European Journal of Operational Research 240 (2015) 213–219

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Optimal appointment scheduling in continuous time: The lag order approximation method

Wouter Vink [a], Alex Kuiper [b], Benjamin Kemper [b,c,*], Sandjai Bhulai [d]

[a] McKinsey & Company, Amstel 344, 1017 AS Amsterdam, The Netherlands
[b] Institute for Business and Industrial Statistics, University of Amsterdam, Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands
[c] EY Transaction Advisory Services, Antonio Vivaldistraat 150, 1083 HP Amsterdam, The Netherlands
[d] Stochastic Operations Research, Department of Mathematics, VU University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 25 June 2012
Accepted 19 June 2014
Available online 1 July 2014

Keywords:
Appointment scheduling
Heuristics
Lag order approximation method
Utility functions

ABSTRACT

We study appointment scheduling problems in continuous time. A finite number of clients are scheduled such that a function of the waiting time of clients, the idle time of the server, and the lateness of the schedule is minimized. The optimal schedule is notoriously hard to derive within reasonable computation times. Therefore, we develop the lag order approximation method, that sets the client's optimal appointment time based on only a part of his predecessors. We show that a lag order of two, i.e., taking two predecessors into account, results in nearly optimal schedules within reasonable computation times. We illustrate our approximation method with an appointment scheduling problem in a CT-scan area.

1. Introduction

In this paper we study appointment scheduling problems in continuous time. In our setting, we refer to appointment scheduling as the phenomenon in which a service provider is able to schedule arriving clients with the help of an *appointment schedule*; that is, a series of appointment times. The appointment time then offers the client a point in time upon which he or she should actually arrive to receive service.

It may be convenient to present this phenomenon as an appointment scheduling problem in two stages: in the first stage the provider schedules the appointments and in the second stage the server executes the service. In practice, one can imagine that the clients (or jobs) present themselves in random order at the first stage, and request the service provider to schedule them for a service. This paper discusses the decision making process of a service provider, in the first stage, on how to choose the appointment times of N clients that are to be scheduled to the server in the second stage. We develop a new approximation method that is generic in terms of the client's service-time distribution, numerically tractable for large problem instances while offering good performance.

Applications of an appointment scheme to schedule clients can be found in manufacturing (e.g., Wang, 1993), services (e.g., Kemper, Klaassen, & Mandjes, 2014), and health care (e.g., Cayirli & Veral, 2003). The basic setting in our paper, as described in the above, belongs to the so-called – *static* – class of appointment scheduling approaches, in which a finite number of appointments are scheduled prior to the beginning of the actual service, see Cayirli and Veral (2003). The origin of such an approach dates back to the work of Bailey (1952) and Welch and Bailey (1952), and generated substantial interest over the last decades.

Suppose in the first stage, the service provider is given N clients with random service times that are to be scheduled on a certain working day. Furthermore, suppose that the service-time distribution and clients' loss function due to waiting time, as well as the server's loss function, in terms of idle time and possible lateness after the final client (overtime), are known. The goal is then to minimize a convex combination of, possibly weighted, sum of the server's idle time and lateness (overtime), and the client's waiting time. Exact calculations of the optimal appointment times is problematic when there are many clients, since it requires the evaluation of high-dimensional integrals (Denton & Gupta, 2003).

Most of the contributions on appointment scheduling are based on exponential service times, such as in Wang (1999), Kaandorp and Koole (2007), Hassin and Mendel (2008) and Turkcan, Zeng, Muthuraman, and Lawley (2011); or a phase-type distribution for the service times, such as in Wang (1997), Vanden Bosch, Dietz, and Simeoni (1999) and Kuiper, Kemper, and Mandjes (2014). Also,

* Corresponding author at: Institute for Business and Industrial Statistics, University of Amsterdam, Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands. Tel.: +31 6 24994693.

