European Journal of Operational Research 239 (2014) 243–255

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Application of OR

Optimized appointment scheduling

Benjamin Kemper [a,*], Chris A.J. Klaassen [b,c], Michel Mandjes [b,c,d,1]

[a] Institute for Business and Industrial Statistics (IBIS UvA), Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands
[b] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
[c] Eurandom, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
[d] CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 11 April 2013
Accepted 15 May 2014
Available online 2 June 2014

Keywords:
Appointment scheduling
Utility functions
Queues

ABSTRACT

In service systems, in order to balance the server's idle times and the customers' waiting times, one may fix the arrival times of the customers beforehand in an appointment schedule. We propose a procedure for determining appointment schedules in such a D/G/1-type of system by sequentially minimizing the per-customer expected loss. Our approach provides schedules for any convex loss function; for the practically relevant cases of the quadratic and absolute value loss functions appealing closed-form results are derived. Importantly, our approach does not impose any conditions on the service time distribution; it is even allowed that the customers' service times have different distributions.

A next question that we address concerns the *order* of the customers. We develop a criterion that yields the optimal order in case the service time distributions belong to a scale family, such as the exponential family. The customers should be scheduled then in non-decreasing order of their scale parameter.

While the optimal schedule can be computed numerically under quite general circumstances, in steady-state it can be computed in closed form for exponentially distributed service times under the quadratic and absolute value loss function. Our findings are illustrated by a number of numerical examples; these also address how fast the transient schedule converges to the corresponding steady-state schedule.

1. Introduction

In service systems the service provider would like to minimize costs in terms of the server's idle times, while the customers would like to be served with minimal waiting times. To accommodate these goals of the service provider and the customers, for example in case of a dentist and his patients, one may fix the arrival times of the customers beforehand in an appointment schedule.

In this paper we consider such appointment schedules aiming at optimally balancing the idle times of the (single) server and the waiting times of the customers. Indeed, if the system is frequently idle, then it is not functioning in a cost-effective manner for the service provider, whereas if it is virtually always busy, then the customers' waiting times may become substantial. The 'classical' objective is then to minimize the system's risk (in terms of the idle times of the service provider, as well as the waiting times of the clients) by optimally choosing the clients' arrival epochs. Commonly chosen objective functions are of the type, with $\gamma > 0$,

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

where $\gamma = 1$ corresponds to the case of *linear loss* and $\gamma = 2$ to *quadratic loss*; here t_i denotes the appointed arrival time of client i, with I_i the preceding idle time of the server, and with W_i the associated waiting time. (As $t_1 = 0$, the minimum can be taken over t_2 up to t_n; as $I_1 = W_1 = 0$, we can reduce the sum to the contributions related to client $i = 2$ up to n.) Now it is crucial to observe that the random variables I_i and W_i are also affected by the arrival epochs $t_1 = 0, t_2, \ldots, t_{i-1}$ of all previous clients. This explains why solving the above optimization problem is hard: apart from numerical approaches, to the best of our knowledge no manageable characterization for the optimal schedule is known. Ideally, one would like to have a tractable solution for arbitrary loss functions (that is, not just the quadratic one) and general service time distributions, to obtain an approach that can be used across a broad range of application areas, such as health care, manufacturing, and other service systems. The general idea behind our paper is that we propose an alternative to the above 'classical' optimization framework, in

* Corresponding author. Tel.: +31 6 24994693.
E-mail addresses: benjaminkemper@gmail.com (B. Kemper), c.a.j.klaassen@uva.nl (C.A.J. Klaassen), m.r.h.mandjes@uva.nl (M. Mandjes).

[1] Part of this work was done while Michel Mandjes was at Stanford University, Stanford, CA 94305, USA.

http://dx.doi.org/10.1016/j.ejor.2014.05.027
0377-2217/

which this all is possible. The idea to work with loss functions that include both idle time and waiting time has found widespread use in the literature; see, among many other references, for example Ho and Lau (1992), Kaandorp and Koole (2007), and Wang (1999).

There is a sizeable literature on appointment scheduling, but the findings tend to be rather case-specific: often one particular loss function is considered that is appropriate for the application at hand, and in view of numerical tractability exponential or Erlang service times are assumed (Fries & Marathe, 1981; Kaandorp & Koole, 2007; Wang, 1999). Besides, many studies rely on simulation to overcome the inherent computational complexity, and to obtain support for specific heuristics, see for example Brahimi and Worthington (1991), and Rohleder and Klassen (2000). These approaches have clear limitations: it is not *a priori* clear whether an approach that is designed for an application with its specific loss function and service time distribution can be used in other application domains as well. In addition, and more importantly, these approaches do not give the theoretical insight into the nature of optimal schedules.

As pointed out in Mondschein and Weintraub (2003), in order to deal with the opposite interests of the server and the clients, two complementary levels can be distinguished. In the first place, one may facilitate the process environment with features so that waiting time and idle time are either perceived or used differently; note that this is essentially manipulating the 'disutilities' of the server and customers. On another level, one defines a *loss function*, that in some way encompasses the disutilities experienced by both server and customers. Then a schedule needs to be determined that minimizes the expected loss, that is, the *risk*, thus realizing an optimal trade-off between the agents' interests. Our work follows the latter approach.

In this paper we propose a *sequential* optimization approach as a useful and natural alternative to (1). By 'sequential' we refer to an approach that determines the i-th appointment time t_i with the earlier arrival epochs $t_1, \ldots, t_{i-1}$ being known. For instance in the case of a quadratic loss function, the sequential optimization problem yielding t_i (for given $t_1, \ldots, t_{i-1}$) is

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

The idea is that the t_i are determined recursively. Remarkably, it turns out that (2) allows an *explicit* solution: performing the optimization for $i = 1, \ldots, n$ we obtain for this quadratic loss function the optimal schedule

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

with S_j denoting client j's sojourn time, which is defined as the sum of the associated waiting time and service time.

Importantly, the approach sketched above applies to the quite general class of convex loss functions, and to arbitrary service time distributions. It is neither required that clients' service times stem from a single distribution, nor that the clients have the same loss function. Where we find for the quadratic loss function that the optimal arrival epoch equals the sum of the *means* of the sojourn times of the previous customers, for linear loss (that is, the risk function of the i-th customer equalling $\mathbb{E}I_i + \mathbb{E}W_i$) it is the sum of the *medians* of the sojourn times. In practice one often relies on the heuristic that the arrival epochs are chosen in accordance with the sum of the expected *service times* of the previous customers, rather than their sojourn times. In light of the above results, it is concluded that this commonly used strategy is suboptimal, as it does not take into account the expected waiting time.

In situations in which all information about all customers is available *a priori* (i.e., a list of customers to be scheduled, including the distributions of their service times), the logical procedure is to minimize a simultaneous objective function. The applicability of such an approach may severely suffer from the requirement that all this information should be available before a planning can be made: when calling the service provider to make an appointment, customers typically want to hear immediately when they are expected to arrive at the service facility, and they do not want to wait to be assigned an appointment time until the planner has gathered all information needed. In cases the planner does not *a priori* have all information about all customers that are to be scheduled, one would rather use an approach in which the schedule gradually fills, thus making a sequential policy the more natural setup. For this reason, the sequential approach presented in this paper is particularly useful in any situation in which customers should be given an appointment time immediately, which is a very common situation in e.g. various health care situations (a typical example being the situation of a client contacting the dentist to make an appointment).

The sequential appointment scheduling setup that we consider in this paper, can be viewed as a two-stage procedure. Prior to the, say, day that the actual service is performed, service requests arrive. At this first stage, arrival epochs are assigned to these requests (and potentially these epochs are also put in an optimal order). Then there is a second stage, at which the server executes the actual service.

As mentioned above, our paper succeeds in explicitly solving the sequential optimization problem. Earlier papers predominantly focused on approximations of the joint optimization problem, assuming specific loss functions and service distributions, and resorting to numerical techniques or simulation. We have followed our sequential approach for various reasons. (i) An evidently and very substantial advantage of the sequential approach is that it allows explicit, closed form results, and that it, in addition, enables a solution to the problem of finding the optimal order of the n jobs. In relation to this, solving the sequential scheme is computationally significantly less demanding than the simultaneous optimization problem. (ii) In the second place, as argued earlier, our approach naturally fits the situation in which customers sequentially contact the provider to make an appointment (as opposed to the situation in which *a priori* all information is available of all customers to be scheduled). (iii) Some clients may be better off under the sequential scheme, some under the joint scheme, but there is no compelling reason why one of the schemes leads to 'better' schedules. It is realized, though, that the sequential scheme allows full freedom in terms of the choice of the utility functions related to the individual clients. As a consequence, if, for some reason, it is felt that the risk associated to a specific customer is more important, one can adapt her utility function to reflect this.

The main contribution of the paper is the sequential optimization approach for appointment scheduling, as described above. Apart from the nice features that we already mentioned (applicable for a broad class of loss functions, general service time distributions), it is highly flexible, in that it allows the incorporation of various real-life phenomena such as urgent arrivals and 'no-shows'. In addition, we quantify the impact of customers arriving early or late, that is, the impact of small random perturbations with respect to the scheduled arrival epochs.

The above results concern the determination of the optimal arrival epochs, for the situation that the *order* in which the customers are served has been given. A next question concerns the optimal order; this is the second contribution of our work. We prove the appealing result that if all service time distributions concerned stem from a scale family with finite variances, clients should arrive in non-decreasing order of their scale parameter. For instance in the case that the service times obey exponential distributions with mean values $\mu_1^{-1}, \mu_2^{-1}, \ldots$, our ordering result implies that the order

in which the customers should arrive is such that the μ_is decrease (and hence the means, as well as the variances, increase). In this sense, our result is in line with the commonly used heuristic to keep the variability initially as low as possible, see e.g. Lehaney, Clarke, and Paul (1999) and Wang (1999).

