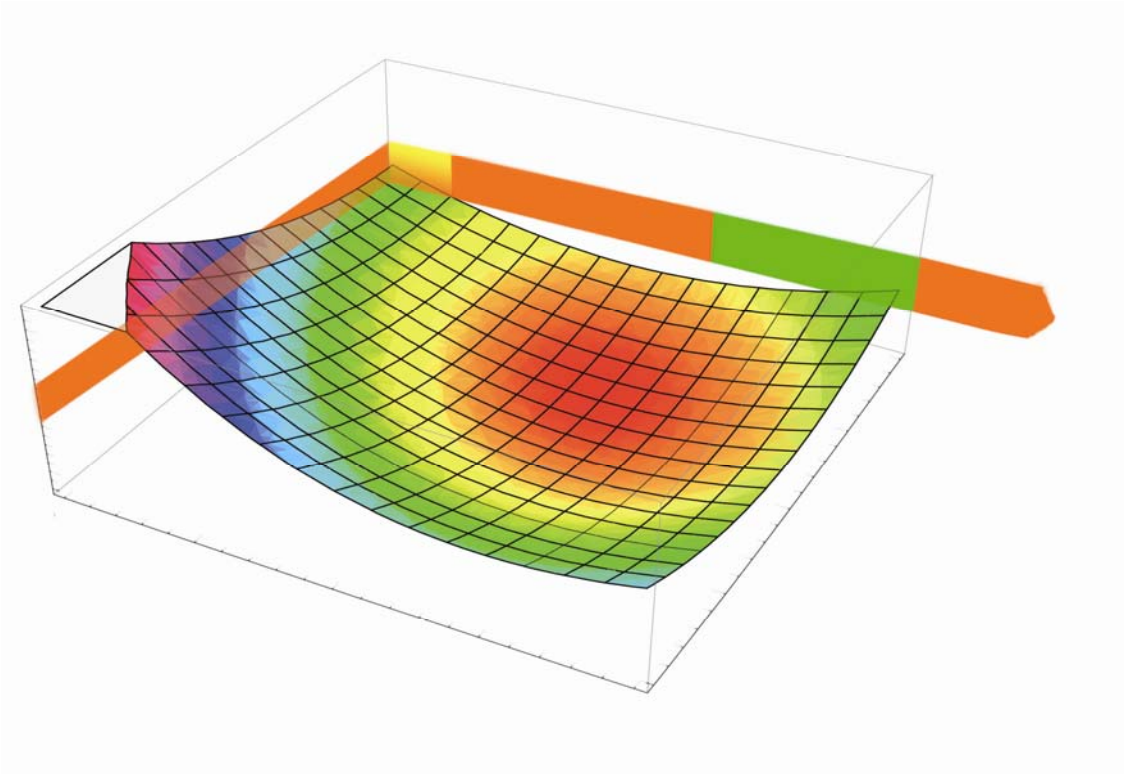
Universiteit Utrecht

Utility-Based Appointment Scheduling in Continuous Time: The Lag Order Approximation Method

W.E.J. Vink

August 2011



IBIS UvA
Instituut voor Bedrijfs- en Industriële Statistiek

Cover illustration by Marcel Buur

Utility-Based Appointment Scheduling in Continuous Time: The Lag Order Approximation Method

W.E.J. Vink
Utrecht University, Department of Mathematics
Institute for Business and Industrial Statistics (IBIS UvA)

August 14, 2011

Master's Thesis

Supervisors: Dr. S. Bhulai, VU University Amsterdam
Dr. K. Dajani, Utrecht University
Drs. B.P.H. Kemper, IBIS UvA
Co-reader: Dr. M.C.J. Bootsma, Utrecht University

“If you are a patient...and wait long enough...
Nothing will happen!”

-WOUTER VINK

Based on:

“If you are patient...and wait long enough...
Nothing will happen!”

-JIM DAVIS, creator of the comic strip "Garfield"

Summary

An appointment schedule aims to achieve a proper balance between different interests: the waiting time for the customers and the waiting time for the service providers, the so called idle time. These interests are truly conflicting; focusing on small idle time results in large waiting times for the customer and vice versa. This thesis considers the area of appointment scheduling for a single server in continuous time with punctual customers. There is a vast amount of literature addressing this topic, however, there is not a generic approach to generate large schedules analytically with general service time distributions. First, we study the characteristics of this problem with exponential service times. Next, we propose an approximation method that enables us to design appointment schedules with general service time distributions. This method is the lag order approximation method, which is the main result of this thesis. Another method is proposed to overcome the dimensionality problem, when designing appointment schedules for a large number of customers. We will present both methods, and investigate their performance and applicability.

Keywords: Appointment Scheduling ⋆ Approximation Methods ⋆ Continuous Time Scheduling ⋆ Equidistant ⋆ General Service Time Distributions ⋆ Heuristic Methods ⋆ Lag Order.

Preface

You are now reading the Master's thesis of Wouter Vink. This thesis is about scheduling customers for a single server. For example, patients for a CT scan or court cases for a court. We will introduce a new approximation method for this problem, improving and enhancing current analytical methods. It reduces computation time and increases applicability.

At the beginning of the second year of my Master in Utrecht, I was investigating the possibilities of doing an internship to write my thesis. Via via, I came in contact with Tjarko de Vree, who wrote his thesis at the Institute for Business and Industrial Statistics of the University of Amsterdam (IBIS UvA). I would like to thank Tjarko for sharing his experiences, which led to my application for an internship at this institute.

Professor Ronald Does, director of IBIS UvA, and Benjamin Kemper, consultant at IBIS UvA, gave me the opportunity to write my master thesis at their company for which I am very thankful. Benjamin was assigned to me as my supervisor. Despite the fact that he was finishing his PhD thesis, he found the time to guide and help me. The complete period at IBIS UvA I was given under good supervision, thanks gozertie! I was lucky Benjamin and Michel Mandjes just discovered a new area of research, in which I was allowed to do my own research. Soon it became clear in which direction my thesis would go and we booked results quite quickly. I thank Benjamin and Michel for the cooperation on this topic and the fruitful discussions.

In an earlier stage I asked Sandjai Bhulai, associate professor at the VU, to be my supervisor, to which he agreed, enthusiastically! I would like to thank Sandjai for his overwhelming positive demeanor, which inspired me and helped me to finish this thesis pleasantly. Also thanks for the thorough explanation of the Hindu religion (or philosophy) and the automatic car parking showcase.

Every good thing comes in three applies for Karma Dajani, senior lecturer at Utrecht University, being my third supervisor. Thanks for supervising me, although you had close to zero time for this. The few meetings we had were very helpful and the comments on my draft versions very useful.

Martin Bootsma was added on the finish line, being the second reader from Utrecht University. I would like to thank you for your time and the unexpected revision, which I received just before my deadline.

I would like to thank Tashi Erdmann for sharing his room with me. You were always willing to converse about whatever topic, helping with a mathematical problem or talking about day-to-day things. Also accepting the mess I made, my occupation of the white board and the music-question almost every day. Thanks for that Tashi, we had a good laugh.

Of course I would like to thank the entire IBIS crew: Atie Buisman, fellow almost-West-Frisian Joran Lokkerbol, Jeroen de Mast and Marit Schoonhoven for welcoming me so friendly and making sure I was part of the team very quickly. I would like to thank Ronald for teaching me in making decisions ('just make them!') and for taking me along on the black belt training for three days.

Finally, I would like to thank my friends Jöbke Janssen, Tomas Molenaars en Hans Westrik for being the best study mates one can wish for, I would not have finished this study without you guys! My family for creating the best boundary conditions for a young man's life, and a special thanks to Tera Pijnacker for standing by me on tough moments, supporting me to get the best out of myself and for being my reasons.

You will now read my master thesis called: 'Utility-Based Appointment Scheduling in Continuous Time: The Lag Order Approximation Method.' Enjoy!

Amsterdam
August 2011 **Wouter Vink**

Contents

List of Figures

Chapter 1

Introduction

Modern health care involves the use of several high cost devices and facilities such as MRI machines, CT scanners and operating rooms. For these facilities, appointment scheduling is vital to ensure a high utilization of the resources and at the same time high quality of service, i.e., short waiting times for the customers.

Consider the problem of scheduling surgeries for patients in an operating room at a hospital. The planning of the surgeries needed to be performed on a particular day, is known in advance. However, the time needed to perform each surgery varies. The resource planner needs to decide in advance the time at which a particular surgery is scheduled, and the duration to assign to that surgery. If on one hand, a relatively small time period is assigned to a surgery, then it is likely that the realized time will exceed this assigned time period, thus delaying the next surgery. The inconvenience, potentially fatal haste and costs, resulting from the delay of both the patients and the staff, constitute extra cost of that surgery. If, on the other hand, the hospital manager assigns an excessively long time period for a surgery, then the surgery may end earlier than expected and the operating room will be left idle until the next appointment. In that case, the hospital again incurs extra costs caused by the under-utilization of the resources in the operating room.

Similar problems could arise in many other operational contexts. For example, if ship-to-shore cranes in a harbor are left idle, they do not yield and their earnings diminish. Then again, if the ships are left waiting for a crane to be available, extra costs are incurred for keeping the boat operating. Also the risk of shipping companies choosing for alternative harbors with smaller

waiting times, should be taken into account. For these reasons a good schedule is necessary.

Many service providers in the medical, legal or financial professions operate on an appointment basis and are usually concerned with both the idle time of the service provider and the waiting times of the customers.

A good appointment schedule should achieve the right trade-off between idle time and waiting time.

Motivation:
This thesis is part of the master Mathematical Sciences at the Utrecht University. It has been written at the Institute for Business and Industrial Statistics (IBIS UvA) in Amsterdam. At this institute I got involved in this topic due to an article on appointment scheduling, written by my supervisor. In this article a new optimization approach is proposed. Further insights of this approach compared to the conventional approaches were needed and this is part what we do in this thesis. Some of the results of this thesis are included in Kemper et al. (2011).

Problem Statement:
An appointment schedule consists of non-random arrival times for customers. These customers arrive at their assigned times and will be served by the server directly if it is available. However, it is possible that the server is still busy while the next customer arrives. Consequentially, the arriving customer has to wait, what we call waiting time. If the service provider is idle before the next customer arrives, the time the server has to wait is called idle time. These idle and waiting time are losses of the appointment system. The objective of scheduling is to find an appointment system for which a particular measure of performance for the system's losses in terms of these waiting and idle time is minimized. In the literature this is referred to as: 'An application of resource scheduling under uncertainty' (Cayirli and Veral, 2003). The resources are the service provider and the customers and the uncertainty originates from the modeling of the service times with a probability distribution. Hence, we work in a probabilistic setting using measures as percentiles of idle and waiting time or expected idle and waiting time. The former measure may be used if ones wishes to assess the quality of schedules with tail expressions. The latter, is the performance measure we use in this thesis, it is in terms of the quantities expected idle time of the server and expected waiting time of the customer. This measure is referred to as, the loss function, since it measures the expected losses of the appoint-

ment system in terms of the aforementioned quantities. The objective is to minimize this loss function by optimally choosing the arrival times of the customers.

The idle time of the server before the arrival of customer i is denoted with I_i, W_i is the waiting time of customer i. An example of a loss function is:

$$\sum_{i=1}^{n} \Big(\mathbb{E}I_i^2 + \mathbb{E}W_i^2 \Big). \tag{1.1}$$

With n the number of customers. Note that this loss function measures the sum of the second moments of the idle and waiting time equally.

Expression (1.1) is an example of a loss function, another choice can be the linear loss function with the option to penalize waiting times more than idle times, or vice versa:

$$\sum_{i=1}^{n} \Big(\alpha \mathbb{E}I_i + (1-\alpha)\mathbb{E}W_i \Big), \qquad \alpha \in [0,1],$$

with α a normalized weight factor.

The arrival time of customer i is denoted with t_i. A crucial observation is that the random variables I_i and W_i are affected by the services and arrivals of preceding customers, and its own arrival. These customers arrive at $t_1, \ldots, t_{i-1}$, and therefore I_i and W_i are functions of $t_1, \ldots, t_i$, including the arrival of customer i itself. Using expression (1.1) we arrive at a typical example of a minimization problem which we investigate in this thesis:

$$\min_{t_1,\ldots,t_n} \sum_{i=1}^{n} \Big(\mathbb{E}I_i^2(t_1, \ldots, t_i) + \mathbb{E}W_i^2(t_1, \ldots, t_i) \Big).$$

This problem is an n-dimensional optimization problem. As n increases, it gets harder to solve. Apart from numerical approaches, to the best of our knowledge no manageable generic characterization for the optimal schedule is known. Ideally, one would like to have a closed form solution for general service time distributions and arbitrary loss functions. Given that such a solution would be at hand, it can be applied across a broad range of application areas; such as health care, service systems, manufacturing, transportation, et cetera.

In a setting where we have exponential service times we are able to give

an explicit characterization for the optimal schedule. However, for general service time distributions no such result exists. In this thesis we investigate the reason for this absence. We will introduce certain approximation methods that are able to design appointment schedules with general service time distributions.

Approach of the Stated Problem
From current literature protrudes two complications to obtain a closed-form solution for large optimal schedules with general service time distributions. The first one is that with general service time distributions it is hard to derive expressions for the waiting time, what is not the case with exponential distributed service times. In that setting, it is possible to derive expressions for the waiting time for customer i in terms of all the previous customers. In other words, it is possible to incorporate the influence of all the predecessors on the waiting time of the current customer. This is possible since the exponential distribution has the memoryless property. This means that the expected remaining service time is equal to the expected service time at any moment in time. In mathematical notation:

$$\mathbb{P}(X > s + t | X > t) = \mathbb{P}(X > s) \text{ for } s, t > 0,$$

with X an exponentially distributed random variable.

The analytical consequence of this is that we only have to compute the probability of a predecessor still being in the system upon the arrival of customer i. If we multiply these probabilities with the expected service times we are done. This property can also implicitly be applied for Erlang distributions, since it can be described as a composition of memoryless stages. Unfortunately, this memoryless property does not apply for any other continuous distribution. Hence, deriving the same expressions for general distribution functions is hard. In our attempt to do so nonetheless, we came up with the lag order approximation method, which comes down to capturing the influence of only a few predecessors on the waiting time of the current customer. Not only do we discuss the specifications, we will also look at the performance in various problem settings of this approximation method and we will address where this method can add value to current conventional methods to derive appointment schedules.

The second complication of appointment scheduling is the dimension of it. As we mentioned before, with one customer more in the schedule, the dimension of the problem increases with one as well. In case we need to serve

the same type of customers, hence equal service time distributions, we observe in the optimal schedule for most of these customers equal interarrival times. We will take advantage of this knowledge by batching customers with the same type. This results in a reduction of the problem dimension and therefore a reduction of computation times.

Note that in this thesis most of the numerical calculations are performed using Mathematica and Matlab. The codes and programs will be made available on the readers request.

Organization of this thesis
This thesis is organized as follows. We start in Chapter 2 with an overview of the current literature, placing our contribution in a proper context using the classification of Westeneng (2007). We continue with the formulation of the mathematical model in Chapter 3. Most of the inspiration for our ideas came from the comparison between the sequential and the simultaneous optimization approach. Chapter 4 shows the results of this comparison, giving a lot of insight in the problem itself and the characteristics of both methods. Then we introduce some approximation methods for this problem in Chapter 5. The application of these methods on the problem in a realistic setting will be done in Chapter 6. Even more realistic will be Chapter 7, where we will show a problem from practice. In Chapter 8 we analyze the sensitivity of (sub-)optimal schedules, extracting some rules for the resource planner on how to deal with the derived results. Finally, in Chapter 9 we discuss the results, the advantages and disadvantages of the approximation methods and some recommendations for future research.

Throughout this thesis τ is defined as $\tau := \frac{C}{r}$, with C the circumference and r the radius of a circle, i.e., $\tau = 2\pi$. See Appendix A.8 for more details on this.

Chapter 2

Literature

In this chapter we present an overview of the current literature. The problem of appointment optimization has generated substantial interest over the last decades. Beginning with the seminal papers of Bailey (1952) and Welch and Bailey (1952), an extensive body of literature on appointment scheduling has been accumulated. State-of-the art surveys are given in Cayirli and Veral (2003) or Westeneng (2007) and more recently, a study of literature focusing on challenges and opportunities in this area of research, Gupta and Denton (2008). The articles in these surveys are mainly related to two medical problems: the scheduling of patients into a clinic, and the scheduling of surgeries into an operating room.

Appointment scheduling can be classified in two broad categories: static and dynamic. In the static case, all decisions must be made prior to the beginning of a session. In the dynamic case, the schedule of future arrivals is revised continuously over the course of the scheduling period, based on the current state of the system (Cayirli and Veral, 2003). For a good example of a paper on dynamic appointment scheduling see Liu et al. (2010). These authors use Markov decision processes to maximize the long-run average net 'reward' for the clinic. They also provide a good overview of the current literature on dynamic appointment scheduling, and a discussion about Open Access systems. The idea of Open Access is to keep your system open and let customers decide when they want to be served.

In this thesis our focus is on the static paradigm.

We classify the papers on static scheduling as follows: there are those that evaluate (potential) schedules, often using simulation, and those that design algorithms to find good schedules, the analytical approach.

Examples of the former can be found in the overview papers of Cayirli and Veral (2003) and Westeneng (2007). An example not included in both is Hutzschenreuter (2005), who uses non-exponential service times.

In the analytical approach one can distinguish papers focusing on continuous time and papers focusing on discrete time. The former deals with finding the optimal interarrival times, whereas the latter deals with the question how many arrivals should be scheduled at each potential arrival moment (with the possibility of zero arrivals included). Some important results for the discrete time, i.e., a finite number of potential arrival moments, include Liao et al. (1993). This paper uses a branch-and-bound method to find the optimal schedule. This, however, works only for small instances. Vanden Bosch et al. (1999) derive upper and lower bounds for the optimal appointment schedule. To show these bounds, they use, what they call, submodularity, what is related to convexity. These upper and lower bound schedules often coincide and can be made starting from a specific schedule. Kaandorp and Koole (2007) propose a method extending these results. Their method gives convergence to the optimal schedule starting from any initial schedule. They also extend the results to different types of patients. Vanden Bosch et al. (1999) uses Erlang service times, whereas Kaandorp and Koole (2007) uses exponential service times.

One of the reasons to study this problem in discrete time is due to dimensionality issues. As the number of slots increases, the problem gets harder to solve, because the minimization algorithm considers every extra slot as an extra dimension to the problem. By making time discrete, the number of possible schedules decreases, and we can use local search algorithms to derive (sub)-optimal schedules, e.g., Kaandorp and Koole (2007). In case of continuous time this dimension problem remains.

This thesis focuses on appointment scheduling in continuous time. Hence, computing an optimal schedule means computing the optimal interarrival times for the customers. Wang (1993) and Gray and Wang (1995) obtained the optimal schedule for arrivals over continuous time by applying nonlinear programs to the cost functions they chose. Wang also proved that the expected waiting times are convex with respect to the arrival time vector in case of general service time distributions (Wang, 1993). Vanden Bosch et al. (1999) obtained this result also in discrete time, while only considering waiting time and server completion time, not idle time.

These results however only consider exponential service times. Robinson and

Chen (2003) do consider a different service time distribution, the generalized lambda distribution. This distribution is very useful to fit to datasets of, for example, realized service times. Using this distribution makes it possible to derive schedules with a realistic service time distribution. Yet, they only managed to do so for schedules with a maximum of 16 customers and only with this specific distribution.

There have been more attempts in continuous time, see Westeneng (2007) for an overview. However, to the best of our knowledge, there has been no successful attempt for large schedules with general service time distributions in continuous time. A recent attempt in continuous time, not mentioned in the overview articles, is Kemper et al. (2011). In their paper, the authors suggest an sequential approach to the problem. By sequential they refer to an approach that determines the i-th appointment time t_i, with the earlier arrival epochs $t_1, \cdots \ldots, t_{i-1}$ being fixed already.