E-mail addresses: wejvink@gmail.com (W. Vink), a.kuiper@uva.nl (A. Kuiper), benjaminkemper@gmail.com (B. Kemper), s.bhulai@vu.nl (S. Bhulai).

http://dx.doi.org/10.1016/j.ejor.2014.06.024
0377-2217/

it is common to assume independent and identically distributed random variables for the service times. It is reasonable to assume that the service-time distribution of the clients are independent, since clients call in at random for an appointment in the first stage. However, in practice the service times often do not follow an exponential distribution, let alone the service-time distributions of the arriving clients are identical (although Wang (1999) allows for service-time distributions with different service rates).

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

An alternative approach to deal with the high-dimensional optimization problem is to impose restrictions, such as equally-spaced interappointment times, see for example Hassin and Mendel (2008). Note that, however, in case of nonidentical service-time distribution it is argued that one should assign different interappointment times to different clients (Wang, 1999).

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

One may deal with different service-time distributions, and trace, in addition, an optimal sequence in case of a small schedule; see Weiss (1990) and Wang (1999), or slightly larger schedules (up to 16 clients) with a generalized lambda distribution, see Robinson and Chen (2003). In practice, however, one often encounters larger schemes, such as a car glass repair service or a dentist practice, which schedule up to 30 appointments per day.

Given the importance and relevance of the problem, and the fact that there is, to the best of our knowledge, no clean solution available, we decided to explore an alternative approach. Our approach is able to deal with general service-time distributions, such as the lognormal (Klassen & Rohleder, 1996) or the Weibull (Babes & Sarma, 1991) as often seen in practice, and larger schedules. In our approach the optimal appointment times depend only on a limited number of clients, that arrived previously to the client's appointment, leading to an optimization problem with reduced dimensionality. For example, we optimize a client's appointment time by minimizing his expected waiting times, corresponding idle times, and lateness of the server, while taking into account the effects of just two preceding clients. We refer to this method as the *lag order approximation method* in which the lag order refers to the number of predecessors taken into account.

The organization of the paper is as follows. In Section 2 we mathematically formulate the problem. The lag order approximation method is then presented in Section 3. The performance of the lag order approximation method is evaluated in Section 4 by studying some numerical examples and a real-life example from a radiology department. The results show that our method needs significantly less computational effort, and is able to derive appointment schedules that are close to optimal. Finally, we conclude and discuss directions for further research in Section 5.

2. Problem statement

Consider a service system at which N clients arrive at specified moments in time, i.e., client n arrives at time t_n with $t_n \in \mathbb{R}^+$ for $n = 1, \ldots, N$. Each client has a service-time requirement, which is denoted by the random variable B_n for client n. The service system has a single server and if upon arrival client n finds the server idle, he immediately starts his service. If the server is busy, then client n awaits his turn until all clients that are scheduled before client n have finished their service. We assume that both the clients and the server are punctual, and we do not allow for no-shows and walk-in clients. For studies that do include these phenomena, although in a different setting, we refer to Kemper et al. (2014) and references therein.

The vector $(t_1, \ldots, t_N)$ is called an appointment schedule for this service system. For a given schedule, we denote by I_n the time that the server has been idle upon start of the service of client n. We denote by W_n the waiting time of client n. Note that, the sojourn time S_n of client n can then be defined by $S_n = W_n + B_n$. In most settings the planning horizon (that is, the time span, T, in which clients can be scheduled) is finite. However, it can happen that after the planning horizon there are still clients that need to be served. We therefore define the lateness L as the overtime that the server has to make in order to finish all services. It is useful to define the *interappointment times* by

$$x_n = t_{n+1} - t_n, \quad n = 1, \ldots, N-1.$$

The idleness I_n can then be written as

$$I_n = \max\{x_{n-1} - S_{n-1}, 0\}, \quad n = 2, \ldots, N. \tag{1}$$

The waiting time W_n is given by

$$W_n = \max\{S_{n-1} - x_{n-1}, 0\}, \quad n = 2, \ldots, N. \tag{2}$$

From (1) and (2) follow $W_n + I_n = |S_{n-1} - x_{n-1}|$ for $n > 1$. The lateness can be expressed as

$$L = \max\{t_N + S_N - T, 0\}. \tag{3}$$

Clearly, it is reasonable to assume that $t_1 = 0$, so that both $W_1 = 0$ and $I_1 = 0$. If $t_1 > 0$ then $W_1 = 0$ as the first client still arrives on an empty system, but $I_1 > 0$ because the service provider has to wait t_1 amount of time. Moreover, it holds that $W_n \cdot I_n = 0$, $n = 1, \ldots, N$.

