Queueing Syst
DOI 10.1007/s11134-014-9398-6

A Computational approach to optimized appointment scheduling

Alex Kuiper · Benjamin Kemper · Michel Mandjes

Received: 24 December 2012 / Revised: 27 October 2013

Abstract Appointment scheduling is prevalent in various healthcare settings. Generally, the objective is to determine a *schedule* (i.e., the sequence of epochs at which the individual patients are asked to appear) that appropriately balances the interests of the patients (low waiting times) and the medical staff (low idle times). In queueing language, the planner is given the distributions of the service times of the individual clients, and then it is his task to determine the arrival epochs of the clients. In this paper, we demonstrate how to generate schedules that have certain optimality properties. As a general principle, we express the performance of a schedule in terms of its associated utility, which incorporates both waiting times and idle times. In a first class of schedules (referred to as the *simultaneous approach*), the arrival epochs are chosen such that the sum of the utilities of all clients as well as the service provider is minimized. In a second class (*sequential approach*), the arrival epoch of the next client is scheduled, given the scheduled arrival epochs of all previous clients. For general service times the numerical evaluation of the optimal schedules is often prohibitive; it essentially requires knowledge of the waiting-time distribution in an appropriately chosen D/G/1 queue. In this paper, we demonstrate that by using the phase-type counterparts of the service-time distributions, it is feasible to efficiently determine an optimized schedule, that is, we obtain accurate results with low computational effort. We do so both for transient scenarios (in which the number of clients is relatively low, so that the interarrival time is not uniform) and stationary scenarios (with many clients, and

A. Kuiper (✉) · B. Kemper
Institute for Business and Industrial Statistics, University of Amsterdam, Plantage Muidergracht 12,
1018 TV Amsterdam, The Netherlands
e-mail: a.kuiper@uva.nl

M. Mandjes
Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 904,
1098 XH Amsterdam, The Netherlands

Published online: 18 April 2014

essentially constant interarrival times). Our approach is backed by several examples, that give insight in the impact of the variability of the service times on the schedule; it also shows the impact of the utility function selected.

Keywords Appointment scheduling · Healthcare · Utility functions · D/G/1 queue · Phase-type distribution

Mathematics Subject Classification 60K25 · 68M20 · 68Q87 · 90B36

1 Introduction

The problem of appointment scheduling is prevalent in various healthcare settings. The objective is to (somehow) optimize the utilities of the agents involved, i.e., the provider of the service (i.e., the medical staff) and the clients (i.e., the patients). It is clear that the server and clients have opposite interests: the service provider's interest is to minimize the amount of server idleness, whereas clients seek to minimize their waiting times. For example, clients arrive at the appointed arrival times at a dentist. Upon arrival, the job (i.e., the client) may have to wait till the server (i.e., the dentist) finishes the work on the previous clients. This situation is favorable for the dentist (no time wasted), but not for the client (waiting time). On the other hand, sometimes the dentist finishes the service for all previous clients, and stays idle till the appointed arrival time of the next client. This is obviously favorable for the client, but less so for the dentist.

A way to balance the interests of the medical staff and the patients is to minimize the "disutilities" experienced by both server and clients. More concretely, an optimized schedule is such that the system's *risk* (the expectation of a loss function that involves both waiting times and idle times) is minimized, thus realizing an optimal trade off between the agents' interests; here "schedule" is understood as the vector of appointed arrival times.

The main objective of this paper is to develop (computationally feasible) techniques for generating optimal schedules. More precisely, in queueing-theoretic terms: the planner is given the distributions of the service times of the individual clients, and then it is his task to determine the corresponding optimal (that is, disutility-minimizing) arrival epochs. A good schedule has the potential to have relatively low cost of service (in terms of idle time), while maintaining a sufficiently high level of quality at the same time (in terms of waiting time). In our work we limit ourselves to the situation in which the *order* of the arrivals is fixed.

Above we mentioned the appointment scheduling problem at a dentist's, but there are many more healthcare-related examples. One could think of scheduling MRI and CT patients; realize that MRI and CT scanners are expensive devices, so that it makes sense to make sure that they are left unused relatively infrequently. Another typical example is scheduling the usage of operating rooms in a hospital or clinic; bear in mind that there are often just a small number of these rooms. In both examples, the patients' waiting times should be taken into account as well: poor scheduling performance may lead to patients choosing other hospitals.

A commonly studied version of the appointment scheduling problem is the following. Assuming a quadratic loss function, finding the optimal schedule requires solving

$$\min_{t_1,\ldots,t_n} \sum_{i=1}^{n} \left(\mathbb{E}I_i^2 + \mathbb{E}W_i^2\right), \tag{1}$$

with t_i denoting the appointed arrival time of client i, with I_i the server's idle time prior to that arrival, and with W_i the waiting time of the i-th client. Importantly, the random variables I_i and W_i are also affected by the arrival epochs $t_1, \ldots, t_{i-1}$ of all previous clients. As a consequence, solving the above optimization problem is typically hard: for general service times no manageable expressions are available for the quantities involved, and, in addition, an n-dimensional optimization needs to be performed. To mitigate these complications, we developed in [10] a *sequential* variant of the above *simultaneous* optimization problem. The "sequential" variant determines the i-th appointment time t_i with all earlier arrival epochs being known, that is $t_1, \ldots, t_{i-1}$. Instead of optimizing over all t_i simultaneously, the optimization problem reduces to optimizing over a single t_i (for given $t_1, \ldots, t_{i-1}$)

$$\min_{t_i} \left(\mathbb{E}I_i^2 + \mathbb{E}W_i^2\right), \quad i = 1, \ldots, n. \tag{2}$$

Since we have n clients to be scheduled, we have to perform n subsequent optimizations to determine all t_i s. Besides a computational advantage, due to the fact that we optimize over a single variable only, this approach has an *explicit* solution:

$$t_1 := 0, \quad \text{and} \quad t_i := \sum_{j=1}^{i-1} \mathbb{E}S_j, \quad i = 2, \ldots, n,$$

with S_j denoting client j's sojourn time. Evidently, our model is a stylized description of reality: patients arrive punctually, show up with certainty, and there is no additional stream of urgent arrivals. Some of these features, however, can be included in the analysis; see, for example, the discussion in [10].

Importantly, it is shown in [10] that this approach applies to not just quadratic loss, but to the more general class of convex loss functions, and to arbitrary service time distributions. In addition, it is neither required that clients' service times stem from a single distribution, nor that the clients have the same loss function.

A first general remark is that there are situations in which the simultaneous approach is the more natural framework, while in other situations the simultaneous approach fits better. In a situation in which all information about all patients is available *a priori* (i.e., a list of patients to be scheduled, including the distributions of their service times), the logical procedure is to minimize a simultaneous objective function. There are situations, however, where patients call the service provider to make an appointment on the same day. In such cases, the schedule gradually fills and there a sequential policy is natural.

The sequential approach has a substantial computational advantage over the simultaneous approach. The optimization is done on a patient-by-patient basis, while in the simultaneous approach there is a single "social welfare" optimization (e.g., the optimal schedule maximizes the aggregate utility of all actors involved, viz., all clients and the server). A crucial question here is how the choice of a specific optimization scheme affects the utilities as perceived by the individual actors: does one of the approaches lead to a schedule that is in favor of some patients but highly disadvantageous for others? Later in the paper, we systematically assess this effect.

In the literature, a wide variety of objective functions have been proposed. We do not advocate the use of a specific loss function; as mentioned earlier, one of the attractive features of the techniques presented in this paper is that we allow for a broad spectrum of possible choices (simultaneous vs. sequential approach, quadratic vs. linear loss, weighing the individual terms the objective function is composed of by specific coefficients). The choice for a particular loss function is very much case-specific, even within the domain of healthcare. Any objective chosen (simultaneous vs. sequential approach, quadratic vs. linear loss, weighing the individual terms the objective function is composed of by specific coefficients) has its own repercussions in terms of the (dis-)utility experienced by the individual actors of the model (the patients and the server).

The main objective of this paper is to study the computational feasibility of the scheduling algorithms described above. They require knowledge of first and second moments of the idle times and waiting times, which are (for general service times) not available. The main idea is to rely on a *two-moment fit*, advocated in, for example, [18], in which the service time under consideration is replaced by a phase-type distribution (of low dimension) with the same mean and variance. These phase-type distributions do allow (semi-)explicit expressions for the utility functions, which can then be optimized over the t_i s. In the sequential case this is a single-dimensional optimization (which has an explicit solution for quadratic and linear loss, as described above); in the simultaneous case it is an n-dimensional optimization.

Consider the case that the service times are independent and identically distributed (i.i.d.). In case the number of clients is relatively low, the interarrival times may substantially vary (realize that the first client finds the system empty with certainty). If, on the contrary, there are many clients, the schedule will tend to steady state: the optimal interarrival times will be essentially constant. By making a connection to appropriately chosen D/G/1 queues (with the service times phase-type), we demonstrate how to determine the corresponding "stationary schedules," both in the simultaneous and sequential approach. Methodologically, our work is related to, for example, [7,8,13,21].

In [7] a generating function approach is relied on to facilitate quick and accurate evaluation of (discrete-time) schedules. The main idea behind [13] is to approximate the service-time distribution, using a fit of the first four moments, by a beta distribution (which has four parameters); then an efficient technique is developed to numerically evaluate the (linear) objective function. The idea of using phase-type distributions has been advocated in [21].

In addition, we mention the related work by Weiss [22], who develops a surprisingly accurate, yet easily implementable, heuristic to approximate the schedule that minimizes a linear cost function (where it is noted that this heuristic can be seen as an example of the sequential approach derived in [10]). Robinson and Chen [17] focus on techniques that facilitate the estimation of the relative cost of the patient waiting time given average queue length and occupation rate. Luo et al. [15] develop, for general cost functions, effective heuristics that also take into account the fact that patients may cancel the appointment, or do not show up.

This paper is organized as follows. In Sect. 2 we mathematically introduce our model, and define the risk functions considered. The approach followed is presented in Sect. 3. Section 4 demonstrates our approach for transient schedules (situations with relatively low numbers of clients, that is), while Sect. 5 considers the stationary counterpart (many clients). In Sect. 6, we discuss the potential and limitations of our approach; in particular, we show that the error due to the phase-type fit is small. Section 7 concludes and suggests ideas for future work.

Various graphs pictorially illustrate a number of interesting effects. We quantify the following features: (i) the convergence of transient schedules (relatively low number of clients) to their stationary counterparts; (ii) the impact of the choice of the risk function on the schedule; (iii) the impact of the service times' variability on the schedule. Also the differences between the simultaneous and sequential approach are studied in greater detail. Evidently, replacing a non-phase-type distribution by its phase-type counterpart introduces an error; a simulation study shows that the impact of this error is negligible.

2 Background and model

The mathematical treatment of the appointment scheduling problem with one server dates back to at least the seminal works of [4] and [23]. Since then, a sizeable number of papers has appeared in the operations research literature. As a general remark, the results in these papers tend to be rather case-specific, in terms of the service-time distribution under consideration as well as the loss function chosen. One often relies on simulations to overcome the inherent computational complexities. Such an approach has clear limitations: it evidently lacks general applicability, and, more importantly, it does not provide us with any structural insights into the nature of the solution. Our aim is, therefore, to develop an approach that works for general service times, general loss functions, and that is numerically feasible.