Below we put our contribution in the classification of Westeneng (2007):

Classification parameter	**Our setting**
Methodology	Analytical
Number of doctors	1
Number of patients per session	n varies, n=100 is shown
Appointment rule	Interval
Patient classification	Service times do not need to be i.i.d.
Scope	One session
Stages	Single stages
Queue discipline	First appointment, first serve
Performance measurement	Patients waiting time, doctors idle time and doctors overtime
Service time distribution	Exponential, log-normal, weibull and general service time distribution
Patient punctuality	Punctual
Walk-ins	None
Doctors' lateness	Punctual
Doctors' interruption level	None

Chapter 3

Model Formulation

This chapter provides a technical introduction to the theory we use throughout this thesis. In Section 3.1, we list the main assumptions and the outline of our research. We continue with the mathematical notation in Section 3.2 and in Section 3.3 we introduce the most common loss functions. In Section 3.4 we present our algorithm that computes optimal schedules. We conclude with the characteristics of two optimization approaches in Sections 3.5 and 3.6. In Chapter 4 we will deepen the understanding of the differences between both.

3.1 Problem Setting

An appointment schedule consists of arrival times for the customers. When a customer arrives, the server is either available or busy. If it is available, it is not utilized and the server is waiting, what is called idle time. Hence, it is the time the server is waiting for the next customer after having served the current customer earlier than expected. If the server is busy, the customer has to wait, this time is called waiting time. The objective of scheduling is to find an appointment system for which a particular measure of performance for the system's losses in terms of these waiting times and idle times is optimized. In some settings also the completion time or system lateness are added to the performance measure of the schedule. The system lateness is the expected time the schedule exceeds the reserved running time of the server.

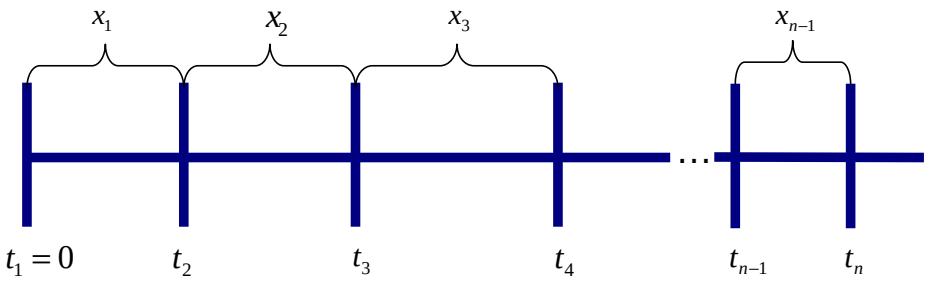


Figure 3.1: The notation used for the time slots and appointment times.

We investigate this problem under certain assumptions and in a setting specific characteristics, which we list below:

1. There is one service provider.
2. The customers are punctual, i.e., customer i arrives at its assigned time t_i.
3. The server is punctual, i.e., if the server is idle and a customer arrives, the server will start serving this customer immediately.
4. All customers show up for their appointment, i.e., the model does not include no-shows.
5. The service times of the customers are modeled by a probability distribution. Unless stated otherwise, the service times for all customers are i.i.d..
6. We do not consider a schedule which allows walk-in customers, e.g., emergency patients.
7. The service order of the customers is first appointment, first serve.

Further on in this thesis we will relax some of these restrictions, adjusting our problem setting to make our results more applicable to practical situations.

3.2 Notation

We consider an appointment schedule with customers arriving at their assigned arrival time, denoted by t_i with $t_1 = 0$ and $t_1 \leq t_2 \leq \cdots \leq t_n$. So

customer 1 arrives and will immediately get served at $t_1 = 0$. We define

$$x_{i-1} := t_i - t_{i-1}, \tag{3.1}$$

to be the slot size to serve customer $i-1$, for $i > 1$. Note that $x_i \geq 0$ for $i \geq 1$. Figure 3.1 gives an illustration on how to interpret notation (3.1). The schedule in the figure starts on the left with the arrival of customer 1 on $t_1 = 0$, next on t_2 customer 2 arrives resulting in the time slot x_1 for customer 1. On t_3 customer 3 arrives and so forth. We will use both x_i, the slot size for customer i, and t_i, the arrival time for customer i, as variables in this thesis. We denote n as the number of customers in the appointment schedule.

Let us now introduce the following random variables:

I_i	Idle time of the server before customer i arrives,
W_i	Waiting time for customer i,
B_i	Service time for customer i,
S_i	Sojourn time for customer i; the time customer i is in the system $(= W_i + B_i)$.

If customer i finds the server idle at his arrival, we can express I_i in terms of random variables of the previous slot:

$$\begin{aligned} I_i &= \max\{t_i - t_{i-1} - (W_{i-1} + B_{i-1}), 0\}, \\ &= \max\{x_{i-1} - S_{i-1}, 0\}. \end{aligned} \tag{3.2}$$

If customer i finds the server busy at his arrival, we can express W_i in terms of random variables of the previous slot:

$$\begin{aligned} W_i &= \max\{W_{i-1} + B_{i-1} - (t_i - t_{i-1}), 0\}, \\ &= \max\{S_{i-1} - x_{i-1}, 0\}. \end{aligned} \tag{3.3}$$

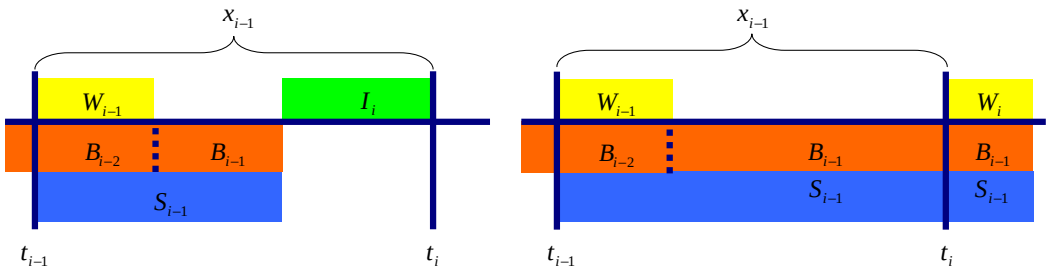


Figure 3.2: Graphical explanation of the relation between the random variables B_i, I_i, W_i and S_i.

Note that $W_1 = I_1 = 0$ and if $W_i > 0$ than $I_i = 0$ and if $I_i > 0$ than $W_i = 0$, for $i > 1$. Expressions (3.2) and (3.3) are due to the Lindley recursion (Lindley, 1952) and are illustrated in Figure 3.2. In this figure time is on the horizontal axis and what is stacked vertically happens at the same time. The left hand side represents the situation for expression (3.2) and the right hand side represents the situation for expression (3.3). In both situations time slot $i-1$ starts with waiting time for customer $i-1$, which is equal to the exceeding service time of customer $i-2$. This waiting time is followed by its service time. Together this is the sojourn time for customer $i-1$. The sojourn time on the left hand side is smaller than the size of the slot and hence we have idle time, which is denoted by I_i. The sojourn time on the right is larger than the size of the slot and hence customer i has to wait on the service of its predecessor. So the time the previous customers exceeds its time slot, is the waiting time for the next customer.

As we explained in the introduction our objective is to assure little waiting time for the customer and the server. In case of simple linear loss, this comes down to designing an appointment schedule with arrival times t_i for the customers such that,

$$\sum_{i=1}^{n} \Big(\mathbb{E}I_i(t_1, \ldots, t_i) + \mathbb{E}W_i(t_1, \ldots, t_i) \Big) \tag{3.4}$$

is minimal, i.e., our problem becomes,

$$\min_{t_2,\ldots,t_n} \sum_{i=2}^{n} \Big(\mathbb{E}I_i(t_1, \ldots, t_i) + \mathbb{E}W_i(t_1, \ldots, t_i) \Big). \tag{3.5}$$

The value of function (3.4) represents the total expected idle time of the server and the total expected waiting time of all the customers together. Note that the sum starts at $i = 2$, since $I_1 = W_1 = 0$. In some cases the idle time of a server is more important than the waiting time of the customer. For example an MRI scanner can cost millions, therefore the importance of this scanner not being idle is larger than reducing the patient's waiting time. For this we introduce normalized weights, α and $(1-\alpha)$, for both random variables:

$$\min_{t_2,\ldots,t_n} \sum_{i=2}^{n} \Big(\alpha\mathbb{E}I_i(t_1, \ldots, t_i) + (1-\alpha)\mathbb{E}W_i(t_1, \ldots, t_i) \Big), \qquad \alpha \in [0, 1].$$

Secondly, the evaluation of the waiting time of a customer does not necessarily show a linear pattern. That means that two customers waiting one hour does not have to be evaluated equally to one customer waiting two hours. Summarizing, the designer of the schedule might not wish to measure both random variables linearly. For these purposes, we introduce nondecreasing continuous functions $g(\cdot)$ and $h(\cdot)$ with $g(0) = h(0) = 0$. In these functions we can include the weights. Hence, we arrive at:

$$\min_{t_2,\ldots,t_n} \sum_{i=2}^{n} \Big(\mathbb{E}[g(I_i(t_1,\ldots,t_i))] + \mathbb{E}[h(W_i(t_1,\ldots,t_i))] \Big).$$

Until this moment we used $t_i, 1 \leq i \leq n$ as variables. For the main part of this thesis we will use $x_i, 1 \leq i \leq n-1$ as variables. Recall that, $x_i = t_{i+1} - t_i$, is the interarrival time for customer i. The general loss function (LF) for an appointment schedule now becomes:

$$LF(x_1,\ldots,x_{n-1}) = \sum_{i=2}^{n} \Big(\mathbb{E}[g(I_i(x_1,\ldots,x_{i-1}))] + \mathbb{E}[h(W_i(x_1,\ldots,x_{i-1}))] \Big).$$

Sometimes we prefer to analyze the loss function per slot, this is noted by LF_i with,

$$\begin{aligned} LF(x_1,\ldots,x_{n-1}) &= \sum_{i=2}^{n} LF_i(x_1,\ldots,x_{i-1}), \\ &= \sum_{i=2}^{n} \Big(\mathbb{E}[g(I_i(x_1,\ldots,x_{i-1}))] \\ &\quad + \mathbb{E}[h(W_i(x_1,\ldots,x_{i-1}))] \Big). \end{aligned} \tag{3.6}$$

Further on in this thesis we do not explicitly denote the dependence of the waiting and idle time on the previous slots. Instead we simply denote them with I_i and W_i.

3.3 Various Loss Functions

3.3.1 General Loss Function

Upon deriving an appointment schedule one has to choose its loss function. To do so, we introduced functions $g(\cdot)$ and $h(\cdot)$ to give weights and powers

to the random variables: idle time and waiting time. However, it depends on the problem approach if one is able to work with the loss function of your choice. In case of simulation one is truly free in choosing non-decreasing functions $g(\cdot)$ and $h(\cdot)$ with $g(0) = 0$ and $h(0) = 0$. In case of the analytical approach things become harder.

It is for example difficult to deal with the case where g and h have different powers. The loss functions that are still applicable in the analytical approach are those where g and h have the same power with the possibility of adding weights. Hence for the analytical approach one may use:

$$LF(x_1, \ldots, x_{n-1}) = \sum_{i=2}^{n} \Big(\alpha \mathbb{E}[I_i^k(x_1, \ldots, x_{i-1})] + (1-\alpha)\mathbb{E}[W_i^k(x_1, \ldots, x_{i-1})] \Big), \quad \alpha \in [0,1], k \in \mathbb{N}_+.$$

Within this class of functions the linear loss function, as we will explain in Section 3.3.3, is very well suited to make a correspondence to the systems costs, this is well illustrated in Vanden Bosch et al. (1999). The quadratic loss function is widely used because of its mathematical attraction, as we will show in Section 3.3.2 and as is illustrated in Kemper et al. (2011). Both type of loss functions can be used with and without weights.

3.3.2 Quadratic Loss Function

A common choice for $g(\cdot)$ and $h(\cdot)$ is the quadratic function, $g(x) = h(x) = x^2$. This function is used in for instance, Schild and Fredman (1961) and Kemper et al. (2011). The loss function is denoted by Q. We obtain the following loss function:

$$LF_i(x_1, \ldots, x_{i-1}) = Q_i(x_1, \ldots, x_{i-1}) = \mathbb{E}[I_i^2] + \mathbb{E}[W_i^2], \quad i > 1. \tag{3.7}$$

Using (3.2) and (3.3) we get:

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

So (3.7) now becomes.

$$\begin{aligned} Q_i(x_1, \ldots, x_{i-1}) &= \mathbb{E}[I_i^2 + W_i^2], \\ &= \mathbb{E}(S_{i-1} - x_{i-1})^2, \\ &= \mathbb{E}S_{i-1}^2 - 2x_{i-1}\mathbb{E}S_{i-1} + x_{i-1}^2. \end{aligned}$$

The objective function we want to minimize becomes:

$$Q(x_1, \ldots, x_{n-1}) = \sum_{j=1}^{n-1} \mathbb{E}[(S_j - x_j)^2]. \tag{3.8}$$

Note that we change index in this expression.

In case one wants to add weights, function (3.7) becomes:

$$\begin{aligned} Q_i(x_1, \ldots, x_{i-1}) &= \mathbb{E}[\alpha I_i^2 + (1-\alpha) W_i^2], \\ &= \alpha \mathbb{E}[I_i^2 + W_i^2] + (1 - 2\alpha)\mathbb{E}W_i^2, \ \ \alpha \in [0, 1]. \end{aligned}$$

We choose to put the weight difference separately for the waiting time, because we do already need to derive expressions for the waiting time.

3.3.3 Linear Loss Function

Another common choice for $g(\cdot)$ and $h(\cdot)$ is the identity function $g(x) = x = h(x)$. This function is used in for instance, Wang (1999),Vanden Bosch et al. (1999) and Kaandorp and Koole (2007). The loss function is denoted by L. We obtain the following loss function:

$$LF_i(x_1, \ldots, x_{i-1}) = L_i(x_1, \ldots, x_{i-1}) = \mathbb{E}I_i + \mathbb{E}W_i, \quad i > 1. \tag{3.9}$$

Here again by (3.2) and (3.3) we get:

$$W_i + I_i = |S_{i-1} - x_{i-1}|, \quad i > 1.$$

Hence, the objective function we want to minimize in this case becomes:

$$L(x_1, \ldots, x_{n-1}) = \sum_{j=1}^{n-1} \mathbb{E}|S_j - x_j|.$$

In case one wants to add weights, function (3.9) becomes:

$$\begin{aligned} L_i(x_1, \ldots, x_{i-1}) &= \mathbb{E}[\alpha I_i + (1-\alpha) W_i], \\ &= \alpha \mathbb{E}[I_i + W_i] + (1 - 2\alpha)\mathbb{E}W_i, \ \ \alpha \in [0, 1]. \end{aligned}$$

Along the same arguments as in the previous section, we choose to put the weight difference separately for the waiting time.

3.3.4 Completion Time

For a schedule it is also important to finish on time. Hence, one does not only wants to minimize expected waiting and idle time but also minimize the expected total completion time of the schedule. This can be done by simply incorporating the total expected completion time in the loss function. This is also done in: Fries and Marathe (1981), Wang (1999), Vanden Bosch et al. (1999) and Kaandorp and Koole (2007). This results in:

$$LF(x_1, \ldots, x_{n-1}) = \sum_{i=2}^{n} \Big(\mathbb{E}[g(I_i)] + \mathbb{E}[h(W_i)] \Big) + \mathbb{E} f\Big(S_n + \sum_{i=1}^{n-1} x_i \Big),$$

with $f(\cdot)$ a nondecreasing continuous function.

Likewise, the systems lateness can be incorporated. This is the expected time the schedule exceeds the reserved operating time of the server:

$$\begin{aligned} LF(x_1, \ldots, x_{n-1}) \;=\; & \sum_{i=2}^{n} \Big(\mathbb{E}[g(I_i)] + \mathbb{E}[h(W_i)] \Big) \\ & + \max\Big(0, f\Big(t_{\text{start}} - t_{\text{end}} + \mathbb{E}S_n + \sum_{i=1}^{n-1} x_i \Big)\Big), \end{aligned}$$

with $t_{\text{start}} = t_1 = 0$, the time the schedule starts and t_{end} the time the service provider and its operating space is reserved for a following session. Exceeding of this time leads to penalties.

3.4 Wang's Extended Algorithm for Sojourn Times

In the next chapter we start to investigate the characteristics of appointment scheduling with exponential service times. To do so, we use the results from Wang (1999). This article studies the problem of sequencing and scheduling n customers for a single-server system and showed that the optimal schedule can be obtained by solving a set of nonlinear equations. We use these results to efficiently derive expressions for the first moment of the sojourn times and extend these results to be able to derive expressions for the second moment of the sojourn times as well (in case $\mu_1 = \cdots = \mu_n = 1$). The algorithm makes use of the memoryless property of the exponential distribution. The

analytical consequences of this property for appointment scheduling is elaborated at the end of Chapter 1.

Following the results, we introduce the extended algorithm from Wang. Let d_n be the cumulative distribution function of the sojourn time for customer n, u a vector of order n with all elements 1 and $P_n(t)$ a probability vector of dimension n, which is introduced as an auxiliary vector. The coordinates in P_n represent the probabilities of a number of predecessors still in service.

Let us define:

$$d_n(t) = 1 - < P_n(t), u >,$$

with $< \cdot, \cdot >$ the dot product. The following expression is to be computed before the algorithm starts:

$$P_1(t) = \exp(M_1 t),$$

with

$$M_n = \begin{pmatrix} -\mu_1 & \mu_1 & & 0 \\ & -\mu_2 & \mu_2 & \\ & \cdots & & \\ 0 & & & -\mu_n \end{pmatrix}.$$

The following step of the algorithm is the computation of $P_n(t)$ for n ranging from 2 to the number of customers in the schedule:

$$P_n(t) = [P_{n-1}(x_{n-1}), d_{n-1}(x_{n-1})] \exp(M_n t).$$

Note that $\exp(M_n t)$ is a n x n matrix and $[., \ldots, .]$ is a row vector.

With these results we are able to derive expressions for the first and second moment of the sojourn time, $\mathbb{E}S_n$ and $\mathbb{E}S_n^2$, which we will need in the upcoming chapters:

$$\begin{aligned} \mathbb{E}S_1 &= \frac{1}{\mu_1}, \\ \mathbb{E}S_i &= [P_{n-1}(x_{n-1}), d_{n-1}(x_{n-1})]\beta_n, \\ \mathbb{E}S_1^2 &= \frac{2}{\mu_1^2}, \\ \mathbb{E}S_i^2 &= [P_{n-1}(x_{n-1}), d_{n-1}(x_{n-1})]\zeta_n, \end{aligned}$$

with:

$$\begin{aligned} \beta_n &= [n, n-1, \ldots, 1], \\ \zeta_n &= [n(n+1), (n-1)n, \ldots, 2]. \end{aligned}$$

This applies in a setting with i.i.d. service times with $\mu = 1$.

For arbitrary μ, β_n represents the sum of the expected service times of customers still in the system:

$$\beta_n = [\sum_{i=1}^{n} \frac{1}{\mu_i}, \sum_{i=1}^{n-1} \frac{1}{\mu_i}, \ldots, \frac{1}{\mu_n}].$$

So far, an expression for ζ_n with arbitrary μ is not yet known. To obtain such an expression turns out to be a hard problem. By hand we are able to derive expressions for the second moment of the waiting time for customer 1 and 2, for arbitrary μ. By some algebra we obtain an expression for ζ_2 and also for the last two coordinates in ζ_n. These results can be found in Appendix A.2. In general, ζ_n does not represent the sum of second moments of service times of customers in the system, but wil be a combination of first and second moments.

The algorithm in this section is written in a Matlab function. The code, including a clarification, can be found in Appendix A.1.