The objective of the appointment scheduling problem is to find a schedule $(t_2, \ldots, t_N)$, or equivalently $(x_1, \ldots, x_{N-1})$, such that a loss function LF, which depends on I_n, W_n, and L, is minimized. Throughout the paper, we assume that LF has the form

$$LF(x_1, \ldots, x_{N-1}) = \sum_{n=2}^{N} [\mathbb{E}f(I_n) + \mathbb{E}g(W_n)] + \mathbb{E}h(L), \tag{4}$$

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

3. The lag order approximation method

In this section we reveal the lag order approximation method in its general form. Basically, the optimal schedule is found through the optimization of (4), that is

$$\min_{x_1, \ldots, x_{N-1}} LF(x_1, \ldots, x_{N-1}) \tag{5}$$

The waiting time of client n is a random variable depending on $x_1, \ldots, x_{n-1}$, i.e., all the predecessors of client n are incorporated, $W_n = W_n(x_1, \ldots, x_{n-1})$. The main idea of the lag order approximation method is to neglect part of the predecessors that influence the waiting time (and idle time and lateness) of the loss function in (4), and express the waiting time for each client n in terms of its K predecessors, where K is the number of lags taken into the

optimization. However, in the beginning there are, of course, less than K predecessors included. Therefore we define $k = \min\{K, n-1\}$, so that the computation of the n-th client's waiting time depends only on $x_{n-k}, \ldots, x_{n-1}$, resulting in $\widetilde{W}_n = W_n(x_{n-k}, \ldots, x_{n-1})$. Similarly, we express the idle times and lateness as $\widetilde{I}_n = I_n(x_{n-k}, \ldots, x_{n-1})$ and $\widetilde{L} = L(x_{n-k}, \ldots, x_{n-1})$. The optimization with respect to this partial information in (4) is called the lag order approximation method of order K, which in essence minimizes

$$\min_{x_1,\ldots,x_{N-1}} LF_K(x_1, \ldots, x_{N-1}) = \sum_{n=2}^{N} \left[\mathbb{E}f(\widetilde{I}_n) + \mathbb{E}g(\widetilde{W}_n)\right] + \mathbb{E}h(\widetilde{L}). \tag{6}$$

Note that $K = N-1$ corresponds with the original optimization problem of (4). The advantage of this approach is that by limiting the dependence on predecessors we are able to use convolution formulas to compute the $\widetilde{W}_n$ and $\widetilde{I}_n$, and the schedule's lateness $\widetilde{L}$.

3.1. Loss functions

In this subsection, we present two loss functions that are commonly used in the literature. The loss function includes the expected waiting times, the expected idle times, and the expected lateness with different weighing factors. Most literature use the same weighing factors for the waiting times and the idleness, see Fries and Marathe (1981) for a detailed discussion on how to choose these factors. In this paper we adopt the same convention but remark that our method does not require it.

One often chooses general polynomial function and sets $f(x) = \alpha_1 x^\lambda$, $g(x) = \alpha_2 x^\lambda$ and $h(x) = \beta x$, where α_1, α_2, $\beta \geqslant 0$, and $\lambda > 0$. However, setting $\alpha_1 = \alpha_2$ and taking $\lambda = 1, 2$ gives us two remarkable insights, which we will refer to as the absolute value loss function and the quadratic loss function. Note that in these cases the idle and waiting times are equally weighted.

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

$$LF(x_1, \ldots, x_{N-1}) = \alpha \sum_{n=1}^{N-1} \mathbb{E}|S_n - x_n| + \beta \mathbb{E}L.$$

This loss function penalizes deviations from the planning (either caused by waiting or by idling) linearly. It has been used (with $\beta = 0$) by, e.g., Wang (1997), Vanden Bosch et al. (1999), Kaandorp and Koole (2007) and Kuiper et al. (2014).

