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/© 2017 Elsevier B.V. All rights reserved.

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 $i = 1, \dots, n$ 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 $B_1, \dots, B_n$ 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,

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

For given weight ω, the optimal schedule is the sequence $\bar{x}_1, \dots, \bar{x}_{n-1}$ that minimizes the objective function $\mathscr{F}^{(\ell)}[x_1, \dots, x_{n-1}]$. 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 $\bar{x}_1, \dots, \bar{x}_{n-1}$ 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:

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

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].

We have chosen to characterize the service-time distributions by fitting a phase-type distribution with the correct first two moments; the values of these moments can be chosen in line with for instance the findings of [5]. This choice is motivated by the fact that it is cumbersome to estimate higher moments, where it is in addition anticipated that those higher moments have only a modest impact on the performance of an appointment schedule (a claim that we later corroborate in Section 4.1).

In line with the literature on scheduling, we represent the first two moments by (i) the mean, and (ii) the *squared coefficient of variation* (ϱ), a unitless quantity that is defined as the ratio of the variance and the square of the mean. We follow the standard procedure, advocated in e.g. [19], to approximate the service time by a mixture of two Erlang random variables if it has a ϱ smaller than 1, and by a hyperexponential random variable if it has a ϱ larger than 1. In more detail, the approximation is constructed as follows.

- In case ϱ is smaller than 1 the service-time distribution is approximated by a mixture of Erlang distributions: it is an Erlang distribution with $K-1$ phases and mean $(K-1)/\mu$ with probability $p \in [0, 1)$, and an Erlang distribution with K phases and mean K/μ with probability $1-p$. It can be verified that the ϱ of this distribution lies in the interval $(1/K, 1/(K-1)]$, for $K \in \{2, 3, \dots\}$. As a result, we can identify unique K, μ, and p such that our mixture of Erlangs has the desired mean and ϱ.
- In the other situation, in which ϱ is larger than 1, the service time is approximated by a hyperexponential random variable, which is constructed as an exponential random variable with mean μ_1^{-1} with probability p, and an exponential random variable with mean μ_2^{-1} with probability $1-p$. By imposing *balanced means* (i.e., $\mu_1 = 2p\mu$ and $\mu_2 = 2(1-p)\mu$ for some $\mu > 0$) one reduces the number of free parameters from three to two, so that for each mean and ϱ a unique hyperexponential distribution can be determined.

In [13] it is explained how to evaluate the objective functions in the case that the service times are of phase-type, relying on earlier results presented in [21]. In Section 4.1 we assess the error due to approximating the service-time distribution by its phase-type counterpart.

3. Stationary schedules

We consider the model introduced in Section 2, where we assume the number of customers n to be relatively large. As the service times are independent and identically distributed, we anticipate that in an optimal schedule the interarrival times of customers scheduled in the middle of a session are about equal, say $\bar{x}$. In this situation, the optimal schedule could be approximated by a schedule in which *all* interarrival times are set to $\bar{x}$, the so-called *stationary schedule*. The main objective of this section is to devise an efficient procedure to identify $\bar{x}$ when the service times stem from the phase-type distributions introduced in Section 2.

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

$$\bar{x} \equiv \bar{x}(\omega, \varrho) = 1 + A(\omega)\varrho^{B(\omega)}, \tag{5}$$

for functions $A(\cdot)$ and $B(\cdot)$ that we can accurately determine. The crucial advantage of the functional form (5) is that only knowledge of the two curves $A(\cdot)$ and $B(\cdot)$ is needed to compute $\bar{x}(\omega, \varrho)$ for any ω and ϱ.

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

3.1. Numerical determination of stationary schedules

Our first approach is to empirically identify the functions $A(\cdot)$ and $B(\cdot)$ featuring in (5). To this end, we have implemented the following least-squares approach, for each $\omega \in \Omega := \{0.1, \dots, 0.9\}$.