The structure of the paper is as follows. Section 2 presents standard scheduling schemes and an overview of the relevant literature; it also further motivates the research reported on in this paper. Section 3 introduces our sequential optimization approach; it includes a proof of the 'mean rule' for quadratic loss and the 'median rule' for linear or absolute value loss. In Section 4 we address the problem of identifying the optimal order of the clients. Section 5 discusses a number of more practical considerations: it is pointed out how to include additional issues such as urgent arrivals and no-shows, it addresses the effect of small random perturbations around the scheduled arrival epochs, and there is a short account of computational issues. Then, in Section 6 we present some numerical examples. Section 7 concludes.

2. Preliminaries: background and literature

As mentioned in the introduction, the server and customers have opposite interests: the service provider wishes to minimize the amount of server idleness and is therefore in favor of a 'dense schedule', whereas customers are interested in minimizing waiting times and hence prefer schedules in which the slots are relatively long. In this section we provide more background on this tradeoff, as well as a brief literature overview.

2.1. Background

Let us consider the following standard scheduling scheme, which is, owing to its simplicity, frequently used in practice. Consider a sequence of jobs, each of random duration, and assume the job durations $B_1, \ldots, B_n$ to be mutually independent. Let job i be the i-th job to be scheduled. Now we define the scheduling scheme $\mathcal{S}$ by setting the arrival epoch of job i, denoted by t_i, equal to the sum of the expected durations of the previous jobs

$$t_1 := 0, \quad \text{and} \quad t_i := \sum_{j=1}^{i-1} \mathbb{E}B_j, \ i = 2, \ldots, n. \tag{3}$$

Due to its simple structure, this standard scheduling scheme is often seen in practice; cf. also Klassen and Rohleder (1996). It has a serious drawback, though: the system becomes essentially a queue with load 1, leading to potentially long waiting times. As a result this scheme might be attractive for the server, but for the customers it is not.

To support this claim, consider for the moment the situation that the B_i are identically distributed as a random variable B, such that strategy $\mathcal{S}$ can be viewed as a D/G/1 queue with load 1 starting empty. The next result shows that $\mathbb{E}W_n$ blows up as $\sqrt{n}$; while this result has appeared in various forms in the literature (cf. e.g. Whitt, 1972, Thm. 4.1), for the sake of completeness we include its proof in Appendix A.

Proposition 2.1. *In a D/G/1 queue with load 1 starting empty, with the service times having finite variance σ^2, the mean waiting time of the n-th customer obeys, as $n \to \infty$,*

$$\frac{\mathbb{E}W_n}{\sqrt{n}} \to \sigma\sqrt{\frac{2}{\pi}}.$$

This result and its proof remain valid in the GI/G/1 setting, with $\sigma^2 := \mathbb{V}\mathrm{ar}\, A + \mathbb{V}\mathrm{ar}\, B$, where A is distributed as the interarrival time.

To remedy the undesirable effect that the mean waiting times explode, one could introduce the 'adapted scheme' $\mathcal{S}_\Delta$, for some $\Delta \geqslant 0$, with

$$t_1 := 0, \quad \text{and} \quad t_i := \Delta \cdot \sum_{j=1}^{i-1} \mathbb{E}B_j, \ i = 2, \ldots, n. \tag{4}$$

Observing that $\mathcal{S}_1 = \mathcal{S}$, the above result on $\mathbb{E}W_n$ relates to the case $\Delta = 1$. Obviously, the server's idle time is reduced compared to $\mathcal{S}$ when picking $\Delta \in [0, 1)$; in the extreme case of $\Delta = 0$, all customers arrive at time 0, thus minimizing the expected idle time at the expense of the clients' waiting time. On the other hand, it is clear that the mean waiting times in the adapted scheme $\mathcal{S}_\Delta$ are reduced relative to $\mathcal{S}$ when picking $\Delta > 1$ at the expense of the expected server's idle time; then the corresponding D/G/1 queue is stable in the sense that it has a proper steady-state distribution.

These observations suggest that one should pick some Δ larger than 1 in order to find a good compromise between the waiting times of the clients and the idle time of the server, as was also recognized by Ho and Lau (1992) and references therein. It is evident, though, that Δ does not uniquely predict the customer's waiting time: a given Δ can lead to a broad range of waiting time distributions, depending on the service time distribution. Indeed, for $\Delta > 1$ deterministic service times lead to zero waiting times, while the waiting times can be quite substantial if the service time distribution has heavy tails. Intuitively, one could anticipate them to increase in the coefficient of variation of the service times. The above reasoning tells us that the schedule should incorporate more detail of the service time distributions than just their means.

To set up the schedule in a sounder way, one could use the concept of 'risk function', which measures the aggregate disutility of the server and client. More specifically, the risk associated with the i-th arrival depends on the distribution of the waiting time W_i of the i-th client, as well as the idle time I_i prior to the arrival of this i-th client. It makes sense to choose non-decreasing loss functions $g(\cdot)$ and $h(\cdot)$ with $g(0) = h(0) = 0$, and to define the *risk* associated with the i-th arrival as

$$R_i^{(g,h)}(t_1, \ldots, t_i) := \mathbb{E}g(I_i) + \mathbb{E}h(W_i);$$

clearly, this i-th risk depends on arrival epochs t_1 up to and including t_i. Note that $g(\cdot)$ and $h(\cdot)$ determine how much weight should be given to the idle and waiting time respectively. In this framework, the optimal schedule then follows from solving the minimization problem

$$\min_{t_1, \ldots, t_n} \sum_{i=1}^{n} (\mathbb{E}g(I_i) + \mathbb{E}h(W_i)). \tag{5}$$

As argued in the introduction, this optimization problem is intrinsically complex, and therefore we propose in the next section to analyze its 'sequential counterpart'. Before we do so, we first give a brief literature overview.

2.2. Literature

The literature on appointment scheduling started with the seminal works of Bailey and Welch, see e.g. Bailey (1952), Bailey (1954), Welch (1964) and Welch and Bailey (1952). They proposed a simple schedule that sets interarrival times equal to the average service time, but starts with *two* arrivals scheduled at time 0. In line with these works, most papers focus on applications of appointment scheduling in healthcare, see Cayirli and Veral (2003) for an extensive overview. We also mention Denton and Gupta (2003), and Mondschein and Weintraub (2003), who discuss a somewhat more general setting.

The usual starting point of this optimization approach concerns the choice of the specific risk. Besides waiting time and idle time, this may include various other performance metrics; see Cayirli and Veral (2003, Table 2) for an overview. It is emphasized that the choice of the loss function and service time distribution is often rather application-specific, and as a consequence of limited use for practitioners in other application domains. This is a distinguishing feature of our work: we allow any user-specific convex loss function, and any client-specific service time distribution. In particular, in our setup the practitioner can pick her or his risk function and apply our approach; evidently, it is beyond the scope of our work to develop guidelines that help choosing a specific loss function for a particular appointment scheduling problem.

Many studies rely on simulation to overcome the inherent analytical and computational complexity, and to obtain support for specific heuristics, such as for example Brahimi and Worthington (1991), and Rohleder and Klassen (2000); in addition we mention Klassen and Rohleder (2004), and Patrick, Puterman, and Queyranne (2008). In Robinson and Chen (2003) the focus is on techniques that facilitate the estimation of the relative cost of the patient waiting time given average queue length and occupation rate. The authors of Begen and Queyranne (2011) devise an efficient scheme to set up schedules for the case that the random service times have discrete support.

There are similarities between appointment scheduling, as discussed in the present paper, and (single) machine scheduling. The main difference between these branches of research is that in appointment scheduling the release dates (in machine scheduling lingo) are to be determined, for a given sequence of jobs. In addition, objective functions used in the machine scheduling literature tend to be quite different from the ones used in the appointment scheduling literature (balancing idle times and waiting times). We refer for an in-depth treatment to the book by Pinedo (2001).

3. Sequential optimization

In our sequential counterpart of (5), we minimize, for each i the risk

$$R_i^{(g,h)}(t_1,\dots,t_i) = \mathbb{E}g(I_i) + \mathbb{E}h(W_i)$$

over t_i, where it is essential that $t_1,\dots,t_{i-1}$ are given; we do so in a recursive manner for $i = 1,\dots,n$. A crucial observation is that I_i and W_i cannot be both positive, and it is therefore natural to introduce the *loss function*

$$\ell(x) = g(-x)\mathbf{1}_{[x<0]} + h(x)\mathbf{1}_{[x>0]}, \quad x \in \mathbb{R},$$

which is non-increasing on $(-\infty,0]$ and non-decreasing on $[0,\infty)$ with $\ell(0) = 0$. In terms of this loss function we may write

$$R_i^{(g,h)}(t_1,\dots,t_i) = \mathbb{E}g(I_i) + \mathbb{E}h(W_i) = \mathbb{E}\ell(W_i - I_i), \quad i = 1,\dots,n,$$

and we define the risk at the i-th arrival with loss function $\ell(\cdot)$ as

$$R_i^{(\ell)}(t_1,\dots,t_i) = \mathbb{E}\ell(W_i - I_i), \quad i = 1,\dots,n. \tag{6}$$

3.1. Schedule for quadratic and linear risk functions

To ease the exposition, we first present our procedure to find the optimal interarrival times for loss functions that are used most frequently in the literature: the absolute value and quadratic loss functions. Then Section 3.2 shows that this approach essentially carries over to the class of convex loss functions.