A common way to reason about an appointment scheduling problem is to define for each arrival i a so-called *risk*. This risk is then the expectation of a loss function that consists of a part reflecting the idle time and a part reflecting the waiting time. A natural choice is $\mathbb{E}g(I_i) + \mathbb{E}h(W_i)$, where it makes sense to choose non-decreasing loss functions $g(\cdot)$ and $h(\cdot)$ with $g(0) = h(0) = 0$. Observe that these risks clearly depend on the arrival epochs t_i and service times B_i; more precisely, the risk associated to the i-th client depends on the arrival epochs t_1 up to t_i and service times B_1 up to B_{i-1}. The optimal schedule corresponding to the simultaneous approach then follows from solving the minimization problem over the arrival epochs solely

$$\min_{t_1,\ldots,t_n} \sum_{i=1}^{n} \left(\mathbb{E}g(I_i) + \mathbb{E}h(W_i)\right), \tag{3}$$

whereas its sequential counterpart minimizes $\mathbb{E}g(I_i) + \mathbb{E}h(W_i)$ over t_i, with $t_1, \ldots, t_{i-1}$ given.

In our paper, we focus on a quadratic and a linear loss function, but, importantly, the setup carries over to any loss function in the class defined above. For a quadratic loss the risk is defined by

$$R_i^{(q,\alpha)}(t_1,\ldots,t_i) := \alpha\mathbb{E}I_i^2 + (1-\alpha)\mathbb{E}W_i^2, \qquad i=1,\ldots,n \qquad \text{and} \qquad \alpha\in(0,1).$$

Due to the well-known *Lindley recursion* [14],

$$I_i = \max\{t_i - t_{i-1} - W_{i-1} - B_{i-1}, 0\} \tag{4}$$

and

$$W_i = \max\{W_{i-1} + B_{i-1} - t_i + t_{i-1}, 0\}. \tag{5}$$

Let $S_i := W_i + B_i$ denote the sojourn time of the i-th client, with distribution function $F_{S_i}(\cdot)$. In addition, define by $x_{i-1} := t_i - t_{i-1}$ the time between the $(i-1)$-st and i-th arrival. Then, with (4) and (5) in mind, we may write the system's risk (in relation to the i-th client) as

$$\begin{aligned} R_i^{(q,\alpha)}(t_1,\ldots,t_{i-1},t_{i-1}+x_{i-1}) : &= \alpha\mathbb{E}I_i^2 + (1-\alpha)\mathbb{E}W_i^2 \\ &= \alpha\mathbb{E}\,(x_{i-1}-S_{i-1})^2\,\mathbf{1}_{x_{i-1}>S_{i-1}} \\ &\quad + (1-\alpha)\mathbb{E}\,(S_{i-1}-x_{i-1})^2\,\mathbf{1}_{x_{i-1}<S_{i-1}}. \end{aligned} \tag{6}$$

This is a nonnegative convex function of x_{i-1}. In the sequel, we specialize to the case of equal weights, that is, $\alpha = \frac{1}{2}$. In that case, the risk related to the i-th client reduces to $\frac{1}{2}\mathbb{E}\,(S_{i-1}-x_{i-1})^2$ (where we can leave out the factor $\frac{1}{2}$ for obvious reasons). For $\alpha \neq \frac{1}{2}$ there is no such a simplification of the expressions. The computation time required to determine optimal schedules does not depend on the choice of α, though: all cases can be evaluated in essentially the same amount of computation time. At the end of Sect. 5.3, we assess the effect of the weights in steady state. This study is done by computing the optimal interarrival times for various α s in (0, 1).

In the case of a linear loss function, the risk associated with the i-th client equals the sum of the expected waiting time and the expected idle time. Again, due to (4) and (5), we obtain, again with $\alpha\in(0,1)$,

$$\begin{aligned} R_i^{(a,\alpha)}(t_1,\ldots,t_{i-1},t_{i-1}+x_{i-1}) : &= \alpha\mathbb{E}I_i + (1-\alpha)\mathbb{E}W_i \\ &= \alpha\mathbb{E}\,(x_{i-1}-S_{i-1})\mathbf{1}_{x_{i-1}>S_{i-1}} \\ &\quad + (1-\alpha)\mathbb{E}\,(S_{i-1}-x_{i-1})\mathbf{1}_{x_{i-1}<S_{i-1}}, \end{aligned} \tag{7}$$

which is a nonnegative convex function of x_{i-1}. Again, we consider in this paper just the case of equal weights, so that the risk related to the i-th client reduces to $\frac{1}{2}\mathbb{E}|S_{i-1}-x_{i-1}|$ (where again we can leave out the factor $\frac{1}{2}$).

3 The phase-type approach

As argued earlier in this paper, the main problem when generating schedules of a realistic size concerns the fact that neither explicit expressions are available for the expected idle times and waiting times (or the corresponding second moments), nor for the distributions of the sojourn times—these are needed to be able to evaluate the objective function (which then needs to be optimized, either sequentially or simultaneously). This section proposes an approach to circumvent this problem, by replacing the service times by a phase-type counterpart (of relatively low dimension). For these approximate service times, we can evaluate the first and second moments of the client's sojourn time and therefore, through (6) and (7), client i's risk associated with I_i and W_i, as it will turn out.

The approach we propose in this paper consists of three steps:

1. Based on the service-time's mean and variance (or, equivalently, the mean and the coefficient of variation), we fit a phase-type distribution.
2. With a recursive procedure we derive, for each client, the sojourn-time distribution (for the fitted phase-type distribution).
3. The phase-type based sojourn-time distribution enables us to evaluate the objective function. Relying on standard numerical packages, we can then solve the simultaneous optimization problem as stated in (3). In the sequential counterpart it suffices to compute the expected value (in case of a quadratic loss) or median (in case of a linear loss) of the sojourn times.

In this section, we provide further details on our approach; in Sects. 4 and 5, we demonstrate the resulting procedure in transient (relatively few clients) and steady-state (relatively many clients) settings.

3.1 Phase-type fit of service-time distribution

In the first step of our approach we use phase-type distributions to fit the service-time distributions in the system under study. It is well-known from the literature that phase-type distributions, that are mixtures and convolutions of exponential distributions (such as mixtures of Erlang distributions, or hyperexponential distributions), are able to approximate any positive distribution arbitrarily accurately, see, for example, [2] and [18].

The reason to use phase-type distributions is twofold. In the first place, due the enforced Markovianity, the resulting system often enables the computation of explicit expressions for various queueing-related metrics, such as the waiting times distribution (where "explicit" typically means in terms of eigenvalues/eigenvectors of an associated eigensystem). In the second place, restricting ourselves to phase-type distribution of a certain dimension, estimating this distribution from data can be done via a (semi-) parametric density estimation procedure.

In our study, we use the idea presented in [18] to match the first and second moment of the service-time distribution, or, equivalently, the mean and the squared coefficient of variation (SCV); the SCV of the random variable X is defined as its variance divided by the square of the mean. In line with [11], we choose to match a mixture of two

Erlang distributions in case the actual service-time distribution has an SCV smaller than 1, and a hyperexponential distribution in case of an SCV larger than (or equal to) 1. More precisely:

- In case SCV < 1, we match the service-time distribution with a mixture of two Erlang distributions with the same scale parameter, denoted as $E_{K-1,K}(\mu;\, p)$. A sample from this distribution is obtained by sampling from an Erlang distribution with K phases and mean K/μ with probability p, and from an Erlang distribution with $K-1$ phases and mean $(K-1)/\mu$ with probability $1-p$. Its n-th moment is given by

$$\mathbb{E}\left[E_{K-1,K}^{n}\right] = p\frac{(K+n-2)!}{(K-2)!}\frac{1}{\mu^{n}} + (1-p)\frac{(K+n-1)!}{(K-1)!}\frac{1}{\mu^{n}},$$

  with $p \in [0, 1]$. The corresponding SCV equals

$$\frac{K-(1-p)^2}{(K+p-1)^2},$$

  which lies between $1/K$ and $1/(K-1)$ for $K \in \{2, 3, \ldots\}$. We can thus uniquely identify an $E_{K-1,K}(\mu;\, p)$ distribution matching the first two moments of the target distribution, as long as SCV < 1.
- In case SCV ≥ 1, we match the service-time distribution with a specific type of the hyperexponential distribution, viz., a mixture of two exponential distributions, to be denoted by $H_2(\boldsymbol{\mu};\, p)$, with $\boldsymbol{\mu} = (\mu_1, \mu_2)$. Its n-th moment is given by

$$\mathbb{E}\left[H_2^{n}\right] = p\frac{n!}{\mu_1^{n}} + (1-p)\frac{n!}{\mu_2^{n}}.$$

  We impose the additional condition of *balanced means*, see Eq. (A.16) in [18]; that is, we require $\mu_1 = 2p\mu$ and $\mu_2 = 2(1-p)\mu$ for some $\mu > 0$. The corresponding SCV equals $(2p(1-p))^{-1}$, which is larger than (or equal to) 1. It can be verified that

$$p = \frac{1}{2}\left(1 \pm \sqrt{\frac{\text{SCV}-1}{\text{SCV}+1}}\right).$$

  It is readily checked that for the special case SCV = 1, the fit results in an exponential distribution (with $p = \frac{1}{2}$).

3.2 Recursive procedure to derive sojourn time distribution

We now present a procedure to compute the sojourn-time distribution of any specific client, in case the service times are of phase type. We specialize to mixtures of Erlangs (i.e., $E_{K-1,K}(\mu;\, p)$) and hyperexponentials (i.e., $H_2(\boldsymbol{\mu};\, p)$), as these are the ones we fitted our “actual” service-time distributions to. It is noted, though, that the procedure works for any phase-type distribution; see, for example, [21]. In the sequel, we assume that the service times are i.i.d., but the procedure can be extended to independent, *non*-identically distributed phase-type service times, at the expense of rather involved notation.

A phase-type distribution is characterized by an $m \in \mathbb{N}$, an m-dimensional vector $\boldsymbol{\alpha}$ with nonnegative entries adding up to 1, and $\boldsymbol{S} = (s_{ij})_{i,j=1}^{m}$ an $(m \times m)$-dimensional matrix such that $s_{ii} < 0$, $s_{ij} \geq 0$ and $\sum_{j=1}^{m} s_{ij} \leq 0$ for any $i \in \{1, \dots, m\}$.

- In case SCV < 1, we use an $E_{K-1,K}(\mu; p)$ distribution (as explained in Sect. 3.1). Then $m = K$, and the vector $\boldsymbol{\alpha}$ such that $\alpha_1 = 1$ and $\alpha_i = 0$ for $i = 2, \dots, K$. In addition $s_{ii} = -\mu$ for $i = 1, \dots, K$ and $s_{i,i+1} = -s_{ii} = \mu$ for $i = 1, \dots, K-2$, while $s_{K-1,K} = (1-p)\mu$; all other entries are 0.
- In case SCV ≥ 1, we use a $H_2(\boldsymbol{\mu}; p)$ distribution (as explained in Sect. 3.1). Then $m = 2$, and $\alpha_1 = p = 1 - \alpha_2$. Also, $s_{ii} = -\mu_i$, for $i = 1, 2$, while the other two entries of $\boldsymbol{S}$ equal 0.

For more background on phase-type distributions, see [1].