- For $\varrho \in \mathscr{R} = \{0.2, 0.3, \dots, 3.0\}$ we numerically determine $\bar{x}(\omega, \varrho)$, as follows. For any given value of x, we can evaluate the objective function (requiring the computation of first and/or second moments of the idle and waiting times) relying on the machinery for phase-type service times of [13]. Then we use standard numerical software to minimize the resulting objective function over x.
- We then observe that (5) entails that

  $$\log\big(\bar{x}(\omega, \varrho) - 1\big) = \log A(\omega) + B(\omega) \log \varrho.$$

  This means that we can use the method of least squares to determine, for any ω, $A(\omega)$ and $B(\omega)$, based on the 29 data points determined in the first step.

The resulting curves, say $A_{\rm num}(\cdot)$ and $B_{\rm num}(\cdot)$, are depicted in Fig. 1–2. The fit is excellent: in the case of a linear objective function the coefficient of determination is at least $R^2 = 0.9998$ for any $\omega \in \Omega$, and for a quadratic objective function at least $R^2 = 0.9988$. The mixed linear–quadratic objective functions give similar results.

3.2. Analytical derivation of stationary schedules

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

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

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

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

We now explain how to evaluate $\bar{x}(\omega, 1)$. We start by identifying the interarrival time $\bar{x}^{(\ell)}(\omega, 1)$ corresponding to the case of the weighted-linear objective function. As pointed out in [11], the distribution of the stationary waiting-time $W \equiv W(x)$ is given by

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

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

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

with $f_B(z) = e^{-z}$ denoting the service-time density. In other words, S has an exponential distribution with mean $1/(1-\sigma_x)$. Recalling

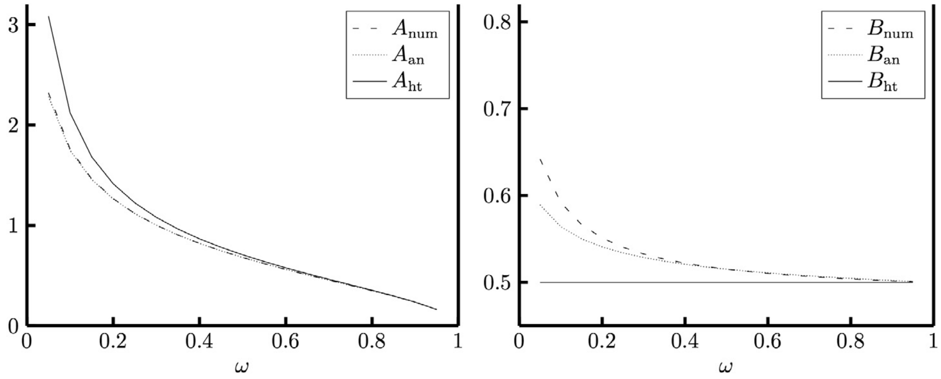

Fig. 1. The curves $A(\omega)$ (left panel) and $B(\omega)$ (right panel) for the weighted-linear objective function. The dashed curves use the numerical approach described in Section 3.1, the dotted curves the analytical approach in Section 3.2, and the solid curves correspond to the heavy-traffic approach in Section 3.3.

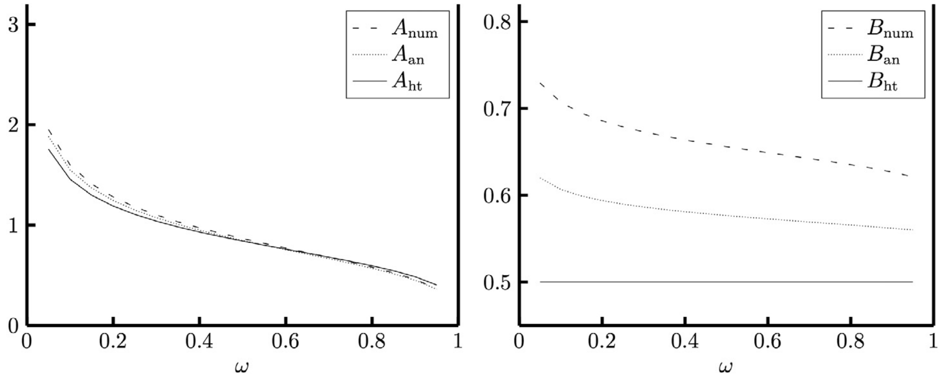

Fig. 2. The curves $A(\omega)$ (left panel) and $B(\omega)$ (right panel) for the weighted-quadratic objective function. The dashed curves use the numerical approach described in Section 3.1, the dotted curves the analytical approach in Section 3.2, and the solid curves correspond to the heavy-traffic approach in Section 3.3.