Quadratic loss function. A simple (that is, non-weighted) quadratic loss function is defined by

$$R_i^{(v)}(t_1,\dots,t_i) := \mathbb{E}I_i^2 + \mathbb{E}W_i^2, \quad i = 1,\dots,n.$$

Due to the well-known Lindley recursion Lindley (1952),

$$I_i = \max\{t_i - t_{i-1} - W_{i-1} - B_{i-1}, 0\} \tag{7}$$

and

$$W_i = \max\{W_{i-1} + B_{i-1} - t_i + t_{i-1}, 0\}. \tag{8}$$

Let $S_i := W_i + B_i$ denote the sojourn time of the i-th customer, 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 (7) and (8) in mind, we may write

$$W_i^2 + I_i^2 = (S_{i-1} - x_{i-1})^2 \tag{9}$$

and so the system's risk (in relation to the i-th client) reads

$$R_i^{(v)}(t_1,\dots,t_{i-1},t_{i-1}+x_{i-1}) = \mathbb{E}(S_{i-1} - x_{i-1})^2. \tag{10}$$

The following result is an immediate consequence of the general approach that will be presented in Section 3.2. We give two proofs: the first one is elementary and insightful; the second one has the flavor of the approach of Section 3.2.

Proposition 3.1. *Let the job durations $B_1,\dots,B_n$ be independent nonnegative random variables with finite second moment. Define the schedule $\mathcal{V}$ through*

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

For the simple quadratic loss function, the schedule $\mathcal{V}$ sequentially minimizes the risk (10).

Proof 1. In view of

$$S_i = W_i + B_i \leqslant \sum_{j=1}^{i} B_j, \tag{11}$$

the sojourn time S_i has finite second moments. Observe that, with $W_1 = 0$, $I_1 = 0$, (7) and (8) hold. This immediately implies that the maxima in Eqs. (7) and (8) vanish in

$$I_i^2 + W_i^2 = (W_{i-1} + B_{i-1} - t_i + t_{i-1})^2 = (S_{i-1} - t_i + t_{i-1})^2.$$

Now note

$$\mathbb{E}(S_{i-1} - t_i + t_{i-1})^2 = \mathbb{V}\mathrm{ar}\,S_{i-1} + (\mathbb{E}S_{i-1} - t_i + t_{i-1})^2.$$

Consequently, for given t_{i-1}, the risk of customer i equals

$$\min_{t_i} R_i^{(v)}(t_1,\dots,t_i) = \min_{t_i} \mathbb{E}(S_{i-1} - t_i + t_{i-1})^2 = \mathbb{V}\mathrm{ar}S_{i-1},$$

where the minimum is attained for $t_i = t_{i-1} + \mathbb{E}S_{i-1}$. Hence the optimal interarrival time $x_{i-1}^\star$ is $\mathbb{E}S_{i-1}$, in agreement with schedule $\mathcal{V}$. □

Proof 2. Again, observe that $W_1 = 0$ and $I_1 = 0$, and

$$I_i^2 + W_i^2 = (S_{i-1} - t_i + t_{i-1})^2.$$

Minimize, for given t_{i-1}, the risk at the arrival of client i:

$$\min_{t_i} R_i^{(v)}(t_1,\dots,t_i) = \min_{t_i} \mathbb{E}(S_{i-1} - t_i + t_{i-1})^2 = \min_{x} \mathbb{E}(S_{i-1} - x)^2.$$

Since we deal with a nonnegative convex loss function in x, the first order condition (use 'Leibniz's rule') yields the optimal interarrival time for the $(i-1)$-st client, and consequently also the optimal arrival time. We have to solve

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}(S_{i-1} - x)^2 = -2\int_0^\infty (s - x)\mathrm{d}F_{S_{i-1}}(s) = 0,$$

which gives us the optimal interarrival time for the $(i-1)$-st client $x_{i-1}^\star = \mathbb{E}S_{i-1}$. □

Note that the latter proof is reminiscent of that featuring in the well known *newsvendor problem*, see for instance Hopp and Spearman (1995). Evidently, it can be used to other loss functions as well. Next, we consider the case of the absolute value as loss function.

Absolute value loss function. Consider the simple (that is, non-weighted) linear loss function: the risk associated with the i-th customer equals the sum of the expected waiting time and idle time. Again, due to (7) and (8), we obtain

$$R_i^{(u)}(t_1,\dots,t_{i-1},t_{i-1}+x_{i-1}) := \mathbb{E}I_i + \mathbb{E}W_i = \mathbb{E}|S_{i-1}-x_{i-1}|, \tag{12}$$

which is based on the absolute value, a nonnegative convex loss function.

Proposition 3.2. *Let the job durations $B_1,\dots,B_n$ be independent nonnegative random variables with finite first moments. Define the schedule $\mathcal{U}$ through*

$$t_1 := 0, \quad \text{and} \quad t_i := \sum_{j=1}^{i-1} F_{S_j}^{-1}\left(\frac{1}{2}\right), \quad i=2,\dots,n,$$

For the simple linear loss function, the schedule $\mathcal{U}$ sequentially minimizes the risk (12).

Proof. In view of (11), S_i has finite first moment. By 'Fubini' we have to minimize

$$\begin{aligned}\mathbb{E}|S_{i-1}-x| &= \int_0^x \int_s^x \mathrm{d}y \quad \mathrm{d}F_{S_{i-1}}(s) + \int_x^\infty \int_x^s \mathrm{d}y \quad \mathrm{d}F_{S_{i-1}}(s)\\ &= \int_0^x \int_0^y \mathrm{d}F_{S_{i-1}}(s) \quad \mathrm{d}y + \int_x^\infty \int_y^\infty \mathrm{d}F_{S_{i-1}}(s) \quad \mathrm{d}y\\ &= \int_0^x F_{S_{i-1}}(y)\mathrm{d}y + \int_x^\infty (1-F_{S_{i-1}}(y))\mathrm{d}y.\end{aligned}$$

Note that the derivative

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}|S_{i-1}-x| = 2F_{S_{i-1}}(x)-1,$$

exists for all x at which $F_{S_{i-1}}$ is continuous, and changes sign at $F_{S_{i-1}}^{-1}(\frac{1}{2})$. This implies that we should take the optimal interarrival time $x^\star$ for the $(i-1)$-st customer equal to a *median* of S_{i-1}, that is, $x_{i-1}^\star = F_{S_{i-1}}^{-1}(\frac{1}{2})$, as claimed. □

Interestingly, we conclude that the absolute value loss function leads to interarrival times equaling a *median* of the sojourn times, whereas a quadratic loss function leads to interarrival times equaling the *mean* of the sojourn times. There is a connection with statistical estimation theory: there one obtains the (sample) median when imposing the absolute value as loss function and the mean absolute deviation as risk, whereas the (sample) mean is found when imposing the square as loss function and the mean square error as risk.

It is noted that the above approach, which is essentially based on Leibniz's rule, carries over to more general loss functions. We present the resulting general approach in Section 3.2.

3.2. Schedule for convex loss functions

We now present our sequential optimization approach for convex loss functions, which contains the cases dealt with in Section 3.1. The approach borrows elements from statistical decision theory; see e.g. Ferguson (1967) or Bickel and Doksum, 2001, chap. 10.

As observed before, due to (7) and (8),

$$W_i - I_i = W_{i-1} + B_{i-1} - t_i + t_{i-1} = S_{i-1} - x_{i-1} \in \mathbb{R}, \tag{13}$$

so that we can define the general risk (to be minimized over x_{i-1}) by

$$R_i^{(\ell)}(t_1,\dots,t_i) := \mathbb{E}\ell(S_{i-1}-x_{i-1}), \quad x_{i-1} = t_i - t_{i-1}. \tag{14}$$

If the loss function is convex, then $x_{i-1}^\star$ can be found by solving the first order condition, as explained in Lemma B.1 in full detail. As for the quadratic and linear case, we can set up a sequentially optimized scheme, in which the arrival epochs can be determined recursively. More precisely, our main result Theorem 3.3 states how to generate the optimal schedule for any nonnegative convex loss function $\ell(\cdot)$, and for any sojourn time distribution function $F_{S_i}(\cdot)$.

Theorem 3.3. *Let $\ell(\cdot)$ be a nonnegative convex loss function on $\mathbb{R}$ with $\ell(0)=0$. Let the job durations $B_1,\dots,B_n$ be independent nonnegative random variables such that*

$$\mathbb{E}\ell\left(\sum_{i=1}^{n-1} B_i + x\right) < \infty \tag{15}$$

holds for all positive x. Let

$$R_i^{\ell}(t_1,\dots,t_i) = \mathbb{E}\ell(W_i - I_i) = \mathbb{E}\ell(S_{i-1}-x_{i-1}), \quad x_{i-1} = t_i - t_{i-1}, \tag{16}$$

be the risk associated with the i-th customer.

Define the schedule $\mathcal{W}$ through

$$t_1 := 0 \quad \text{and} \quad t_i := \sum_{j=1}^{i-1} x_j^\star, \quad i=2,\dots,n,$$

where $x_j^\star$ is a nonnegative value at which $\mathbb{E}\ell'(S_j - x)$ changes sign or vanishes; if such a value does not exist $x_j^\star$ is set equal to ∞. The schedule $\mathcal{W}$ sequentially minimizes the risk (16).

Proof. In view of (11) and (15) we have

$$\mathbb{E}\ell(S_j - x) \leqslant \mathbb{E}\ell\left(\sum_{i=1}^{n-1} B_i - x\right) < \infty, \quad x < 0,$$

$$\mathbb{E}\ell(S_j - x) \leqslant \mathbb{E}\ell(S_j) + \ell(-x) \leqslant \mathbb{E}\ell\left(\sum_{i=1}^{n-1} B_i\right) + \ell(-x) < \infty, \quad x \geqslant 0.$$

Consequently, Lemma B.1 may be applied. Note that $\mathbb{E}\ell'(S_j - x)$ is nonincreasing in x and that it is nonnegative at $x=0$ in view of $S_j \geqslant 0$ as stated. It follows that $x_j^\star$ may be chosen to be nonnegative. □

If the loss function $\ell(\cdot)$ is not identically 0, but vanishes on the negative half line, and S_j is not a bounded random variable, then $x_j^\star$ has to be chosen ∞. At the other extreme, if the loss function $\ell(\cdot)$ is not identically 0, but vanishes on the positive half line, $x_j^\star$ may be chosen equal to 0. These cases correspond to the situation that idle times do not matter, and the situation that waiting times are irrelevant, respectively. See also the next subsection.