Next, we briefly describe the algorithm, presented in [21], that determines the clients' sojourn-time distributions. To this end, define the bivariate process $\{N_i(t), K_i(t), t \geq 0\}$ for client $i = 1, \dots, n$. Here, $N_i(t)$ is the number of clients present in front of the i-th arriving clients, t time units after her arrival; obviously $N_i(t) \in \{i, \dots, i-1\}$. The second component, $K_i(t) \in \{1, \dots, m\}$, represents the phase of the client in service t time units after the arrival of the i-th client, where $N_i(t) = 0$ refers to the case that the last arriving client is in service. We also introduce the probabilities, for $t \geq 0, i = 1, \dots, n,\ j = 0, \dots, i-1$, and $k = 1, \dots, m$,

$$p_{j,k}^{(i)}(t) = \mathbb{P}\left(N_i(t) = j, K_i(t) = k\right).$$

In addition, the following vector (of dimension mi) plays a crucial role:

$$\begin{aligned}\boldsymbol{P}_i(t) := \Big(&p_{i-1,1}^{(i)}(t), \dots, p_{i-1,m}^{(i)}(t), p_{i-2,1}^{(i)}(t), \dots, p_{i-2,m}^{(i)}(t),\\ &\dots, p_{0,1}^{(i)}(t), \dots, p_{0,m}^{(i)}(t)\Big).\end{aligned}$$

The sojourn-time distribution of the i-th client can be computed from $\boldsymbol{P}_i(t)$ through the identity, with $\boldsymbol{e}_{mi}$ an all-one vector of dimension mi,

$$F_i(t) := \mathbb{P}(S_i \leq x) = 1 - \sum_{j=0}^{i-1} \sum_{k=1}^{m} p_{j,k}^{(i)}(t) = 1 - \boldsymbol{P}_i(t)\boldsymbol{e}_{mi},$$

Considering the first client, to arrive at $t_1 = 0$, it is standard that $\boldsymbol{P}_1(t) = \boldsymbol{\alpha} \exp(\boldsymbol{S}t)$ (which is an m-dimensional object). Concerning the second client, arriving x_2 after the first client, it can be argued that

$$\boldsymbol{P}_2(t) = (\boldsymbol{P}_1(x_2), \boldsymbol{\alpha} F_1(x_2)) \exp(\boldsymbol{S}_2 t), \quad t \geq 0$$

which is an object of dimension $2m$; here, with $\boldsymbol{s} := -\boldsymbol{S}\boldsymbol{e}_m$ and $\mathbf{0}_{m,m}$ an $(m \times m)$ all-zero matrix,

$$\boldsymbol{S}_2 := \begin{pmatrix} \boldsymbol{S} & \boldsymbol{s}\boldsymbol{\alpha} \\ \mathbf{0}_{m,m} & \boldsymbol{S} \end{pmatrix}.$$

The sojourn-time distributions of the other clients can be found recursively in a similar manner. To this end, define the matrix $\boldsymbol{T_i}$ of dimension $(i-1)m \times m$ through

$$\boldsymbol{T_i} := (\mathbf{0}_{m,m}, \mathbf{0}_{m,m}, \dots, \mathbf{0}_{m,m}, \boldsymbol{s}\boldsymbol{\alpha})^T;$$

also

$$\boldsymbol{S}_i := \begin{pmatrix} \boldsymbol{S}_{i-1} & \boldsymbol{T}_i \\ \mathbf{0}_{m,(i-1)m} & \boldsymbol{S} \end{pmatrix}.$$

Then the vector $\boldsymbol{P}_i(t)$ (dimension mi) can be found from $\boldsymbol{P}_{i-1}(t)$ (dimension $m(i-1)$) by the recursion

$$\boldsymbol{P}_i(t) = (\boldsymbol{P}_{i-1}(x_{i-1}), \boldsymbol{\alpha} F_{i-1}(x_{i-1})) \exp(\boldsymbol{S_i t}), \quad t \geq 0.$$

Realize that in our examples the matrix $\boldsymbol{S}$ is upper triangular (in the hyperexponential case in fact even diagonal), and hence so are the matrices $\boldsymbol{S}_i$. As a consequence, the eigenvalues can be read off from the diagonal. This property facilitates easy computation of the matrix exponent $\exp(\boldsymbol{S_i t})$; in case of the $E_{K-1,K}(\mu; p)$ all eigenvalues are μ.

3.3 Optimal schedule for sequential and simultaneous approach

Above we explained how to approximate any distribution on $[0, \infty)$ by a phase-type distribution of relatively low dimension (either a mixture of Erlang distributions or a hyperexponential distribution, depending on the value of the SCV), and how to compute the corresponding sojourn-time distributions. The next step is to use these findings to determine optimal schedules, for the sequential and simultaneous optimization approach, and for quadratic and the linear loss functions, as in [10].

3.3.1 The sequential optimization approach

In the sequential optimization approach, we optimize for each arriving client i the corresponding risk. This means that we minimize the expected loss over t_i, for given values of $t_1 (= 0), \dots, t_{i-1}$. In suggestive notation, we are faced with the optimization program

$$\min_{t_i} R(t_i \mid t_{i-1}, \dots, t_1) = \min_{t_i} (\mathbb{E}g(I_i) + \mathbb{E}h(W_i)).$$

As we argued earlier, to solve this sequential optimization problem, we only need to know the sojourn-time distribution of the previous arrival, S_{i-1}, given given $t_1, \ldots, t_{i-1}$ [10]. We now show in greater detail how this works for the *weighted-linear* and the *weighted-quadratic* loss function.

Weighted-linear loss function. Let the risk for each arrival be a weighted expected linear loss over the idle time and waiting time, i.e.,

$$\min_{t_i} R^{(a,\alpha)}(t_i|t_{i-1}, \ldots, t_1) = \min_{t_i} \alpha \mathbb{E} I_i + (1-\alpha)\mathbb{E} W_i, \quad i = 1, \ldots, n, \quad \alpha \in (0, 1).$$

Given (7) we may write for $i = 2, \ldots, n$ and again for $\alpha \in (0, 1)$

$$\min_{x_{i-1}} \alpha \mathbb{E}\ (x_{i-1} - S_{i-1})\, \mathbf{1}_{x_{i-1} > S_{i-1}} + (1-\alpha)\mathbb{E}\ (S_{i-1} - x_{i-1})\, \mathbf{1}_{x_{i-1} < S_{i-1}},$$

where the interarrival time x_{i-1} equals $t_i - t_{i-1}$.

Then the optimal interarrival time $x^\star_{i-1}$ can be found by solving the first-order equation

$$\alpha F_{S_{i-1}}(x) - (1-\alpha)\left(1 - F_{S_{i-1}}(x)\right) = F_{S_{i-1}}(x) - 1 + \alpha = 0.$$

This leads to the optimal schedule

$$t^\star_1 := 0 \quad \text{and} \quad t^\star_i := \sum_{j=1}^{i-1} F^{-1}_{S_j}(1-\alpha), \quad i = 2, \ldots, n.$$

For $\alpha = \frac{1}{2}$, we obtain that client i is supposed to arrive after a time that equals the sum of the *medians* of the sojourn times of all previous clients.

Weighted-quadratic loss function. Let the risk for each arrival be a weighted expected quadratic loss over the idle time and waiting time

$$\begin{aligned}\min_{t_i} R^{(q,\alpha)}(t_i|t_{i-1}, \ldots, t_1) = \min_{t_i} \alpha \mathbb{E} I_i^2 \\ +(1-\alpha)\mathbb{E} W_i^2, \quad i = 1, \ldots, n, \qquad \alpha \in (0, 1).\end{aligned}$$

Given (6) we may write for $i = 2, \ldots, n$ and again for $\alpha \in (0, 1)$, with $x_{i-1} = t_i - t_{i-1}$,

$$\min_{x_{i-1}} \alpha \mathbb{E}\ (x_{i-1} - S_{i-1})^2\, \mathbf{1}_{x_{i-1} > S_{i-1}} + (1-\alpha)\mathbb{E}\ (S_{i-1} - x_{i-1})^2\, \mathbf{1}_{x_{i-1} < S_{i-1}}.$$

As above, the optimal interarrival time $x^\star_{i-1}$ follow from the first-order equation, which now reads

$$\alpha(x - \mathbb{E} S_{i-1}) - (1-2\alpha)\int_x^\infty \mathbb{P}(S_{i-1} > s)\mathrm{d}s = 0.$$

For $\alpha = \frac{1}{2}$ we obtain the optimal schedule

$$t_1^\star := 0 \quad \text{and} \qquad t_i^\star := \sum_{j=1}^{i-1} \mathbb{E}S_j, \quad i = 2, \ldots, n.$$

This means that for $\alpha = \frac{1}{2}$ we obtain that client i is supposed to arrive after a time that equals the sum of the *means* of the sojourn times of all previous clients.

3.3.2 The simultaneous optimization approach

In case of a simultaneous optimization approach we set the optimal schedule that jointly minimizes

$$\min_{t_1,\ldots,t_n} R(t_1, \ldots, t_n) = \min_{t_1,\ldots,t_n} \sum_{i=1}^{n} \left(\mathbb{E}g(I_i) + \mathbb{E}h(W_i)\right).$$

It is known that this joint optimization has in general no tractable solution, as was the case in the sequential approach; only in case of an exponential service-time distribution and a linear loss function it has a tractable solution, see [21]. Therefore, we rely on numerical analysis software to find the optimal schedule.

In the next sections, we present numerical examples that feature schedules generated by our approach. Section 4 concentrates on the situation of a relatively low number of clients, whereas Sect. 5 uses results for the steady-state of the D/G/1 queue (with phase-type service times) to analyze the situation of a relatively high number of clients. In Sect. 6, we discuss the potential and limitations of our approach; in particular, we show that the error due to the phase-type fit is small.

4 Optimal scheduling in a transient environment

If the number of clients in the schedule, n, is relatively high, and their service times are i.i.d., then one will obtain schedules with more or less constant interarrival times. This section presents results for optimal schedules that relate to the opposite case, i.e., situations in which the number of clients is relatively low; particularly at the beginning of the schedule (and in the simultaneous approach also at the end) it is expected that the optimal interarrival times will substantially vary. Our experiments show that for the loss functions and the range of SCVs that we consider, this “transient effect” has significant impact up to, say, $n = 25$ clients.

Normalizing time such that the mean service time equals 1, we use four different SCVs (viz., SCV $\in \{0.1225, 0.7186, 1.0000, 1.6036\}$); these can be considered typical values in services and healthcare, see [11]. The latter three values are also used in [21] to compute optimal schedules for. We added SCV $= 0.1225 = 0.35^2$ to be consistent with healthcare literature, where the typical range for CVs is from 0.35 up to 0.85, as Cayirli and Veral identified in their extensive literature survey [5].

Based on our approach, as proposed in Sect. 3, we first find the corresponding phase-type service-time distribution, then we derive for each arrival the sojourn-time distribution, and finally we compute the optimal schedule (for the sequential and simultaneous approach, with linear and quadratic loss functions).

- We model an SCV $= 0.1225 < 1$ with an $E_{K-1,K}(\mu;\, p)$ distribution with parameters $K = 9$ (realize that SCV $\in [\frac{1}{9}, \frac{1}{8}]$), $\mu = 8.3958$ and $p = 0.6042$.
- We model an SCV $= 0.7186 < 1$ with an $E_{K-1,K}(\mu;\, p)$ distribution with parameters $K = 2$, $\mu = 1.6003$ and $p = 0.3997$. The resulting parameters are $\boldsymbol{\alpha} = (1, 0)$ and

$$\boldsymbol{S} = \begin{pmatrix} -1.6003 & 0.9606 \\ 0 & -1.6003 \end{pmatrix}.$$

- We model an SCV $= 1$ with an exponential distribution with parameter $\mu = 1$.
- We model an SCV $= 1.6036$ with a $H_2(\boldsymbol{\mu};\, p)$ distribution under the condition of *balanced means* and with parameters chosen by the matching method explained above. The resulting parameters are $\boldsymbol{\alpha} = (p, 1 - p) = (0.7407, 0.2593)$ and

$$\boldsymbol{S} = \begin{pmatrix} -1.4815 & 0 \\ 0 & -0.5185 \end{pmatrix}.$$

Note that for all cases the mean service time is given by ($\boldsymbol{e_m}$ is the all-ones vector)

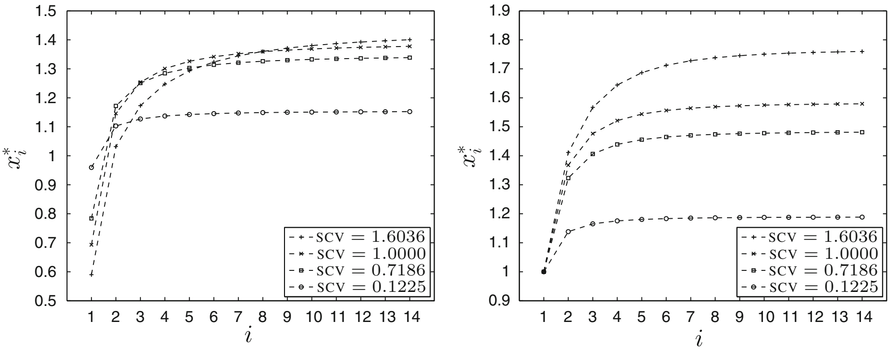
$$-\boldsymbol{\alpha}\boldsymbol{S}^{-1}\boldsymbol{e_m} = 1,$$

as desired.