that, due to (4), $\mathbb{E}I = x - \mathbb{E}B = x - 1$, the following objective function needs to be minimized with respect to $x \geqslant 1$:

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

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

$$(1-\omega)\frac{\sigma_x'}{(1-\sigma_x)^2} + \omega = 0. \qquad (6)$$

From the definition of σ_x, it directly follows that

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

From (6), after straightforward algebra, we find that $\bar{x}^{(\ell)}(\omega, 1)$ equals $-\log\sigma^{(\ell)}(\omega)/(1-\sigma^{(\ell)}(\omega))$, where $\sigma^{(\ell)}(\omega)$ is the unique solution in $[0, 1]$ of

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

Let $\mathscr{W}(\cdot) : [e^{-1}, \infty) \mapsto [-1, \infty)$ denote one of the two real branches of the Lambert W-function, i.e., $\mathscr{W}(x)$ is the largest solution for w in the equation $we^{w} = x$; see e.g. [6]. It follows that $\sigma^{(\ell)}(\omega) \in [0, 1]$ can be written as

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

We eventually obtain

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

A similar procedure can be followed for the other objective functions we consider. For completeness we also show how $A_{\mathrm{an}}^{(\mathrm{q})}(\omega)$ can be found; the mixed linear–quadratic objective functions can be handled analogously, and therefore are not discussed. First observe that $\mathbb{E}W^2 = 2\sigma_x/(1-\sigma_x)^2$. Also,

$$\mathbb{E}I^2 = \mathbb{E}(S-x)^2 - \mathbb{E}W^2 = \mathbb{E}(W+B-x)^2 - \mathbb{E}W^2, \qquad (7)$$

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

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

so that we eventually obtain

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

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

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

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

$$\omega\left(\frac{(\log\sigma)^2 + 2\log\sigma + 2(1-\sigma)}{(1-\sigma)^2}\right) + (1-\omega)\frac{2\sigma}{(1-\sigma)^2}. \qquad (8)$$

With $\sigma^{(\mathrm{q})}(\omega)$ the $\sigma \in [0, 1]$ that minimizes (8), it thus follows that

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

We are thus left with determining the function $B_{\mathrm{an}}(\cdot)$. Note that the case $\varrho = \frac{1}{2}$ corresponds to the D/E$_2$/1 queue for which analytic expressions are available, too; see e.g. [8, p. 109]. Once we have identified $\bar{x}(\omega, \frac{1}{2})$, we can find $B_{\rm an}(\omega)$ by solving

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

Realizing that we have derived $A_{\rm an}(\omega)$ above, it now follows that

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

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

$$f(\tau) := \left(1 - \frac{\tau}{2}\right)^2 = e^{-\tau x} =: g(\tau).$$

Essentially from (i) $f(\infty) = \infty$ and $g(\infty) = 0$, (ii) $f(2) = 0$ and $g(2) > 0$, (iii) $f(0) = g(0) = 1$, and (iv) $f'(0) = -1 > -x = g'(0)$, it follows that the above equation has two positive roots, one of which lies between 0 and 2 while the other is larger than 2. Call these roots $\tau_1 \equiv \tau_{1,x}$ and $\tau_2 \equiv \tau_{2,x}$. Then $\mathbb{P}(W > y) = c_1 e^{-\tau_1 y} + c_2 e^{-\tau_2 y}$, where $c_1 \equiv c_{1,x}$ and $c_2 \equiv c_{2,x}$ solve

