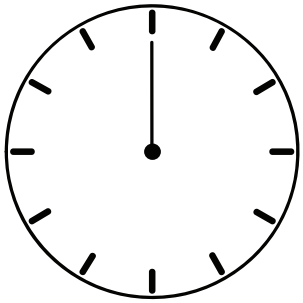
11. Curriculum Vitae

Alex Kuiper is geboren op 20 december 1987 te Amsterdam. Hij groeide op in Landsmeer en ging in 2000 naar het Montessori Lyceum Amsterdam. In 2006 begon hij met zijn bachelor wiskunde aan de Universiteit van Amsterdam (UvA). In 2009 rondde hij zijn bachelor met honours af en startte hij een double degree programma bestaande uit de masters Stochastics & Financial Mathematics en Econometrics.

In het najaar van 2010 ging Alex voor een half jaar op uitwisseling naar de University of Calgary. Hij voltooide zijn masterperiode met een onderzoeksstage bij het Instituut voor Bedrijfs en Industriële Statistiek van de Universiteit van Amsterdam (IBIS UvA) onder begeleiding van dr. Benjamin Kemper en prof. dr. Michel Mandjes. Gedurende zijn masterperiode was hij actief betrokken in het wiskunde curriculum van de UvA als tutor en werkcollegedocent.

In februari 2013 trad Alex bij IBIS UvA in dienst. Hij verricht daar diverse werkzaamheden. Zo doet hij onderzoek naar het optimaliseren van afsprakenroosters met als voornaamste toepassingsgebied de gezondheidszorg onder begeleiding van prof. dr. Michel Mandjes en prof. dr. Jeroen de Mast. Dit proefschrift is het resultaat daarvan. Daarnaast is hij als adviseur verantwoordelijk voor het opleiden en begeleiden van projectleiders, die zich willen ontwikkelen tot Lean Six Sigma Green of Black Belt. Tot slot is hij docent geweest bij verschillende vakken binnen de Amsterdam Business School, namelijk Management Research Methods 1 & 2, Quantitative Methods en Operations & Process Management.

Naast deze activiteiten haalt Alex zijn ontspanning uit bordspellen en sport. Zo is hij een fervent judoka, in bezit van de zwarte band (2^e dan) en vervult hij op vrijwillige basis de rol van secretaris in het bestuur van de judoclub. Daarnaast is hij ´s winters regelmatig op de ijsbaan te vinden.



12. Acknowledgments

Completing my dissertation has been a milestone in my life. When I look back on this period of time I realize that I am indebted to all the people that have contributed to the research project. First of all, my promoters, *Michel Mandjes* and *Jeroen de Mast*, who have helped me with setting out new research directions, proposing models and reporting results. I believe that without their help this dissertation would not have come into being.

Michel, I thank you for the stimulating research environment. I have enjoyed our various discussions about both the research as well as off-topic subjects. Sometimes you had to slow me down as I rushed to various conclusions, helping me to be more precise both verbally and in writing. Jeroen, I am impressed by your conceptual reasoning and your excellent language skills. You helped me to set the bar high in various projects and, more importantly, you coached me to clear that bar.

I am thankful to *Ronald Does* for supporting my development as a consultant at IBIS UvA. Together we taught our prestigious Lean Six Sigma courses several times. During these sessions I appreciated the freedom you gave me and your confidence in my capabilities. I am grateful that *Benjamin Kemper* contacted me in 2012 for a position at IBIS UvA that combines consultancy with scientific research. It has turned out to be a truly challenging and fruitful combination for me. I also thank the co-authors *Sandjai Bhulai* and *Wouter Vink*; and *Ruben Brokkelkamp*, who was an exemplary master student.

Furthermore, I am thankful to my current and former colleagues at IBIS UvA: *Marit Schoonhoven*, *Tashi Erdmann*, *Inez Zwetsloot*, *Thomas Akkerhuis*, *Rob Goedhart* and *Atie Buisman* for the inspiring work environment that I have experienced. Especially with Thomas and Inez with whom I had a lot of fun, such as having dinner together, going to Kriterion, having drinks at Café Koosje, or a union of the events just described.

I want to express my gratitude to all my friends and family for their support and love. The many social distractions during my time as a Ph.D. student were very welcome. For instance, the spiritful evenings with a group of friends from secondary education: *Job*, *Mira*, *Rik* and *Tom*. I also enjoyed the Risk sessions with the friends I met in my first year of mathematics: *Adrian*, *Dirk*, *Noach*, *Sander* and *Sjoerd*. I thank *Jermo* for the sportive and moral support on and off the ice skating rink.

Finally, I would like to thank my mother, *Ypie*, for her everlasting faith in me; *Carlo* for being an attentive brother and for all the fun moments in Landsmeer; and, of course, *Mariska* for her great support as my caring, helpful and loving companion.

Alex Kuiper
Amsterdam, May 2016