To ease the exposition, we have so far assumed that the loss functions are uniform in i, that is, equal for any customer. Inspection of the above theorem shows that this is by no means necessary. The result straightforwardly extends to risk functions of the type $R_i^{(\ell)}(t_1,\dots,t_{i-1}+x_{i-1}) := \mathbb{E}\ell_i(S_{i-1}-x_{i-1})$, that is, the function $\ell_i(\cdot)$ is client-specific.

3.3. Weighted standard loss function

The loss functions of Section 3.1 can be generalized in the sense that we could relax the equally weighing restriction. As argued in e.g. Ho and Lau (1992), it is sometimes justified to weigh the server's idle time in a different manner than the client's waiting time. We here consider both a weighted linear and weighted quadratic loss function.

A weighted-linear loss function. Let the risk be a weighted sum of the idle time and waiting time, $\beta\mathbb{E}I_i + \gamma\mathbb{E}W_i$ for non-negative β, γ. Without loss of generality we may concentrate on risks of the form

$$R_i^{(u,\alpha)}(t_1,\ldots,t_i) := \alpha\mathbb{E}I_i + (1-\alpha)\mathbb{E}W_i, \quad i = 1,\ldots,n, \quad \alpha \in (0,1).$$

Note that for $\alpha \downarrow 0$ this risk minimizes only the client's waiting time. This results in a schedule that favors the clients by making the interarrival times excessively long, thus generating substantial idle times for the server. For $\alpha \uparrow 1$ the risk minimizes the idle times of the server, which is similar to setting $\Delta = 0$ in (4): all customers are to arrive at time 0, resulting in long waiting times for the clients.

The optimal interarrival time $x_{i-1}^\star$ can be found by solving the equation (cf. the proof of Proposition 3.2)

$$\alpha F_{S_{i-1}}(x) - (1-\alpha)\big(1 - F_{S_{i-1}}(x)\big) = F_{S_{i-1}}(x) - 1 + \alpha = 0,$$

for $i = 2,\ldots,n$. By Theorem 3.3 this leads to the optimal schedule

$$t_1 := 0 \quad \text{and} \quad t_i := \sum_{j=1}^{i} F_{S_{i-1}}^{-1}(1-\alpha), \quad i = 2,\ldots,n.$$

For $\alpha = 1/2$ it is easily seen that this schedule equals the optimal scheme of Proposition 3.2, as desired. Note that $\alpha/(1-\alpha)$ may be viewed as the ratio between the cost of idle time and the cost of waiting time. Guidelines so as how to choose α are given by Fries and Marathe (1981).

A weighted-quadratic loss function. Here we consider a loss function that is of the form

$$R_i^{(v,\alpha)}(t_1,\ldots,t_i) := \alpha\mathbb{E}I_i^2 + (1-\alpha)\mathbb{E}W_i^2,$$

for $i = 1,\ldots,n$ and $0 \leqslant \alpha \leqslant 1$. Applying Theorem 3.3, we obtain that the optimal interarrival time x_{i-1} solves

$$\alpha(x - \mathbb{E}S_{i-1}) - (1-2\alpha)\int_x^\infty \mathbb{P}(S_{i-1} > s)\mathrm{d}s = 0, \tag{17}$$

which for $\alpha = 1/2$ reduces to the scheme of Proposition 3.1, as desired. We present an example involving a weighted-quadratic loss function in Section 6.

4. Optimal ordering

After having dealt with the optimal schedule for a given order of the clients, the obvious next question is: how should the order of arriving clients be chosen? This question will be addressed in this section.

A commonly used heuristic is that the service times are put in increasing order of variance. The underlying idea is that the variability (in terms of waiting times and idle times) in a D/G/1 system, is exclusively caused by the variability of the service times. When putting the clients with low variability (in their service times) early in the schedule, the uncertainty for clients arriving later is reduced. In this section we study the ordering issue, by deriving a result that confirms the above heuristic. Related results were presented in Wang (1999) for the case of exponentially distributed jobs.

Before going to the main result of this section, we present a simple covariance inequality due to Chebyshev, see Hardy, Littlewood, and Pólya (1934, pp. 43–44) for an overview. Note that this result is known as Chebyshev's algebraic inequality, see Mitrinović and Vasić (1974), and has been rediscovered several times later, see Jogdeo (1977) and references therein.

Lemma 4.1. *Let $s(\cdot)$ be a non-decreasing function and let X be a random variable such that $\mathbb{E}X^2 < \infty$ and $\mathbb{E}s^2(X) < \infty$ hold. Then $\mathbb{C}\mathrm{ov}(s(X),X) \geqslant 0$ holds. This inequality is strict, if $s(\cdot)$ is strictly increasing and X is non-degenerate.*

Proof. Note that $[s(X) - s(\mathbb{E}X)][X - \mathbb{E}X] \geqslant 0$ holds a.s. and that we have $\mathbb{C}\mathrm{ov}(s(X),X) = \mathbb{E}([s(X) - s(\mathbb{E}X)][X - \mathbb{E}X])$. □

The main contribution of this section is the following. Consider n customers with independent service times $B_1,\ldots,B_n$, and let B_i be distributed as $\sigma_i B$ for $i = 1,\ldots,n$, where we assume $\sigma_1 \leqslant \sigma_2 \leqslant \ldots \leqslant \sigma_n$. Let π be a permutation of $\{1,\ldots,n\}$. The corresponding permutation $(B_{\pi(1)},\ldots,B_{\pi(n)})$ of the service times $(B_1,\ldots,B_n)$ that sequentially minimizes the risks, is the identical permutation $\pi(i) = i, i = 1,\ldots,n$. More precisely, we have the following result; see Appendix C for a proof.

Theorem 4.2. *Let $R_i^{(\ell)}(t_1,\ldots,t_i) := \mathbb{E}\ell(W_i - I_i)$ be the risk corresponding to a non-negative convex loss function $\ell(\cdot)$ with $\ell(0) = 0$, and let $\sigma_1 \leqslant \sigma_2 \leqslant \ldots \leqslant \sigma_n$ be positive numbers. In addition, let for all i, all $\sigma > 0$, and all $x \in \mathbb{R}$ the expectations $\mathbb{E}\ell(W_i + \sigma B - x)$, $\mathbb{E}|\ell'(W_i + \sigma B - x)|^2$, and $\mathbb{E}B^2$ be finite. Furthermore, for any permutation π, let $R_i^{(\ell)}(\pi)$ be the risk from (14) sequentially minimized by the schedule $\mathcal{W}$ from Theorem 3.3 for $i = 1,\ldots,n$, when the service times B_i are distributed as $\sigma_{\pi(i)}B, i = 1,\ldots,n$.*

If $\ell'(\cdot)$ is continuous or if the random service time B has a density with respect to Lebesgue measure, then the identical permutation $\pi(i) = i$ sequentially minimizes the risk $R_i^{(\ell)}(\pi)$ at the i-th arrival, $i = 1,\ldots,n$.

Since the σ_i are scale parameters, this theorem confirms the intuitive idea that the clients should be put in increasing order of variance. For the special case of the B_i having exponential distributions with parameters $\lambda_1,\ldots,\lambda_n$, it implies that the order should be such that the λ_i decrease with i; that is, the one with lowest variance (and mean) should be served first. In Wang (1999) partial proofs were given for a related result for the special case of exponential service times.

5. Extensions, robustness, computational issues

In this section we focus on a number of issues that directly relate to implementing our approach in a practical setting. The first subsection covers a number of practically relevant extensions. Then, in Section 5.2 we argue that our scheme has desirable robustness properties with respect to small deviations from the scheduled arrival epochs. This section concludes with a brief account of computational issues.

Before being able to set up a schedule, the service provider has to decide on which objective function he or she wants to use. It is not the objective of this paper to advise service providers on the specific loss function that should be chosen: this is a strongly situation-dependent issue, and reflects the specific policy the service provider has in terms of the grade of service that should be offered to customers. Our methodology is particularly useful when assessing the effect of choosing specific types of loss functions, e.g. linear or quadratic; see for instance Fig. 3 in Section 6, in which this effect has been visualized. In addition, the same figure shows how it facilitates the assessment of the schedule's sensitivity as a function of the weight α.

5.1. Additional issues: urgent arrivals, no-shows, and multiple servers

In this section we present a number of extensions of the scheduling scheme developed in the previous sections. We focus on two specific complications, cf. Cayirli and Veral (2003): an additional stream of customers that has to be handled with priority, and the impact of no-shows. Although we discuss the two complications for the case of quadratic loss functions only, the results can be generalized to any convex loss function, in the spirit of Lemma B.1 and Theorem 3.3.

Urgent arrivals. A common approach is to model urgent arrivals by adding a random process, see for example Rising, Baron, and Averill (1973), and Swisher, Jacobsen, Jun, and Balci (2001). Consider the model presented in Section 3, but let there be an additional Poisson stream of customers that has to be handled with priority – if the server is busy upon arrival of such an 'urgent customer', the job in service is finished before the server starts serving the urgent customer(s). Let urgent customers arrive according to a Poisson process of rate λ, and let their service requirements $J_1, J_2, \ldots$ be i.i.d. random variables distributed as a generic random variable J.

Under these additional urgent arrivals and quadratic loss functions, the scheduling scheme $\mathcal{V}$ could be adapted to

$$t_i = t_{i-1} + \mathbb{E}W_{i-1} + \mathbb{E}B_{i-1} + \lambda(t_i - t_{i-1}) \cdot \mathbb{E}J,$$

leading to

$$t_1 := 0, \quad \text{and} \quad t_i := \frac{1}{1-\lambda\,\mathbb{E}J} \sum_{j=1}^{i-1} \left(\mathbb{E}W_j + \mathbb{E}B_j\right),\ i = 2, \ldots, n.$$

Note that we should necessarily have $\lambda\ \mathbb{E}J < 1$, as otherwise the second 'regular' job would never be scheduled.

