Operations Research Letters 45 (2017) 549–555

Contents lists available at ScienceDirect

Operations Research Letters

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

Optimal stationary appointment schedules

Alex Kuiper [a], Michel Mandjes [a,b,*], Jeroen de Mast [a]

[a] *Amsterdam Business School, University of Amsterdam, The Netherlands*
[b] *Korteweg–de Vries Institute for Mathematics, University of Amsterdam, The Netherlands*

ARTICLE INFO

Article history:
Received 1 May 2017
Received in revised form 6 July 2017
Accepted 1 August 2017
Available online 30 August 2017

Keywords:
Appointment scheduling
Queueing
Phase-type distributions
Heavy traffic

ABSTRACT

A prevalent operations research problem concerns the generation of appointment schedules that effectively deal with variation in e.g. service times. In this paper we focus on the situation in which there is a large number of statistically identical customers, leading to an essentially equidistant ('stationary') schedule. We develop a powerful approach that minimizes an objective function incorporating the service provider's idle times and the customers' waiting times. Our main results concern easily computable, or even closed-form, approximations to the optimal schedule with a near-perfect fit. In addition, accurate explicit heavy-traffic approximations are provided, which, as we argue, can be considered as *robust*.

1. Introduction

Providers of service systems, e.g. in healthcare, are confronted with two opposite interests: on the one hand there is a need to control (or even reduce) costs, on the other hand, there is great pressure to improve service quality. On an operational level this amounts to avoiding excessive waiting times, whereas at the same time the utilization level at which the staff works should be kept sufficiently high; healthcare-related references are e.g. [1,9,10]. In operations research these conflicting interests are typically managed by using appropriate *appointment schedules*. The adequate design of such schedules is challenging due to various unpredictable factors as pointed out by [5].

In this paper we develop schedules for the situation that the randomness is caused by uncertainties in the service times, relying on techniques that originate in queueing theory. The focus is on the situation that (i) the service times of the individual customers are independent and statistically identical, and (ii) the number of customers to be scheduled is large. In this setting, we optimize an objective function that incorporates the system's utilization level (through the provider's idle time) as well as the customers' waiting times; these components are weighted with factors ω and $1-\omega$, respectively (for some $\omega \in (0,1)$). As the number of customers is large, the resulting optimal schedule is equidistant, and the queue is effectively behaving as a D/G/1 system in stationarity. Typically, already for relative small numbers of customers the thus obtained *stationary schedule* provides an accurate approximation for settings with finitely many customers; see e.g. [11, Fig. 2].

Early references on stationary schedules are [15,18]; a more recent paper in which stationary queues feature in an appointment scheduling setting is [17], where the focus lies on estimating the service provider's preferred value of ω. A general procedure that determines the interarrival time $\bar{x}$ of the optimal stationary schedule (for any given service-time distribution and any weight ω, that is), however, is still lacking; the development of such a procedure is the objective of this paper.

In line with a commonly used procedure in the appointment scheduling literature, we characterize the service-time distribution by its first two moments. Without loss of generality, we may normalize time such that the mean service time is 1, and we denote by ϱ the corresponding squared coefficient of variation. Our objective is to show that $\bar{x}$ follows (by good approximation) the functional form $1+A(\omega)\varrho^{B(\omega)}$. This functional form has the crucial advantage that knowledge of the functions $A(\cdot)$ and $B(\cdot)$ (which are both functions from $(0,1)$ to $(0,\infty)$) suffices to determine $\bar{x}$.

The main contributions of the paper are the following. We present three approaches to identify the optimal interarrival time. (i) In the first approach we approximate the service times by their phase-type counterpart, and determine $A(\omega)$ and $B(\omega)$ by numerical approximation. (ii) In the second approach we use explicit knowledge about the special cases that the service times have an exponential or Erlang(2) distribution, leading to a semi-explicit approximation for $A(\omega)$ and $B(\omega)$. (iii) The third approach, particularly accurate when ω is close to 1, is based on a heavy-traffic approximation, and yields a closed-form expression for $A(\omega)$ and $B(\omega)=\frac{1}{2}$. In addition we assess the impact of approximating a general service-time distribution by its phase-type counterpart.

* Corresponding author at: Amsterdam Business School, University of Amsterdam, The Netherlands.
E-mail addresses: a.kuiper@uva.nl (A. Kuiper), m.r.h.mandjes@uva.nl (M. Mandjes), jdemast@outlook.com (J. de Mast).

http://dx.doi.org/10.1016/j.orl.2017.08.002
0167-6377/

We conclude the paper by showing that the robust schedule (to be used when only the first two moments of the service-time distribution are available) coincides with the one based on the heavy-traffic approximation.

The paper is organized as follows. Section 2 discusses our model and preliminaries (such as those on the phase-type approximation). In Section 3 our approximations for the optimal stationary schedule are derived. Section 4 discusses the impact of the phase-type approximation and robust schedules.