$$\frac{c_1}{1 - \tau_1/2} + \frac{c_2}{1 - \tau_2/2} = 1 \quad \text{and} \quad \frac{c_1}{(1 - \tau_1/2)^2} + \frac{c_2}{(1 - \tau_2/2)^2} = 1,$$

i.e.,

$$c_1 = \frac{\tau_2}{\tau_2 - \tau_1}\left(1 - \frac{\tau_1}{2}\right)^2 = \frac{\tau_2 e^{-\tau_1/x}}{\tau_2 - \tau_1}$$
$$\text{and} \quad c_2 = \frac{\tau_1}{\tau_1 - \tau_2}\left(1 - \frac{\tau_2}{2}\right)^2 = \frac{\tau_1 e^{-\tau_2/x}}{\tau_1 - \tau_2}.$$

Realizing that again $\mathbb{E}I = x - 1$, we are to minimize

$$\varphi^{(\ell)}[x] = \omega\,\mathbb{E}I + (1 - \omega)\mathbb{E}W = \omega(x - 1) + \frac{(1 - \omega)c_{1,x}}{\tau_{1,x}} + \frac{(1 - \omega)c_{2,x}}{\tau_{2,x}}$$

in order to obtain $\bar{x}^{(\ell)}(\omega, \frac{1}{2})$. (Semi-)closed-form expressions cannot be derived now, but it takes a straightforward numerical routine to perform the minimization.

The resulting curves $A_{\rm an}(\cdot)$ and $B_{\rm an}(\cdot)$ are depicted in Figs. 1–2. Regarding the $A_{\rm an}(\cdot)$ curve, the fit is still remarkably good; it is also seen that the curve is at most 1% off the curve determined by the approach of Section 3.1. The $B_{\rm an}(\cdot)$ curve in the linear case is still rather precise, in the quadratic case the performance is slightly worse; note that due to the (fine) scales chosen in the picture, the difference may seem more substantial than it actually is. The overall performance remains rather good, as reflected in the R^2 which measures for any $\omega \in \Omega$ the discrepancy between the $\bar{x}(\omega, \frac{1}{2})$ as predicted by the analytical approach presented in this subsection and the corresponding true values. In the case of a weighted-linear objective function for any $\omega \in \Omega$ the R^2 is at least 0.9914. For the weighted-quadratic objective function the fit somewhat degrades: the R^2 ranges from 0.91 for $\omega = 0.1$ to 0.96 for $\omega = 0.9$.

3.3. *Heavy traffic derivation of stationary schedules*

In this subsection we provide a theoretical justification for the use of schedules based on the form (5). More specifically, we show that this relation is *exact* in the heavy-traffic regime, i.e., the regime in which x is only slightly larger than the mean service time, which we normalized to 1.

It is well-known that $(x - 1)W(x)$ for $x \downarrow 1$ converges to a random variable that is distributed exponentially with mean $\frac{1}{2}\,\mathbb{V}{\rm ar}\,B$ [3, Section X.7]. As a consequence, one can approximate $W(x)$ by an exponential distribution with mean μ_x^{-1}, with $\mu_x = 2(x - 1)/\varrho$. Observe that due to our normalization $\mathbb{E}B = 1$ we have $\mathbb{V}{\rm ar}\,B = \varrho$. It is anticipated that for ω close to 1, the optimal interarrival times are relatively short as the system is relatively indifferent with respect to waiting times, and therefore one could expect that in this regime heavy-traffic approximations lead to accurate predictions.

Thus, the weighted-linear objective function reads

$$\varphi^{(\ell)}_{\rm ht}[x] = \omega(x - 1) + (1 - \omega)\frac{\varrho}{2(x - 1)}.$$

The corresponding minimization allows an explicit solution:

$$\bar{x}^{(\ell)}_{\rm ht}(\omega, \varrho) = 1 + A^{(\ell)}_{\rm ht}(\omega)\varrho^{B^{(\ell)}_{\rm ht}(\omega)}, \quad \text{with } A^{(\ell)}_{\rm ht}(\omega) = \sqrt{\frac{1 - \omega}{2\omega}},$$
$$B^{(\ell)}_{\rm ht}(\omega) = \frac{1}{2}.$$