3.5 Simultaneous Optimization Approach

The simultaneous approach is an often used optimization approach. For most problems one intends to find the global minimum, implicitly one will use the simultaneous optimization approach to do so. In this approach all the variables are simultaneously varied to obtain the global minimum of your objective function. In our problem setting this implies that all the interarrival times are simultaneously varied such that the objective function attains its global minimum. Our objective function is the general loss function, a sum of loss functions per slot. Let us recall expression (3.6):

$$LF(x_1, \ldots, x_{n-1}) = \sum_{i=2}^{n} LF_i(x_1, \ldots, x_{i-1}).$$

This function is convex in all its variables, see Appendix A.5. Hence, there is one local minimum, which is the global minimum. We call this minimum $\hat{x}$, being the vector with the optimal interarrival times $[\hat{x}_1, \ldots, \hat{x}_{n-1}]$. In $\hat{x}$ the derivative in each variable is equal to zero, i.e.,

$$\frac{d}{dx_i} LF(\hat{x}_1, \ldots, \hat{x}_{n-1}) = 0, \qquad \text{for } i = 1, \ldots, n-1. \tag{3.10}$$

The loss function per slot is a function of the previous slots and hence does not depend on all the variables. For that reason the derivative of (3.10) with respect to x_i results in:

$$\frac{d}{dx_i}\sum_{j=2}^{n} LF_j(x_1,\ldots,x_{j-1}) = \frac{d}{dx_i}\sum_{j=i}^{n} LF_j(x_1,\ldots,x_{j-1}).$$

Summarizing the previous expressions, the following holds:

$$\begin{aligned}
\frac{d}{dx_1}\sum_{j=2}^{n} LF_j(x_1,\ldots,x_{j-1})|_{x=\hat{x}} &= 0\\
\frac{d}{dx_2}\sum_{j=3}^{n} LF_j(x_1,\ldots,x_{j-1})|_{x=\hat{x}} &= 0\\
\vdots \qquad\qquad & \quad \vdots\\
\frac{d}{dx_{n-1}} LF_n(x_1,\ldots,x_{n-1})|_{x=\hat{x}} &= 0
\end{aligned}$$

Hence, we need to solve the following system of equations:

$$\begin{aligned}
\frac{d}{dx_1}\sum_{j=2}^{n} LF_j(x_1,\ldots,x_{j-1}) &= 0,\\
\frac{d}{dx_2}\sum_{j=3}^{n} LF_j(x_1,\ldots,x_{j-1}) &= 0,\\
\vdots \qquad\qquad & \quad \vdots\\
\frac{d}{dx_{n-1}} LF_n(x_1,\ldots,x_{n-1}) &= 0.
\end{aligned} \tag{3.11}$$

Solving this system yields the exact, global minimum.

Now we will take a closer look at the optimization characteristics of the sequential approach.

3.6 Sequential Optimization Approach

The sequential optimization approach is proposed in Kemper et al. (2011). Following this approach, we first compute the interarrival time for customer

1 by minimizing the loss function for slot 1. So

$$\min_{x_1} LF_2(x_1).$$

Resulting in $\tilde{x}_1$. Secondly, we compute the interarrival time for customer 2 by minimizing the loss function for slot 2, given the previously computed $\tilde{x}_1$. Hence,

$$\tilde{x}_2 = \min_{x_2} LF_3(\tilde{x}_1, x_2).$$

We continue this process until we arrive at customer n.

$$\tilde{x}_{n-1} = \min_{x_{n-1}} LF_n(\tilde{x}_1, \ldots, \tilde{x}_{n-2}, x_{n-1}).$$

If one would substitute the expressions in each other, one arrives at the following cumbersome expression:

$$\min_{x_{n-1}} LF_n(\arg\min_{x_1} LF_2(x_1), \arg\min_{x_2} LF_3(\arg\min_{x_1} LF_2(x_1), x_2), \ldots, x_{n-1}).$$

The loss function for slot i, $LF_i(\tilde{x}_1, \ldots, \tilde{x}_{i-1}, x_i)$, is convex in x_i, see Appendix A.5. Hence, in order to solve this minimization problem, the following system of equations needs to be solved sequentially:

$$\begin{aligned}
&\frac{d}{dx_1} LF_2(x_1) = 0,\\
&\frac{d}{dx_2} LF_3(\tilde{x}_1, x_2) = 0,\\
&\vdots \qquad\qquad\qquad \vdots\\
&\frac{d}{dx_{n-1}} LF_n(\tilde{x}_1, \ldots, \tilde{x}_{n-2}, x_{n-1}) = 0.
\end{aligned}$$

Note that the main difference with the simultaneous approach is that now we have $n-1$ minimization problems of 1 dimension.

For linear and quadratic loss functions, the sequential approach yields optimal interarrival times:

$$\tilde{x}_i = \mathbb{E}S_i, \qquad \text{in case of a quadratic loss function;} \quad (3.12)$$

$$\tilde{x}_i = F_{S_i}^{-1}(\frac{1}{2}), \qquad \text{in case of a linear loss function,} \quad (3.13)$$

with $F_{S_i}^{-1}$ the inverse of the distribution function of the sojourn time of customer i. Both results are derived in respectively Section 4.1.3 and 4.2.3. For more details, see Kemper et al. (2011).

In the next chapter we have a closer look at both optimization approaches in a setting with i.i.d. exponential(μ=1) service times.

Chapter 4

Analytic Approaches: Simultaneous versus Sequential

In this section, we compare the performance of the sequential approach with the simultaneous approach. The difference between these optimization approaches is that the sequential approach does not take future customers into account, while the simultaneous approach does. As a consequence the resulting schedules differ, which we elaborate in this chapter.
We compare both optimization approaches with i.i.d. exponentially distributed service times with $\mu = 1$, unless stated otherwise.

4.1 Quadratic Loss Function

We start our comparison with the unweigthed quadratic loss function. Recall this function from page 15:

$$Q(x_1, \ldots, x_{n-1}) = \sum_{j=1}^{n-1} \mathbb{E}[(S_j - x_j)^2].$$

4.1.1 A Schedule with Three Customers

First we consider the case where we have three customers with, $\mu_1 = \mu_2 = \mu_3 = 1$. The first customer arrives at $t_1 = 0$. According to Wang's extended

algorithm we obtain:

$$\begin{aligned}\mathbb{E}S_1 &= 1,\\ \mathbb{E}S_1^2 &= 2,\\ \mathbb{E}S_2 &= 1+e^{-x_1},\\ \mathbb{E}S_2^2 &= 2+4e^{-x_1}.\end{aligned}$$

Using these, the expressions for the loss function per slot becomes:

$$\begin{aligned}Q_2(x_1) &= \mathbb{E}S_1^2 + x_1^2 - 2x_1\mathbb{E}S_1 = 2 + x_1^2 - 2x_1,\\ Q_3(x_1,x_2) &= \mathbb{E}S_2^2 + x_2^2 - 2x_2\mathbb{E}S_2 = 2 + 4e^{-x_1} + x_2^2 - 2x_2(1+e^{-x_1}).\end{aligned}$$

According to the sequential approach $\tilde{x}_1$ is given by the solution of

$$\frac{d}{dx_1}Q_2(x_1) = 0.$$

Hence, $2x_1 - 2 = 0$, so,

$$\tilde{x}_1 = 1.$$

$\tilde{x}_2$ is given by the solution of

$$\frac{d}{dx_2}Q_3(\tilde{x}_1, x_2) = 0.$$

Hence, $2x_2 - 2(1+e^{-\tilde{x}_1}) = 0$, so

$$\tilde{x}_2 = 1 + e^{-\tilde{x}_1} = 1 + e^{-1} \approx 1.37. \tag{4.1}$$

Note that we could have derived these values from (3.12) much faster.

The total loss function is $Q(x_1,x_2) = Q_2(x_1) + Q_3(x_1,x_2)$ which you can see in Figure 4.1. The plot shows the convexity of $Q(x_1,x_2)$ and we see a global minimum. We now minimize according to the simultaneous approach to obtain this global minimum. So we take the derivative in both variables equal to zero and solve the following system of equations:

$$\begin{aligned}\frac{dQ}{dx_1} &= 2x_1 - 2 - 4e^{-x_1} + 2x_2e^{-x_1} = 0,\\ \frac{dQ}{dx_2} &= 2x_2 - 2 - 2e^{-x_1} = 0.\end{aligned}$$

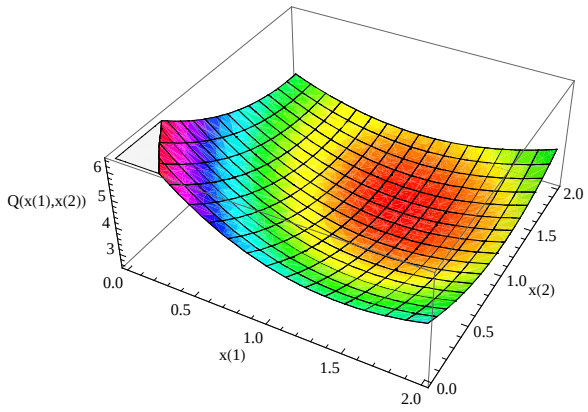


Figure 4.1: Plot of the loss function as a function of x_1 and x_2. Quadratic loss, n=3, i.i.d exponential(μ=1) service times.

The second equation leads to:

$$x_2 = 1 + e^{-x_1}. \tag{4.2}$$

Note that this is the same expression as in the sequential case (4.1), as we could have expected from (3.11) and (3.12). So the difference between both optimization approaches, in this example, is the choice for the size of x_1. In the sequential case this value is already known and equal to 1. For the simultaneous case we still have to compute this value. We substitute expression (4.2) in $Q(x_1, x_2)$ and arrive at a convex function of only one variable, x_1, see Figure 4.2. One can see that the optimal value of x_1 in the simultaneous case is larger than $\tilde{x}_1$. This results, as the figure illustrates, in a lower value of the loss function (Q). These results are summarized in Table 4.1. This table shows the results for both optimization approaches. Also one can see that for the losses per slot, $Q_2(\hat{x}_1) > Q_2(\tilde{x}_1)$ but $Q_2(\hat{x}_1, \hat{x}_2) < Q_3(\tilde{x}_1, \tilde{x}_2)$ what results in $Q(\hat{x}_1, \hat{x}_2) < Q(\tilde{x}_1, \tilde{x}_2)$. We will continue the comparison with a larger appointment schedule in the next section.

	x_1	x_2	$Q_2(x_1)$	$Q_3(x_1, x_2)$	$Q(x_1, x_2)$
sim	1.21	1.30	1.04	1.51	2.55
seq	1.00	1.37	1.00	1.60	2.60

Table 4.1: The differences between the sequential and simultaneous approach for the quadratic loss function, $n = 3$, $\mu_1 = \mu_2 = \mu_3 = 1$

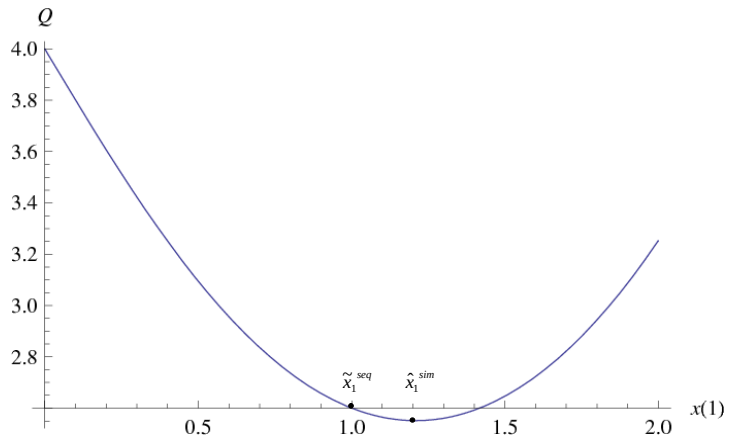


Figure 4.2: Plot of the loss function $Q(x_1, 1 + e^{-x_1})$ as a function of x_1. Quadratic loss, n=3, i.i.d exponential(μ=1) service times.

$\hat{x}_1$	$\hat{x}_2$	$\hat{x}_3$	$\hat{x}_4$	$\hat{x}_5, \ldots, \hat{x}_{74}$	$\hat{x}_{75}$	$\hat{x}_{76}$	$\hat{x}_{77}$	$\hat{x}_{78}$	$\hat{x}_{79}$
1.36	1.70	1.78	1.82	1.85	1.81	1.79	1.75	1.66	1.41

$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_4$	$\tilde{x}_5$	$\tilde{x}_6$	$\tilde{x}_7$	$\tilde{x}_8$	$\tilde{x}_9, \ldots, \tilde{x}_{79}$
1.00	1.36	1.48	1.52	1.54	1.56	1.57	1.57	1.58

Table 4.2: The actual value of the slots of the schedules from Figure 4.3. Quadratic loss, n=80, i.i.d. exponential(μ=1) service times.

4.1.2 More Customers, Scaling the Problem

In the previous section we analyzed the difference between the two optimization approaches in a setting with only three customers. We now analyze the difference with more customers. For this we use Wang's extended algorithm, see Wang (1999) and Section 3.4. With this we are still able to solve this minimization problem analytically and exact. We use Matlab to do this. The code for this extended algorithm can be found in Appendix A.1. The code computes a value of the quadratic loss function for a certain appointment schedule, this schedule is presented as a vector $\vec{x}$ with interarrival times x_i. Next, we are able to minimize this function using `fminsearch`, a multidimensional unconstrained nonlinear minimization function using the Nelder-Mead simplex method. D'Errico (2005), extended this function to a

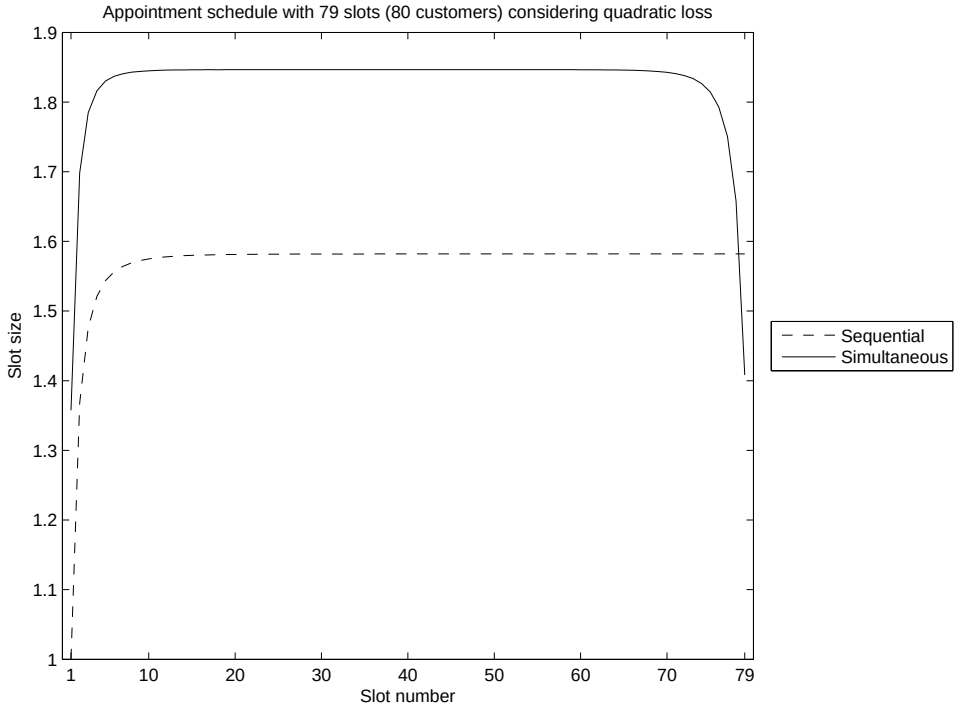


Figure 4.3: The slot sizes of an appointment schedule, optimized using the simultaneous and the sequential approach. Quadratic loss, n=80, i.i.d. exponential(μ=1) service times.

constrained version, which we used for our multidimensional optimization problem.

Figure 4.3 shows the problem optimized, using the sequential and the simultaneous approach. We see the slot size of the various slots on the y-axis and the slot number on the x-axis. You can see that the slot sizes derived by the sequential approach are smaller than the ones derived by the simultaneous approach, except for the final few slots. After customer 10 the slot sizes of the sequential approach stabilize to a value of approximately 1.58. The slot sizes of the simultaneous approach also stabilizes after customer 10 to a value of approximately 1.84, however, towards the end of the schedule, the slot sizes decrease. This results in a dome shaped graph and this phenomenon illustrates what we stated before, that the simultaneous approach

takes future arrivals into account. At first, it takes future services into account, resulting in larger slots. Towards the end of the schedule there are less future arrivals, resulting in a decrease of slot size. This is in contrast to the sequential approach where we do not see this behavior. In the next section we come back to the values of stabilized slot sizes.

Next, we analyze the differences between both approaches on the value of the objective function, which is the value of function (3.8), $Q(x_1, \ldots, x_{n-1})$. This difference is more interesting since it is the systems loss we want to minimize. In Figure 4.4 we show these values for a range of problem dimensions for both optimization approaches. For all the considered problem dimensions, the objective of simultaneously derived schedules is smaller than for sequentially derived schedules. The percentage difference between both is also shown in the figure. This value seems to stabilize at approximately 20%. In Section 4.3 we will prove that as the dimension increases we trully have convergence to this percentage.

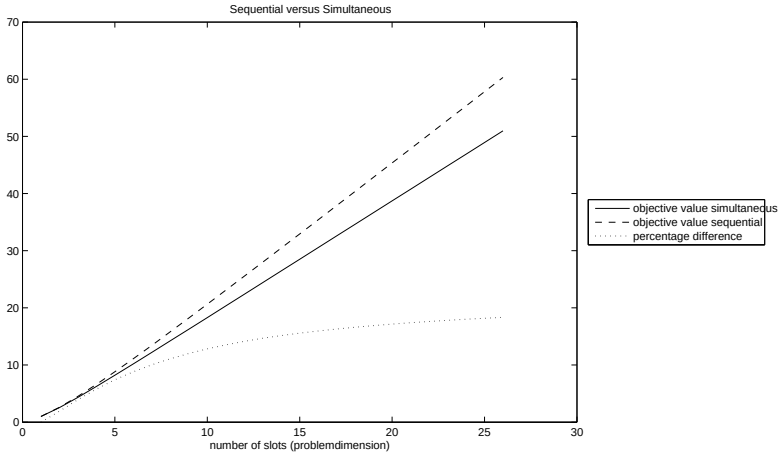


Figure 4.4: The objective values of both optimization approaches and their percentage differences. Quadratic loss, $0<n<28$, i.i.d exponential(μ=1) service times.

4.1.3 Limiting Properties

In the previous section, we have seen that in case of large schedules, (i.e., many customers) the size of the majority of the slots is equal to approxi-

mately 1.58 for the sequential approach and equal to approximately 1.84 for the simultaneous approach. In this section we prove these values.