No-shows. As argued in, e.g., Hassin and Mendel (2008) and Kaandorp and Koole (2007), the impact of no-shows may be substantial. To analyze this effect, let δ_i be the indicator that customer i actually shows up, independently of the job sizes and other customers showing up or not, where δ_i equals 1 with probability p_i and 0 else. This means that the service time B_i is replaced by $B_i\delta_i$. It is readily checked that in this model we would obtain the schedule (under quadratic loss functions)

$$t_1 := 0, \quad \text{and} \quad t_i := \sum_{j=1}^{i-1} \left(\mathbb{E}W_j + p_j\,\mathbb{E}B_j\right),\ i = 2, \ldots, n.$$

Other loss functions can be dealt with similarly. Note that W_i and hence $\mathbb{E}W_i$ are influenced by no-shows.

Multiple server setting. Our approach can be extended to the case of multiple servers (say s). In this setting the service provider sequentially decides to schedule the next appointment to the server that contributes the least expected loss to the system. This can be implemented by the following iterative procedure. Suppose the i-th customer enters, and the previous $i-1$ have been assigned an arrival time, and a specific server by whom they will be served. For customer i it is then computed what his optimal arrival time would be, for each of the s possible servers. Then customer i is assigned to the queue with the lowest risk contribution of this customer.

Note that, due to our generic sequential approach, the servers need not be identical, since the approach can deal with distinct service time distributions (which could depend on both client and server). Also remark that in the procedure sketched above it has not been taken into account that each server, in case it idles earlier than expected, could potentially serve customers that are waiting at other queues. In principle, however, the procedure can be adapted to that setting, too, by using a D/G/s-type of queue, which is computationally more involved.

Objective function. It is evident that the objective function should include features that involve both the clients' interests and the server's interests, and a natural choice is to build the objective function around waiting times and idle times, as we do in this paper. The incorporation of other metrics on top of these is not always possible. Consider for instance the *makespan* (or *session end time*), equalling

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

This can be incorporated in the simultaneous approach with linear objective function, but this cannot be done in the simultaneous approach with other objective functions, and not in the sequential approach either. A similar remark applies to the *facility overtime*. This metric, considered in e.g. Kaandorp and Koole (2007), is defined as the positive part of the difference between the session end time and some scheduled end time T:

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

5.2. Impact of small perturbations

In practice the clients' arrival epochs will slightly deviate from the scheduled epochs. In this subsection we assess the impact of these perturbations, showing how to adapt the schemes identified in Section 3.1. Further results on 'almost deterministic arrival processes' are given in e.g. Araman and Glynn (2012), while there is a strong relation to the analysis of the effect of *jitter* in communication networks as well Roberts et al. (1996, chap. 3).

The setup considered in this subsection, is that a particular client, say the i-th, arrives not necessarily on time. We first study how this affects his optimal arrival epoch, and then we comment on the impact on customers arriving later, that is, customers $i+1, i+2, \ldots$. We consider the situation that the perturbation around the scheduled arrival time is substantially smaller than the typical job durations. This is realistic in practice; think of a dentist whose check-ups last for instance 15 up to 20 minutes, while clients may be one or two minutes early or late. The consequence is that customer $i+1$ does not take over customer i.

In the sequel we let $N_i^{(\varepsilon)}$ be the perturbation around the arrival of the i-th customer. $N_i^{(\varepsilon)}$ could have for instance a Normal distribution with mean 0 and variance ε^2, or an alternative distribution on $\{-\varepsilon, \varepsilon\}$ (each with probability $\frac{1}{2}$).

It is not hard to check that in the case of quadratic loss the optimal interarrival time $x_{i-1}^\star$ now minimizes

$$\mathbb{E}\left(S_{i-1} - x_{i-1} - N_i^{(\varepsilon)}\right)^2,$$

whereas in the case of absolute value loss it minimizes

$$\mathbb{E}\left|S_{i-1} - x_{i-1} - N_i^{(\varepsilon)}\right|.$$

We now analyze both cases separately. For ease we leave out the index, as this is constant throughout the analysis. We let $x^\star(\varepsilon)$ be the optimal interarrival time in the ε-perturbed situation; $x^\star \equiv x^\star(0)$ is therefore the solution we have identified in Section 3.1.

Quadratic loss function. It is trivial to observe that $x^\star(\varepsilon) = \mathbb{E}S - \mathbb{E}N^{(\varepsilon)}$, which reduces to $x^\star(\varepsilon) = \mathbb{E}S$ due to $\mathbb{E}N^{(\varepsilon)} = 0$. We conclude that we obtain the same solution as in the non-perturbed case.

Linear loss function. In self-evident notation, we have that

$$x^\star(\varepsilon) = F^{-1}_{S-N^{(\varepsilon)}}\left(\frac{1}{2}\right),$$

so that we need to solve

$$\frac{1}{2} = \int_{-\infty}^{x^\star(\varepsilon)} F_S(x^\star(\varepsilon) - y) f_{N^{(\varepsilon)}}(-y)\mathrm{d}y;$$

here and in the sequel $f_X(\cdot)$ denotes the density (assumed to exist) of the random variable X. Now expand, under the obvious regularity properties,

$$F_S(x^\star(\varepsilon) - y) \approx F_S(x^\star) + f_S(x^\star)(x^\star(\varepsilon) - x^\star - y) + \frac{f_S'(x^\star)}{2}(x^\star(\varepsilon) - x^\star - y)^2.$$

Putting $x^\star(\varepsilon) = x^\star + \kappa_1\varepsilon + \kappa_2\varepsilon^2 + O(\varepsilon^3)$ and using that $F_S(x^\star) = \frac{1}{2}$, routine calculations yield the equation (neglecting terms of the order ε^3 and higher)

$$f_S(x^\star)\int_{-\infty}^{\infty}(\kappa_1\varepsilon+\kappa_2\varepsilon^2-y)f_{N^{(\varepsilon)}}(-y)\mathrm{d}y+\frac{f_S'(x^\star)}{2}\times\int_{-\infty}^{\infty}(\kappa_1\varepsilon+\kappa_2\varepsilon^2-y)^2f_{N^{(\varepsilon)}}(-y)\mathrm{d}y=0,$$

which reduces to

$$f_S(x^\star)(\kappa_1\varepsilon+\kappa_2\varepsilon^2)+\frac{f_S'(x^\star)}{2}(\kappa_1^2+1)\varepsilon^2=0$$

(where we used that the first moment of $N^{(\varepsilon)}$ is 0 and the second moment ε^2). As this holds for any ε small, we find that $\kappa_1 = 0$ and $\kappa_2 = -f_S'(x^\star)/(2f_S(x^\star))$. We conclude that a stochastic perturbation of the interarrival times of the order ε, leads to *just a change in the optimal schedule in the order of* ε^2.

The consequence of the above is that, regarding the optimal arrival time for the i-th customer, the scheme we identified is robust with respect to perturbations in the arrival process. We now briefly consider the impact on the customers arriving after customer i; for ease we focus on the case of quadratic loss. To this end, observe that the departure time of the i-th customer only changes (with respect to the non-perturbed situation) if she is late *and* the queue is empty when she arrives. As a consequence, the optimal arrival epoch of customer $i+1$ in the perturbed case should equal the arrival epoch in the non-perturbed case increased by a small positive quantity; this quantity is the product of the queue being empty upon the arrival of the i-th customer, multiplied by

$$\int_0^\infty y f_{N^{(\varepsilon)}}(y)\mathrm{d}y,$$

which is essentially linear in ε (for small ε). In the same way, the optimal arrival epoch of customer $i+2, i+3, \ldots$ can be determined, but the formulas do not offer any additional insight. In the situation described above, in which the departure of the i-th customer is delayed, the perturbed system matches again with the non-perturbed system as soon as the perturbed system idles.

5.3. Computational aspects

Our sequential approach requires the availability of a computational procedure to evaluate the customers' sojourn-time distributions. While such a computation is feasible e.g. for the case of (not necessarily identical) exponentially distributed service times, no explicit results are available for general service time distributions. To overcome this problem, a widely used approach is to approximate the service times by their *phase-type* counterparts, a class of distributions that allow fairly explicit computational procedures; see e.g. Asmussen (2003). More specifically, following an idea presented in e.g. Tijms (1986), we could replace each customer's service time distribution by a phase-type distribution with the same mean and variance; for the situation that this distribution corresponds to a coefficient of variation (defined as the ratio of the standard deviation and the mean) smaller than one, a mixture of Erlang distributions can be used, whereas if the coefficient of variation is larger than one, we fit a hyperexponential distribution. For the queue with (not necessarily evenly spaced) deterministic arrivals and such phase-type service times, algorithms can be devised to evaluate the sojourn-time distributions of the individual customers. These algorithms are of a recursive nature: the sojourn-time distribution of the i-th customer can be computed from that of the $(i-1)$-st customer. We refer to e.g. Kuiper, Kemper, and Mandjes (2014) for a systematic validation of this procedure, as well as related computational features.

In practical situations, additional requirements will be imposed on the schedule: lunch breaks should be included, all slots should be a multiple of Δ (for instance, 5) minutes, etc. Regarding the latter issue, a pragmatic option would be to round off all the t_is to a multiple of the granularity Δ. Another option would be to optimize the sequential objective function

$$\mathbb{E}\ell(S_{i-1} - t_i + t_{i-1}),$$

over $t_i \in \Delta\mathbb{N}$ (for known $t_{i-1} \in \Delta\mathbb{N}$).

6. Examples and numerical experiments

Above we presented a method to determine the optimal interarrival time given the sojourn time distribution of the previous jobs, for any given convex loss function. To illustrate this method, we discuss a set of examples. Although the method works for all service time distributions, we consider the exponential case for its attractive computational properties; extensions to phase-type service time distributions are feasible, as discussed in Section 5.3.