The weighted-quadratic objective function requires more care. Recall (7); the objective function equals

$$\omega\,\mathbb{E}I^2 + (1 - \omega)\mathbb{E}W^2 = \omega\,\mathbb{E}(W + B - x)^2 + (1 - 2\omega)\mathbb{E}W^2.$$

Using standard properties of the exponential distribution, we have the heavy-traffic approximation

$$\mathbb{E}W^2 = \int_0^\infty \mu_x e^{-\mu_x z} z^2 {\rm d}z = \frac{2}{\mu_x^2} = \frac{\varrho^2}{2(x - 1)^2}.$$

Likewise, with $f_B(\cdot)$ the density of B,

$$\mathbb{E}(W + B - x)^2 = \int_0^\infty \int_0^\infty f_B(y)\,\mu_x e^{-\mu_x z}(y + z - x)^2 {\rm d}z\,{\rm d}y,$$

which by an elementary computation turns out to equal $\mathbb{E}W^2 + (x - 1)^2$. We thus obtain the objective function

$$\varphi^{(\rm q)}_{\rm ht}[x] = \omega(x - 1)^2 + (1 - \omega)\frac{\varrho^2}{2(x - 1)^2},$$

being minimized by

$$\bar{x}^{(\rm q)}_{\rm ht}(\omega, \varrho) = 1 + A^{(\rm q)}_{\rm ht}(\omega)\varrho^{B^{(\rm q)}_{\rm ht}(\omega)}, \quad \text{with } A^{(\rm q)}_{\rm ht}(\omega) = \sqrt[4]{\frac{1 - \omega}{2\omega}},$$
$$B^{(\rm q)}_{\rm ht}(\omega) = \frac{1}{2}.$$

Observe that in both the weighted-linear case and the weighted-quadratic case, we find that in heavy traffic the optimal interarrival times have the shape $1 + A(\omega)\sqrt{\varrho}$, but interestingly, for the mixed objective functions we obtain different structures:

$$\bar{x}^{(\ell{\rm q})}_{\rm ht}(\omega, \varrho) = 1 + \sqrt[3]{\frac{1 - \omega}{\omega}}\varrho^{2/3}$$
$$\text{and} \quad \bar{x}^{({\rm q}\ell)}_{\rm ht}(\omega, \varrho) = 1 + \sqrt[3]{\frac{1 - \omega}{4\omega}}\varrho^{1/3}.$$

The resulting curves for the weighted-linear case and the weighted-quadratic cases, say $A_{\rm ht}(\cdot)$ and $B_{\rm ht}(\cdot)$, can be found in Figs. 1–2. The overall fit is remarkably good, and for ω approaching 1 even excellent.

4. Discussion

In this section we first assess the impact of the phase-type approximation. We also show 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 of Section 3.3.

4.1. Impact of phase-type approximation

In this subsection we assess the impact of replacing the service-time distribution by its phase-type counterpart as we proposed in Section 2.4. We present experiments with Weibull and lognormal service times (in line with earlier healthcare-related studies, see e.g. [5]). Importantly, the Weibull and lognormal distributions do *not* belong to the class of phase-type distributions. Both distributions are characterized by two parameters, which are chosen such that the mean equals 1, whereas ϱ is varied in the experiment. It is stressed that there are no explicit results for D/G/1 systems with Weibull or lognormal service times, and therefore we have to resort to simulation: we first estimate the objective function as a function of the interarrival time x, and then optimize it over x. We show that the resulting stationary schedule virtually coincides with the one obtained using the phase-type service times with the same first two moments.

More specifically, we have used the following procedure to estimate the objective function as a function of x. In each simulation run we sampled $M = 100\,000$ service times. We used this (single) run to estimate the distribution of the stationary waiting time $W(x)$ (for all $x \geqslant 1$ on a fine grid), by simulating the queue with interarrival times x and service times $B_1, \dots, B_M$. The estimates of the stationary waiting-time distributions are further improved by repeating this experiment 1000 times. Then we minimize the objective function; in the output presented below this is the weighted-linear objective function $\varphi^{(\ell)}$, but the other objective functions give comparable results.