We have seen that if n increases the number of slots with the same size increases, for both simultaneously and sequentially derived schedules. So for large n most slots have equal size. Hence, for n large enough, the total loss function can be approximated by:

$$Q(x_1, \ldots, x_n) = \sum_{i=1}^{n} \mathbb{E}(S_i - x_i)^2 \approx n\mathbb{E}(S(x) - x)^2. \tag{4.3}$$

Clearly, convexity is preserved since (4.3) is a sum of convex functions. We take the derivative and set it equal to zero to obtain an expression for the optimal slot size:

$$\frac{d}{dx} n\mathbb{E}(S(x) - x)^2 = n(\frac{d}{dx}\mathbb{E}S_x^2 + 2x - 2\mathbb{E}S(x) - 2x\frac{d}{dx}\mathbb{E}S(x)) = 0. \tag{4.4}$$

The scheduling setting we study in this thesis can be seen as a queue with deterministic arrivals, exponential service times and 1 server, i.e., a $D|M|1$ queue. From Tijms (1986) we know that for the steady state of such a $D|M|1$ queue we have:

$$\mathbb{E}S(x) = \frac{1}{\mu(1-\rho_x)}, \tag{4.5}$$

$$\mathbb{E}S^2(x) = \frac{2}{\mu^2(1-\rho_x)^2}. \tag{4.6}$$

Where $\rho := \rho_x$ is the occupation rate of the queue, with $\rho \in (0, 1)$ the unique solution of:

$$e^{-\mu(1-\rho)x} = \rho, \tag{4.7}$$

with x the slot size of the stabilized system. More details can be found in Appendix A.4

For n large we may assume that our system is in steady state. With steady state we mean that optimal interarrival times do not change if n increases. This assumption is supported by the results from Section 4.1.2, so we may use the aforementioned results from Tijms, expressions (4.5) and (4.6). Next, to obtain a solution for equation (4.4) we need the derivatives of these results:

$$\frac{d}{dx}\mathbb{E}S(x) = \frac{\rho_x^{'}}{\mu(1-\rho_x)^2}, \tag{4.8}$$

$$\frac{d}{dx}\mathbb{E}S^2(x) = \frac{4\rho_x^{'}}{\mu^2(1-\rho_x)^3}. \tag{4.9}$$

The previous expressions indicate that we need an expression for $\rho_x^{'}$. By implicit differentiation of (4.7) we arrive at:

$$\rho_x^{'} = \frac{\mu\rho_x(\rho_x - 1)}{1 - \mu\rho_x x}. \tag{4.10}$$

Via (4.7) we can express x in terms of ρ_x:

$$x = \frac{\log \rho_x}{\mu(\rho_x - 1)}. \tag{4.11}$$

Now we can substitute (4.8), (4.9), (4.10) and (4.11) into (4.4):

$$\begin{aligned}
0 &= \frac{4\rho_x^{'}}{\mu^2(1-\rho_x)^3} + 2x - 2\frac{1}{\mu(1-\rho_x)} - 2x\frac{\rho_x^{'}}{\mu(1-\rho_x)^2}, \\
&= \frac{4\mu\rho_x(\rho_x - 1)}{(1-\mu\rho_x x)\mu^2(1-\rho_x)^3} + 2x - 2\frac{1}{\mu(1-\rho_x)} - 2x\frac{\mu\rho_x(\rho_x-1)}{(1-\mu\rho_x x)\mu(1-\rho_x)^2}, \\
&= \frac{4\rho_x}{(1-\rho_x+\rho_x\log\rho_x)\mu(\rho_x-1)} + \frac{2\log\rho_x + 2}{\mu(\rho_x-1)} + \frac{2\rho_x\log\rho_x}{(1-\rho_x+\rho_x\log\rho_x)\mu(\rho_x-1)}, \\
&= \frac{2}{\mu(1-\rho_x)(1-\rho_x+\rho_x\log\rho_x} \left[1 + \log\rho_x + \rho_x(1 + \log\rho_x + \log^2\rho_x)\right].
\end{aligned}$$

This equation implies that:

$$0 = \rho_x + (1 + \log\rho_x)(1 + \rho_x \log\rho_x).$$

Note that this expression does not contain μ, so the optimal value of $\hat{\rho}_x$ is independent of the value of μ. Solving this equation leads to $\hat{\rho}_x \approx .25$ and hence by (4.11) we get:

$$\hat{x} = \frac{\log\rho_x}{\mu(\rho - 1)} \approx \frac{1.85}{\mu}.$$

For the sequential optimization approach we obtain a different result. In this approach the derivatives in equation (4.4) are zero. Solving this yields the result (3.12) on page 20,

$$\tilde{x} = \mathbb{E}S(x).$$

Substituting (4.5) and (4.11) into this result we get:

$$\frac{\log\rho_x}{\mu(\rho_x - 1)} = \frac{1}{\mu(1-\rho_x)},$$

which solves for $\tilde{\rho}_x = e^{-1}$. Note that this value is also independent of μ. For the slot size we get:

$$\tilde{x} = \frac{\log \tilde{\rho}_x}{\mu(\tilde{\rho}_x - 1)} = \frac{e}{\mu(e-1)}\Big|_{\mu=1} \approx 1.58.$$

4.1.4 Sequential with Simultaneous Start

To understand the difference between both optimization approaches even better we will derive a schedule starting with simultaneously derived slots followed by sequentially derived slots, see Figure 4.5. For clients $1, \ldots, 10$ we use the interarrival times from a simultaneous optimized problem ($n = 33$). After customer 10 we schedule the customers according to the sequential approach. The result is illustrative for the differences between both approaches. As we start the sequential approach, the interarrival time drops and continues with the same interarrival times as in the sequential approach, see dotted line in Figure 4.5. This happens because it does not take future customers into account. Slot x_{10} is computed according to the sequential approach. Hence $\tilde{x}_{10} = \mathbb{E}S_{10}(\hat{x}_1, \ldots, \hat{x}_9)$. The slots 1-9, simultaneously derived, are larger than they would be when sequentially derived. This results in:

$$\mathbb{E}S_{10}(\hat{x}_1, \ldots, \hat{x}_9) < \mathbb{E}S_{10}(\tilde{x}_1, \ldots, \tilde{x}_9).$$

4.2 Linear Loss Function

We start our comparison with the unweigthed linear loss function. Recall this function from page 15

$$L(x_1, \ldots, x_{n-1}) = \sum_{j=1}^{n-1} \mathbb{E}|S_j - x_j|.$$

4.2.1 A Schedule with Three Customers

We consider the case where we have three customers and $\mu_1 = \mu_2 = \mu_3 = 1$. The first customer arrives at $t_1 = 0$. According to Wang's algorithm the

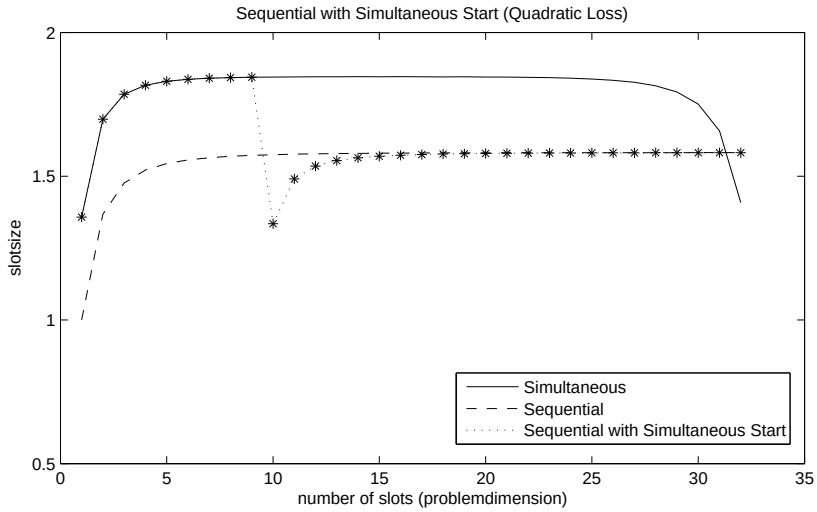


Figure 4.5: First 9 clients according to simultaneous approach continued by sequential approach. Quadratic loss, n=33, i.i.d exponential(μ=1) service times.

loss function per slot becomes:

$$\begin{aligned} L_2(x_1) &= \mathbb{E}|S_1 - x_1| = -1 + x_1 + 2e^{-x_1}, \\ L_3(x_1, x_2) &= \mathbb{E}|S_2 - x_2| = -1 + x_2 + 2e^{-x_2} - e^{-x_1} + 2(1 + x_2)e^{-x_1 - x_2}. \end{aligned}$$

According to the sequential approach $\tilde{x}_1$ is given by the solution of

$$\frac{d}{dx_1} L_2(x_1) = 0.$$

Hence, $1 - 2e^{-x_1} = 0$, so,

$$\tilde{x}_1 = \log 2 \approx .69.$$

The solution of $\tilde{x}_2$ is given by :

$$\frac{d}{dx_2} L_3(\tilde{x}_1, x_2) = 0.$$

Hence, $1 - 2e^{-\tilde{x}_1 - x_2}(e^{\tilde{x}_1} + x_2) = 0$, so

$$\tilde{x}_2 = -2 - \text{product logarithm}[-1, -1/e^2] \approx 1.15.$$

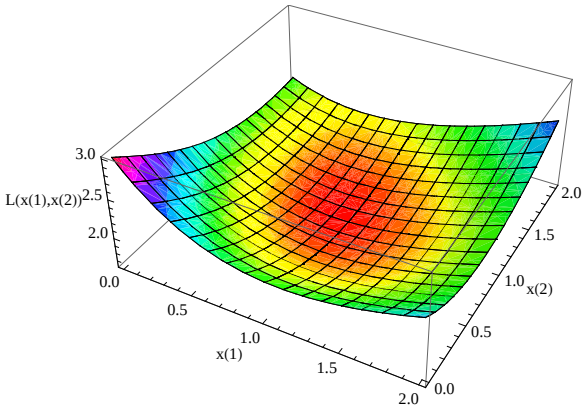


Figure 4.6: Plot of the loss function as a function of x_1 and x_2. Linear loss, n=3, i.i.d exponential(μ=1) service times.

This result is obtained with Mathematica, a mathematical software package. The product logarithm is also called the Omega function or the Lambert W function. It is defined as the solution for w in $z = we^w$, with $z \in \mathbb{R}$ Note that we could have derived these values from (3.13) much faster.

The total loss function is $L(x_1, x_2) = L_2(x_1) + L_3(x_1, x_2)$, which you can see in Figure 4.6. The plot shows again the convexity of $L(x_1, x_2)$ and we see a global minimum. We will now minimize according to the simultaneous approach to obtain this global minimum. So we take the derivative in both variables equal to zero and solve the following system of equations:

$$\begin{aligned} \frac{dL}{dx_1} &= e^{-x_1-x_2}(e^{x_2}(e^{x_1}-1)-2-2x_2) = 0, \\ \frac{dL}{dx_2} &= 1-2e^{-x_1-x_2}(e^{x_1}+x_2) = 0. \end{aligned}$$

Solving this system leads to

$$\hat{x}_1 = .89, \qquad \hat{x}_2 = 1.05.$$

The results are summarized in Table 4.3. This table shows the results for both optimization approaches. Also one can see that for the losses per slot, $L_2(\hat{x}_1) > L_2(\tilde{x}_1)$ but $L_2(\hat{x}_1, \hat{x}_2) < L_3(\tilde{x}_1, \tilde{x}_2)$ what results in $L(\hat{x}_1, \hat{x}_2) < L(\tilde{x}_1, \tilde{x}_2)$. We will continue the comparison with a larger appointment schedule in the next section.

	x_1	x_2	$L_2(x_1)$	$L_3(x_1, x_2)$	$L(x_1, x_2)$
sim	.89	1.05	.71	.93	1.64
seq	.69	1.15	.69	.96	1.66

Table 4.3: The differences between the sequential and simultaneous approach in resulting appointment schedules for the linear loss function, n=3, μ_1=μ_2=μ_3=1

4.2.2 More Customers, Scaling the Problem

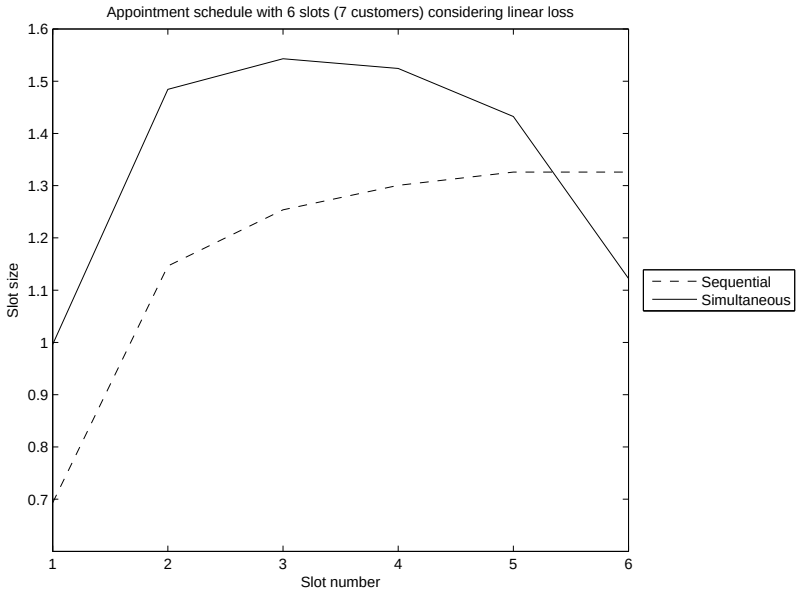


Figure 4.7: Slot sizes of an appointment schedule, optimized using the simultaneous and the sequential approach. Linear loss, n=7, i.i.d exponential(μ=1) service times.

As in the quadratic case we can also scale the problem in the linear case. We used Matlab to solve this minimization problem.

As an example we optimized a schedule for 7 customers in Figure 4.7. You can see that the slot sizes derived by the sequential approach are smaller than the ones derived by the simultaneous approach, except for the final few slots. Again we see the dome shape for the interarrival times, as we

saw with the quadratic loss. Again, the slot size of the sequential approach stabilizes as n increases. The size of this slot is approximately 1.39. The slot size of the simultaneous approach stabilizes to a value of approximately 1.68. Again this applies for the majority of the slots in the middle of the schedule. We come back to the stated values in the next section.

4.2.3 Limiting Properties

In the previous section we have seen that in case of large schedules, (i.e., many customers) the size of the majority of the slots is equal to approximately 1.58 for the sequential approach and equal to approximately 1.84 for the simultaneous approach.

We have seen that if n increases the number of slots with the same size increases as well, for both simultaneously and sequentially derived schedules. So for large n most slots have equal size. Hence for n large enough, the total loss function can be approximated by:

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}|S_i - x_i| &\approx n\mathbb{E}|S(x) - x|, \\ &= n\left[\int_x^\infty (s-x) f_{S(x)}(s)ds + \int_0^x (x-s) f_{S(x)}(s)ds\right] \quad (4.12)\end{aligned}$$

Kemper et al. (2011) tells us that the distribution of the sojourn time is

$$F_{S(x)}(s) = 1 - e^{-\mu(1-\rho_x)s}. \quad (4.13)$$

So the density of the sojourn time is

$$f_{S(x)}(s) = \frac{d}{ds} F_{S(x)}(s) = \mu(1-\rho_x)e^{-\mu(1-\rho_x)s}.$$

Substituting this together with (4.11) and $\rho_x \in (0,1)$, the unique solution of (4.7), in equation (4.12), we get:

$$n\left[\frac{1 - 2e^{-\mu(1-\rho_x)x} - \mu(1-\rho_x)x}{\mu(\rho_x - 1)}\right] = n\left[\frac{1 - 2\rho_x + \log\rho_x}{\mu(\rho_x - 1)}\right].$$

This again is a convex function and so we set the derivative equal to zero. By using (4.10) and (4.11) we get:

$$\begin{aligned} 0 &= \frac{d}{dx}\frac{1-2\rho_x+\log\rho_x}{\mu(\rho_x-1)}, \\ &= -\frac{(1+\rho_x(-2+\log\rho_x))\rho_x'}{\mu(\rho_x-1)^2\rho_x}, \\ &= \frac{1+(-2+\log\rho_x)\rho_x}{\mu(\rho_x-1)(1-\rho_x+\rho_x\log\rho_x)}. \end{aligned}$$

This equation implies that:

$$1 = (2-\log\rho_x)\rho_x.$$

Note that this expression does not contain μ, so the optimal value of $\hat{\rho}_x$ is independent of value of μ. Solving this equation leads to:

$$\hat{\rho}_x = \frac{-1}{\text{product logarithm}(-1, \frac{-1}{e^2})} \approx .32,$$

with the product logarithm as in the previous section.

Hence by (4.11) we get:

$$\hat{x} = \frac{1.68}{\mu},$$

For the sequential optimization approach we obtain a different result. In this approach, setting the derivative of expression (4.12) equal to zero yields, by an application of Leibniz's rule, result (3.13) on page 20,

$$\tilde{x} = F^{-1}(1/2).$$

Substituting (4.11) and rewriting (4.13), we get for this result,

$$\frac{\log\rho_x}{\mu(\rho_x-1)} = \frac{\log(1-\frac{1}{2})}{\mu(\rho_x-1)},$$

which solves for $\tilde{\rho}_x = \frac{1}{2}$. Note that this is independent of μ. So we get:

$$\tilde{x} = \frac{\log\tilde{\rho}_x}{\mu(\tilde{\rho}_x-1)} = \frac{2\log 2}{\mu}\Big|_{\mu=1} \approx 1.39.$$

4.3 Loss Function Value Comparison

In case of the quadratic loss function, Figure 4.4 already indicated a convergence of the percentage difference to approximately 20%. In this section we prove that the percentage difference of the objective values of both optimization approaches converges to 20% as the schedule size increases. This applies to both the quadratic loss function as the linear loss function. We use the results from the previous sections:

Quadratic loss function:

$$\begin{aligned}\frac{n\mathbb{E}(S(\tilde{x})-\tilde{x})^2}{n\mathbb{E}(S(\hat{x})-\hat{x})^2} &= \frac{\mathbb{E}(S(\tilde{x})-\tilde{x})^2}{\mathbb{E}(S(\hat{x})-\hat{x})^2},\\ &= \frac{2+2\log\tilde{\rho}_x+\log^2\tilde{\rho}_x}{\mu^2(\tilde{\rho}_x-1)}\Big/\frac{2+2\log\hat{\rho}_x+\log^2\hat{\rho}_x}{\mu^2(\hat{\rho}_x-1)},\\ &= \frac{2+2\log\tilde{\rho}_x+\log^2\tilde{\rho}_x}{\tilde{\rho}_x-1}\frac{\hat{\rho}_x-1}{2+2\log\hat{\rho}_x+\log^2\hat{\rho}_x}.\end{aligned}$$

Observe that this fraction is independent of μ, recall that $\tilde{\rho}_x = e^{-1}$ and $\hat{\rho}_x = .25$ are also independent of μ. Hence,

$$\frac{n\mathbb{E}(S(\tilde{x})-\tilde{x})^2}{n\mathbb{E}(S(\hat{x})-\hat{x})^2} \approx 1.23.$$

Linear loss function:

$$\begin{aligned}\frac{n\mathbb{E}|S(\tilde{x})-\tilde{x}|}{n\mathbb{E}|S(\hat{x})-\hat{x}|} &= \frac{\mathbb{E}|S(\tilde{x})-\tilde{x}|}{\mathbb{E}|S(\hat{x})-\hat{x}|},\\ &= \left[\frac{1-2\tilde{\rho}_x+\log\tilde{\rho}_x}{\mu(\tilde{\rho}_x-1)}\right]\Big/\left[\frac{1-2\hat{\rho}_x+\log\hat{\rho}_x}{\mu(\hat{\rho}_x-1)}\right],\\ &= \frac{1-2\tilde{\rho}_x+\log\tilde{\rho}_x}{(\tilde{\rho}_x-1)}\frac{(\hat{\rho}_x-1)}{1-2\hat{\rho}_x+\log\hat{\rho}_x}.\end{aligned}$$

Again, observe that this fraction is independent of μ, recall that $\tilde{\rho}_x = \frac{1}{2}$ and $\hat{\rho}_x = .32$ are also independent of μ. Hence,

$$\frac{n\mathbb{E}|S(\hat{x})-\hat{x}|}{n\mathbb{E}|S(\hat{x})-\hat{x}|} \approx 1.21.$$

So, for both types of loss functions the difference between the sequential approach and the simultaneous approach is over 20%, independent of μ.

4.4 Conclusions

For the comparison between both optimization approaches studied in this chapter, we can conclude that in a i.i.d. exponential setting, the sequential approach yields over 20% more expected loss than the simultaneous approach. This applies for both the unweighted quadratic and unweighted linear loss function and is independent of the value of μ.

An important aspect of appointment scheduling is the computation time for a schedule. The sequential approach is able to construct a schedule in only a few seconds. While for the simultaneous approach this is a matter of minutes in the exponential case and a matter of hours with other service time distributions. Hence, with the simultaneous approach you derive schedules with a lower value of the loss function than as with the sequential approach, but the latter approach creates schedules much faster with the price of having more expected loss for the system. Especially for large schedules the computation time becomes substantial. We come back to this remark in the upcoming chapters.