We first consider 'steady-state schedules': if all jobs stem from the same distribution, then the schedules prescribe that the customers should arrive equidistantly in time. We denote the risk per customer in the steady-state for loss function $\ell(\cdot)$ at interarrival time x by $R^{(\ell)}(x)$. We present closed-form optimal interarrival times for the various loss functions introduced above. Then we verify the legitimacy of the use of steady-state results, which is particularly relevant in case the number of jobs is relatively low. In the third example, we compare our sequential approach with the simultaneous optimization program (1). Finally, we study the impact of the weight α.

Example 6.1. In this example we consider the effect of scheduling policies $\mathcal{U}$ and $\mathcal{V}$ by considering the situation of i.i.d. service times, and the number of jobs n being large. Our goal is to compute the limiting interarrival time for both scheduling policies.

We assume that the service times are exponential with mean $1/\mu$, so that the queue under consideration is an D/M/1. Let x be the interarrival time between two subsequent jobs; it is evident (cf. Proposition 2.1) that we should have x larger than the average service requirement $1/\mu$. Then the distribution of the steady-state waiting time W is given through Asmussen (2003) and Tijms (1986).

$$\mathbb{P}(W > y) = \sigma_x e^{-\mu(1-\sigma_x)y}, \quad y > 0,$$

where $\sigma \equiv \sigma_x$ is the unique solution in $(0,1)$ of $e^{-\mu(1-\sigma)x} = \sigma$. By straightforward calculus, with B exponentially distributed with mean $1/\mu$, we obtain

$$G(y) := \mathbb{P}(W + B \leqslant y) = 1 - e^{-\mu(1-\sigma_x)y}, \quad y > 0.$$

(i) First consider the absolute value loss function and strategy $\mathcal{U}$. It follows directly that

$$G^{-1}\left(\frac{1}{2}\right) = \frac{\log 2}{\mu(1-\sigma_x)}.$$

We find for the optimal interarrival time $x^\star = G^{-1}(1/2)$

$$\sigma_{x^\star} = \frac{1}{2} \quad \text{and} \quad x^\star = \frac{2\log 2}{\mu}.$$

Note that in case of a weighted-linear loss function the optimal x solves

$$G^{-1}(1-\alpha) = \frac{-\log\alpha}{\mu(1-\sigma_x)},$$

yielding

$$\sigma_{x^\star} = \alpha, \quad x^\star = \frac{1}{\mu}\cdot\frac{-\log\alpha}{1-\alpha}, \quad \text{and} \quad R^{(u,\alpha)}(x^\star) = \frac{-\alpha\log\alpha}{\mu(1-\alpha)}.$$

For $\alpha \uparrow 1$, the optimal $x^\star$ converges to $1/\mu$. This results in a stable queue with large waiting times for the clients, due to the heavy weight imposed on idle times in the risk.

(ii) Let us now focus on the quadratic loss function and policy $\mathcal{V}$. It is easily verified that

$$\mathbb{E}W + \mathbb{E}B = \frac{\sigma_x}{\mu(1-\sigma_x)} + \frac{1}{\mu} = \frac{1}{\mu(1-\sigma_x)}.$$

Straightforward calculations now reveal that, with $x^\star$ being the optimal interarrival time,

$$\sigma_{x^\star} = \frac{1}{e}, \quad \text{and} \quad x^\star = \frac{1}{\mu}\cdot\frac{e}{e-1}.$$

As $e/(e-1) \approx 1.5820$ and $2\log 2 \approx 1.3863$, we conclude from the above that under the quadratic loss function the scheduling is somewhat more conservative than under the linear loss function.

Finally, we consider the weighted-quadratic loss function. The use of (17) and the constraint on σ_x yield the equation

$$\alpha(x-\mathbb{E}S) + (1-2\alpha)\left(\frac{-e^{-\mu(1-\sigma_x)x}}{\mu(1-\sigma_x)}\right) = \frac{-\alpha(1+\log\sigma_x) - (1-2\alpha)\sigma_x}{\mu(1-\sigma_x)} = 0,$$

which is equivalent to $\psi(\sigma_x) = 0$ with

$$\psi(\sigma) = \alpha(1+\log\sigma) + (1-2\alpha)\sigma, \quad 0 < \sigma \le 1, \tag{18}$$

strictly increasing, $\lim_{\sigma\downarrow 0}\psi(\sigma) = -\infty$, and $\psi(1) = 1-\alpha \geqslant 0$. It follows that (18) has a unique solution $\sigma_{x^\star} \in (0,1]$, and we get

$$x^\star = \frac{-\log\sigma_{x^\star}}{\mu(1-\sigma_{x^\star})} \quad \text{and} \quad R^{(v,\alpha)}(x^\star) = \frac{1}{\alpha}\left(\frac{\alpha+(1-2\alpha)\sigma_{x^\star}}{\mu(1-\sigma_{x^\star})}\right)^2.$$

The optimal interarrival times $x^\star$ for the various optimization schemes exhibit different sensitivities with respect to the weight α, as will be considered in detail in Example 6.4. ♢

Example 6.2. In this example we analyze the speed of convergence of the various scheduling schemes by considering the situation of i.i.d. service times exponentially distributed with mean 1, and the number of jobs n being relatively small. For each scheme we analyze the speed of convergence; that is, we investigate the difference between the sequentially optimized interarrival time and the asymptotic regime as studied above.

The schemes $\mathcal{U}$ and $\mathcal{V}$ analyzed in this example are based on the ordinary (i.e., $\alpha = 1/2$) linear and quadratic loss functions. The optimal interarrival times $x_1^\star, x_2^\star, \ldots$ are numerically determined, by first evaluating the distributions of the sojourn times S_i. This is facilitated by an algorithm proposed by Pegden and Rosenshine (1990) to find the distribution of the number of customers $N(t_i)$ in the system, just prior to the time of the i-th arrival; obviously, $N(t_i) \in \{0,\ldots,i-1\}$ and $i \in \{1,\ldots,n\}$. It requires an elementary verification to observe that

$$\mathbb{P}(N(t_i) = \ell \,|\, N(t_{i-1}) = k) = \begin{cases} e^{-x_i}\dfrac{x_i^{k+1-\ell}}{(k+1-\ell)!}, & \text{if } \ell \in \{1,\ldots,k+1\}, \\ \displaystyle\sum_{\ell'=k+1}^{\infty} e^{-x_i}\frac{x_i^{\ell'}}{\ell'!}, & \text{if } \ell = 0. \end{cases}$$

Evidently, the sojourn time of the i-th customer is the convolution of $N(t_i)+1$ exponential random variables, each of which has mean 1.

Based on our findings in the previous example, we expect the quadratic scheme $\mathcal{V}$ to be slightly more defensive for $\alpha = 1/2$. As can be seen in Fig. 1 the optimal values for $x^\star$ are increasing in the job number; that is, the first jobs are scheduled 'tighter' than the jobs later on in the schedule (which is due to the fact that the first customers are facing less uncertainty).

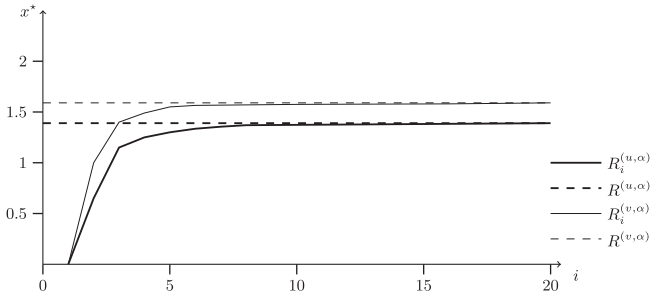

Fig. 1. Speed of convergence for the linear and quadratic schemes. The figure shows the scheduled interarrival times as a function of the customer number, as well as their limiting value; $\mu = 1$.

Table 1
The optimal interarrival times for the different job numbers in schemes $\mathcal{U}$ and $\mathcal{V}$; $\mu = 1$.

Scheme			$D\{x_\infty^\star - x_i^\star\}$ (%)
$\mathcal{U}$	$x_5^\star$	1.3245	4.10
	$x_{10}^\star$	1.3673	1.37
	$x_{20}^\star$	1.3814	0.36
	$x_\infty^\star$	1.3863	0.00
$\mathcal{V}$	$x_5^\star$	1.5438	2.41
	$x_{10}^\star$	1.5749	0.45
	$x_{20}^\star$	1.5813	0.05
	$x_\infty^\star$	1.5820	0.00

From Table 1 we conclude that the transient scheme converges rather fast to the stationary scheme. In our example, the relative difference between the optimal interarrival of a job and the steady-state interarrival, which we denote here by $D\{x_\infty^\star - x_i^\star\}$, is smaller than 5% for jobs scheduled after the 4-th arrival. Therefore, the use of the steady-state optimal interarrival times $x_\infty^\star$ for all jobs reduces the expected waiting time for the jobs early in the schedule (but at the expense of increasing the server's idle time). The example indicates that simple heuristics, in the spirit of "schedule the first five jobs at 95% of the steady-state interarrival time and the rest of the jobs at steady-state interarrival time", are close to the optimum and easily applicable for practitioners. ♢

Example 6.3. We now compare the output of our sequential approach with that of the simultaneous program (1). In the latter approach one obtains the clients' optimal arrival times through the *simultaneous* optimization

$$\min_{t_1,\ldots,t_n} \sum_{i=1}^{n} \mathbb{E}(W_i - I_i)^2. \tag{19}$$

As mentioned in the introduction, this simultaneous approach is numerically typically harder than our sequential counterpart, the most substantial advantage of the latter scheme being that only single-dimensional optimizations need to be performed. For the special case of exponential service times, the objective function in (19) can be evaluated once we have a procedure to compute the number of customers present at t_1 up to t_n (due to the memoryless property), and this can be done by an algorithm developed in Wang (1999). Below we compare the simultaneous and sequential scheme, in terms of both their steady-state and transient properties.