The experiments have revealed that there is almost a perfect fit, in terms of the maximum difference between the optimal interarrival time based on the actual service-time distribution and the one based on the phase-time fit, when varying ϱ between 0.1 and 1.5. For Weibull service times this maximum difference is 2.7% for $\omega = 0.1$, but this sharply drops when ω increases, and is only 0.067% for $\omega = 0.9$. For lognormal service times we see the same pattern, with the maximum difference decreasing from 3.0% for $\omega = 0.1$ to 0.093% for $\omega = 0.9$.

4.2. Robust schedules

In a recent paper by Mak et al. [14] the setup is studied in which there is *limited information* on the shape of the service-time distribution available; a specific example of this concerns the case in which only the mean $\mathbb{E}B$ and the variance $\mathbb{V}\text{ar}\,B$ are given (for instance due to the fact that there are only few historical data available).

Following the line of reasoning in [14], the optimization problem to be solved in this 'robust setting' becomes (for the linear objective function)

$$\min_{x \geqslant 1} \max_{B \in \mathscr{B}(\varrho)} \varphi^{(\ell)}[x],$$

where $\mathscr{B}(\varrho)$ is the set of non-negative distributions with mean 1 and squared coefficient of variation ϱ. Now observe that $\mathbb{E}I = x - 1$, and hence does not depend on ϱ. The maximization over $B \in \mathscr{B}(\varrho)$ therefore amounts to maximizing $\mathbb{E}W$ over this set of distributions.

Kingman [12] showed, for the specific case of the D/G/1 queue, that for any $B \in \mathscr{B}(\varrho)$,

$$\mathbb{E}W \leqslant \frac{\varrho}{2(x-1)}.$$

In addition, Trengove [20] proved that this upper bound is *tight*; see also [7]. This means that one can construct distributions $B \in \mathscr{B}(\varrho)$ such that the upper bound in the above inequality can be approached arbitrarily closely. The crucial consequence of this fact is that the robust optimization problem reduces to

$$\min_{x \geqslant 1} \omega(x-1) + (1-\omega)\frac{\varrho}{2(x-1)},$$

which is precisely the optimization problem that we encountered in Section 3.3 for the heavy-traffic regime (and which allowed an explicit solution). We conclude that the solution to the robust optimization problem coincides with the heavy-traffic solution that we obtained in Section 3.3.

Table 1
Stationary schedules with $\omega = 0.8$ and various ϱ values.

ϱ	$\bar{x}^{(\ell)}_{\text{num}}$	$\bar{x}^{(\ell)}_{\text{an}}$	$\bar{x}^{(\ell)}_{\text{ht}}$
0.2	1.155	1.155	1.158
0.5	1.246	1.246	1.250
1.0	1.349	1.349	1.353
2.0	1.495	1.495	1.500

This paper has studied the case of fixed interarrival times. An extension could concern situations in which the clients' arrival epochs may slightly deviate from the scheduled epochs. The resulting setting relates to the 'almost deterministic arrival processes' that were analyzed in e.g. Araman and Glynn [2]. In addition, there is a strong connection to the analysis of the effect of *jitter* in communication networks; see e.g. Roberts et al. [16, Ch. III].

4.3. Example

We end this section by an example that demonstrates how our findings can be applied. Suppose we use the weighted-linear objective function, and we consider a situation with parameter values suggested in [5]: $\omega = 0.8$ and $\varrho = 0.5$. We assume that the mean service time is 10 min.