The results from this chapter made us more aware of the dynamics of appointment scheduling. This awareness consists of the consequences of whether taking future arrivals into account and the derivation of the waiting times capturing the influence of all the previous customers. The latter knowledge we intend to exploit in the upcoming chapters to overcome the analytical complications of appointment scheduling with general service time distributions. Furthermore, the large number of optimal interarrival times with the same size in a schedule (when service times are i.i.d.) justified the usage of limiting results. This property motivated us to introduce an approximation method to overcome the problem with dimensionality, which you can find in the next chapter.

Chapter 5

Approximation Methods for Appointment Scheduling

In this chapter we will introduce approximation methods to overcome the complications we stated in the introduction, these are, the dimensionality and the analytical problems with general service time distributions. The methods can be used while considering the simultaneous optimization approach, which is the approach we consider throughout this chapter, unless stated otherwise.

5.1 Equidistant Appointment Times

In the case of a schedule with i.i.d service times, we have seen that the slot sizes of the customers in the centre of the optimal schedule are equal. Hence, if n is large most of the customers will have the same interarrival time in the optimal schedule. So generally speaking, we consider an n dimensional optimization problem, which has an optimal solution with most of the variables having the same optimal value. Perhaps we can exploit this to reduce the dimension of the problem by letting those slots with the same optimal size correspond to only one variable. One radical choice even, is if we take for example $x_1 = x_2 = \cdots = x_{n-1}$. In this case we reduce the dimension of the problem from n to 1. Note that this approximation method only works if the customers have the same service time distribution and we approach the problem simultaneously. This radical choice results in schedules with equidistant appointment times, hence the name of this approximation method. Recall the results from previous chapter and notice the similarities

with Sections 4.1.3 and 4.2.3, where we approached the limiting schedule with one single variable as well.

For small n this heuristic affects the performance considerably more than for larger schedules, but still it is not substantial as Figure 5.1 illustrates. In this figure we show the value of the quadratic loss function of optimal schedules divided by the value of schedules derived with the equidistant method. The figure shows that as n grows, the ratio converges to 1 and is never larger than 1.02 (2%).

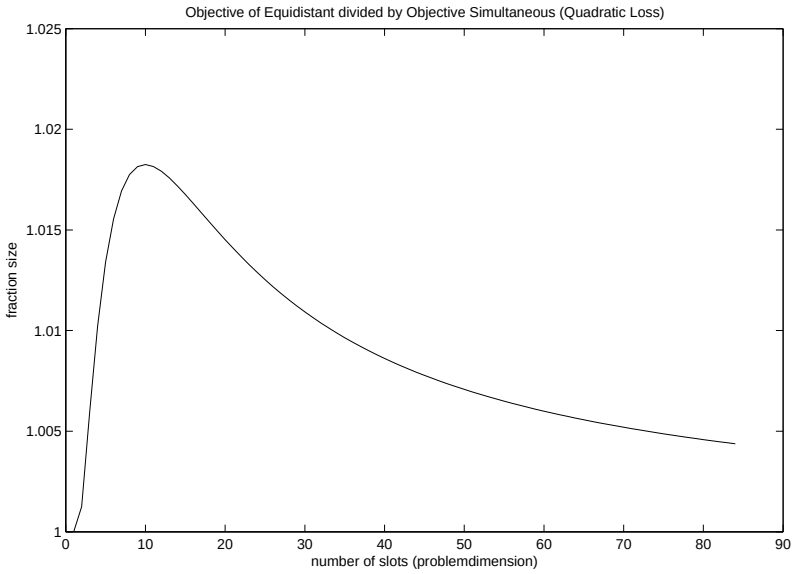


Figure 5.1: The fractional difference in objective of the equidistant approach and the non-equidistant. Quadratic loss, $0<n<85$, i.i.d exponential(μ=1) service times.

5.2 Lag Order

In this section we introduce the main result of this thesis. The Lag order approximation method. In the first part of this thesis we considered the waiting time for a certain customer as a function of the service times of all the previous customers. Hence we considered the waiting time expression:

$$\mathbb{E}h(W_i(x_1, \ldots, x_{i-1})). \tag{5.1}$$

The key observation is that customers i's waiting time depends on all the time slots and service times of preceding customers. Additionally, the further away a predecessor, the smaller his influence on the waiting time, illustrated by Figure 5.2. The figure shows the sojourn time of a customer from the center of an simultaneously optimized schedule. This sojourn time consists of its service time and waiting time. The waiting time by itself, consists of waiting time caused by the customers first predecessor, caused by the customers second predecessor and so forth, until we arrive at 100%. As the figure illustrates the waiting time caused by preceding customers declines as this customer is earlier in the schedule, until the influence is neglegible. Also one can discard this influence even before it is neglegible. Cutting of this influence until a certain level is the basic idea of this approximation method. Summarizing:

We do *not* create expressions of the waiting time in terms of all the previous customers, but *only* the first few. So the expression for lag order k of (5.1) would be:

$$\mathbb{E}h(W_i(x_{i-k}, \ldots, x_{i-1})).$$

As a result this approximation method enables the use of other service time distributions besides the exponential. In case of the exponential service times we have seen that it is possible to derive expressions for the waiting times considering the influence of all the predecessors. This is very hard in case of other service time distributions, since they do not have the memoryless property, as we explained in Chapter 1. So in case of other service times distributions we arrive at difficult expressions and evaluations of complex integrals. With a decrease of the influence of previous customers these expressions become more attractive, without losing too much performance in terms of the value of the loss function, as you will see in Chapter 6.

As we mentioned before, we call this approximation method the lag order method as we consider the lags of influence on the waiting time up to a certain order. The various lag orders are defined as follows.

- Lag order I: there is no influence of previous customers on the current customer waiting time
- Lag order II: there is only influence of first previous customer on the current customers waiting time
- Lag order III: there is influence of first two previous customers on the current customers waiting time

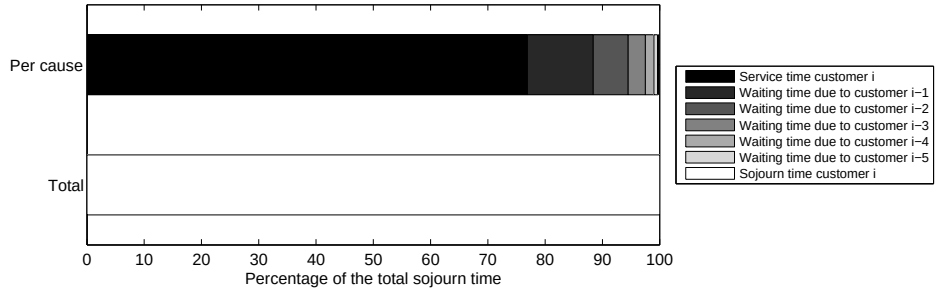


Figure 5.2: Percentage influence on the sojourn time of a customer from the center of an optimized schedule. Service times exponentially ($\mu = 1$) distributed and we considered the quadratic loss function with $n = 30$.

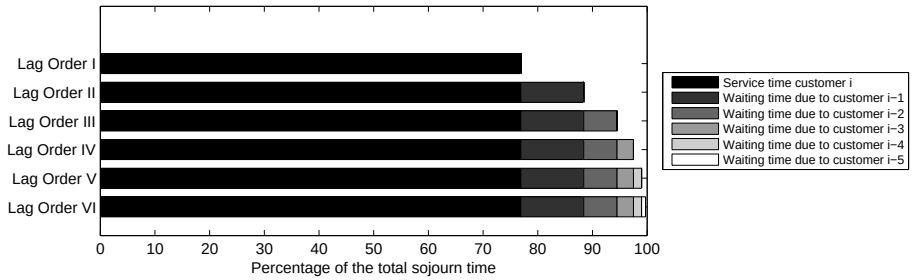


Figure 5.3: Explanation of the lag order terminology.

- Lag order k: there is influence of first $k-1$ previous customers on the current customers waiting time
- Full lag order: The normal setting, without using the approximation method, i.e., full influence.

The various lag orders are illustrated in Figure 5.3. One can see the same values as in Figure 5.2, but now we discard the level of influence of preceding customers.

To give an impression of the effect of different orders we apply it to both optimization approaches. The results can be found in Figure 5.4 and Table 5.1. One sees that as the order increases the schedules converge to a limiting schedule, the full lag order schedule. How close every order approximates the full lag order is subject of Chapter 6, Table 5.1 gives an impression already. It shows the objective values of various lag order schedules, and

one can see the convergence for both optimization approaches, as the order increases the objective value of that schedule converges to the objective of the full lag order.

Notice the kinks in both figures. In Figure 5.4(a) one can see the kink moving to the right as the lag order increases, this is caused by the neglected waiting time which results in an optimistic next interarrival time. In the sequential approach the kink represents the moment the lag order starts being active (before the kink the order is bigger than the number of customers already scheduled). The same reasoning applies for the simultaneous approach.

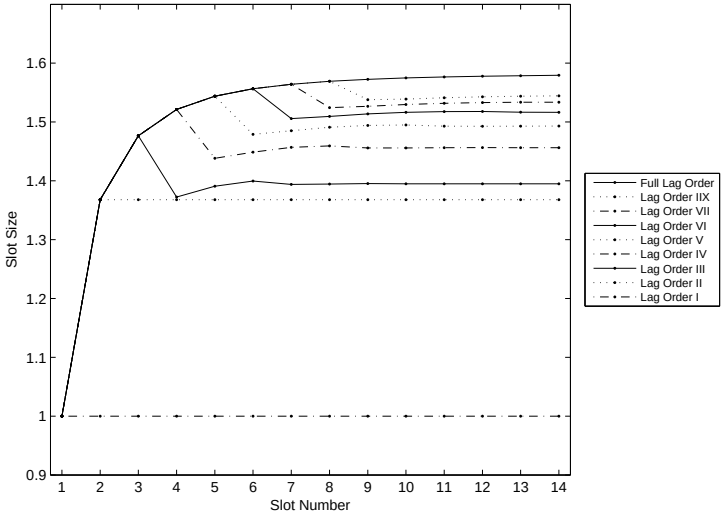
lag order	I	II	III	IV	V	Full
sim	47.6	22.9	19.5	18.6	18.4	18.3
seq	47.6	24.3	23.3	21.8	21.2	20.6

Table 5.1: The objective value of various lag order using both the simultaneous and the sequential optimization approach. Quadratic loss function, n=11, i.i.d. exponential(μ=1) service times.

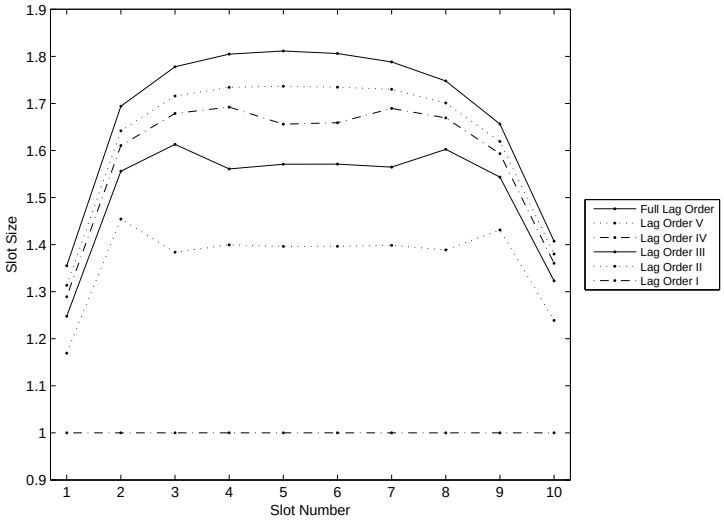
5.3 Hybrid Method

In an appointment scheduling problem with i.i.d. general service times we may apply both proposed approximation methods at the same time. By applying both methods, one is able to derive large appointment schedules with general service times, this we may call the hybrid method.

In the next chapters we will apply the proposed approximation methods to scheduling problems which have the aforementioned complications, general service times and dimensionality.



(a) Sequentially optimized schedules for various lag orders. Quadratic loss, n=15, i.i.d exponential(μ=1) service times.



(b) Simultaneously optimized schedules for various lag orders. Quadratic loss, n=11, i.i.d exponential(μ=1) service times.

Figure 5.4: The lag order approach applied to exponential distributed service times considering quadratic loss.

Chapter 6

Realistic Service Time Distributions

Most analytical studies use exponential service times in order to make their methods tractable, like we did as well in Chapter 4. However empirical evidence shows that this assumption is too restrictive and unrealistic (Cayirli and Veral, 2003). Hence, real life data of service times do not fit to the exponential distribution, see also Brown et al. (2005). In this chapter we replace the exponential distribution by a general distribution for the service times, enabling a more realistic setting. We are however unable to generate expressions for the waiting times with general service times. So with a general service time distribution we can not derive optimal schedules analytically. Therefore, we use the lag order approximation method from Section 5.2.

In this chapter we work with the quadratic loss function. Other loss functions are possible as well, however they will require the derivation of different moments of idle and waiting time than show below.

From the previous chapters we need:

$$\begin{aligned}
\mathbb{E}W_{i+1}^2 + \mathbb{E}I_{i+1}^2 &= \mathbb{E}(S_i - x_i)^2, \\
&= \mathbb{E}S_i^2 + x_i^2 - 2x_i\mathbb{E}S_i, \\
&= \mathbb{E}(B_i + W_i)^2 + x_i^2 - 2x_i\mathbb{E}(B_i + W_i), \\
&= \mathbb{E}B_i^2 + \mathbb{E}W_i^2 + 2\mathbb{E}B_i\mathbb{E}W_i + \\
&\quad x_i^2 - 2x_i(\mathbb{E}B_i + \mathbb{E}W_i). \qquad (6.1)
\end{aligned}$$

6.1 Lag Order Waiting Times

Expression (6.1) indicates that, for the quadratic loss function, we need expressions for the first and the second moment of the waiting time for general service time distributions. Below we derive these expression for lag order I, II and III.

Lag order I:
In this order the waiting time of customer i does not depend on preceding customers, hence,

$$\begin{aligned} \mathbb{E}W_i &= 0, \\ \mathbb{E}W_i^2 &= 0. \end{aligned}$$

For this order, expression (6.1) becomes:

$$\mathbb{E}W_{i+1}^2 + \mathbb{E}I_{i+1}^2 = \mathbb{E}B_i^2 + x_i^2 - 2x_i\mathbb{E}B_i.$$

Lag order II
In this order the waiting time of customer i depends on the service time of customer $i-1$. This implies that,

$$\mathbb{E}W_i = \int_{x_{i-1}}^{\infty} (s - x_{i-1}) f_{B_{i-1}}(s) ds, \tag{6.2}$$

$$\mathbb{E}W_i^2 = \int_{x_{i-1}}^{\infty} (s - x_{i-1})^2 f_{B_{i-1}}(s) ds, \tag{6.3}$$

with $f_{B_i}(s)$ the probability density function (pdf) of the service time of customer i.

To derive expressions (6.2) and (6.3), we assume that the service of the previous customer starts at its assigned arrival time, t_{i-1}.

Lag order III
In this order the waiting time of customer i depends on the service time of customer $i-1$ and $i-2$. For ease of notation and also to keep the expressions readable, we chose to derive these expressions for $i = 3$. For more details

see Appendix A.6. Recall that the service times are independent.

$$\begin{aligned}
\mathbb{E}[W_3] &= \int_0^{\infty} \mathbb{E}[W_3|B_1 = u_1] f_{B_1}(u_1) du_1, \\
&= \int_0^{t_2} \mathbb{E}[W_3|B_1 < t_2] f_{B_1}(u_1) du_1 + \int_{t_2}^{t_3} \mathbb{E}[W_3|t_2 < B_1 < t_3] f_{B_1}(u_1) du_1 \\
&\quad + \int_{t_3}^{\infty} \mathbb{E}[W_3|t_3 < B_1] f_{B_1}(u_1) du_1, \\
&= \int_0^{x_1} f_{B_1}(u_1) du_1 \int_{x_2}^{\infty} (u_2 - x_2) f_{B_2}(u_2) du_2 \\
&\quad + \int_{x_1}^{x_1+x_2} \int_{x_1+x_2-u_1}^{\infty} (u_2 - (x_1 + x_2 - u_1)) f_{B_2}(u_2) du_2 f_{B_1}(u_1) du_1 \\
&\quad + \int_{x_1+x_2}^{\infty} (u_1 - (x_1 + x_2) + \mathbb{E}B_2) f_{B_1}(u_1) du_1,
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[W_3^2] &= \int_0^{x_1} f_{B_1}(u_1) du_1 \int_{x_2}^{\infty} (u_2 - x_2)^2 f_{B_2}(u_2) du_2 \\
&\quad + \int_{x_1}^{x_1+x_2} \int_{x_1+x_2-u_1}^{\infty} (u_2 - (x_1 + x_2 - u_1))^2 f_{B_2}(u_2) du_2 f_{B_1}(u_1) du_1 + \\
&\quad \int_{x_1+x_2}^{\infty} \left[(u_1 - (x_1 + x_2))^2 + 2(u_1 - (x_1 + x_2)) \mathbb{E}B_2 + \mathbb{E}B_2^2 \right] f_{b_1}(u_1) du_1.
\end{aligned}$$

To derive expressions $\mathbb{E}[W_3]$ and $\mathbb{E}[W_3^2]$ in lag order III, we assume that the service of the two customers earlier, in this case customer 1, starts at its assigned arrival time, $t_1 = 0$. This is of course the case for customer 1. But in this lag order it will be generally assumed that customer $i-2$ starts getting service at t_{i-2}, while deriving the waiting time expression for customer i.

6.2 Log-Normal Distribution

A commonly used distribution for service times in a realistic setting is the log-normal distribution, see Cayrili et al. (2006) and Klassen and Rohleder (2004). The pdf of the log-normal distribution with parameters μ and σ is given by:

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{\tau}} e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}}, x > 0.$$

We will generate an appointment schedule using this distribution. Without loss of generality we choose $\mathbb{E}B_i = \mathrm{Var}B_i = 1$, like in the exponential setting of Chapter 4. Hence, $\mathbb{E}B_i^2 = 2$. The first and second moment are given through, $e^{\mu+\sigma^2/2}$ and $\mathbb{E}B_i^2 = e^{2(\mu+\sigma^2)}$. Solving this leads to:

$$\sigma = \sqrt{\log 2}, \qquad \mu = -\frac{\log 2}{2}.$$

We derive the required expressions from the previous section using manipulations by hand and Mathematica. We implement these in a Matlab code and we use `fminsearchbnd` to solve the optimization problems. Table 6.1(a) and 6.1(b) show the results. First, we applied the lag order to a scheduling problem with 21 customers and second to a problem with 31 customers. The tables show results for the various lag orders. Simulated optimal, from the first column, means that we use `fminsearchbnd` in combination with our simulation program, as explained in Appendix A.7. This approach results in the optimal schedule. One can find in the second column the most frequent slot size of the derived schedules, in the third column the value of the loss function obtained by simulation, see Appendix A.7. The simulated value of the total loss has confidence interval of less than 1%. The forth column (Δ Opt) shows the percentage difference of the specific lag order with the simulated optimal schedule. The value of Δ Opt is a indication of the performance of the specific lag order. Finally the last column shows the computation time for that schedule, the E stands for the power of ten.

The coefficient of variation(CV), which is the standard deviation divided by the mean ($CV = \sigma/\mu$), is a commonly used measure for the variability of service times. Empirical studies report CV values that range from approximately .35 to .85 (Cayirli and Veral, 2003). In the previous situation we had $CV = 1$. We choose $CV = .5$ and keep without loss over generality the first moment equal to 1. For the parameters this implies:

$$\sigma = \sqrt{\log \frac{5}{4}}, \qquad \mu = -\frac{\log \frac{5}{4}}{2}.$$

According to what one would expect, the slots become smaller, but also the lag order performs better. This is due to the lower variance, which causes smaller expected waiting times and with that comes smaller influence of preceding customers on the waiting time, what results in a better performance for a specific lag order. The results can be found in Table 6.1(c). The objective value of the lag order III schedule has a error of less than 1% with respect to the simulated optimal schedule.

