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