2. Model and approach

In this section we first sketch the model considered in this paper by casting the appointment scheduling problem in a queueing-theoretic framework. The model is introduced for a finite population of customers, and then it is argued what queueing system is obtained when the number of customers grows large. We also describe how the service times are approximated by an appropriately chosen phase-type counterpart.

2.1. Preliminaries

We model the situation as a single-server queueing model. Customers $i = 1, \ldots, n$ arrive at or before their scheduled arrival time t_i, with $t_1 = 0$, where n is the number of customer to be seen in a single session; in this paper we primarily focus on the situation that n is large. We consider the situation in which the customers have appointments with a specific service provider, who therefore acts as a single server. We assume that the service times $B_1, \ldots, B_n$ are i.i.d. random variables. We define by W_i the net *waiting time* of the ith customer, that is, the time in between her scheduled arrival and the moment she receives service, where we set $W_1 = 0$. Define I_i as the *idle time* prior to the ith customer's arrival, with $I_1 = 0$. It is a standard result that, by virtue of the Lindley recursion, with $x_i = t_{i+1} - t_i$ (the interarrival time), the I_i can be determined recursively:

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

likewise,

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

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

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

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

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

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

2.2. Objective function

In our approach the schedules are generated so as to optimize a specific objective function consisting of the customers' waiting times and server's idle time. Weighting the relative importance of idle and waiting times by $\omega \in (0, 1)$, this performance degradation is expressed by the so-called weighted-linear objective function: for a customer population of size n,

$$\mathscr{F}^{(\ell)}[x_1, \ldots, x_{n-1}] = \omega \sum_{i=1}^{n} \mathbb{E}I_i + (1-\omega) \sum_{i=1}^{n} \mathbb{E}W_i. \tag{3}$$

For given weight ω, the optimal schedule is the sequence $\bar{x}_1, \ldots, \bar{x}_{n-1}$ that minimizes the objective function $\mathscr{F}^{(\ell)}[x_1, \ldots, x_{n-1}]$. Define $\overline{W}(\omega) = \sum_{i=1}^{n} \mathbb{E}W_i$ and $\bar{I}(\omega) = \sum_{i=1}^{n} \mathbb{E}I_i$ as the mean total waiting and idle time of the optimal schedule $\bar{x}_1, \ldots, \bar{x}_{n-1}$ for the weight ω. Generally, when ω approaches 1 (i.e., the situation in which the value of the objective function is essentially determined by the idle times only), $\overline{W}(\omega)$ explodes. Vice versa, when ω approaches 0 the contribution of the mean total idle time experienced by the server, i.e., $\bar{I}(\omega)$, increases sharply.

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

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

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

2.3. Stationarity

In this paper we focus on the situation that the B_i are governed by a single distribution, while we let n grow large. When the customers arrive equidistantly with interarrival time x, the distribution of the waiting time is uniquely defined through the distributional fixed point equation, cf. Eq. (1),

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

The resulting queueing system is of the D/G/1 type, which does not allow explicit solutions in general. In e.g. the cases of exponential and Erlang(2) service times, however, the stationary waiting-time distribution can be given in (semi-)closed-form; these results will facilitate the generation of accurate approximations of the optimal schedule, as we demonstrate in Section 3.

We now point out that the first moment $\mathbb{E}I$ can easily be found. Dividing (2) by n, taking expectations, and considering the limit when $n \to \infty$, we conclude that

$$\mathbb{E}I = x - \mathbb{E}B. \tag{4}$$

In the stationary setting the weighted-linear objective function equals

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

which is now a function of the (constant) interarrival time x only. The goal is to find the minimizer $\bar{x}$. It is easily seen that such a minimizer uniquely exists (and is larger than $\mathbb{E}B$), due to the fact that the objective function is convex. To this end, observe that $\mathbb{E}I$ is linear in x, whereas it is known that $\mathbb{E}W$ is convex in x.

The stationary version of the weighted-quadratic objective function evidently reads

$$\varphi^{(\mathrm{q})}[x] = \omega\, \mathbb{E}I^2 + (1-\omega)\mathbb{E}W^2.$$

The 'mixed' stationary objective functions $\varphi^{(\ell\mathrm{q})}$ and $\varphi^{(\mathrm{q}\ell)}$ are defined in a self-evident manner.

2.4. Phase-type fit

Unfortunately, for general service times B no analytical procedures are available to determine the above objective functions. We remedy this by replacing the actual service times by their so-called *phase-type counterparts*. The rationale behind this approach is the well-known fact that phase-type distributions are capable of approximating any positive distribution with arbitrary accuracy; see e.g. [4]. The resulting queueing system allows (semi-)explicit computation of the objective function, as pointed out in e.g. [13].