6.3 Weibull Distribution

Another commonly used distribution for service times in a realistic setting is the Weibull distribution, see Liu and Liu (1998) and Babes and Sarma (1991). The pdf of the Weibull distribution is given by:

$$f(x;\lambda,k) = \frac{k}{\lambda}\Big(\frac{x}{\lambda}\Big)^{k-1} e^{-(x/\lambda)^k}, \qquad x > 0.$$

Again we choose for a realistic coefficient of variation and a first moment equal to 1. We choose $CV \neq \frac{1}{2}$ but $CV = \sqrt{8/\tau - 1} \approx 0.52$, since this choice reduces the Matlab code for the resulting expressions from Section 6.1 substantial. The results can be found in Table 6.1(d). We find lag order II performing having an error of 3.4% with respect to the simulated optimal schedule.

6.4 Conclusions

In this chapter we applied the lag order approximation method to scheduling problems with realistic service time distributions. We derived expressions for lag order I, II and III for arbitrary service time distributions. The results show that as the lag order increases the performance of the suboptimal schedule converges to the performance of the optimal schedule.

If one takes a closer look at Table 6.1, one can see that the performance of a certain lag order depends on two parameters. Which are, the number of customers and the coefficient of variation. For this, one has to investigate upfront what lag order suffices for the specific schedule one has to construct.

In the next Chapter we intend to put this process into practice with a problem from practice.

(a) Various lag orders compared with simulated optimal schedule. Quadratic loss, n=21, i.i.d log-normal(CV=1) service times.

Computation type	Most frequent slot size	Total loss (simulated)	Δ Opt	CPU time (s)
Simulated optimal	1.80	45	–	5E3
lag order I	1	190	322%	0
lag order II	1.38	74	64%	3
lag order III	1.78	49	9%	4.3E3

(b) Various lag orders compared with simulated optimal schedule. Quadratic loss, n=31, i.i.d log-normal(CV=1) service times.

Computation type	Most frequent slot size	Total loss (simulated)	Δ Opt	CPU time (s)
Simulated optimal	1.79	72	–	15E3
lag order I	1	425	490%	0
lag order II	1.38	129	79%	16
lag order III	1.78	78	8%	12E3

(c) Various lag orders compared with simulated optimal schedule. Quadratic loss, n=31, i.i.d log-normal(CV=.5) service times.

Computation type	Most frequent slot size	Total loss (simulated)	Δ Opt	CPU time (s)
Simulated optimal	1.79	15.2	–	3E3
lag order I	1	102	577%	0
lag order II	1.19	23.5	55%	4
lag order III	1.49	15.3	<1%	4E3

(d) Various lag orders compared with simulated optimal schedule. Quadratic loss, n=31, i.i.d weibull(CV=.52) service times.

Computation type	Most frequent slot size	Total loss (simulated)	Δ Opt	CPU time (s)
Simulated optimal	1.28	20.4	–	4E3
lag order I	1	108	427%	0
lag order II	1.20	21.1	3.4%	2
lag order III	1.18	20.9	2.4%	2E3

Table 6.1: Results of the lag order approximation method with realistic service time distributions. All schedules are simultaneously optimized. The simulated values of the total loss have a confidence interval of 1%.

Chapter 7

Real Life Appointment Schedule

In this chapter we consider a real life scheduling problem. We fit a distribution to a dataset and use this distribution to design appointment schedules. We use both approximation methods from Chapter 5 as well as the hybrid method.

7.1 CT Scan

To illustrate the methods introduced in the previous chapters, we discuss the scheduling process in a computed tomography (CT) scan department. The example results from a lean six sigma project from IBIS UvA at the Deventer Hospital, described in Mast et al. (2011). A CT scan is a medical imaging method, used in the diagnostic phase of a healthcare process. The patients are typically scheduled between 8am−1pm on workdays, and are treated in their scheduled order. In general, 20 patients are scanned during this time period. Since the investment of the CT scan is around 1.3 million euros and you need personnel (1 operator) to operate this machine we will weight the idle time of this machine higher than the waiting time of a patient. The decision for these weights are to be made by the schedule designer or a manager. Guidelines so as how to choose these weights can be found in Fries and Marathe (1981). We choose our weights by considering the running costs of the CT scan and the costs of a patient waiting. We take 6 years depreciation time for the CT scan, which implies .2 million costs per year. We estimate the costs of one operator on .1 million. Hence, the total

running costs are .3 million per year. We estimate a patient's income also on .1 million a year and assume that the time a patient is waiting, he would have been able to yield income as well. So we obtain at a cost ratio of 3:1 for idle time and waiting time. Taking this into account we arrive at the following weighted loss function:

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

With the expression on the right hand side we can continue, this now becomes:

$$\sum_{i=2}^{20} \frac{3}{4}\Big(\mathbb{E}B_{i-1}^2 + \frac{1}{3}\mathbb{E}W_{i-1}^2 + \mathbb{E}W_{i-1}(2\mathbb{E}B_{i-1} - 2x_{i-1}) + x_{i-1}^2 - 2x_{i-1}\mathbb{E}B_{i-1} \Big)$$
$$-\frac{1}{2}\mathbb{E}W_n^2 + \frac{1}{2}\mathbb{E}W_1^2.$$

Note that $\mathbb{E}W_1^2 = 0$. This is the total loss function and will be used as objective function to derive appointment schedules.

7.2 Fitting Data for Service Time

De Vree (2010) did part of his research on this CT scan area. This research investigates the process of the CT scan and for this data of the service times were acquired, which were made available to us. The service time is defined as the time a patient spends in the scan area. The dataset contains 93 realizations to which we are able to fit a distribution. We used Minitab, a statistics software package, to fit this service time distribution. We tested 12 distribution functions, with only one good fit as result. To illustrate that the exponential distribution is not realistic, we tried to fit this distribution to the dataset, see Figure 7.1(b). One can see that the exponential distribution does not fit to the dataset, also indicated by the p-value being smaller than .003. For the log-normal distribution we obtained a good distributional fit, as indicated by a p-value .37, see Figure 7.1(a). For the scale and location parameter we obtained:

$$\mu = 2.4, \qquad \sigma = .58.$$

The first and second moment of this distribution are:

$$\mathbb{E}B_{\text{ct}} = 13, \qquad \mathbb{E}B_{\text{ct}}^2 = 238.$$

So the mean service time is approximately 13 minutes. The values of these moments result in a coefficient of variation of, CV = .63, which is inside the interval mentioned on page 47 and given in Cayirli and Veral (2003).

7.3 Making a Schedule

Since we weight idle time more than waiting time, we can expect a schedule with slot sizes close to the expected service time and some maybe even smaller. We applied lag order II approximation method. The resulting schedule can be found in Table 7.1(a). The second arrival is scheduled after 12.7 minutes, which is smaller than the expected service time of 13 minutes.

(a) n=19, the slots of the optimized appointment schedule for the CT scan using the lag order II method in minutes.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
12.7	16.3	15.1	15.4	15.3	15.3	15.3	15.3	15.3	15.3
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	
15.3	15.3	15.3	15.3	15.3	15.4	15.1	16	13.1	

(b) n=19, the slots of the optimized appointment schedule for the CT scan using the lag order II method in minutes. We have incorporated the lateness of the schedule in the loss function as well.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
12.4	16.1	14.8	15.2	15.1	15.1	15.2	15.1	15.1	15.1
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	
15.1	15.1	15.1	15.1	15.1	15.2	14.9	15.8	12.9	

(c) n=19, the slots of the optimized appointment schedule for the CT scan using the lag order II method in minutes. We have incorporated the lateness of the schedule in the loss function as well.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
14.6	16.2	17.0	17.1	17.2	17.2	17.2	17.2	17.2	17.2
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	
17.2	17.2	17.2	17.2	17.2	17.1	17.0	16.0	15.0	

Table 7.1: Appointment schedules for the CT scan area.

Computation type	Most frequent slot size	Total loss (simulated)	Δ Opt	CPU time (s)
Simulated optimal	17.2	1660	–	12E3
Lag order II	15.1	1940	17%	4.3
Hybrid method	14.9	1990	20	1.7
Equidistant simulation	17	1670	<1%	5E2
Hospital's schedule	15	2000	20%	-

Table 7.2: The hospitals's current schedule and lag order II compared with the simulated optimal. Quadratic weighted loss with lateness, n=19, i.i.d. log-normal(CV=.63) service times. The simulated values of the total loss have a confidence interval of 1%

In this schedule we did not consider lateness of the system, which is the total expected completion time minus 5 hours, since we schedule from 8am-1pm. If the CT scan is still operating after 1 pm, the operator will be paid double and the CT scan is not available for its next session. In our loss function we had a weight of $\frac{3}{4}$ for the idle time, so now we take twice this weight for the systems lateness which results in a weight of $\frac{6}{4}$. The loss function becomes:

$$LF = \sum_{i=2}^{20} \left(\frac{3}{4}\mathbb{E}I_i^2 + \frac{1}{4}\mathbb{E}W_i^2 \right) + \frac{6}{4} \max \left(0, \mathbb{E}B_{20} + \mathbb{E}W_{20} - 300 + \sum_{i=1}^{19} x_i \right).$$

The resulting schedule for lag order II, can be found in Table 7.1(b). While comparing both, one can see that penalizing lateness results in smaller slots, as one might have expected.

7.3.1 Different Optimal Schedules

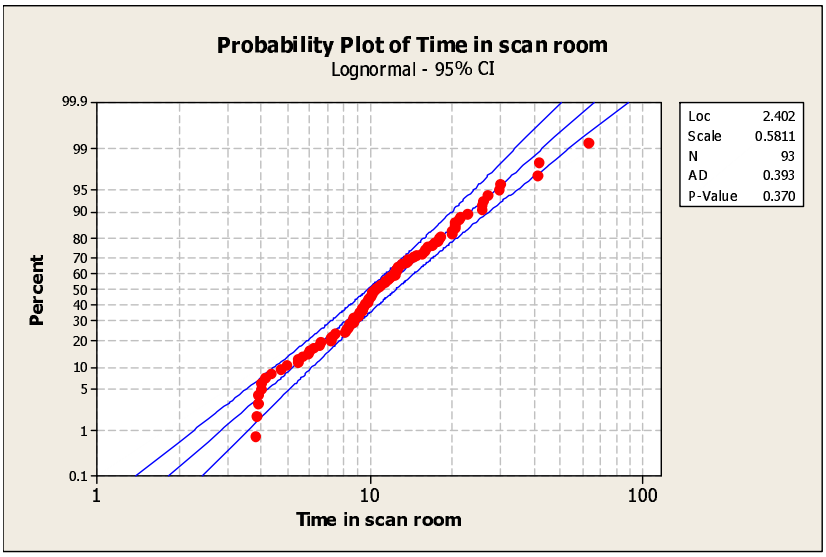
We try to use other methods or combine methods to optimally derive an appointment schedule. The hybrid method from Section 5.3 with lag order II yields a schedule with equidistant appointment times of 14.9 minutes. Another combined method, applying simulation and equidistant at the same time, yields a schedule with equidistant appointment times of 17.0 minutes. See Appendix A.7 for further details on simulation. Finally, we use again the simulated optimal, also explained in Appendix A.7. This will yield the

optimal schedule and can be found in Table 7.1(c). If one compares these slot sizes with the lag order II sizes in Table 7.1(b), one can see that most of the slots from simulated optimal are approximately 2 minutes larger.

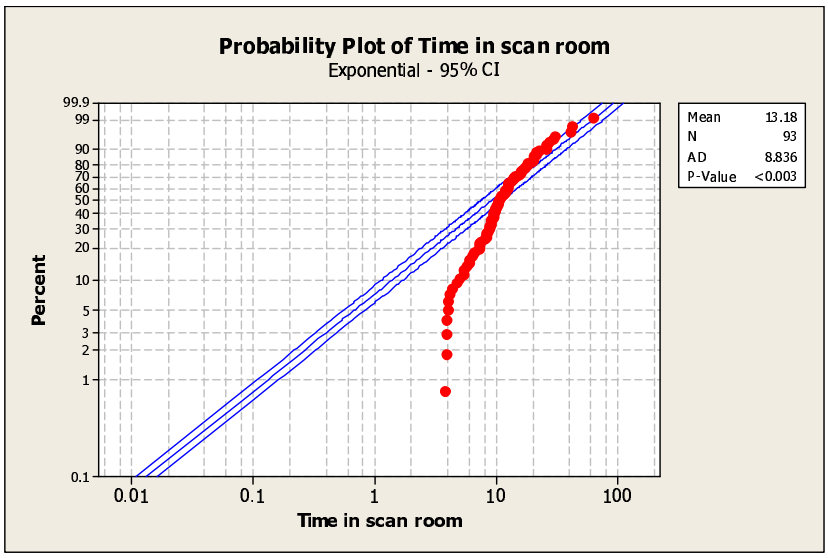
The current schedule of use in the hospital is one with equidistant appointment times of 15 minutes. We compare all these (sub-)optimal schedules with each other, see the results in Table 7.2. In this table one can see that lag order II performs similar to the hybrid method and the current schedule of the hospital. The combi method performs substantial better.

7.4 Conclusions

This chapter shows that the lag order approximation method and the equidistant method are easy to apply in a practical situation. One can conclude by looking at the derived schedules, that lag order II does not suffice, since this schedule has an error 17% with respect to the simulated optimal schedule. The equidistant method combined with simulation turnes out to be very strong. The computation time for this combi method is 20 times smaller than for the simulated optimal, while the error is 1% with respect to the simulated optimal schedule. Looking at the results in Table 7.1, one may come up with even another optimization strategy. First derive the lag order II schedule, and then optimize this schedule using simulation, adjusting every slot together with the same value. We will explain and support this strategy in the next chapter.



(a) Probability plot for time spend in the CT scan room. We fit the service time with a log-normal distribution.



(b) Probability plot for time spend in the CT scan room. As you can see the exponential distribution does not fit the dataset.

Figure 7.1: Two probability plots for the CT scan data fitted to a distribution. The 95% confidence intervals are also shown.

Chapter 8

Sensitivity Analysis

In this chapter we show how sensitive the schedules are to small changes (perturbations) in the slot sizes. We do this for global optimal schedules and for sub-optimal schedules. The latter is even more interesting since it gives some insight how the sub-optimal schedules relate to the optimal schedules. Note that throughout this chapter the schedules are derived using the simultaneous optimization approach.

8.1 Two Different Perturbation Analyses

We introduce two sorts of sensitivity analyses:

- Changing every slot in the schedule at the same time, called parallel perturbation analysis (PPA):

  $$[\tilde{x}_1+\delta,\ldots,\tilde{x}_{n-1}+\delta] \qquad \text{with} \qquad \delta \in \left[-\min_{i\in\{1,\ldots,n-1\}} \tilde{x}_i, \min_{i\in\{1,\ldots,n-1\}} \tilde{x}_i\right].$$

- Varying only one single slot in the schedule, called single perturbation analysis (SPA):

  $$[\tilde{x}_1,\ldots,\tilde{x}_{i-1},\tilde{x}_i+\delta_i,\tilde{x}_{i+1},\ldots,\tilde{x}_{n-1}],$$

  with $\delta_i \in [-\tilde{x}_i,\tilde{x}_i]$ for $i = 1,\ldots,n-1$.

For every δ we compute the expected loss of the perturbated schedule and hence arrive at a graph which displays the sensitivity. In PPA analysis we

put δ on the x-axis, whereas in the SPA analysis we put the slot size on the x-axis, i.e., $\tilde{x}_i + \delta_i$.

We start the analysis with a global optimal schedule:

$$\tilde{x} = [\tilde{x}_1, \ldots, \tilde{x}_{n-1}].$$

Recall that this is called the full lag order in our terminology.

We look at the effect of these perturbations in our objective function, the loss function and in this case the quadratic loss function. The results for the global optimal schedule can be found in Figure 8.2(b)(PPA) and 8.1(f)(SPA). We analyzed the quadratic linear loss function with i.i.d. exponential($\mu = 1$) service times. The * in the PPA and SPA figures represent the objective value of the schedule that is being analyzed. In the SPA analysis you can also see the current value of the slot on the x-axis. The figures confirm that the schedule is optimal, because we see the * in both schedules in the minimum of the graphs.

We can apply the same analysis to sub-optimal schedules. For this we derived sub-optimal schedules with lag order I, II and III, IV and V and did both a PPA and an SPA. The results for PPA for all the orders can be found in Figure 8.1. What this figure illustrates is the convergence of the orders to the optimal schedule. One can see the * getting closer to the minimum of the graph as the lag order increases. For SPA we only show the slots of lag order II in Figure 8.2(a). In this figure one can clearly see the suboptimality of the schedule, since the * is not in the minimum of the graphs. This applies for the SPA graph of every slot. Recall that lag order I results in a schedule with interarrival times equal to 1. Also recall the dome shape of the global optimal schedule. If we look at the SPA graph of slot 1 and 20 we see that the current slot values are closer to the optimal value for that slot than the values of the other slots, for example slot 10. This is because the first and last slot in the global optimal schedule are the smallest slots in the schedule (≈ 1.4) and hence closer to one than slot 10 (≈ 1.84). The SPA for the remaining lag orders can be found in Appendix A.9.

8.2 Conclusions

In the SPA of the optimal appointment schedule (Figure 8.2(b)), we obtain convex graphs in every slot. In these graphs the optimal interarrival time is the minimum value. Hence, we can conclude that we computed the optimal

value in every dimension. In case of sub-optimal schedules, in this case a certain lag order, we see that the chosen slot sizes are indeed not optimal.

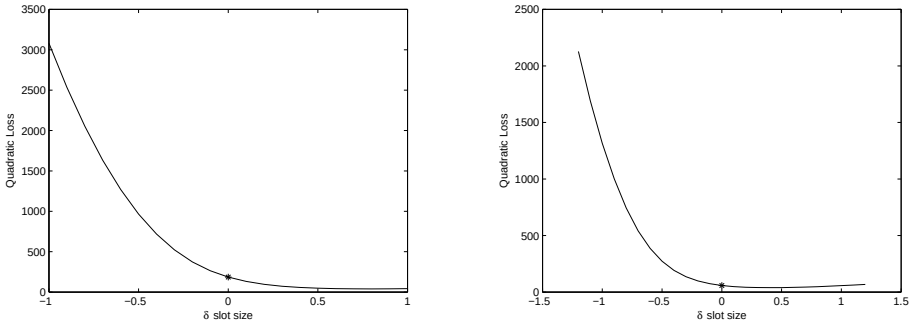
In SPA, one investigates the sensitivity on perturbations in one interarrival time. For a single slot, if we perturb the value of this slot with one time unit, i.e., by 50%, we obtain a difference in loss of 4%. For the other analysis (PPA), we look at the sensitivity for the entire appointment schedule. If we decrease every interarrival time with .4 time units, i.e., an average change of 22% for a slot, this results in an increase of loss by 40%. Likewise, if we increase every interarrival time with .4 time units, the loss increases by 20%.

These observations bring us to some practical rules for designers of appointment schedules.

If you derive a (sub)-optimal schedule you should be careful with changing the size of all the slots together. This might be done on ground of practical reasons, for example rounding up every slot size to a multiple of 5 minutes for convenience. Our results show that the loss function is sensitive to such changes. However, in case of a lag order schedule, i.e., a sub-optimal, this sensitivity is mainly on the left hand side of the schedule. This means that the schedule is more sensitive to decreasing all the slots together, than for increasing all the slots together, i.e., rounding up might improve the performance.

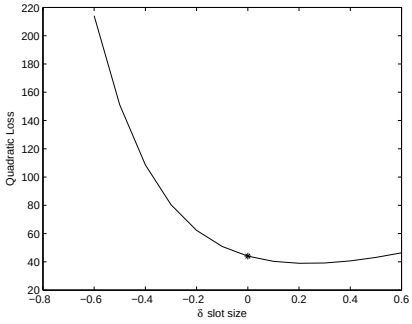
For the sensitivity per slot, one can see that if you change only one time slot the loss of the system does not increase substantially. this might be practical if the designer of a schedule knows upfront that a particular patient might take more time than expected. If he is right with his judgment, he improved the schedule himself. If he is not right, we showed that he will not incur substantial extra loss of the system. In general, if one derived a schedule using the lag order approximation method, the time slots are more sensitive to a decrease in size than to an increase, see Figures 8.1 and A.1.