The numerical method pointed out in Section 3.1 can be applied as follows. From Fig. 1, we see that $A_{\text{num}}(0.8) = 0.349$ and $B_{\text{num}}(0.8) = 0.504$, leading to $\bar{x}^{(\ell)}_{\text{num}}(0.8, 0.5) = 1 + 0.349 \cdot 0.5^{0.504} = 1.246$, and hence the optimal interarrival time is $10 \cdot 1.246 \approx 12.5$ min. The analytical method of Section 3.2 gives the same value for $\bar{x}_{\text{an}}(0.8, 0.5)$ (up to three digits). The approach of Section 3.3, based on the heavy-traffic regime (which coincides with the optimal robust schedule), yields nearly the same result:

$$\bar{x}^{(\ell)}_{\text{ht}}(0.8, 0.5) = 1 + \sqrt{1/8} \cdot 0.5^{0.5} = 1.250,$$

leading to an optimal arrival time of 12.5 min as well.

For the sake of completeness we performed the same calculations for other values of ϱ, keeping ω at 0.8; see Table 1.

Acknowledgments

The authors would like to thank Ruben Brokkelkamp for coding assistance, Madelon de Kemp for bringing [7,20] to our attention, and Gideon Weiss for various valuable discussions. We also thank an anonymous reviewer for his/her thoughtful feedback.

References

[1] R. Anderson, F. Camacho, R. Balkrishnan, Willing to wait?: The influence of patient wait time on satisfaction with primary care, BMC Health Serv. Res. 7 (1) (2007) 7–31.

[2] V. Araman, P. Glynn, Fractional Brownian motion with $H < \frac{1}{2}$ as a limit of scheduled traffic, J. Appl. Probab. 49 (2012) 710–718.

[3] S. Asmussen, Applied probability and queues, in: Stochastic Modelling and Applied Probability, Springer-Verlag, New York, NY, USA, 2003.

[4] S. Asmussen, O. Nerman, M. Olssen, Fitting phase-type distributions via the EM algorithm, Scand. J. Stat. 23 (4) (1996) 419–441.

[5] T. Çayırlı, E. Veral, Outpatient scheduling in health care: A review of literature, Prod. Oper. Manage. 12 (4) (2003) 519–549.

[6] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, D. Knuth, On the Lambert W-function, Adv. Comput. Math. 5 (1) (1996) 329–359.

[7] D. Daley, A. Kreinin, C. Trengove, Inequalities concerning the waiting-time in single-server queues: a survey, Queueing Relat Models (1992).

[8] L. Goddard, Mathematical Techniques of Operational Research, in: International Series of Monographs on Pure and Applied Mathematics, Pergamon, Oxford, United Kingdom, 1963.

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

[10] X. Huang, Patient attitude towards waiting in an outpatient clinic and its applications, Health Serv. Manage. Res. 7 (1) (1994) 2–8.

[11] B. Kemper, C. Klaassen, M. Mandjes, Optimized appointment scheduling, European J. Oper. Res. 239 (1) (2014) 243–255.

[12] J. Kingman, Some inequalities for the GI/G/1 queue, Biometrika 49 (1962) 315–324.

[13] A. Kuiper, B. Kemper, M. Mandjes, A Computational approach to optimized appointment scheduling, Queueing Syst. 79 (1) (2015) 5–36.

[14] H. Mak, Y. Rong, J. Zhang, Appointment scheduling with limited distributional information, Manage. Sci. 61 (2) (2015) 316–334.

[15] A. Mercer, A queueing problem in which the arrival times of the customers are scheduled, J. R. Stat. Soc. Ser. B Stat. Methodol. (1960) 108–113.

[16] J. Roberts, U. Mocci, Virtamo , Broadband network teletraffic –Performance evaluation and design of broadband multiservice networks: Final report of action COST 242, 1996.

[17] L. Robinson, R. Chen, Estimating the implied value of the customer's waiting time, Manuf. Serv. Oper. Manag. 13 (1) (2011) 53–57.

[18] A. Soriano, Comparison of two scheduling systems, Oper. Res. 14 (3) (1966) 388–397.

[19] H. Tijms, Stochastic modelling and analysis — a computational approach, in: Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, Chichester, UK, 1986.

[20] C. Trengove, Bounds for the mean waiting times in queues, M. Sc. Thesis, Univ. of Melbourne, 1978.

[21] P. Wang, Optimally scheduling N customer arrival times for a single-server system, Comput. Oper. Res. 24 (8) (1997) 703–716.