Quadratic loss function. The quadratic loss function LF penalizes the deviation from the schedule quadratically instead of linearly. This can be achieved by taking.

$f(x) = g(x) = \alpha x^2$ and $h(x) = \beta x^2$. Since $W_n^2 + I_n^2 = (S_{n-1} - x_{n-1})^2$ for $n > 1$, the loss function reduces to

$$LF(x_1, \ldots, x_{N-1}) = \alpha \sum_{n=1}^{N-1} \mathbb{E}(S_n - x_n)^2 + \beta \mathbb{E}L.$$

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

3.2. Technical background of the lag order procedure

To compute the solution $x_1, \ldots, x_{N-1}$ through the lag order approximation method of order K, which we will refer to as lag order K throughout the rest of the paper, we use the following derivations. In case of lag order 0 the sojourn time distribution for each client is equal to his service time, i.e., $S_n = B_n$. Obviously, the interdependence between interappointments is removed in this way. This is different in the case of lag order I. Then the sojourn time distribution are linked through the waiting time of the previous client $n > 1$ (for $n = 1$ we use the lag order 0: $S_1 = B_1$)

$$S_n = B_n + W_n \approx B_n + \widetilde{W}_n(x_{n-1}) = B_n + \max\{B_{n-1} - x_{n-1}, 0\}. \tag{7}$$

In lag order II the interdependence increases to the waiting times of the two patients before client $n > 2$ (for $n = 2$ we use (7))

$$\begin{aligned} S_n &= B_n + W_n \approx B_n + \widetilde{W}_n(x_{n-2}, x_{n-1}) \\ &= B_n + \max\{B_{n-1} + \max\{B_{n-2} - x_{n-2}, 0\} - x_{n-1}, 0\}. \end{aligned} \tag{8}$$

Since we have now deduced approximation of the client specific sojourn times we can implement those in (1)–(3) to compute the waiting and idle times, and the lateness respectively. The algorithms to get to optimal lag order K solutions are written in `MATLAB R2012b`. This program finds the optimal interarrival times that minimizes the lag order K approximation of either the linear loss function or the quadratic loss function, for any service-time distribution, exploiting `MATLAB R2012b`'s built-in minimization routines.

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

1. The lag ordered loss functions are implemented in a `for loop`, using the lag order 0 for the second client; lag order I for the third client, c.f. (7); and the lag order II for the remaining clients, c.f. (8).
2. The sum of all losses is computed using `MATLAB`'s adaptive Simpson quadrature routines (tolerance is 10^{-5}), over the different lag ordered loss functions.
3. Finally, we jointly optimize the aggregate over $N-1$ interappointment times with tolerance level of 10^{-4} and a start vector consisting of ones with dimension $N-1$.

3.3. Technical background for generating LF values

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

In this paper we study the lag order approach for a broad range of settings, that is, for various service-time distributions, and for both linear and quadratic loss functions. In this section we explore the approach for exponential service times. Then, in Section 4, we study the lognormal and Weibull service-time distributions. For all results generated in the various studies in this paper, the simulation study concludes when the estimated values of the loss function exhibit a confidence interval of 1‰ of the estimates.

3.4. Example with exponentially distributed service times

We illustrate the results of our lag order approximation method for a system with $N = 11$ clients in Fig. 1. Here, we assume that the service times of the clients are independent and exponentially distributed (i.i.d.) with parameter $\mu_n = 1$ for $n = 1, \ldots, N$. The system operates under the quadratic loss function, i.e., the quadratic LF with $\alpha = 1$ and $\beta = 0$. Fig. 1 displays the optimal slot sizes per arriving client for various lag orders. The figure supports the conclusion that optimal scheduling according to lag order 0 corresponds to setting each interappointment time equal to the average service time, which is in essence a $D/M/1$ queue with load 1.