One may take these rules into account in the predetermination phase of an appointment schedule or while a schedule is already running, in order to decrease the realization of the systems loss.

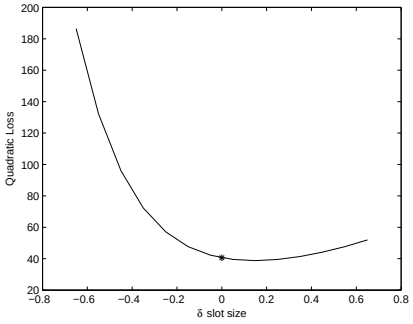


(a) Lag Order I

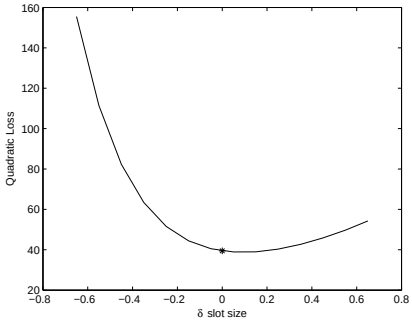
(b) Lag Order II



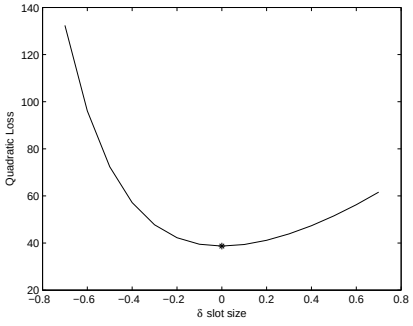
(c) Lag Order III



(d) Lag Order IV

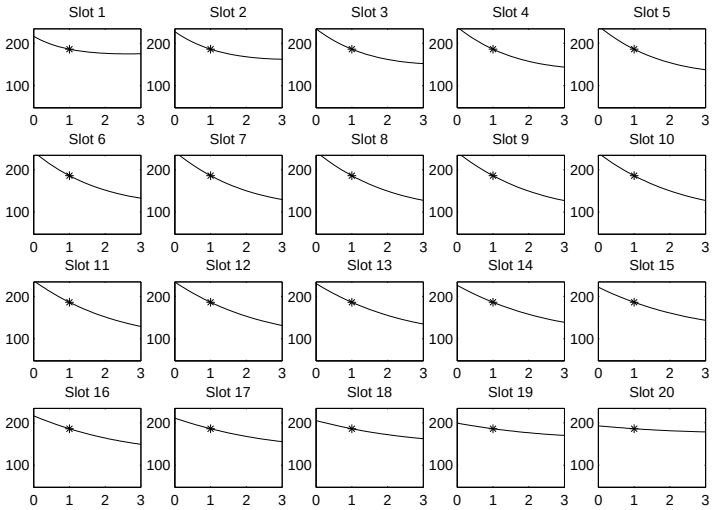


(e) Lag Order V

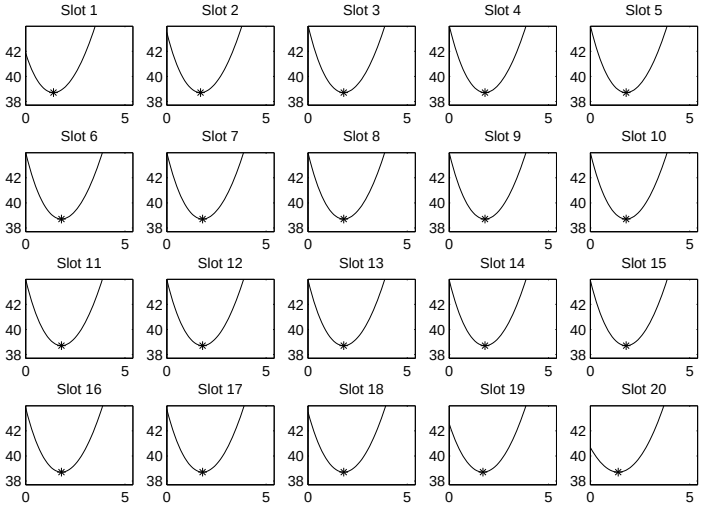


(f) Full Lag Order

Figure 8.1: Sensitivity analysis of various lag order-optimal schedules. Quadratic loss, n=21, i.i.d. exponential (μ=1) service times. The star is the actual value of the schedule. On the x-axis the value, the interarrival times are parallel varied with. The quadratic loss on the y-axis. From lag order I, see Figure (a), to full lag order, see Figure (f).



(a) Lag Order I



(b) Full Lag Order

Figure 8.2: Sensitivity analysis of lag order I and optimal (full lag order) schedule per slot. Quadratic loss, n=21, i.i.d. exponential (μ=1) service times. The star is the actual value of the interarrival time. On the x-axis the value of the interarrival time with the quadratic loss on the y-axis.

Chapter 9

Conclusions and Future Work

We finish this thesis by summarizing the conclusions from the chapters. Also, we come up with recommendations for applying the results stated in this thesis and finally we give proposals for future research.

9.1 Conclusions

Sequential versus Simultaneous

We did a thorough investigation on the difference between the two optimization approaches; sequential and simultaneous. In the derivation of the global minimum, the simultaneous optimization approach is often implicitly used, as we explained in Section 3.5. In Section 4.1.1 and 4.2.1 we showed the computation of the global minimum explicitly in case of a schedule with three customers. In both sections we also showed that the sequential optimization approach does not yield the global minimum, which is confirmed in the other sections of that chapter.

The sequential approach does not optimize the total loss function itself, but sequentially the loss functions per slot. It starts by choosing x_1 optimally, only considering the expected loss in the first slot. The size of this slot is fixed and the approach continues with choosing x_2 optimally, only considering the expected loss in the second slot, given $x_1 = \tilde{x}_1$. This sequential way of optimizing reduces an n-dimensional optimization problem to n problems of 1-dimension.

In Section 4.3 we compared the objective values of large schedules (i.e., schedules with n large) for both optimization approaches. We showed that

the value of the loss function after applying the sequential approach was 20% higher than the value of the global simultaneous optimum. This result was obtained for both the quadratic and linear loss function with i.i.d exponential service times and independent of the value of μ.

Taking these results into account one may call the sequential optimization approach a heuristic or approximation method for the hard global optimization problem. Since this approach reduces the computation time, while losing performance in terms of the value of the objective function. Hence, this approach is a powerful method to quickly determine the optimal arrival times, also online, when arrivals did occur already.

Approximation Methods

Lag order Approximation Method

In this thesis we propose two approximation methods, the lag order approximation method for the waiting time and the equidistant appointment times. For the former method, we showed the expressions of the first and second moment of the waiting time for lag order I, II and III, for general service time distributions, with the possibility to extend this to higher orders.

The results of the lag order method, as you can see in Tables 5.1 and 6.1, are very promising. In a realistic setting with i.i.d log-normal service times, lag order III has an error of 1% with respect to the global simultaneous optimum. Below we list the advantages and disadvantages of the lag order method that we propose.

Advantages

- The schedule designer is flexible in choosing any service time distribution with a density. The distributions do not necessarily have to be identical for every customer.

- If the full order method is at hand, e.g., in case of exponential service times, the lag order method takes less computation time to derive schedules. The price for this, are higher expected losses, in the sense of expected waiting times. However, we can approximate any full order with a certain lag order to come arbitrary close to the optimal schedule (performance wise). This can be investigated for a particular problem before deriving the schedules on a daily basis.

- The lag order method can be used for the calculation of an initial schedule in a simulation program. This will save the simulation method time compared to starting with a random schedule.
- The lag order method can be applied together with the equidistant method.
- The lag order method can be applied for both optimization approaches we used throughout this thesis.

Disadvantages

- Before you want to apply the lag order approach for your specific problem, you have to investigate which order performs well enough. As we have seen in Section 6.2, this depends strongly on the coefficient of variation of the service time distribution.
- By applying the lag order approach, you lose performance. This is in the sense that you get more expected waiting time, since you neglect part of these waiting times in the derivation of that schedule.

Equidistant Method

The equidistant method is only applicable when all customers agree to the same service time distribution. This method reduces the computation time substantially, especially when considering large schedules. In Chapter 7, we see that since the hospital already uses equidistant appointment times, this method is very useful. In that problem setting the equidistant method has an error of 1% with respect to the simulated optimal schedule.

Below we list the advantages and disadvantages of the equidistant method that we propose.

Advantages

- The method reduces computation time, since it converts an n-dimensional problem into a 1-dimensional problem.
- In case i.i.d service times the equidistant method can be used for the calculation of an initial schedule in a simulation program. This saves the simulation method time.
- The equidistant method can be applied together with the lag order method.

Disadvantages

- The method requests i.i.d. service times.

- By applying the equidistant method, you lose performance. This is because you do not obtain a dome shaped graph of the interarrival times, which is the optimal shape. Hence, the interarrival times derived with the equidistant method, are larger at the beginning and the end of the schedule, than what is optimal.

  This is because you loose the advantages of the dome shape, hence, the interarrival times at the beginning and the end of the schedule are bigger than optimal.

Real Life Schedule

In Chapter 7 we showed that our theory is applicable in practice. We used both approximation methods and combined them with simulation to derive an optimal schedule. This chapter takes you through the steps from practice to theory. The result is that we found an equidistant schedule that has an error of 1% with respect to the simulated optimal schedule.

Sensitivity Analysis

In Chapter 8 we showed how sensitive (sub-)optimal schedules are to perturbations of the slot size, both parallel and single perturbed. The practical rules we extracted from these analyses are.

- If one has a global optimal schedule at hand, we recommend not to increase nor decrease all the slots of the schedule together.

- If one has a lag order optimal schedule at hand, we recommend to increase all slots together and recommend not to decrease all the slots together.

- If one wishes to change the size of a single slot this does not affect the performance substantially.

9.2 Recommendations

Summarizing the conclusions from this thesis we come up with some recommendations for using the theory and methods proposed in this thesis.

- In case of large schedules and i.i.d. customers we would recommend to use the equidistant approach for the interarrival times in the middle of the schedule and no approximation method for the beginning and the end of the schedule. This results in a faster method than using simulation and the performance will be close to optimal.
- If you consider a problem with i.i.d. general service times, we recommend to use the lag order method for the beginning and the end of the schedule and the equidistant for the center. Further simulation might improve the schedule, but this depends on the lag order that is applied.
- In case of customers with different general service time distributions, we recommend using the lag order approach first to continue by using a simulation method. The lag order result will function as an initial schedule.
- In case of customers with different exponential service time distributions, we recommend to use the lag order approach up to an order which approaches the optimal schedule well enough. This takes less computation time than using the full lag order.
- Take the scheduling rules from Chapter 8 into account while you are in the predetermination phase of a schedule or when a schedule is already running.
- If you consider to use a lag order schedule as initial schedule for simulation, one can obtain an ever better initial schedule by increasing all the slots together. See the scheduling rules in Subsection 9.1 and the analysis in Chapter 8 for further details.

9.3 Future Work

The list below states problems or areas in appointment scheduling that are worthwhile investigating. This came about while writing and working on this thesis.

- A general expression for ζ_n in the algorithm for Wang (1999) is not yet known, see Section 3.4. A general expression would extend Wang's extended algorithm to second moment sojourn times for arbitrary μ.
- The specific advantages of applying the lag order method as a start schedule for simulation methods.
- How to incorporate the relaxation of certain assumptions in the model. Assumptions such as no-shows, servers punctuality, customers punctuality, server interruption level, et cetera. How does the lag order method perform in these realistic settings?
- The differences between the sequential and simultaneous optimization approach while considering general service times.
- What can we say about tail expressions, regarding the lag order approximation method? The percentiles can be computed while again neglecting influence of predecessors on the waiting time. Further research is needed.
- The speed of convergence of the lag order approximation method to the full lag order.
- Is there a relation between the Bailey-Welch rule (two patients arriving at $t = 0$) and the dome shape of the optimal schedule?

Acknowledgements

Regarding the mathematical content I would like to thank Sandjai Bhulai, Karma Dajani, Benjamin Kemper and Michel Mandjes for sharing their knowledge, their advice, fruitful discussions, their time and their confidence in a good result. Next to that I would like to thank Benjamin and Michel for their acknowledgements in Kemper et al. (2011).

Affiliation

Wouter Vink
Ina Boudier Bakkerlaan 69 II
3582 VV Utrecht, The Netherlands
Telephone: +31626722979
E-mail: wejvink@gmail.com
Student number: 3117715

References

Asmussen, S. (2003). *Applied Probability and Queues, Applications of Mathematics*, Volume 51. Springer-Verlag, New York, second edition, Stochastic Modelling and Applied Probability.

Babes, M. and G. Sarma (1991). Out-patient queues at the lbn-rochd health center. *Operational Research Society 42*(10), 845–855.

Bailey, N. (1952). A study of queues and appointment systems in hospital outpatient departments, with special reference to waiting times. *Journal Royal Statistical Society 14*, 185199.

Begen, M., R. Levi, and M. Queyranne (2011). A sampling-based approach to appointment scheduling. Submitted.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, H. Zeltyn, and L. Zhao (2005). Statistical analysis of a telephone call center; a queueing-science perspective. *Journal of the American Statistical Association 100(469)*, 36–50.

Cayirli, T. and E. Veral (2003). Outpatient scheduling in health care: a review of literature. *Production and Operations Management 12*(4), 519–549.

Cayrili, T., E. Veral, and H. Rosen (2006). Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science 9*, 47–58.

De Vree, T. (2010). Improving ct scan scheduling and queues with non-overtaking. Master's thesis, University of Utrecht.

D'Errico, J. (2005). fminsearchbnd an extension to fminsearch. http://www.mathworks.com/matlabcentral/fileexchange/8277-fminsearchbnd Latest update 2011.

Fries, B. and V. Marathe (1981). Determination of optimal variable-sized multiple-block appointment systems. *Operations Research 29(2)*, 324–345.

Gray, W. and P. Wang (1995). On the computation of optimal schedules in a single-server system. *Journal of the Chinese Institute of Industrial Engineering 14*, 245–252.

Gupta, D. and B. Denton (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions 40(9)*, 800–819.

Hutzschenreuter, A. (2005). Queueing models for outpatient appointment scheduling. Master's thesis, University of Ulm.

Kaandorp, G. and G. Koole (2007). Optimal outpatient appointment scheduling. *Health Care Management Science 3*, 217–229.

Kemper, B., C. Klaassen, and M. Mandjes (2011). Utility-based appointment scheduling. Submitted.

Klassen, K. and T. Rohleder (2004). Outpatient appointment scheduling with urgent clients in a dynamic, multi-period environment. *International Journal of Service Industry Management 15*(2), 167–186.

Liao, C., C. Pegden, and M. Rosenshine (1993). Planning timely arrivals to a stochastic production or service system. *IIE Transactions 25*, 36–73.

Lindley, D. (1952). The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society 48*, 277–289.

Liu, L. and X. Liu (1998). Block appointment systems for outpatients clinics with multiple doctors. *Journal of the Operational Research Society 49*, 1254–1259.

Liu, N., Z. S., and V. Kulkarni (2010). Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing & Service Operations Management 12(2)*, 347–364.

Mast, J., B. Kemper, R. Does, M. Mandjes, and Y. Bijl (2011). Process improvement in healthcare: Overall resource efficiency. *Quality and Reliability Engineering International*. In press.

Robinson, L. and R. Chen (2003). Scheduling doctor's appointments: Optimal and empirically-based heuristic policies. *IIE Transactions 35*, 298–307.

Schild, A. and I. Fredman (1961). On scheduling tasks with deadlines and non-linear loss functions. *Management Science 7*(3), 280–285.

Tijms, H. (1986). *Stochastic Modelling and Analysis: A Computational Approach*. Chichester: John Wiley & Sons.

Vanden Bosch, P., D. Dietz, and S. J.R. (1999). Scheduling customer arrivals to a stochastic service system. *Naval Research Logistics 46*.

Wang, P. (1993). Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics 40*, 345360.

Wang, P. (1999). Sequencing and scheduling n customers for a stochastic server. *European journal of operational research 119*, 729–738.

Welch, J. and N. Bailey (1952). Appointment systems in hospital outpatient departments. *The Lancet 15(3)*, 1105–1108.

Westeneng, J. (2007). Outpatient appointment scheduling. Master's thesis, University of Twente.

Appendix

A.1 Matlab Code

The algorithm proposed by Wang (1999) and its extension, obtained by ourselves, is used to compute the objective value of an appointment schedule with exponential($\mu = 1$) service times. The condition of $\mu = 1$ is because of the absence of a general expression for ζ_n, see Appendix A.2. The input argument of this function is an appointment schedule represented by a vector of interarrival times. The output of this function is the quadratic loss of this schedule. This function as such is minimized by `fminsearchbnd` to compute the optimal schedule with minimal quadratic loss.

```
function [obj] = wangextended(x)
%x is vector with interarrival times
n = length(x);  % number of slots
vector = zeros(n,n);
P = zeros(n,n);
d = zeros(n,1);
w = zeros(n,1);
wkw = zeros(n,1);
z = 1100; zeta = zeros(1,z); zeta(z) = 2;
for i = 1:z-1
   zeta(z-i) = zeta(z-i+1) + (i+1)*2;
   %the zeta vector is being constructed
end

w(1) = 1; %first moment of the sojourn time for customer 1
wkw(1) = 2; %second moment of the sojourn time for customer 1
P(1,:) = exp(-x(1));
d(1) = 1 - P(1,1:1);
```

```
for i = 2:n
matrix = zeros(i,i);
    for j = 1:i
    matrix = matrix + diag(ones(1,i-j+1) * x(i)^(j-1) *...
     exp(-x(i))/factorial(j-1) ,j-1);
    end

vector(i,1:i) = [P(i-1,1:i-1), d(i-1)];
w(i) = vector(i,1:i) * [i:-1:1]';
wkw(i) = vector(i,1:i) * zeta(z-i+1:z)';
P(i,1:i) = vector(i,1:i) * matrix;
d(i) = 1 - P(i,1:i) * ones(i,1);
end

obj = 2 + x(1)^2 -2 * x(1) * 1;
loss(1) = obj; %Q_1, Quadratic Loss for slot 1.

for i = 2:n
    obj = obj + wkw(i) + x(i)^2 -2 * x(i) * w(i);
    loss(i) = wkw(i) + x(i)^2 -2 * x(i) * w(i);
    %loss(i) is the quadratic Loss for slot i.
end

end
```

As mentioned in the code already, x is a vector of interarrival times. So, if you give this function an appointment schedule it computes its quadratic objective with exponential service times ($\mu_i = 1$). This function is now optimized using the function in Matlab called `fminsearch`. This function does not fulfill all our needs, since the elements of x are larger or equal to zero. So, we work with a slightly different function called `fminsearcbnd`, see Appendix A.3.

A.2 A General Expression for ζ_n

For exponential service times with arbitrary μ we can derive the expressions for the waiting time of customer 1 and 2 by hand. By some algebra we

obtain the following results with respect to a general expression for ζ_n.

$$\zeta_2 = [\frac{2}{\mu_2^2} + \frac{2}{\mu_1^2} + 2\frac{1}{\mu_2}\frac{1}{\mu_1}, \frac{2}{\mu_2^2}]$$

$$\begin{aligned} \zeta_n(n) &= \frac{2}{\mu_n^2} \\ \zeta_n(n-1) &= \frac{2}{\mu_n^2} + \frac{2}{\mu_{n-1}^2} + 2\frac{1}{\mu_n}\frac{1}{\mu_{n-1}} \end{aligned}$$

It is important to obtain the full expression for this, since this will enable us to broaden the extended algorithm from Wang (1999) to arbitrary μ.