Steady-state. Informally, in case of a quadratic loss function and for large n, to compute the steady-state interarrival time for the simultaneous approach, we are to evaluate

$$\min_{x_1,\ldots,x_{n-1}} \sum_{i=1}^{n} \mathbb{E}(W_i - I_i)^2 \approx n \cdot \min_x \mathbb{E}(S(x) - x)^2,$$

where the random variable $S(x)$ corresponds to a steady-state sojourn time in a D/G/1 queue with interarrival time x. It is seen that the optimal steady-state interarrival time, say $x^\star_{\text{sim}}$, follows from the first order condition

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

Given that $\mathbb{E}S(x) = 1/(\mu(1-\sigma_x)), \mathbb{E}S^2(x) = 2/(\mu(1-\sigma_x))^2$, and that σ_x is the unique solution in $(0,1)$ of $e^{-\mu(1-\sigma_x)x} = \sigma_x$, the first order condition yields the equation

$$\frac{2}{\mu(1-\sigma_x)}\left[\frac{2\sigma_x + \sigma_x \log \sigma_x}{\sigma_x - 1 - \sigma_x \log \sigma_x} - \log \sigma_x - 1\right] = 0,$$

or $\quad \sigma_x + (1 + \log \sigma_x)(1 + \sigma_x \log \sigma_x) = 0.$

In case $\mu = 1$ we numerically find $x^\star_{\text{sim}} = 1.847$. For the sequential approach we found that the steady-state optimal interarrival time equals $x^\star_{\text{seq}} = 1.582$. Consequently, the simultaneous approach yields longer optimal interarrival times (hence in the clients' favor, and disadvantageous to the server).

Transient. In case the number of jobs n is relatively small, we are able to numerically analyze the optimal transient interarrival times relying on Wang's algorithm Wang (1999). Wang's algorithm enables us to find the distribution of the number of customers in the system at the arrival times t_1 up to t_n, and therefore to evaluate the objective function for a given $t_1, \ldots, t_n$. Then a numerical minimization procedure is used to determine the optimal transient interarrival times.

Our findings are depicted in Fig. 2, together with the steady-state result as well as the results of the sequential approach from the previous example. We observe that all jobs, except for the last one, are scheduled less tight with the simultaneous approach than with the sequential approach. Furthermore, for the sequential approach the optimal interarrival times are increasing and converge towards the steady-state optimal interarrival time $x^\star_{\text{seq}}$, whereas for the simultaneous approach the optimal interarrival times are increasing in the first arrivals and decreasing in the last arrivals, being close to $x^\star_{\text{sim}}$ in the middle part. ♢

Example 6.4. In the last experiment we study the impact of the weight parameter α in the weighted loss function $\alpha\mathbb{E}I_i^\gamma + (1-\alpha)\mathbb{E}W_i^\gamma$, where $\gamma = 1$ corresponds to the linear case, and $\gamma = 2$ to the quadratic case. The weight parameter $\alpha \in (0,1)$ is a tuning parameter when balancing the interests of the service provider: for α close to 0, waiting times should be prevented, leading to relatively high values of $x^\star$, while for α close to 1 the schedule is such that idle times are prevented from happening, leading to relatively low values of $x^\star$.

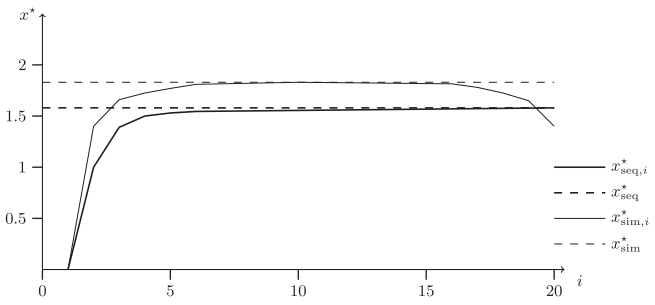

Fig. 2. The optimal interarrival times for the sequential and simultaneous approach in case of a quadratic loss function; $\mu = 1$.

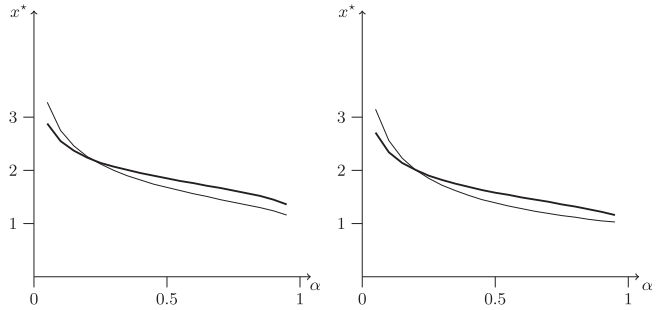

Fig. 3. The stationary optimal interarrival times for the simultaneous (left) and sequential approach (right) for linear (thin lines) and quadratic (thick lines) loss functions, as a function of α; $\mu = 1$.

Fig. 3 numerically assesses how α affects the schedule, for the situation of identically distributed exponential service times. Restricting ourselves to just the steady-state schedule, we compute the optimal interarrival times $x^\star$ for both linear and quadratic loss, and for both the sequential and the simultaneous objective function. ♢.

7. Conclusion and outlook

In appointment scheduling, rules are needed that assure a good trade-off between quality (in terms of the customer's waiting time) and cost (in terms of the server's idle time). In this paper we presented a technique to generate such rules.

More specifically, these rules are based on an approach that sequentially minimizes risks and which can be used to determine a schedule, for any convex loss function and service time distribution. In this framework, one should schedule jobs in the order of increasing variances, for convex loss functions with scale families of service time distributions. Also, the scheduling rules presented here can be extended to cover real-life phenomena such as no-shows, urgent arrivals, and the effect of small perturbations of the arrival epochs.

We demonstrated the approach by four representative examples. In the first we considered a system with a large number of customers, so that the system can be effectively replaced by its steady-state version. In case of exponential service times there are closed-form expressions for the steady-state schedule, whereas the transient schedule can be determined relatively easily relying on basic standard mathematical software; we do so for (possibly weighted) linear and quadratic loss functions. The numerical output illustrates the impact of the choice of the loss function on the interarrival times. In the second example we show how fast the transient schedule converges to the steady-state schedule. Numerical experiments indicate that simple heuristics perform well. In the third example we compare our approach with the joint approach that was described in the introduction. The last example numerically assess the impact of the weight parameter α.

The methodology presented in this paper can be used across a broad range of application areas, such as health care, manufacturing, and other service systems, in situations that primarily focus on setting up an appointment schedule. In addition, it may shed light on *provisioning* issues (that is, decide how many jobs can be scheduled per server per working day, or, similarly, how many servers should be allocated on a working day to process the scheduled appointments). It is clear that our approach can be used to check whether it is realistic to schedule a given number of n appointments on a working day (based on knowledge of the service time distribution of each job), and in this way it can be used to support provisioning and staffing decisions.

A next step in this branch of research would be a study in more detail of both the sequential and the simultaneous approach for a D/G/1 system and various loss functions. These issues, as well as the extension to more complex queueing networks, such as tandem or parallel queues, are topics for future research.

Acknowledgments

We thank Mark Boersma (University of Amsterdam), Rhonda Righter (University of California, Berkeley), and Gideon Weiss (University of Haifa) for useful comments, and Wouter Vink and Alex Kuiper (IBIS UvA, University of Amsterdam) for their valuable research assistance. The second author would like to thank the Court of Justice at Maastricht for suggesting the problem as a consultation project for IBIS UvA, and its judges for useful discussions.

Appendix A. Waiting time

A.1. Proof of Proposition 2.1

Let A_j be the j-th customer's random interarrival time. Our analysis relies on the Spitzer–Baxter identities Asmussen (2003, pp. 229–232), see also Feller (1971, chap. 7 and chap. 18). In view of Asmussen (2003, Prop. 4.5), we have to study, with $x^+ := \max\{x, 0\}$,

$$\begin{aligned}\frac{\mathbb{E}W_n}{\sqrt{n}} &= \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{1}{k}\mathbb{E}\left(\left(\sum_{j=1}^{k}(B_j - A_j)\right)^+\right)\\ &= \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{1}{k}\int_0^\infty \mathbb{P}\left(\left(\sum_{j=1}^{k}(B_j - A_j)\right)^+ > y\right)\mathrm{d}y\\ &= \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{1}{k}\left(\int_0^\infty \mathbb{P}\left(\sum_{j=1}^{k}(B_j - A_j) > y\sqrt{k}\sigma\right)\mathrm{d}y\right)\sigma\sqrt{k}\\ &= \frac{\sigma}{\sqrt{n}}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k}}\mathcal{I}_k,\end{aligned}$$

where

$$\mathcal{I}_k := \int_0^\infty \mathbb{P}\left(\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\frac{B_j - A_j}{\sigma} > y\right)\mathrm{d}y.$$

By Chebyshev's inequality the integrand is bounded, as follows:

$$\mathbb{P}\left(\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\frac{B_j - A_j}{\sigma} > y\right) \leqslant \min\left\{1, \frac{1}{y^2}\,\mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\frac{B_j - A_j}{\sigma}\right)\right\}.$$

Therefore, we have

$$\mathcal{I}_k \leqslant \int_0^\infty (1 \wedge \frac{1}{y^2})\mathrm{d}y = \int_0^1 \mathrm{d}y + \int_1^\infty \frac{1}{y^2}\mathrm{d}y = 2.$$

By dominated convergence and the central limit theorem, this yields

$$\mathcal{I}_k \underset{k\to\infty}{\to} \int_0^\infty (1 - \Phi(y))\mathrm{d}y = \frac{1}{\sqrt{2\pi}}.$$

Subsequently, note

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k}}\mathcal{I}_k = \int_0^1 \frac{\mathcal{I}_{\lceil nx\rceil}}{\sqrt{\lceil nx\rceil/n}}\mathbf{1}_{\left[x\leqslant 1-n^{-1}\right]}\mathrm{d}x.$$