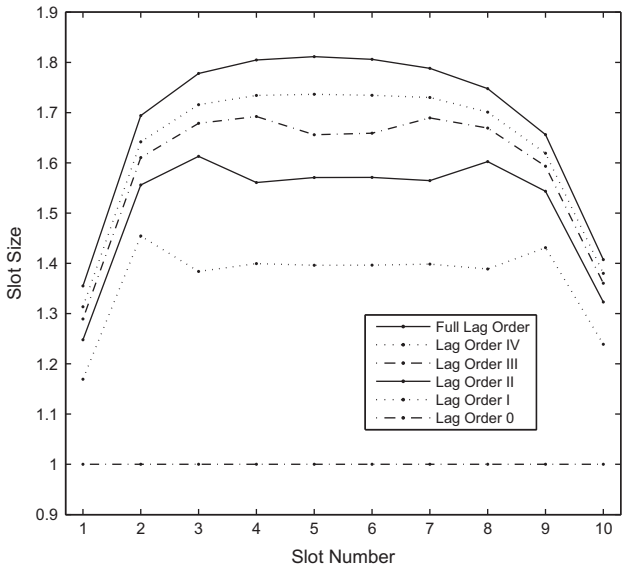


Fig. 1. Optimal slot sizes for the lag order approximation method with quadratic loss. $N = 11$, i.i.d. exponential ($\mu = 1$) service times.

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

The full lag order corresponds to a lag order of $N - 1$, i.e., all information is taken into account in deciding upon the slot sizes. The results of the full lag order correspond to the results of Wang (1997), where we extended the algorithm to handle quadratic loss functions. The results show that as the lag order increases, the appointment schedule converges to the optimal schedule. Fig. 1 also shows that the lag order is 'optimistic' in its next planned appointment time due to negligence of the waiting time, which is in favor of the service provider. As seen in the figure, the slot sizes increase, then decrease, and then increase and decrease again, hence exhibiting a non-unimodal pattern. However, as the lag order increases, this effect diminishes. In fact, the optimal appointment schedule (the full lag order) does not show this behavior, but instead follows a so-called dome shape which is a well-known pattern in the literature on appointment scheduling, see, e.g., Kaandorp and Koole (2007).

Table 1 summarizes the results of the lag order method in practice by displaying the value of the loss function for the first two lag orders including the full lag order obtained by extensive simulation. We see that increasing the lag order reduces the expected total loss of the system, which approaches the loss in the full lag order.

4. The lag order approximation method in practice

In this section we will apply the lag order approximation method to a realistic appointment scheduling problems. In the previous section, we were able to compute the optimal schedule for exponentially distributed service times. However, empirical evidence shows that this setting is too restrictive and unrealistic, see, e.g., Cayirli and Veral (2003). Commonly seen service-time distributions in realistic settings are the Weibull and the lognormal distribution. We study these distributions in Section 4.1. Then, in Section 4.2 we apply our approach to a real-life example of a CT-scan in a hospital.

4.1. Practical service times

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

In Tables 2 and 3 we show the results of the various lag orders to this problem. The results are compared to an optimal schedule derived through simulation only, as described in the background paragraph of Section 3.

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

From the tables we conclude that the application of the lag order approximation method results in a small loss of quality of the appointment schedule. In any case, linear or quadratic loss and either for Weibull or lognormal service-time distribution, our approach with lag order 0 generates schedules that are at least about 63% from the optimal LF value, and hence the lag order 0 is not useful in designing optimal appointment schedules which is in line with the results from the example in previous section (i.e., exponential service distribution). In case of a linear loss function (and either lognormal or Weibull service-time distribution), a lag order I approximation generates schedules that are within 25% from the optimal LF value, and that are more than 10, even 100 times in case of a quadratic loss function, faster to construct in terms of (computation) time, c.f. Table 1.

Furthermore, from the tables we see that the lag order II generates schedules that are reasonably close to the optimal LF value (around 2–7%). However, the computation time may not be sufficiently smaller in the case of linear loss, while with quadratic losses the lag order approximation method results in a reduction of computation time: about 3 times faster when CV = 0.85, up to about 20 times fast in case of CV = 0.35. We note that the code used for the application of the lag order method is programmed straightforwardly, see Section 3.

Table 1
Optimization results of the lag order approximation method compared with simulation. $N = 11$, i.i.d. exponential (CV = 1) service times.