The sojourn-time distribution of each individual client is then found by performing the second step of our approach, as explained in Sect. 3. Next, based on these sojourn-time distributions we compute the optimal interarrival times $x_i^\star$ through both the sequential approach and the simultaneous approach. Both approaches are studied in case of an equally weighted linear loss function and an equally weighted quadratic loss function. In our experiments for the simultaneous case, we study various schedule sizes ($n = 5, 10, \ldots, 25$ arrivals).

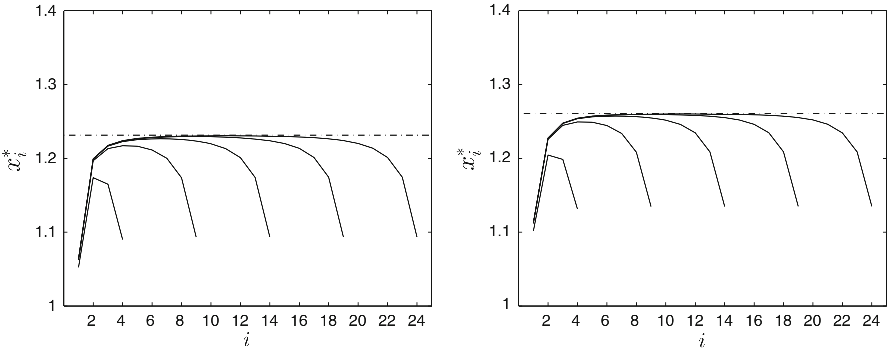
The results for the sequential approach are shown in Fig.1. From these figures we observe that in case of a linear loss that in case SCV > 1 the interarrival times in the beginning of the scheme (up to the 5-th arrival) are smaller than the interarrival times for the cases SCV $= 1$ or SCV < 1. Later in the schedule (from arrival 10 onwards) the optimal interarrival times are rather similar in size (that is, the curves for different SCVs are close together), but increasing in the value of the SCV. In case of a quadratic loss, only the first and second interarrival time are close together, but from arrival 3 onwards the interarrival times differ substantially; again they are increasing in the SCV, as expected. Overall, for any SCV the quadratic loss yields larger optimal interarrival times than the linear loss.

The five lines in Figs. 2, 3, 4, and 5 show the optimal schedules for $n = 5, 10, \ldots, 25$ arrivals under simultaneous optimization. From these figures we observe two interesting features. First, the simultaneous approach leads to schemes for which the optimal



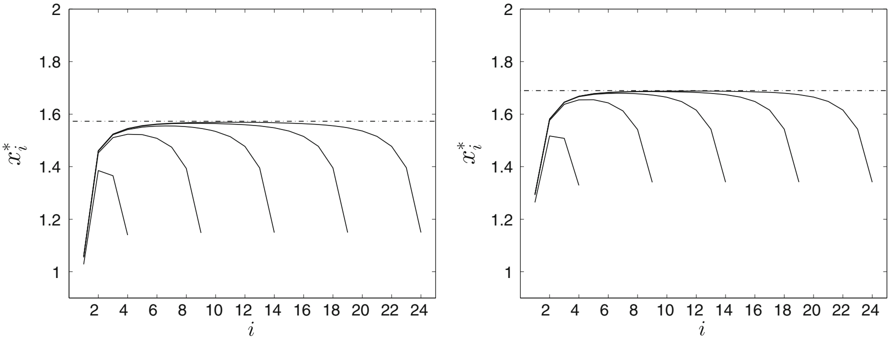
(a) Linear loss, phase-type fit. **(b)** Quadratic loss, phase-type fit.

Fig. 1 The optimal schedule in $x_i^\star$ s by sequential optimization for different SCVs



(a) Linear loss, phase-type fit. **(b)** Quadratic loss, phase-type fit.

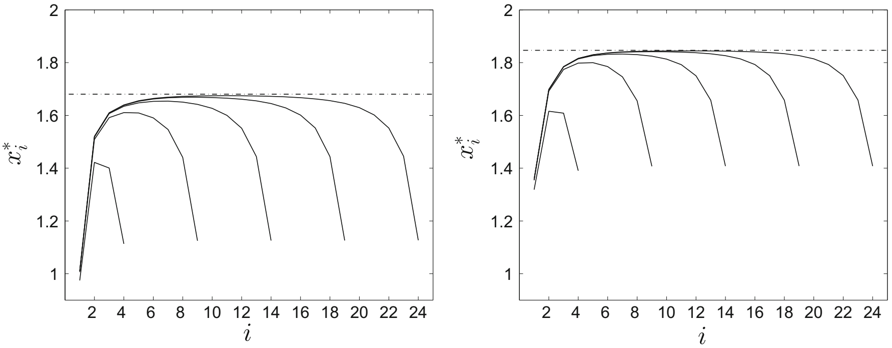
Fig. 2 The optimal schedules in $x_i^\star$ s by simultaneous optimization for SCV $= 0.1225 < 1$



(a) Linear loss, phase-type fit. **(b)** Quadratic loss, phase-type fit.

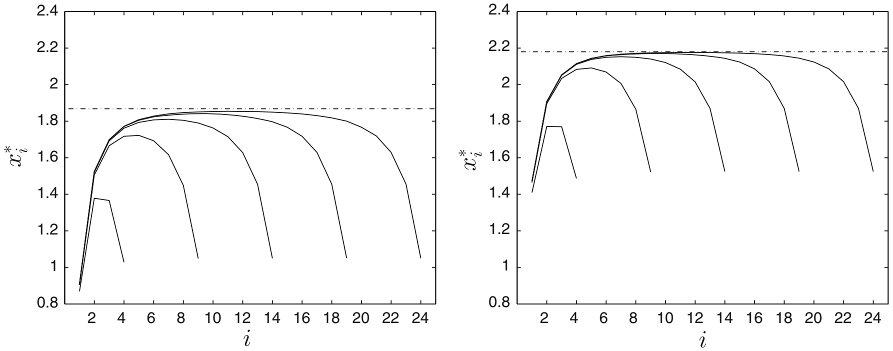
Fig. 3 The optimal schedules in $x_i^\star$ s by simultaneous optimization for SCV $= 0.7186 < 1$

interarrival times increase in the beginning and decrease towards the end of the scheme. The short interarrival times in the beginning of the schedule are essentially due to the fact that there the risk of waiting is relatively low; the short interarrival times at the



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

Fig. 4 The optimal schedules in $x_i^\star$ s by simultaneous optimization for SCV $= 1$



(a) Linear loss, phase-type fit. **(b)** Quadratic loss, phase-type fit.

Fig. 5 The optimal schedules in $x_i^\star$ s by simultaneous optimization for SCV $= 1.6036 > 1$

end can be explained from the fact that, despite a potentially substantial risk of high waiting times, there are few patients suffering from this (e.g., the last patient having a large service time does not affect the waiting time of any other patients). In the middle part the interarrival times are nearly constant indicating that the system is not affected by any start- or end-of-session effects. The steady-state solution, the top horizontal line, is added in each case. For all SCVs the system seems to converge fast to the steady state. This justifies considering the steady-state solution in which all transient effects are neglected; see Sect. 5 for more results.

The pattern described above is the so-called *dome shape*, which is also found in related literature. In [8,20], the expected waiting times and expected session-end time are minimized, while [16,19] minimize the combination of expected waiting and idle times. In [9] also expected overtime is added to the latter minimization problem; for a more detailed discussion on the effect of overtime on the schedule, see Sect. 6.4.

It should be noted that in case of linear loss minimizing expected session-end time, i.e., the "makespan," is equivalent to minimizing the sum of all expected idle times. Furthermore, optimal interarrival times computed by the recursive beta distribution approximation, as advocated in [13], show a dome shape pattern as well. In Sect. 6.3, we further compare this method with the phase-type approach. In addition, in case of a linear loss function the dome shape pattern is also found when minimizing expected quadratic waiting and idle times, see [10].

Second, we observe that the interarrival times tend to increase in the value of the SCV, and, again, for any of the SCVs the quadratic loss leads to larger interarrival times than linear loss.

5 Optimal scheduling in steady-state environment

In the previous section, we showed that our approach enables us to derive optimal interarrival times for different levels of SCV, for both the sequential and simultaneous optimization, and for various risk functions and scheme sizes. Note that we chose the *equally weighted* linear and quadratic loss (that is, $\alpha = \frac{1}{2}$), which we continue to do in this section, apart from Sect. 5.3 in which also the effect of α on the optimal schedule will be studied. The primary goal of this section is to analyze the case of a large number of clients with i.i.d. service times. In this situation, the schedules will have constant interarrival times, and we will show in detail how to determine these.

To study the steady-state interarrival times, given the value of the service-time distribution's SCV, we need to derive the steady-state sojourn-time distribution of the corresponding D/G/1 queue, with the service times having either a mixture of Erlang or hyperexponential distribution. We first show how we derive the steady-state sojourn-time distribution for various SCVs; then we model the optimal interarrival time as a function of the SCV for both sequential and simultaneous optimization using the loss functions mentioned above.

We first point out that the optimality condition, that determines the optimal interarrival times $x^\star$, depends on the choice of the specific case (simultaneous vs. sequential, linear vs. quadratic). This optimality condition is a relation that involves both the distribution of the steady-state sojourn time S and $x^\star$. To this end, we first observe that (use the Lindley recursion, in conjunction with the fact that it cannot be that both W_i and I_i are positive)

$$\mathbb{E}g(I_i) + \mathbb{E}h(W_i) = \ell(S_{i-1} - x_{i-1}),$$

with $\ell(\cdot)$ defined through

$$\ell(x) := g(-x)\mathbf{1}_{\{x<0\}} + h(x)\mathbf{1}_{\{x\geq 0\}}.$$

In case of the sequential optimization approach, [10] proves the (conceivable) result that for any convex loss function the optimal interarrival time solves

$$\frac{\mathrm{d}\mathbb{E}\ell(S-x)}{\mathrm{d}x} = 0;$$

in the transient case we have to take the sojourn-time distributions of the individual clients, whereas in the steady-state case we have to take the stationary sojourn-time distribution. In special cases this representation leads to appealing relations: for equally weighted loss functions we obtain for linear loss the median of the sojourn time, i.e., $x^\star = F_S^{-1}\left(\frac{1}{2}\right)$, and for quadratic loss the mean of the sojourn time, i.e., $x^\star = \mathbb{E}S$.