A.3 Multidimensional Minimization Function `fminsearchbnd`

In order to compute optimal schedules we use the Matlab routine `fminsearchbnd`. This routine is an extension to the build in routine `fminsearch`. It is made by John D'Errico, (D'Errico, 2005) and approved by Mathworks. We explain here how this routine is able to extend the standard routine.

Since `fminsearch` does not allow bounds and constraints, the trick is to insert a wrapper function around the user supplied objective function. The arguments are identical to that which `fminsearch` uses, with as extra arguments a set of bounds. Let $LB, UB, x, z \in \mathbb{R}^n$. There are several classes of bounds and constraints one might consider. Simple lower bound:

$$LB(i) \leq x(i).$$

Upper bounds:

$$x(i) \leq UB(i).$$

Dual constraints

$$LB(i) \leq x(i) \leq UB(i).$$

The bounded variables are transformed such that `fminsearch` itself sees a fully unconstrained problem. For example, in the case of a variable bounded on the lower end by $LB(i)$, the used transformation is:

$$x(i) = LB(i) + z(i)^2.$$

The variable $z(i)$ is fully unconstrained, but since the square of $z(i)$ is always non-negative (for $z \in \mathbb{R}$), $x(i)$ must necessarily be always greater than or equal to $LB(i)$. Likewise, a pure upper bound constraint is implemented as

$$x(i) = UB(i) - z(i)^2.$$

Clearly, $x(i)$ in this case can never rise above $UB(i)$. And finally, the dual bounded variable is handled by a trigonometric transformation:

$$x(i) = LB(i) + (UB(i) - LB(i)) * (\sin(z(i)) + 1)/2.$$

In this last case, the requirement that $LB(i) \leq x(i) \leq UB(i)$ is absolutely enforced.

The tolerances on the parameters are not fully translated yet. The nonlinear transformations mean that `fminsearch` itself will see only the transformed parameters, not the parameters in their real domain. The Matlab routine `fminsearchbnd` is an overlay to `fminsearch` itself, hence, there is no simple way to provide explicit control over the variable tolerances without re-writing `fminsearch`.

You cannot provide general linear/nonlinear equality or inequality constraints, as are provided by `fmincon`, or `lsqlin`. Only simple bound constraints are allowed.

A.4 Steady State of Queue

In the Sections 4.1.3 and 4.2.3 we consider the situation of i.i.d. jobs, and the number of jobs being large. In this case service times are exponentially distributed with parameter μ. Thus the queue we have is an D|M|1 queue. Let x be the interarrival time between two customers. Note that we need $x > \frac{1}{\mu} (= \mathbb{E}B)$, such that the occupation rate (ρ) is smaller than 1. The distribution of the steady-state waiting time, W, is given through Tijms (1986) and Asmussen (2003):

$$\mathbb{P}(W < y) = 1 - \rho(x)e^{-(1-\rho(x))y/\mathbb{E}B}, \qquad y > 0. \tag{A.1}$$

This result is for G|M|1 queues, with ρ the occupation rate as stated before. In our case $\mathbb{E}B = 1/\mu$ and $\rho := \rho(x)$, with $\rho \in (0, 1)$ the unique solution of

$$e^{-\mu(1-\rho)x} = \rho. \tag{A.2}$$

Now we are able to compute $\mathbb{E}S$:

$$\begin{aligned} \mathbb{E}S &= \mathbb{E}B + \mathbb{E}W, \\ &= \frac{1}{\mu} + \int_0^\infty y\rho(x)\mu(1-\rho(x))e^{-\mu(1-\rho(x))}dy, \\ &= \frac{1}{\mu(1-\rho(x))}. \end{aligned} \tag{A.3}$$

Similarly:

$$\mathbb{E}S^2 = \mathbb{E}B^2 + \mathbb{E}W^2 + 2\mathbb{E}B\mathbb{E}W = \frac{2}{\mu^2(1-\rho(x))^2}. \tag{A.4}$$

These results are used to derive the limiting properties of the D|M|1 queue with a particular loss function.

A.5 Convexity

The main result from the paper of Kaandorp and Koole (2007) is that the objective function is multimodular, i.e., a L-convex function. In other words a local optimum in the objective function is a global optimum. The objective function they worked with consisted of the weighted sum of the average expected patient waiting time, the idleness of the doctor during the session length, and the tardiness. The tardiness is the probability that the session exceeds the planned finishing time multiplied by the average excess. For further details we refer to Kaandorp and Koole (2007). Begen et al. (2011) has also proven convexity for the objective function, i.e., the loss function. The consequence of this convexity is that a local minimum is a global minimum. Below we sketch the results from these papers using intuitive arguments. We will sketch the result of the loss function having a unique minimum.

We give our sketch for the weighted loss function, with $0 < \alpha < 1$ and equal moments of idle and waiting time. Hence for:

$$\begin{aligned} LF(x_1,\ldots,x_{n-1}) &= \sum_{i=2}^{n} \Big(\alpha\mathbb{E}[I_i^k(x_1,\ldots,x_{i-1})] + (1-\alpha)\mathbb{E}[W_i^k(x_1,\ldots,x_{i-1})]\Big), \\ & \qquad k \in \mathbb{N}_+. \end{aligned}$$

Since both $\alpha > 0$ and $(1-\alpha) > 0$, it suffices to show for the unweighted case.

For arbitrary arrival times the following holds:

$$\begin{aligned}
\lim_{t_i \to \infty} \mathbb{E}I_i^k(t_1, \ldots, t_i) &= \infty, \\
\lim_{t_i \to \infty} \mathbb{E}W_i^k(t_1, \ldots, t_i) &= 0, \\
\lim_{t_i \searrow t_{i-1}} \mathbb{E}I_i^k(t_1, \ldots, t_i) &= \mathbb{E}I_i^k(t_1, \ldots, t_{i-1}, t_{i-1}) = 0, \\
\lim_{t_i \searrow t_{i-1}} \mathbb{E}W_i^k(t_1, \ldots, t_i) &= \mathbb{E}(W_{i-1}(t_1, \ldots, t_{i-1}) + B_{i-1})^k.
\end{aligned}$$

Hence,

$$\begin{aligned}
\lim_{t_i \to \infty} \mathbb{E}I_i^k(t_1, \ldots, t_i) + \mathbb{E}W_i^k(t_1, \ldots, t_i) &= \infty, \\
\lim_{t_i \searrow t_{i-1}} \mathbb{E}I_i^k(t_1, \ldots, t_i) + \mathbb{E}W_i^k(t_1, \ldots, t_i) &= \mathbb{E}I_i^k(t_1, \ldots, t_{i-1}, t_{i-1}) \\
&\quad + \mathbb{E}W_i^k(t_1, \ldots, t_{i-1}, t_{i-1}), \\
&= \mathbb{E}(W_{i-1}(t_1, \ldots, t_{i-1}) + B_{i-1})^k.
\end{aligned}$$

To arrive at the same time as your predecessor is certainly not optimal. If we would take $t_i = t_{i-1} + \epsilon$ instead of $t_i = t_{i-1}$ we get:

$$\begin{aligned}
\mathbb{E}I_i^k(t_1, \ldots, t_{i-1}, t_{i-1}) + \mathbb{E}W_i^k(t_1, \ldots, t_{i-1}, t_{i-1}) &> \mathbb{E}I_i^k(t_1, \ldots, t_{i-1}, t_{i-1} + \epsilon) \\
&\quad + \mathbb{E}W_i^k(t_1, \ldots, t_{i-1}, t_{i-1} + \epsilon).
\end{aligned}$$

Recall that, $\mathbb{E}I_i^k(t_1, \ldots, t_{i-1}, t_i) + \mathbb{E}W_i^k(t_1, \ldots, t_{i-1}, t_i) = \mathbb{E}[(S_{i-1} - x_{i-1})_+^k]$. This is again equal to $\mathbb{E}[(S_{i-1} + t_{i-1} - t_i)_+^k]$.

So, using the previous equation, we know that $\exists t^*, t_{i-1} < t^* < \infty$, such that:

$$\begin{aligned}
\mathbb{E}[(S_{i-1} + t_{i-1} - t_{i-1})_+^k] &> \mathbb{E}[(S_{i-1} + t_{i-1} - t^*)_+^k] \\
\mathbb{E}[(S_{i-1} + t_{i-1} - \infty)_+^k] &> \mathbb{E}[(S_{i-1} + t_{i-1} - t^*)_+^k]
\end{aligned}$$

This shows already that there is an optimal arrival time for patient i such that the objective is minimal. Also by continuity and convexity of the power function, one can imagine that this arrival time is unique. Rewriting the previous equations we get:

$$\frac{\mathbb{E}[(S_{i-1} + t_{i-1} - t_{i-1})_+^k] + \mathbb{E}[(S_{i-1} + t_{i-1} - \infty)_+^k]}{2} > \mathbb{E}[(S_{i-1} + t_{i-1} - t^*)_+^k].$$

This applies for one specific t^*. This is not a proof for convexity, but one can imagine that, by continuity and convexity of the power function, $\mathbb{E}[(S_{i-1} + t_{i-1} - t_i)_+^k]$ has only one minimum.
So we can conclude that $LF_i(x_1, \ldots, x_{i-1})$ as a function of x_{i-1}, has a unique minimum. Along the same arguments we also have that $LF_j(x_1, \ldots, x_j)$, as a function of x_{i-1}, has a unique minimum, for $j > i-1$. We can apply this argument for the other arrivals as well.

The intuition behind this sketch shows the dynamics of appointment scheduling. We refer to the aforementioned articles in this section for complete proofs of convexity and hence, the existence of a unique minimum.

A.6 Lag Order III Expressions

The complete derivation of the expressions of the first and second moment of the waiting times for lag order III with general service time distributions (not necessarily i.i.d.) is shown here. We use these expressions in Chapter 6.

$$\begin{aligned}
\mathbb{E}[W_3] &= \int_0^\infty \mathbb{E}[W_3|B_1 = u_1] f_{B_1}(u_1) du_1, \\
&= \int_0^{t_2} \mathbb{E}[W_3|B_1 < t_2] f_{B_1}(u_1) du_1 + \int_{t_2}^{t_3} \mathbb{E}[W_3|t_2 < B_1 < t_3] f_{B_1}(u_1) du_1 \\
&\quad + \int_{t_3}^\infty \mathbb{E}[W_3|t_3 < B_1] f_{B_1}(u_1) du_1,
\end{aligned}$$

$$\begin{aligned}
&= \int_0^{t_2} \left(\int_0^\infty \mathbb{E}[W_3|B_1 < t_2, B_2 = u_2] f_{B_2}(u_2) du_2 \right) f_{B_1}(u_1) du_1 \\
&\quad + \int_{t_2}^{t_3} \left(\int_0^\infty \mathbb{E}[W_3|t_2 < B_1 < t_3, B_2 = u_2] f_{B_2}(u_2) du_2 \right) f_{B_1}(u_1) du_1 \\
&\quad + \int_{t_3}^\infty \left(\int_0^\infty \mathbb{E}[W_3|t_3 < B_1, B_2 = u_2] f_{B_2}(u_2) du_2 \right) f_{B_1}(u_1) du_1,
\end{aligned}$$

$$
\begin{aligned}
&= \int_0^{t_2} \left(\int_0^{\infty} (u_2 - t_3)^+ f_{B_2}(u_2) du_2 \right) f_{B_1}(u_1) du_1 \\
&\quad + \int_{t_2}^{t_3} \left(\int_0^{\infty} (u_2 - (t_3 - u_1))^+ f_{B_2}(u_2) du_2 \right) f_{B_1}(u_1) du_1 \\
&\quad + \int_{t_3}^{\infty} \left(\int_0^{\infty} ((u_1 - t_3)^+ + u_2) f_{B_2}(u_2) du_2 \right) f_{B_1}(u_1) du_1,
\end{aligned}
$$

$$
\begin{aligned}
&= \int_0^{t_2} \left(\int_{t_3 - t_2}^{\infty} (u_2 - (t_3 - t_2)) f_{B_2}(u_2) du_2 \right) f_{B_1}(u_1) du_1 \\
&\quad + \int_{t_2}^{t_3} \left(\int_{t_3 - u_1}^{\infty} (u_2 - (t_3 - u_1)) f_{B_2}(u_2) du_2 \right) f_{B_1}(u_1) du_1 \\
&\quad + \int_{t_3}^{\infty} \left(\int_0^{\infty} ((u_1 - t_3)^+ + u_2) f_{B_2}(u_2) du_2 \right) f_{B_1}(u_1) du_1,
\end{aligned}
$$

$$
\begin{aligned}
&= \int_0^{x_1} f_{B_1}(u_1) du_1 \int_{x_2}^{\infty} (u_2 - x_2) f_{B_2}(u_2) du_2 \\
&\quad + \int_{x_1}^{x_1 + x_2} \int_{x_1 + x_2 - u_1}^{\infty} (u_2 - (x_1 + x_2 - u_1)) f_{B_2}(u_2) du_2 f_{B_1}(u_1) du_1 \\
&\quad + \int_{x_1 + x_2}^{\infty} (u_1 - (x_1 + x_2) + \mathbb{E}B_2) f_{B_1}(u_1) du_1,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[W_3^2] &= \int_0^{t_2} \left(\int_{t_3 - t_2}^{\infty} (u_2 - (t_3 - t_2))^2 f_{B_2}(u_2) du_2 \right) f_{B_1}(u_1) du_1 \\
&\quad + \int_{t_2}^{t_3} \left(\int_{t_3 - u_1}^{\infty} (u_2 - (t_3 - u_1))^2 f_{B_2}(u_2) du_2 \right) f_{B_1}(u_1) du_1 \\
&\quad + \int_{t_3}^{\infty} \left(\int_0^{\infty} ((u_1 - t_3)^+ + u_2)^2 f_{B_2}(u_2) du_2 \right) f_{B_1}(u_1) du_1,
\end{aligned}
$$

$$
\begin{aligned}
&= \int_0^{x_1} f_{B_1}(u_1) du_1 \int_{x_2}^{\infty} (u_2 - x_2)^2 f_{B_2}(u_2) du_2 \\
&\quad + \int_{x_1}^{x_1 + x_2} \int_{x_1 + x_2 - u_1}^{\infty} (u_2 - (x_1 + x_2 - u_1))^2 f_{B_2}(u_2) du_2 f_{B_1}(u_1) du_1 \\
&\quad + \int_{x_1 + x_2}^{\infty} \left[(u_1 - (x_1 + x_2))^2 + 2(u_1 - (x_1 + x_2)) \mathbb{E}B_2 + \mathbb{E}B_2^2 \right] f_{B_1}(u_1) du_1.
\end{aligned}
$$

A.7 Simulation

To be able to evaluate schedules with service time distributions other than the exponential distribution we use simulation. In the function we display here, the input argument is an appointment schedule, the output is the loss of this schedule. This can be the (weigthed) quadratic or linear loss, but can also include the systems completion time. The mayor advantage of simulation is the freedom to choose any service time distribution and also any loss function. In the function the user is free to change the number of iterations. The current value of 100000 is used to evaluate schedules. This comes down to a confidence interval of $\approx 1\%$ for the schedules we evaluated.

One can also use this function to derive the optimal appointment schedule by minimizing it with `fminsearchbnd`. This is what we call 'simulated optimal' in Chapters 6 and 7. For this we use less iterations, 10000. This is because every iteration of `fminsearchbnd` needs 10000 iterations of the simulation function. Due to less simulation iteration the confidence interval grows and `fminsearchbnd` is less effective. There is a trade off between the number of simulation iterations and the number of `fminsearchbnd` iterations.

For this function we use a lag order sub optimal schedule as a start value for `fminsearchbnd`.

```
function [obj]=simulation(x)
%this function simulates an appointment schedule
%and computes the quadratic loss function

k = length(x); %aantal klanten
it = 100000; % aantal simulaties

eb2 = 5/4;mu = -log(eb2)/2;sigma = sqrt(log(eb2));
kappa = 2;lambda = 1/gamma(3/2);

leeg=0;
for j=1:it

propgenerator=zeros(1,k);
for i =1:k
    %propgenerator(i)=exprnd(1);
    %propgenerator(i) = lognrnd(mu,sigma);
    propgenerator(i) = wblrnd(lambda,kappa);
```

```
end

idle(1)=0;wait(1)=0;

for i=2:k+1
idle(i)=max(x(i-1)-(wait(i-1)+propgenerator(i-1)),0);
wait(i)=max(wait(i-1)+propgenerator(i-1)-x(i-1),0);
%recall the expressions from section 3.2
end

for i =1:k
     simul(1,i) = idle(i+1)^2;
     simul(2,i) = wait(i+1)^2;
     %simul(3,i) = simul(1,i) + simul(2,i);
end

leeg(j) = sum(simul(1,1:k)) + sum(simul(2,1:k));

end

%kies hier uit mean of confidence interval
obj = mean(leeg);
%obj(1)= mean(leeg)-1.98*std(leeg)/sqrt(it);
%obj(2)= mean(leeg)+1.98*std(leeg)/sqrt(it);

end
```

A.8 Tau

This thesis supports the Tau Manifesto, τ is defined as:

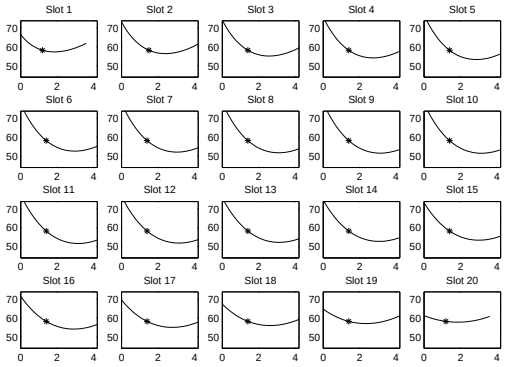
$$\tau := \frac{C}{r},$$

with C the circumference and r the radius of a circle. Hence, τ is related to π via:

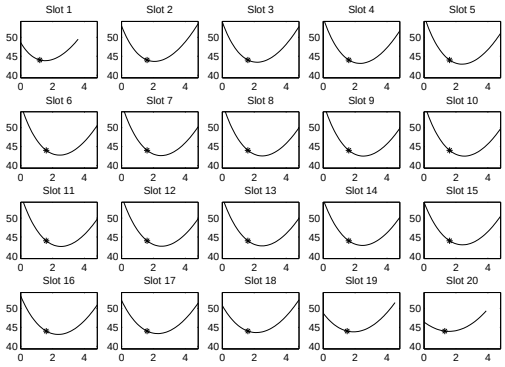
$$\pi = \frac{\tau}{2}.$$

The advantages of the usage of τ over π are that the former is more intuitive for young children when explaining radians. When using τ a quarter of a circle is $\frac{1}{4}\tau$ instead of the non-logic $\frac{1}{2}\pi$. Next, it is more common to find 2π in equations and expressions than π all by itself. See www.tauday.com for more information on the Manifesto.

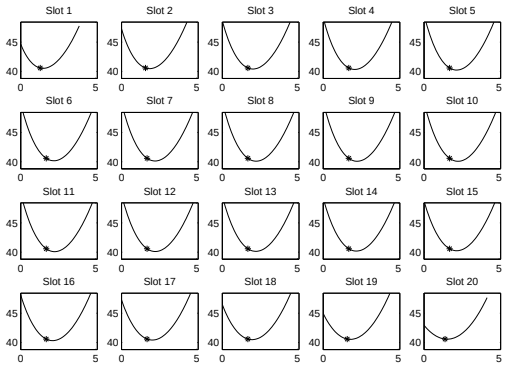
A.9 Sensitivity Analysis



(a) Lag Order II



(b) Lag Order III



(c) Lag Order IV

Figure A.1: Sensitivity analysis of lag order II, III and IV schedules per slot. Quadratic loss, n=21, i.i.d. exponential (μ=1) service times. The star is the actual value of the interarrival time. On the x-axis the value of the interarrival time and the quadratic loss on the y-axis.

Universiteit Utrecht

IBIS UvA

Instituut voor Bedrijfs- en Industriële Statistiek