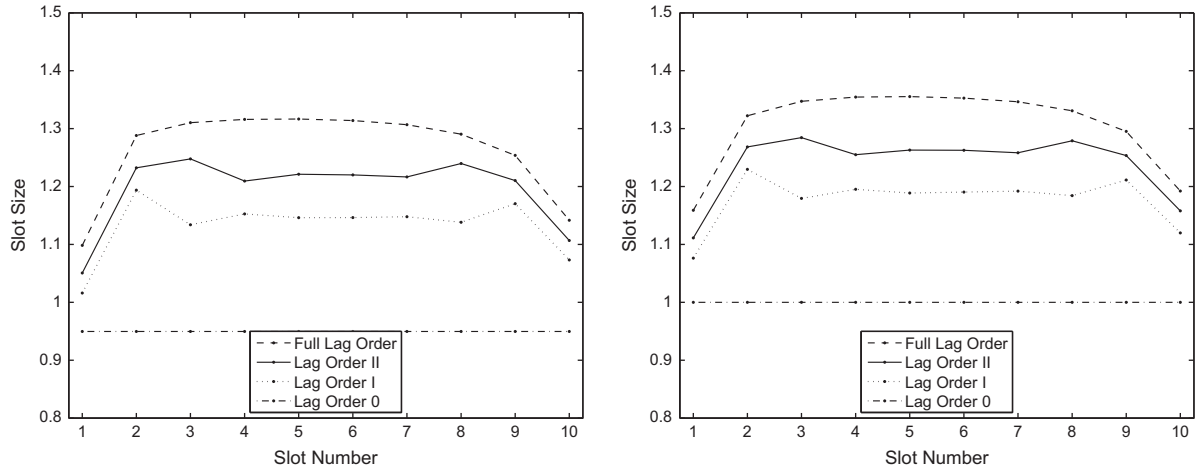
Method	Linear loss			Quadratic loss		
	Value LF	Δ Opt	Time (seconds)	Value LF	Δ Opt	Time (seconds)
Lag order 0	22.220	111.1%	1.2E−1	47.627	160.1%	4.7E−2
Lag order I	12.720	20.8%	1.1E1	22.918	25.2%	2.6E1
Lag order II	11.109	5.5%	3.9E2	19.476	6.4%	1.2E3
Simulation	10.526	–	5.0E2	18.311	–	5.9E3

Table 2
Optimization results of the lag order approximation method compared with the optimal values generated by simulation. $N = 11$, i.i.d. Weibull service times.

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

Table 3
Optimization results of the lag order approximation method compared with the optimal values generated by simulation. $N = 11$, i.i.d. lognormal service times.

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



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

Fig. 2. Optimal slot sizes for the lag order approximation method and simulation. $N = 11$, i.i.d. Weibull distributed service times, with mean 1 and CV = 0.5.

In Figs. 2 and 3 we illustrate the schedules derived by the lag order methods and the simulated optimal in both a Weibull and lognormal setting with coefficient of variation equal to 0.5. We see that the typical shapes of the lag order methods does not depend on actual distribution nor loss function.

In the following section we apply our approximation approach to a realistic example, wherein the clients loss function