In case of the simultaneous optimization approach, we are to evaluate, for large n,

$$\min_{x_1,\ldots,x_n} \sum_{i=1}^{n} \mathbb{E}\ell(S_i(x) - x_i) \approx n \cdot \min_x \mathbb{E}\ell(S(x) - x);$$

we write $S(x)$ rather than just S to emphasize the fact that the sojourn times depend on the interarrival time x. The optimal interarrival time then follows from the first-order condition

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}\ell(S(x) - x) = 0.$$

For linear loss this yields the condition

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}|S(x) - x| = \frac{\mathrm{d}}{\mathrm{d}x}\left(\int_x^\infty (t - x) f_{S(x)}(t)\,\mathrm{d}t + \int_0^x (x - t) f_{S(x)}(t)\mathrm{d}t\right) = 0, \quad (8)$$

whereas for quadratic loss we obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}\left((S(x) - x)^2 = \frac{\mathrm{d}}{\mathrm{d}x}\left(\mathbb{E}S(x)^2 - 2x\mathbb{E}S(x) + x^2\right) = 0. \quad (9)$$

The above formula suggests that the linear case requires knowledge of the distribution function of $S(x)$, but, interestingly, just $\mathbb{E}S(x)$ is needed. This can be seen as follows. Note that

$$\sum_{i=1}^{n} (\mathbb{E}I_i + \mathbb{E}W_i) = \sum_{i=1}^{n} ((\mathbb{E}I_i + \mathbb{E}B_i) + (\mathbb{E}W_i + \mathbb{E}B_i)) - 2\sum_{i=1}^{n} \mathbb{E}B_i.$$

Now realize that, in addition to $W_i + B_i = S_i$, we also have that (recognize the "makespan")

$$\sum_{i=1}^{n} (I_i + B_i) = t_n + S_n.$$

Realizing that the value of $\sum_{i=1}^{n} \mathbb{E}B_i$ does not affect the optimization, we conclude that minimizing the linear loss is equivalent to minimizing $\sum_{i=1}^{n} \mathbb{E}S_i + t_n + \mathbb{E}S_n$. Because $t_n \approx (n-1)x$, we are to minimize $\mathbb{E}S(x) + x$.

5.1 Steady state results in case SCV $= 1$

We illustrate the steady-state results in case of SCV $= 1$ here, since it leads to nice explicit results. Based on results of a G/M/1 queue [18], we have the following expression for the sojourn-time distribution in case SCV $= 1$:

$$\mathbb{P}(S \leq x) = 1 - e^{-\mu(1-\sigma_x)x}, \quad x \geq 0, \tag{10}$$

where $\sigma_x \in (0, 1)$ solves $\sigma_x = e^{-(\mu-\mu\sigma_x)x}$.

Results for a sequential approach. In case the loss function is assumed linear, we solve $x^\star = F_S^{-1}(1/2)$. We find

$$x = F_S^{-1}\left(\frac{1}{2}\right) = \frac{\log 2}{\mu(1-\sigma_x)} \quad \text{and} \quad F_S(x) = \frac{1}{2} = 1 - \sigma_x,$$

leading to an optimal schedule with interarrival times

$$x^\star = \frac{2\log 2}{\mu} \approx \frac{1.3862}{\mu}.$$

For the case of quadratic loss we solve

$$x = \mathbb{E}S = \frac{1}{\mu(1-\sigma_x)} \quad \text{and} \quad \log \sigma_x = -1,$$

and obtain

$$x^\star = \frac{e}{\mu(e-1)} \approx \frac{1.5820}{\mu}.$$

These limiting results are in line with those corresponding to the transient schemes in Fig. 1a—for schedules of more than, say, 15 clients, the middle part of the schedule is close to the steady-state schedule. We observe that in this sequential setup quadratic loss leads to larger optimal interarrival times than linear loss.

Results for a simultaneous approach. To obtain the steady-state results in the simultaneous case and linear loss we solve the first order condition (8):

$$\begin{aligned} &\tfrac{\mathrm{d}}{\mathrm{d}x}\left(\int_x^\infty (t-x) f_{S(x)}(t)\,\mathrm{d}t + \int_0^x (x-t) f_{S(x)}(t)\,\mathrm{d}t\right) \\ &= \tfrac{\mathrm{d}}{\mathrm{d}x}\tfrac{1-2e^{-\mu(1-\sigma_x)x}-\mu(1-\sigma_x)x}{\mu(\sigma_x-1)} = -\sigma_x' \tfrac{1+\sigma_x(\log\sigma_x-2)}{\mu(\sigma_x-1)^2\sigma_x} \\ &= \tfrac{1+(\log\sigma_x-2)\sigma_x}{\mu(\sigma_x-1)(1-\sigma_x+\sigma_x\log\sigma_x)} = 0, \end{aligned}$$

where we used that

$$\sigma_x' = \frac{\mu\sigma_x(\sigma_x-1)}{1-\mu\sigma_x x} \quad \text{and} \quad x = \frac{\log\sigma_x}{\mu(\sigma_x-1)}.$$

This equation is solved for $\sigma_x \approx 0.32$, and we obtain

$$x^\star \approx \frac{1.6803}{\mu}.$$

The case of quadratic loss can be dealt with analogously; now the first-order condition (9) needs to be solved. We eventually obtain

$$x^\star \approx \frac{1.8466}{\mu}.$$

Again these limiting results align well with the results of large transient schemes, see Fig. 1b. We observe that, as in the sequential approach, in the simultaneous approach quadratic loss leads to larger optimal interarrival times than linear loss.

5.2 Steady-state results in case SCV $\neq 1$

As pointed out in Sect. 3, in the first step of our approach we fit a phase-type distribution (with the right mean and SCV) to our service-time distribution. The special case of SCV $= 1$ (i.e., exponentially distributed service times) was dealt with in the previous subsection; now we focus on the cases in which SCV $\neq 1$.

5.2.1 Steady-state analysis for the $\mathrm{D}/E_{K-1,K}/1$ *queue*

As presented in Sect. 3.1, we use the $E_{K-1,K}(\mu, p)$ distribution to approximate service-time distributions with an SCV between $1/K$ and $1/(K-1)$, for $K \in \{2, 3, \ldots\}$. We analyze the resulting $\mathrm{D}/E_{K-1,K}/1$ queue through the sequence $(N_0, N_1, \ldots)$ with $N_0 = 0$ (the system starts empty), and N_i referring to the number of phases in the system just after the i-th arrival. These phases are exponentially distributed with mean $1/\mu$.

First observe that $(N_0, N_1, \ldots)$ follows a (discrete-time) Markov chain. It is elementary to express the transition probabilities $p_{m,n} = \mathbb{P}(N_{i+1} = n \mid N_i = m)$ in terms of the parameters K, p, μ, and the (constant) interarrival time x. The steady-state distribution of N now follows from, with the matrix $\boldsymbol{P} = (p_{m,n})_{m,n=0}^{\infty}$ denoting the transition matrix,

$$\boldsymbol{a} = \boldsymbol{a}\boldsymbol{P}; \tag{11}$$

in addition the normalization constraint $a_0 + a_1 + a_2 + \ldots = 1$ needs to be imposed. Based on the limiting probabilities $\boldsymbol{a}$, we are in a position to derive the steady-state sojourn time distribution and its moments, and hence we can deal with the various first-order conditions of Sect. 5. In order to solve (11), we need to truncate the state space to $\{0, \ldots, M\}$; from the fact that the a_n decay roughly exponentially in n (with a decay rate that can be evaluated explicitly), it is not hard to select an appropriate value for M. Generally speaking, we saw in this SCV < 1 regime that the choice $M = 10 + K$ works well in nearly all situations.

Now with the sojourn time distribution

$$\mathbb{P}(S \leq t) = \mathbb{P}(W + B \leq t) = \int_0^t F_W(t-u) f_B(u)\, \mathrm{d}u \tag{12}$$

and the vector $\boldsymbol{a}$ that solves (11), we may write

$$\begin{aligned}
\mathbb{P}(S \leq t) &= a_0 F_B(t) + \sum_{m=1}^{M} a_m \int_0^t \mu \frac{(\mu(t-u))^{m-1}}{(m-1)!} e^{-\mu(t-u)} f_B(u)\, \mathrm{d}u, \\
\mathbb{E}S &= \mathbb{E}W + \mathbb{E}B = \sum_{m=0}^{M} a_m \left(p \frac{m+K-1}{\mu} + (1-p) \frac{m+K}{\mu} \right), \\
\mathbb{E}S^2 &= \mathbb{E}W^2 + 2\mathbb{E}W\, \mathbb{E}B + \mathbb{E}B^2 \\
&= \sum_{m=1}^{M} a_m \frac{m(m+1)}{\mu^2} + 2 \sum_{m=1}^{N} a_m \left(p \frac{m}{\mu} \frac{K-1}{\mu} + (1-p) \frac{m+1}{\mu} \frac{K}{\mu} \right) \\
&\quad + \left(p \frac{K(K-1)}{\mu^2} + (1-p) \frac{(K+1)K}{\mu^2} \right).
\end{aligned}$$

5.2.2 Steady-state results for the $\mathrm{D}/H_2/1$ *queue*

Mimicking the procedure described in Sect. 5.2.1, we now sketch a procedure to generate the steady-state sojourn-time distribution of a $\mathrm{D}/H_2/1$ system, so as to cover the case SCV > 1. To do so, we analyze the queue through the sequence $((N_0, K_0), (N_1, K_1), \ldots)$ with $(N_i, K_i) = (m, k)$ meaning that the number of patients in the system just after the i-th arrival is m, and the phase of the client in service is k; if $k = 1$ the patient in service is served with rate μ_1, and if $k = 2$ with rate μ_2. Evidently, $(N_i, K_i) \in \{1, 2, \ldots\} \times \{1, 2\}$.

Again we truncate the state-space (in terms of the number of clients) to M, generate the transition probability matrix $\boldsymbol{P}$, and solve (11). Define $a_{m,k}$ as the steady-state probability of m clients in the system immediately after an arrival epoch, jointly with the phase of the client in service being k. We then evaluate (12), which leads to the following expressions. For the distribution function we obtain, with $(H_{j,2})_j$ a sequence of i.i.d. samples from a $H_2(\boldsymbol{\mu}; p)$ distribution, and $B^{(k)}$ having an exponential distribution with mean $1/\mu_k$ $(k = 1, 2)$,

$$P(S \leq t) = a_{0,0} F_B(t) + \sum_{m=1}^{M} \sum_{k=1}^{2} a_{m,k} \int_0^t \mathbb{P}\left(\sum_{j=1}^{m-1} H_{j,2} + B^{(k)} < t - u \right) f_B(u)\, \mathrm{d}u.$$

This expression can be evaluated further, realizing that, with D following a binomial distribution with parameters $m-1$ and p,

$$\sum_{j=1}^{m-1} H_{j,2} \overset{\mathrm{d}}{=} \sum_{i=1}^{D} B_i^{(1)} + \sum_{i=1}^{m-1-D} B_i^{(2)},$$

where $B_i^{(k)}$ are i.i.d. copies of $B^{(k)}$. For the corresponding first and second moment we obtain (with $\mathbb{E}H_2 = p/\mu_1 + (1-p)/\mu_2$)

$$\begin{aligned}
\mathbb{E}S &= \sum_{m=1}^{M} \left\{ a_{m,1}\left(m\mathbb{E}H_2 + \frac{1}{\mu_1}\right) + a_{m,2}\left(m\mathbb{E}H_2 + \frac{1}{\mu_2}\right)\right\} + \mathbb{E}H_2,\\
\mathbb{E}S^2 &= \sum_{m=1}^{M} \left\{ a_{m,1} \sum_{j=0}^{m} \binom{m}{j} \left\{ \frac{(j+1)(j+2)}{\mu_1^2} + 2\frac{(j+1)(m-j)}{\mu_1\mu_2} \right.\right.\\
&\qquad \left. + \frac{(m-j)(m-j+1)}{\mu_2^2} \right\}\\
&\quad \left. + a_{m,2} \sum_{j=0}^{m} \binom{m}{j} \left\{ \frac{j(j+1)}{\mu_1^2} + 2\frac{j(m-j+1)}{\mu_1\mu_2} + \frac{(m-j+1)(m-j+2)}{\mu_2^2} \right\}\right\}\\
&\quad + 2\sum_{m=0}^{M} \left\{ a_{m,1}\left(m\mathbb{E}H_2 + \frac{1}{\mu_1}\right) + a_{m,2}\left(m\mathbb{E}H_2 + \frac{1}{\mu_2}\right)\right\} \mathbb{E}H_2\\
&\quad + \left(\frac{2p_1}{\mu_1^2} + \frac{2p_2}{\mu_2^2}\right).
\end{aligned}$$

Evidently, by letting M grow large we get arbitrarily close to the true vector of stationary probabilities. We validated that the choice of $M = 25$ works well for the range of SCV $\in (0, 3)$ when α equals $\frac{1}{2}$. However, when α is closer to 1 the truncation level M should be suitably increased.

5.3 Computational results in a steady-state environment

In this section, we studied the optimal interarrival time as a function of the service-time distribution's SCV $\in (0, 3)$. We did this for the sequential and the simultaneous optimization approach, in case of both an (equally weighted) linear loss function and an (equally weighted) quadratic loss function. From these results, that are depicted in Fig. 6, we conclude that the steady-state optimal interarrival time is increasing in the SCV for any of the four scenarios considered, as expected. In line with earlier findings, we observe that for each approach and for any SCV $\in (0, 3)$ the quadratic loss function yields larger optimal interarrival times than the linear loss function. Furthermore, for any SCV $\in (0, 3)$ the sequential approach yields smaller optimal interarrival times, for both quadratic and linear loss. Loosely speaking, this says that the sequential approach favors the service provider, since smaller interarrival times lead to smaller idle times.

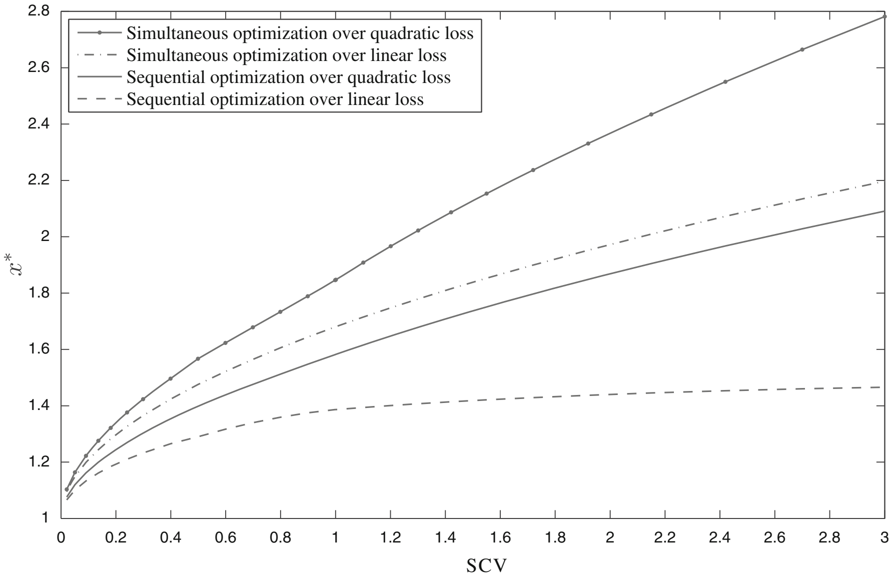


Fig. 6 An overview of the optimal steady-state interarrival times $x^\star$ for four different optimization settings as a function of the SCV, where we take $M = 25$

Given an arbitrary SCV, we now consider the effect of the weight parameter, α. Since an increasing α results in more weight assigned to the practitioner's time, the interarrival times will decrease. Indeed, this is observed from in Fig. 7, where we plotted the dependence of the steady-state solutions resulting from the various optimization programs. We did these computations for SCV $= 0.5625$, that is, the coefficient of variation (CV) equalling 0.75. This value is in the range of common CVs, that is, CV $\in (0.35, 0.85)$, as concluded by [5]. For other SCVs similar graphs can be generated.

We observe that the solution curves of the simultaneous optimization and its sequential counterpart have a similar shape. When α tends to 1, the occupation rate approaches 1, so that we need to increase the truncation level M to reliably compute the steady-state solution. For this reason, we set M to 50 when generating Fig. 7.

6 Discussion

In this section, we systematically study different aspects of the schedules we developed. (i) In the first place we consider the *robustness* of our approach (both steady state and transient), so as to assess the effect of replacing generally distributed nonnegative service times by their phase-type counterparts. (ii) Secondly, we compare our approach with the approach based on the characteristics of the beta distribution introduced by Lau and Lau [13]. (iii) Furthermore, we briefly discuss the effect of overtime in two transient settings. (iv) Also, we provide an account of the computational effort (in terms of computation time) for the various approaches. (v) We conclude this section

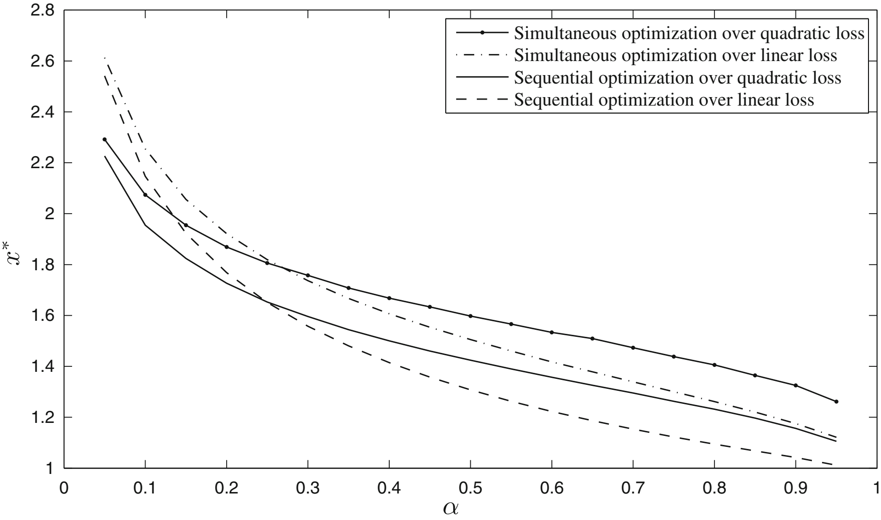


Fig. 7 An overview of the optimal steady-state interarrival times $x^\star$ for four different optimization settings as a function of α, where we take $M = 50$

with a comparison of the sequential and simultaneous optimization approach, in terms of the disutilities perceived by the individual agents.

6.1 Robustness of phase-type approach in steady state

To study the robustness of our approach for the optimal steady-state interarrival times, as presented in the previous section, we apply our approach to a D/G/1 setting in which the service-time distribution is non-phase-type. We here concentrate on the Weibull distribution and the log-normal distribution, as often seen in practice [3,12,19]. In our study, we assume again that the SCV $= 0.5625$ (contained in the interval identified by [5]).

Our study is set up as follows. We consider the following 2-parameter distributions:

- the Weibull distribution, with density

$$\frac{kx^{k-1}}{\lambda^k}e^{-(\frac{x}{\lambda})^k}$$

with parameters $k \approx 1.3476$, and $\lambda \approx 1.0902$, and
- the log-normal distribution, with density

$$\frac{1}{x\sqrt{2\pi\sigma^2}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$

with parameters $\mu = -\frac{1}{2}\log 1.5625$ and $\sigma = \sqrt{\log 1.5625}$,

Table 1 The Monte Carlo optimal steady-state interarrival times and risk in case of log-normal service times compared with our approach and with an approach based on exponential service times

Setting	$\tilde{x}$	$\lvert\tilde{x} - x^{\star}\rvert$	$\lvert\tilde{x} - x^{e}\rvert$	$\tilde{R}$	$\lvert\tilde{R} - R^{\star}\rvert$	$\lvert\tilde{R} - R^{e}\rvert$
Sim. & quad.	1.6661	0.0631	0.1804	1.2866	0.0190	0.1075
Sim. & lin.	1.5085	0.0033	0.1718	0.8680	0.0002	0.0464
Seq. & quad.	1.4398	0.0156	0.1422	1.6147	0.0646	0.2937
Seq. & lin.	1.2749	0.0326	0.1114	1.0826	0.0724	0.1726

Table 2 The Monte Carlo optimal steady-state interarrival times and risk in case of Weibull service times compared with our approach and with an approach based on exponential service times

Setting	$\tilde{x}$	$\lvert\tilde{x} - x^{\star}\rvert$	$\lvert\tilde{x} - x^{e}\rvert$	$\tilde{R}$	$\lvert\tilde{R} - R^{\star}\rvert$	$\lvert\tilde{R} - R^{e}\rvert$
Sim. & quad.	1.5946	0.0084	0.2519	1.0307	0.0005	0.2099
Sim. & lin.	1.5058	0.0007	0.1745	0.8260	0.0001	0.0488
Seq. & quad.	1.4223	0.0019	0.1597	1.2395	0.0051	0.2078
Seq. & lin.	1.3138	0.0063	0.0725	0.9542	0.0114	0.0878

which both lead to SCV $= 0.5625$.

For all four scenarios (sequential or simultaneous approach, and quadratic or linear loss) we determined the optimal interarrival times by simulation, as follows. For a given steady-state interarrival time x, we simulate the queueing system using 100,000 patients (with a "warm-up" corresponding to 1,000 patients), to estimate the value of the loss function for this specific x. In the loop around this routine, we identify the x that minimizes the loss; this is done using MATLAB's minimization routine. We perform this optimization 100 times, and estimate the "real" optimal interarrival time and risk, $\tilde{x}$ and $\tilde{R}$, by the average of the optimal interarrival times of the 100 individual experiments.

We compare these results with the optimal interarrival times resulting form our phase-type based technique with the SCV, i.e., 0.5625. In Table 1, we compare for the log-normal service-time distribution both the optimal interarrival time and the risk per client in steady state. The values resulting from the phase-type-based approach are denoted by $x^{\star}$ and $R^{\star}$. Finally, x^{e} and R^{e} refer to an approach where one assumes exponential service times instead (with mean 1 and SCV $= 1$, that is). In a similar way, in Table 2 we compare for the Weibull service-time distribution the optimal interarrival time and the risk per client in steady state.

From Tables 1 and 2, we observe that the phase-type approximation has just a modest impact on the accuracy of the the optimal interarrival time and risk. It also shows that the naïve approach of assuming exponential distributed service times (thus completely ignoring the effect of the SCV differing from 1) leads to large deviations from the optimal scheme.

6.2 Robustness of phase-type approach in transient environment

To study the robustness of the phase-type approach in a transient environment, we considered the same service-time distributions as used in Sect. 6.1, i.e., Weibull and

Table 3 The Monte Carlo optimal times and risk in simultaneous optimization of linear risk in a transient environment with log-normal or Weibull service times compared with our approach and with an approach based on exponential service times

Setting	Log-normal service times			Weibullian service times		
i	$\tilde{x}_i$	$\lvert\tilde{x}_i - x_i^\star\rvert$	$\lvert\tilde{x}_i - x_i^e\rvert$	$\tilde{x}_i$	$\lvert\tilde{x}_i - x_i^\star\rvert$	$\lvert\tilde{x}_i - x_i^e\rvert$
1	1.0101	0.0546	0.0031	1.0739	0.0093	0.0634
2	1.3546	0.0543	0.1625	1.4188	0.0100	0.0983
3	1.4282	0.0322	0.1789	1.4652	0.0062	0.1418
4	1.4551	0.0232	0.1796	1.4808	0.0049	0.1539
5	1.4702	0.0158	0.1767	1.4879	0.0041	0.1590
6	1.4762	0.0137	0.1775	1.4918	0.0041	0.1620
7	1.4773	0.0130	0.1766	1.4916	0.0046	0.1622
8	1.4748	0.0128	0.1751	1.4898	0.0045	0.1602
9	1.4666	0.0147	0.1751	1.4834	0.0049	0.1583
10	1.4530	0.0189	0.1740	1.4739	0.0042	0.1531
11	1.4312	0.0231	0.1694	1.4579	0.0047	0.1427
12	1.3911	0.0321	0.1606	1.4283	0.0059	0.1234
13	1.3066	0.0461	0.1364	1.3606	0.0080	0.0823
14	1.0884	0.0535	0.0379	1.1525	0.0106	0.0262
Total risk	$\tilde{R}$	$\lvert\tilde{R} - R^\star\rvert$	$\lvert\tilde{R} - R^e\rvert$	$\tilde{R}$	$\lvert\tilde{R} - R^\star\rvert$	$\lvert\tilde{R} - R^e\rvert$
	5.6083	0.0093	0.2201	5.5264	0.0003	0.1637

log-normal. We took $n = 15$ patients to be scheduled resulting in 14 interarrival times (the x_i s). Again, we ran Monte Carlo simulation experiments to determine the optimal interarrival times and associated *total risk* (defined as the aggregate of the risks of all individual patients). In this case, however, the simulations were more involved than in the steady-state counterpart. For a given schedule $(x_1, \ldots, x_{14})$, we estimate the loss (by using 100, 000 repetitions). Then we apply MATLAB's optimization procedure to identify the schedule that minimizes the loss. This optimization is performed 100 times; we estimate the "real" optimal interarrival times and total risk ($\tilde{x}_1, \ldots, \tilde{x}_{14}$ and $\tilde{R}$) by the average of the 100 individual schedules.