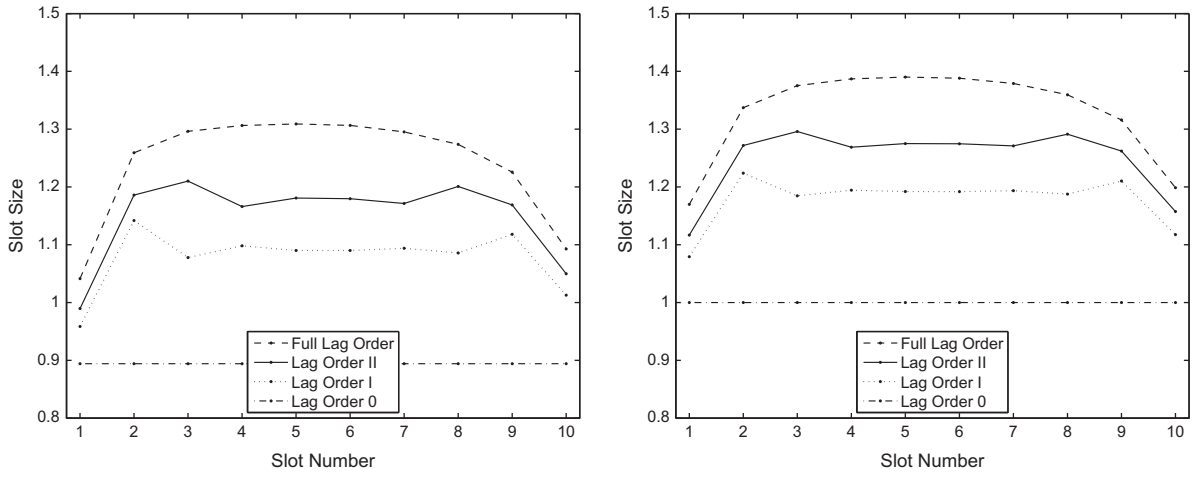


Fig. 3. Optimal slot sizes for the lag order approximation method and simulation. $N = 11$, i.i.d. lognormal distributed service times, with mean 1 and CV = 0.5.

Table 4
The hospital's current schedule and lag orders I and II compared with simulation. Quadratic weighted loss with lateness, $N = 20$, i.i.d. lognormal (CV = 0.63) service times.

Method	Value LF	Δ Opt	Time (seconds)
Hospital's schedule	2.000	20%	–
Lag order I	1.940	17%	4.3
Lag order II	1.788	7.7%	2E2
Equidistant simulation	1.670	<1%	5E2
Simulation	1.660	–	12E3

is quadratic. The service-time distribution follows a Weibull distribution.

4.2. Application to a CT-scan area

Let us consider a real-life scheduling problem in a CT-scan area, with the following typical parameters: $N = 20,\ T = 300$ (min). As loss function we choose the quadratic loss function with weights $\alpha_1 = 0.75,\ \alpha_2 = 0.25$, and $\beta = 1.5$. Thus, we have

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

We obtained service-time data from the Deventer Hospital described in De Mast, Kemper, Does, Mandjes, and Van der Bijl (2011). The best fit to this data results in a lognormal distribution with scale parameter $\mu = 2.4$ and location parameter $\sigma = 0.58$. The coefficient of variation is 0.63.

In order to compare our method, we state several approximation approaches in Table 4. The table includes the hospital's current schedule and the (full lag order) simulated optimal schedule. For each approach in our study the table reports the values of the loss function, the difference between the LF values of the approach and that of the simulated optimal value, and the computer's CPU time of the simulation study.

We observe that, given the loss function, the current schedule differs about 20% compared to the simulated optimal value of the LF, that is, the total loss of the system. Also, we see that the lag order I and lag order II differ about 17% and 7.7% from the optimal simulated outcome of the LF. The advantage of the lag order method is that it significantly gains in computation time compared to the simulated optimal value. Finally, we observe that an equidistant approximation method gets close to the simulated optimal results, but it takes our computer 2.5 times more CPU effort to get to these results.

5. Conclusion

In this paper we study the problem of appointment scheduling of N clients with a finite planning horizon. The clients are punctual and can be considered as jobs having random service times. However, we do not allow for no-shows and walk-in clients. We develop a lag order approximation method that minimizes a function of the waiting time of the clients, the idle time of the server, and the lateness of the schedule. The approximation method with a lag order I yields near-optimal schedules (about 20% from the optimal loss value level) within sufficiently smaller computation times. The lag order II approximation yields schedules that are about 5% from the optimal LF value, and can be up to 20 times faster than simulations in case of a quadratic loss function and when the CV (of the service-time distributions) is relatively small (i.e., CV = 0.35).

There are a few interesting directions for further research. A logical next step would be to extend the lag order approximation method to allow for no-shows and walk-in clients. Including both presents opportunities, since no-shows create gaps in the schedule which in turn can be potentially filled by walk-ins. Managing both at the same time is therefore a challenging endeavour.

We studied a group of identical CT-scan patients, but one could also study a setting for which there are several patient groups, see for example Creemers, Belën, and Lambrecht (2012).