In Table 3, we compare the simulation results with the phase-type approach and the assumption of exponential service times in case of optimization with a linear loss function, while in Table 4 we do the same in case of optimization with a quadratic loss function. Similar to Sect. 6.1, the values resulting from the phase-type-based approach are denoted by $x_i^\star$ and $R^\star$, whereas x_i^e and R^e refer to the optimal arrival times and risk assuming exponential service times.

As in Sect. 6.1 we see that the phase-type approach results in a significant gain, in terms of the total risk, compared to the results obtained when assuming exponential service times. We did not include simulations related to the sequential optimization, since these are only affected by a start-of-session effect resulting in rapid convergence to steady state, as seen in Fig. 1. Therefore these simulations are redundant.

Table 4 The Monte Carlo optimal times and risk in simultaneous optimization of quadratic risk in a transient environment with log-normal or Weibull service times compared with our approach and with an approach based on exponential service times

Setting	Log-normal service times			Weibullian service times		
i	$\tilde{x}_i$	$\lvert\tilde{x}_i - x_i^\star\rvert$	$\lvert\tilde{x}_i - x_i^e\rvert$	$\tilde{x}_i$	$\lvert\tilde{x}_i - x_i^\star\rvert$	$\lvert\tilde{x}_i - x_i^e\rvert$
1	1.2672	0.0099	0.0897	1.2550	0.0044	0.1019
2	1.5221	0.0124	0.1753	1.5087	0.0041	0.1887
3	1.5955	0.0309	0.1878	1.5597	0.0056	0.2236
4	1.6244	0.0413	0.1895	1.5762	0.0073	0.2378
5	1.6388	0.0481	0.1878	1.5831	0.0078	0.2435
6	1.6442	0.0505	0.1875	1.5859	0.0078	0.2458
7	1.6461	0.0521	0.1865	1.5862	0.0082	0.2464
8	1.6452	0.0528	0.1851	1.5846	0.0078	0.2457
9	1.6397	0.0513	0.1847	1.5804	0.0082	0.2440
10	1.6281	0.0473	0.1850	1.5730	0.0080	0.2401
11	1.6084	0.0418	0.1834	1.5598	0.0069	0.2321
12	1.5711	0.0327	0.1788	1.5329	0.0057	0.2170
13	1.4937	0.0184	0.1636	1.4723	0.0036	0.1851
14	1.3033	0.0039	0.1047	1.2994	0.0020	0.1085
Total risk	$\tilde{R}$	$\lvert\tilde{R} - R^\star\rvert$	$\lvert\tilde{R} - R^e\rvert$	$\tilde{R}$	$\lvert\tilde{R} - R^\star\rvert$	$\lvert\tilde{R} - R^e\rvert$
	8.1236	0.0364	0.5801	6.7542	0.0012	0.9764

6.3 Comparison with the approach by Lau and Lau

Instead of using phase-type distributions to compute optimal schedules, one can opt for using a recursive method based on the beta distribution, see [13]. To use this approach, four parameters are needed, which can be picked by matching the first four moments of the patients' service-time distributions. In Table 5 we compare for both methods the optimized schedules in terms of arrival times, expected waiting times and expected idle times per patient. The patients' ($n = 20$) service times are i.i.d. with mean 1, variance 0.25, skewness 1, and kurtosis 4; the risk per patient to be minimized is $R_i^{(a,10/11)}$, i.e., $\alpha = \frac{10}{11}$. These settings are chosen such that they match the problem considered by [13]. To compare the *total risk* found by [13], denoted by $\mathbb{E}C_s$, the risk per patient R_i can be scaled arbitrarily, not affecting the optimal schedule, cf. Eq. (7), (noting that $R_1 = 0$)

$$\sum_{i=2}^{20} R_i^{(a,10/11)} = \frac{10}{11}\left(\sum_{i=2}^{20} \mathbb{E}I_i + \frac{1}{10}\sum_{i=2}^{20} \mathbb{E}W_i\right) = \frac{10}{11}\mathbb{E}C_s.$$

We find that the phase-type fit approach based on the first two moments $\mu = 1$, SCV $= 0.25$ gives nearly identical results, in terms of the optimal schedule and the corresponding waiting and idle times. Furthermore, in case of a linear loss function

Table 5 The optimized schedules for the beta distribution approach and the phase-type fit approach

Method	Beta distribution approach			Phase-type approach		
Patient (i)	Arrival times	$\mathbb{E}W_i$	$\mathbb{E}I_i$	Arrival times	$\mathbb{E}W_i$	$\mathbb{E}I_i$
2	0.542	0.477	0.022	0.535	0.489	0.024
5	3.395	0.792	0.068	3.424	0.780	0.069
10	8.603	0.969	0.072	8.635	0.951	0.077
15	13.785	1.146	0.070	13.815	1.127	0.065
20	18.467	1.698	0.017	18.514	1.644	0.021
$\sum \mathbb{E}W_i$ or $\sum \mathbb{E}I_i$		19.514	1.139		19.165	1.160
Total risk	2.810			2.798		

The optimized schedules minimize the total risk $\sum R_i^{(a,10/11)}$ as defined in Eq. (7)

the phase-type fit approach uses explicit expressions for the expected idle and waiting times, so that it should perform at roughly the same speed as the method by Lau and Lau.

The major strength of the phase-type approach is that it requires just the first two moments of the service-time distribution, which tends to be sufficient to determine the optimal schedule (see the discussion in [5][Sect. 2.5]). In addition, estimating higher moments such as skewness and kurtosis is relatively hard, in the sense that a large sample size is needed to obtain an estimate with low variance.

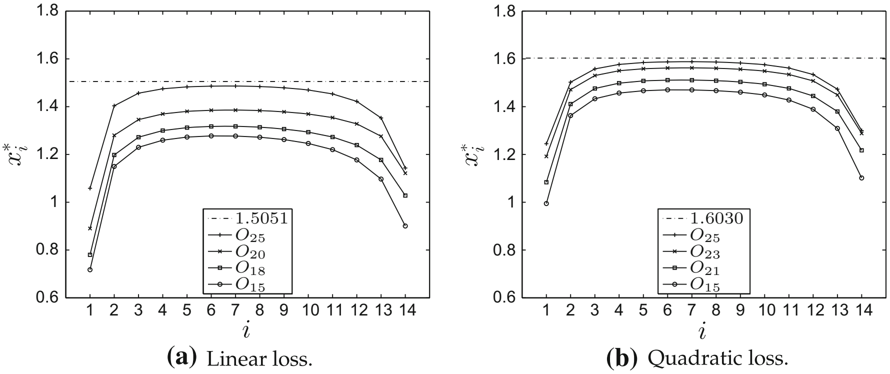
6.4 The effect of overtime on the schedule

In our optimization problem, we only considered the minimization of risk in terms of waiting and idle time. This allowed us to study the difference between the sequential and simultaneous approach, and for both cases how they are affected by the SCV. Another performance measure in healthcare, which could be modeled easily, is the so called overtime O. Overtime is defined as the actual session-end time SET minus the scheduled end time T, that is

$$O := \max\{\textsc{set} - T, 0\} = \max\left\{\sum_{i=1}^{n}(I_i + B_i) - T, 0\right\}.$$

To stress that O depends on the value of T, we have added a subscript, that is, we write O_T. To study the effect of overtime we extend the simultaneous optimization approach with expected overtime. We focus on linear risk, in a schedule of $n = 15$ patients (cf. Eq. (7)), i.e., we consider

$$\min_{t_1,\ldots,t_n} \sum_{i=1}^{15} R_i + \beta\mathbb{E}O_T = \min_{t_1,\ldots,t_n} \sum_{i=1}^{15} (\alpha\mathbb{E}I_i + (1-\alpha)\mathbb{E}W_i) + \beta\mathbb{E}O_T.$$



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

Fig. 8 The effect of overtime on the schedule with simultaneous optimization over linear and quadratic loss, $n = 15$, with the corresponding steady-state solutions

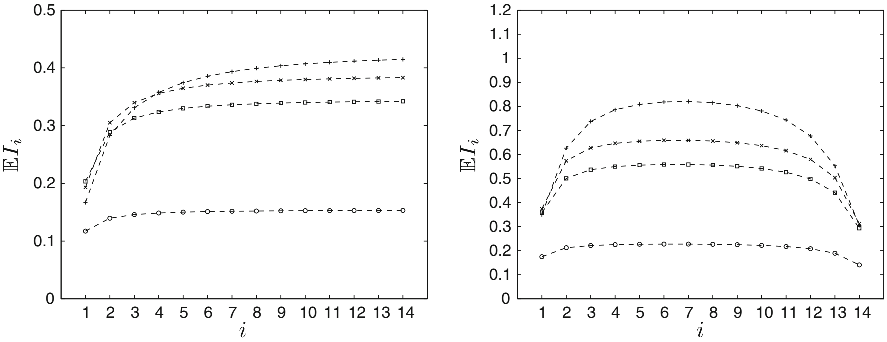
Take $\alpha = 0.5$ (equal weights) and $\beta/\alpha = 1.5$, which models the situation in which overtime is valued roughly 50 % higher than idle time [6]. In Fig. 8, we see the influence of overtime on the schedule; here the service times are chosen by the phase-type approach, so as to generate a distribution with mean 1 and SCV = 0.5625 as in Sect. 6.1. We consider for both linear as quadratic loss four cases, where $T \geq 15$ varies. (Special case is O_{15} so that in order to have no overtime all patients should be served in their expected service time, that is, a queue with load 1.) Indeed, we see that the schedule gets tighter when the scheduled session-end time decreases. Including overtime has a similar effect as assigning a higher weights to the idle times in the risk function, viz., result in tighter schedules. Remark that when T tends to infinity we are in the case of our original models optimized in Sect. 4.

6.5 Computational effort of the various numerical approaches

We now give a brief account of the computational effort required to evaluate the schedules, and further describe how our code has been set up. A general remark is that, for obvious reasons, determining steady-state schedules is substantially less expensive than determining transient schedules. In our numerical experiments, we generated transient schedules of up to 25 clients. All programming was done in MATLAB, benefiting from its built-in function for determining roots, its minimization routine, and its numerical integration routine; as a result the code that had to be developed is relatively "light weight." The most complicated cases (25 clients, SCV > 1) took a few minutes, but usually the computation time was considerably less.

The structure of the code is as follows. Here x is the steady state interarrival time, whereas $\boldsymbol{x} = (x_1, \ldots, x_{n-1})$ is the transient schedule.

1. Determine the phase-type fit (hyperexponential or Erlang mixture) for given mean and SCV.
2. The corresponding loss function is computed as follows.



(a) Linear loss, sequential approach.

(b) Linear loss, simultaneous approach.

Fig. 9 The optimal idle times by sequential and simultaneous approach for the various SCVs, in case of a linear loss

(i) Regarding the steady state, for a given x, the equilibrium probabilities are found through the embedded Markov chain, choosing the truncation level suitably. These probabilities yield the steady-state distribution of amount of work in the system before the arrival of a patient, i.e., the waiting time, see Sect. 5. Then one computes the steady-state sojourn time distribution by evaluating the convolution of the waiting time and service time.

(ii) In the transient case one uses the recursive method outlined in Sect. 3 to evaluate the sojourn-time distribution for given $\boldsymbol{x}$.

We now evaluate the loss function of our choice (sequential or simultaneous approach, and quadratic or linear loss).

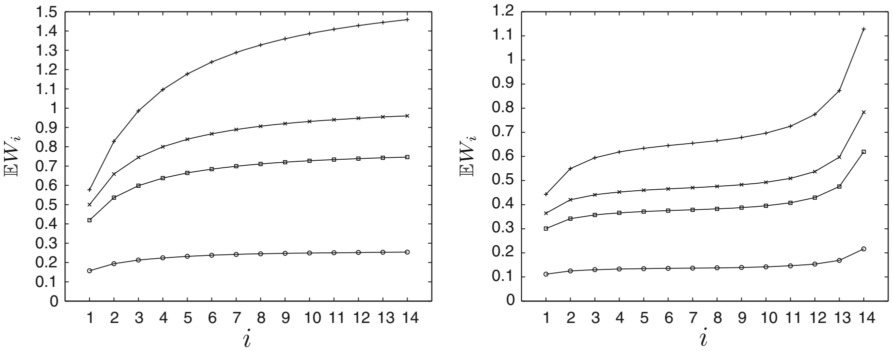
3. Given the loss function, we perform the minimization (In the sequential approach this is implemented by solving the first order condition).

Obviously, the computational effort can be substantially reduced by tailoring the software more directly to our specific needs, e.g., by using 3rd generation programming environments (such as C++). Also, a significant reduction of the computational effort can be achieved by using optimal values of a previously calculated, “nearby” scenario as starting values when determining a next schedule; this idea can be exploited for instance when generating optimal schedules for a range of SCVs.

6.6 Comparison of sequential and simultaneous approaches

In this section, we study the expected waiting time and idle time associated with each individual client, so as to compare the impact of the approach chosen (i.e., sequential vs. simultaneous). In Figs. 9 and 10 we do so for linear loss, whereas Fig. 11a, b relates to quadratic loss. The lines of the figures in this section are labeled as in Fig. 1; that is, the crosses refer to an SCV $= 1.6036$, the blocks to an SCV $= 1.0000$, and the rounds to an SCV $= 0.7186$. In all experiments, we focus on $n = 15$ clients and hence 14 interarrival times, but other values of n show very similar behavior.

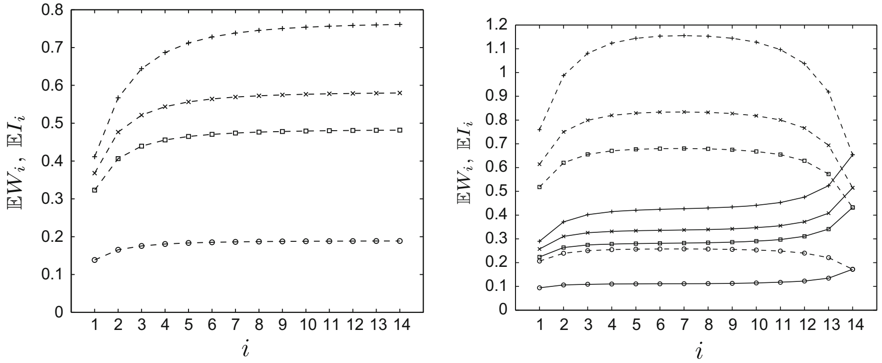
Figure 9a, b shows the idle times for each arrival for the sequential (Fig. 9a) and simultaneous (Fig. 9b) optimization approach, with linear loss. From these results we



(a) Linear loss, sequential approach.

(b) Linear loss, simultaneous approach.

Fig. 10 The optimal waiting times by sequential and simultaneous approach for the various SCVs, in case of a linear loss



(a) Quadratic loss, sequential approach.

(b) Quadratic loss, simultaneous approach; top curves are mean idle times, bottom curves are mean waiting times.

Fig. 11 The optimal waiting times by sequential and simultaneous approach for the various SCVs, in case of a quadratic loss

observe that the mean idle times in the sequential approach are in general smaller than those in the simultaneous approach. Furthermore, the patterns of the mean idle times resonate the patterns of the optimal individual interarrival times—see Fig. 1a for the sequential approach, and Figs. 2a, 3a, 4a, and 5a for the simultaneous approach.

Next, Fig. 10a, b show the mean waiting times for both approaches, with linear loss. From these results we observe that the individual waiting times are larger in case of the sequential approach. This means that, together with the results of the idle times, we conclude that the sequential approach favors the server. Furthermore, we observe that the individual waiting times are more variable for the simultaneous approach than for the sequential approach; this salient feature illustrates the difference in 'fairness' between both schemes.

Finally, we discuss the mean idle and waiting times for quadratic loss, as shown in Fig. 11a, b. From the sequential results of Fig. 11a, we observe that for each arrival the mean idle time equals the mean waiting time. This follows from the risk in (6), in case $\alpha = \frac{1}{2}$, and its corresponding first order condition. The optimal interarrival time follows from $\mathbb{E}(S_{i-1} - x_{i-1}) = 0$ for clients $i = 2, \ldots, n$, entailing that $x^\star$ is chosen so that $\mathbb{E}I_i = \mathbb{E}W_i$.

From the simultaneous results of Fig. 11b, we again conclude that the mean idle times are larger than for the sequential approach; at the same time, the mean waiting times are smaller. From this observation it is seen that also for quadratic loss the sequential approach favors the server. Also, we see that the *dome shape* is reflected in the pattern of the mean idles times; cf. Figs. 2b, 3b, 4b, and 5b; and the mean waiting times of each individual arrival are more variable than for the sequential approach, in line with what we observed for linear loss. The fact that the mean idle time equals the mean waiting time for the final arrival essentially follows from the fact that the final arrival is "sequentially" scheduled, since no subsequent client is to be scheduled.

7 Conclusions and directions for future work

This paper demonstrated how to optimally generate appointment schedules. In our procedure, we replace general service-time distributions by their phase-type counterparts, and then (either sequentially or simultaneously) optimize a (dis-)utility function. The procedures are backed by a series of numerical experiments, that also shed light on the impact of the utility function and the service-times' variability (expressed in terms of the squared coefficient of variation, SCV) on the optimal interarrival times.

The numerics evidence the feasibility of the proposed procedure. At the same time we empirically assessed its robustness; in particular it was shown that replacing non-phase-type distributions (Weibull, log-normal) by phase-type distributions, based on a two-moment fit, hardly affects the optimal schedule.

There are various directions for future research. (i) In the first place the setup can be made more realistic, that is, more in line with specific conditions in healthcare settings. For instance, ideally schedules should be flexible in terms of their capacity to deal with urgent additional clients. This requires insight into the possibility to adapt the schedule *on the fly.* (ii) In the second place one could think of situations with multiple servers, in which it also needs to be determined to which server each client should be assigned. In addition, it would be interesting to study settings in which clients have to undergo multiple (rather than just one) services. (iii) In numerical examples, we considered the situation of all clients having the same service-time distribution. It is readily checked, though, that the modeling framework does not require such a uniformity: all computations can be performed for heterogeneous service times as well.

In those heterogeneous situations the ordering issue plays a role; intuitively one would think that it makes sense to schedule clients with small variances first. It was already proven that for the sequential approach and the service times stemming from a single scale family (that is, B_i is distributed as $\sigma_i U$ for nonnegative σ_i and U some nonnegative random variable) the clients should be ordered in increasing order of variance, see [10]; for other situations, however, no rigorous results have been found so far.

Acknowledgments We are grateful to Gideon Weiss (Haifa University), Chris Klaassen (University of Amsterdam), and Dieter Fiems (Ghent University) for providing useful comments.

References

1. Asmussen, S.: Applied Probability and Queues, Applications of Mathematics. Stochastic Modelling and Applied Probability, vol. 51, 2nd edn. Springer-Verlag, New York (2003)
2. Asmussen, S., Nerman, O., Olssen, M.: Fitting phase-type distributions via the EM algorithm. Scand. J. Stat. **23**(4), 419–441 (1996)
3. Babes, M., Sarma, G.: Out-patient queues at the lbn-Rochd health center. Oper. Res. Soc. **42**(10), 845–855 (1991)
4. Bailey, N.: A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. J. R. Stat. Soc. Ser. B **14**(2), 185–199 (1952)
5. Cayirli, T., Veral, E.: Outpatient scheduling in health care: a review of literature. Prod. Oper. Manag. **12**(4), 519–549 (2003)
6. Cayirli, T., Yang, K., Quek, S.: A universal appointment rule in the presence of no-shows and walk-ins. Prod. Oper. Manag. **21**(4), 682–697 (2012)
7. De Vuyst, S., Bruneel, H., Fiems, D.: Fast evaluation of appointment schedules for outpatients in health care. Proc. ASMTA **2011**, 113–131 (2011)
8. Hassin, R., Mendel, S.: Scheduling arrivals to queues: a single-server model with no-shows. Manag. Sci. **54**(3), 565–572 (2008)
9. Kaandorp, G., Koole, G.: Optimal outpatient appointment scheduling. Health Care Manag. Sci. **10**(3), 217–229 (2007)
10. Kemper, B., Klaassen, C., Mandjes, M.: Utility-based appointment scheduling. IBIS UvA Working Paper. http://www1.fee.uva.nl/pp/bin/1261fulltext.pdf 2011–12 (2011)
11. Kemper, B., Mandjes, M.: Mean sojourn times in two-queue fork–join systems: bounds and approximations. OR Spectr. **34**(3), 723–742 (2012)
12. Klassen, K., Rohleder, T.: Scheduling outpatient appointments in a dynamic environment. J. Oper. Manag. **14**(2), 83–101 (1996)
13. Lau, H., Lau, A.: A fast procedure for computing the total system cost of an appointment schedule for medical and kindred facilities. IIE Trans. **32**(9), 833–839 (2007)
14. Lindley, D.: The theory of queues with a single server. Math. Proc. Camb. Philos. Soc. **48**(2), 277–289 (1952)
15. Luo, J., Kulkarni, V., Ziya, S.: Appointment scheduling under patient no-shows and service interruptions. Manuf. Serv. Oper. Manag. **14**(4), 670–684 (2012)
16. Robinson, L., Chen, R.: Scheduling doctors' appointments: optimal and empirically-based heuristic policies. IIE Trans. **35**(3), 295–307 (2003)
17. Robinson, L., Chen, R.: Estimating the implied value of the customer's waiting time. Manuf. Serv. Oper. Manag. **13**(1), 53–57 (2011)
18. Tijms, H.: Stochastic Modelling and Analysis—A Computational Approach. Applied Probability and Statistics. Wiley Series in Probability and Mathematical Statistics. Wiley, Chichester (1986)
19. Vink, W., Kuiper, A., Kemper, B., Bhulai, S.: Utility-based appointment scheduling in continuous time: the lag order approximation method (2013) (under review)
20. Wang, P.: Static and dynamic scheduling of customer arrivals to a single-server system. Nav. Res. Logist. **40**(3), 345–360 (1993)
21. Wang, P.: Optimally scheduling *n* customer arrival times for a single-server system. Comput. Oper. Res. **24**(8), 703–716 (1997)
22. Weiss, E.: Models for determining estimated start times and case orderings in hospital operating rooms. IIE Trans. **22**(2), 143–150 (1990)
23. Welch, J., Bailey, N.: Appointment systems in hospital outpatient departments. Lancet **259**(6718), 1105–1108 (1952)