Finally, note that there are many more factors that affect the optimality of appointment schedules. We mention a few examples: variability in the interappointment times, preferences of clients for a particular time of day, skill level of the server. Many of these issues cannot be dealt with in a straightforward manner and require new models. From a more algorithmic perspective, it would be interesting to investigate how the lag order approximation method can improve existing heuristics when combined.

References

Babes, M., & Sarma, G. (1991). Out-patient queues at the Ibn-Rochd health center. *Journal of the 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*(2), 185–199.
Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: A review of literature. *Production and Operations Management, 12*(4), 519–549.

Creemers, S., Belën, J., & Lambrecht, M. (2012). The optimal allocation of server time slots over different classes of patients. *European Journal of Operational Research, 219*(3), 508–521.
De Mast, J., Kemper, B., Does, R., Mandjes, M., & Van der Bijl, H. (2011). Process improvement in healthcare: Overall resource efficiency. *Quality and Reliability Engineering International, 27*(8), 1095–1106.
Denton, B., & Gupta, D. (2003). A sequential bounding approach for optimal appointment scheduling. *IIE Transactions, 35*(11), 1003–1016.
Fries, B., & Marathe, V. (1981). Determination of optimal variable-sized multiple-block appointment systems. *Operations Research, 29*(2), 324–345.
Günal, M., & Pidd, M. (2010). Discrete event simulation for performance modelling in health care: A review of the literature. *Journal of Simulation, 4*(1), 42–51.
Hassin, R., & Mendel, S. (2008). Scheduling arrivals to queues: A single-server model with no-shows. *Management Science, 54*(3), 565–572.
Ho, C.-J., & Lau, H.-S. (1992). Minimizing total cost in scheduling outpatient appointments. *Management Science, 38*(12), 1750–1764.
Kaandorp, G., & Koole, G. (2007). Optimal outpatient appointment scheduling. *Health Care Management Science, 10*(3), 217–229.
Kemper, B., Klaassen, C., & Mandjes, M. (2014). Optimized appointment scheduling. *European Journal of Operational Research, 239*(1), 243–255.
Klassen, K., & Rohleder, T. (1996). Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management, 14*(2), 83–101.
Kuiper, A., Kemper, B., & Mandjes, M. (2014). A computational approach to optimized appointment scheduling. *Queueing Systems*, 1–32. http://dx.doi.org/10.1007/s11134-014-9398-6.
Robinson, L., & Chen, R. (2003). Scheduling doctor's appointments: Optimal and empirically-based heuristic policies. *IIE Transactions, 35*(3), 298–307.
Schild, A., & Fredman, I. (1961). On scheduling tasks with deadlines and non-linear loss functions. *Management Science, 7*(3), 280–285.
Turkcan, A., Zeng, B., Muthuraman, K., & Lawley, M. (2011). Sequential clinical scheduling with service criteria. *European Journal of Operational Research, 214*(3), 780–795.
Vanden Bosch, P., Dietz, D., & Simeoni, J. R. (1999). Scheduling customer arrivals to a stochastic service system. *Naval Research Logistics, 46*(5), 549–559.
Wang, P. (1993). Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics, 40*(3), 345–360.
Wang, P. (1997). Optimally scheduling n customer arrival times for a single-server system. *Computers & Operations Research, 24*(8), 703–716.
Wang, P. (1999). Sequencing and scheduling n customers for a stochastic server. *European Journal of Operational Research, 119*(3), 729–738.
Weiss, E. (1990). Models for determining estimated start times and case orderings in hospital operating rooms. *IIE Transactions, 22*(2), 143–150.
Welch, J., & Bailey, N. (1952). Appointment systems in hospital outpatient departments. *The Lancet, 259*(6718), 1105–1108.
Yang, K., Lau, M., & Quek, S. (1998). A new appointment rule for a single-server, multiple-customer service system. *Naval Research Logistics, 45*(3), 313–326.
Zhu, Z., Heng, B., & Teow, K. (2012). Analysis of factors causing long patient waiting time and clinic overtime in outpatient clinics. *Journal of Medical Systems, 36*(2), 707–713.