The integrand is bounded by $2x^{-1/2}$ with $\int_0^1 2x^{-1/2}\mathrm{d}x = 4$. Consequently, dominated convergence yields

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k}}\mathcal{I}_k \underset{n\to\infty}{\to} \int_0^1 \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{x}}\mathrm{d}x = \sqrt{\frac{2}{\pi}}.$$

It thus follows that the claim of Proposition 2.1 holds. □

Appendix B. Convex loss functions

Lemma B.1. *Let $\ell(\cdot)$ be a nonnegative convex function on $\mathbb{R}$ with $\ell(0) = 0$. Then $\ell(\cdot)$ is a loss function, i.e., it is nonincreasing on $(-\infty, 0]$ and nondecreasing on $[0, \infty)$ with $\ell(0) = 0$. Furthermore, it is absolutely continuous with nondecreasing derivative $\ell'(\cdot)$. Let S be a random variable and let $\mathbb{E}\ell(S - x)$ be finite for all $x \in \mathbb{R}$. Then $\mathbb{E}|\ell'(S - x)|$ is finite for all $x \in \mathbb{R}$, and $\mathbb{E}\ell(S - x)$ is a convex function of x, for which*

$$\inf_{x\in\mathbb{R}} \mathbb{E}\ell(S - x) = \lim_{x\to-\infty} \mathbb{E}\ell(S - x)$$

holds, or

$$\inf_{x\in\mathbb{R}} \mathbb{E}\ell(S - x) = \lim_{x\to\infty} \mathbb{E}\ell(S - x)$$

holds, or for which the infimum is attained at a value $x^\star$ at which $\mathbb{E}\ell'(S - \cdot)$ changes sign or equals 0.

B.1. Proof of Lemma B.1

The monotonicity of $\ell'(\cdot)$ and the nonnegativity of $\ell(\cdot)$ imply for all $a \leqslant b$

$$\int_a^b |\ell'(y)|\mathrm{d}y \leqslant \ell(b) + \ell(a).$$

Consequently, by 'Fubini',

$$\int_a^b \mathbb{E}|\ell'(S - x)|\mathrm{d}x = \mathbb{E}\int_a^b |\ell'(S - x)|\mathrm{d}x \leqslant \mathbb{E}\ell(S - b) + \mathbb{E}\ell(S - a) < \infty$$

and hence

$$\int_a^b (-\mathbb{E}\ell'(S - x))\mathrm{d}x = \mathbb{E}\ell(S - b) - \mathbb{E}\ell(S - a)$$

holds. Hence, $\mathbb{E}\ell(S - x)$ is absolutely continuous with derivative $-\mathbb{E}\ell'(S - x)$ and therefore convex. The lemma follows. □

Appendix C. Sequential ordering

C.1. Proof of Theorem 4.2

In view of (6)–(14) we have

$$R_i^{(\ell)}(t_1, \ldots, t_i) = \mathbb{E}\ell(W_i - I_i) = \mathbb{E}\ell(W_{i-1} + \sigma_{i-1}B_1 - t_i + t_{i-1}).$$

Consequently, it suffices to show that for the waiting time W_i and the service time random variable B_1

$$\psi(\sigma) := \inf_{x\in\mathbb{R}} \mathbb{E}\ell(W_i + \sigma B_1 - x) \tag{20}$$

is nondecreasing in $\sigma > 0$. We may write

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\sigma}\mathbb{E}\ell(W_i + \sigma B_1 - x) &= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\mathbb{E}(\ell(W_i + (\sigma + \epsilon)B_1 - x)\\ &\quad -\ell(W_i + \sigma B_1 - x)) = \mathbb{E}(\ell'(W_i + \sigma B_1 - x)B_1)\\ &\quad +\lim_{\epsilon\to 0}\frac{1}{\epsilon}\mathbb{E}\left(\int_0^\epsilon [\ell'(W_i + (\sigma + \eta)B_1 - x)\right.\\ &\quad \left.-\ell'(W_i + \sigma B_1 - x)]\mathrm{d}\eta\, B_1\right)\end{aligned} \tag{21}$$

and we may note that the integrand at the right hand side is bounded in absolute value by $|\ell'(W_i + (\sigma + \epsilon)B_1 - x) - \ell'(W_i + \sigma B_1 - x)|$, since $\ell'(\cdot)$ is nondecreasing. It follows that the square of the expression after the limit sign at the right hand side of (21) is bounded (by virtue of 'Cauchy–Schwarz') by

$$\mathbb{E}|\ell'(W_i+(\sigma+\epsilon)B_1-x)-\ell'(W_i+\sigma B_1-x)|^2\,\mathbb{E}B_1^2.$$

If $\ell'(\cdot)$ is continuous, we may conclude by dominated convergence that the limit at the right hand side of (21) vanishes, and hence that

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}\mathbb{E}\ell(W_i+\sigma B_1-x)=\mathbb{E}(\ell'(W_i+\sigma B_1-x)B_1) \tag{22}$$

holds. Since $\ell'(\cdot)$ has at most countably many discontinuities, we may conclude (22) also if $\ell'(\cdot)$ is not continuous, but B_1 has a density with respect to Lebesgue measure.

Fix σ_0, and choose $x_0^\star$ according to Lemma B.1 such that it satisfies

$$\psi(\sigma_0)=\mathbb{E}\ell(W_i+\sigma_0B_1-x_0^\star),\ \mathbb{E}\ell'(W_i+\sigma_0B_1-x_0^\star)=0. \tag{23}$$

First we will consider the case that $\ell(\cdot)$ is *strictly* convex. Then (22), (23), and Chebyshev's strict inequality from Lemma 4.1 yield

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\sigma}\mathbb{E}\ell(W_i+\sigma B_1-x_0^\star)\big|_{\sigma=\sigma_0}&=\mathbb{E}\big(\ell'(W_i+\sigma_0B_1-x_0^\star)B_1\big)\\&=\mathbb{E}\big(\ell'(W_i+\sigma_0B_1-x_0^\star)[B_1-\mathbb{E}B_1]\big)\\&=\mathbb{E}\big(\mathbb{E}\big(\ell'(W_i+\sigma_0B_1-x_0^\star)[B_1-\mathbb{E}B_1]\,\big|\,W_i\big)\big)\\&=\mathbb{E}\big(\mathbb{C}\mathrm{ov}\big(\ell'(W_i+\sigma_0B_1-x_0^\star),B_1\,\big|\,W_i\big)\big)>0,\end{aligned} \tag{24}$$

since $\ell'(\cdot)$ is strictly increasing and B_1 is non-degenerate. It follows that there exists a $\sigma_1<\sigma_0$ such that for all $\sigma\in[\sigma_1,\sigma_0)$ the strict inequality

$$\mathbb{E}\ell(W_i+\sigma B_1-x_0^\star)<\mathbb{E}\ell(W_i+\sigma_0B_1-x_0^\star)$$

holds, which implies

$$\psi(\sigma)<\psi(\sigma_0),\quad \sigma_1\leqslant\sigma<\sigma_0. \tag{25}$$

Furthermore, for all $\sigma_1>0$ and $\sigma_2>0$ there exist $x_1^\star$ and $x_2^\star$ by Lemma B.1, such that by the convexity of $\ell(\cdot)$

$$\begin{aligned}\frac{1}{2}[\psi(\sigma_1)+\psi(\sigma_2)]&=\frac{1}{2}[\mathbb{E}\ell(W_i+\sigma_1B_1-x_1^\star)+\mathbb{E}\ell(W_i+\sigma_2B_1-x_2^\star)]\\&\geqslant\mathbb{E}\ell\left(W_i+\frac{1}{2}(\sigma_1+\sigma_2)B_1-\frac{1}{2}(x_1^\star+x_2^\star)\right)\\&\geqslant\psi\left(\frac{1}{2}(\sigma_1+\sigma_2)\right)\end{aligned} \tag{26}$$

holds, which means that $\psi(\cdot)$ is convex. Consequently, $\psi(\cdot)$ is continuous, which together with (25) proves that $\psi(\cdot)$ is non-decreasing, as may be seen as follows.

Assume $\psi(\cdot)$ would not be non-decreasing. Then there would exist σ_3 and σ_4, $\sigma_3<\sigma_4$, with $\psi(\sigma_3)>\psi(\sigma_4)$. Since $\psi(\cdot)$ is continuous the infimum of it on $[\sigma_3,\sigma_4]$ is attained at σ_0, say. Note $\sigma_3<\sigma_0$ and $\psi(\sigma_3)>\psi(\sigma_0)$. According to (25) there exists a $\sigma_1<\sigma_0$ with $\psi(\sigma)<\psi(\sigma_0)$ for $\sigma_1\vee\sigma_3\leqslant\sigma<\sigma_0$, which is in contradiction with

$$\inf_{\sigma_3\leqslant\sigma\leqslant\sigma_4}\psi(\sigma)=\psi(\sigma_0).$$

Having proved the monotonicity of $\psi(\cdot)$ for strictly convex loss functions, we now consider the case of a general convex loss function $\ell(\cdot)$ that satisfies the conditions of the theorem. For $\epsilon>0$ we define $\ell_\epsilon(x)=\ell(x)+\epsilon x^2$, $x\in\mathbb{R}$. Since W_i is bounded by $B_1+\cdots+B_{i-1}$ and $\mathbb{E}B_1^2$ is finite, the conditions of the theorem are fulfilled for this *strictly* convex loss function $\ell_\epsilon(\cdot)$ as well. Consequently the corresponding function $\psi_\epsilon(\cdot)$ is non-decreasing. Choose $\sigma_5<\sigma_6$. The definition of $\psi_\epsilon(\cdot)$ and its monotonicity yield

$$\psi(\sigma_5)\leqslant\psi_\epsilon(\sigma_5)\leqslant\psi_\epsilon(\sigma_6). \tag{27}$$

Let $x_6^\star$ satisfy $\psi(\sigma_6)=\mathbb{E}\ell(W_i+\sigma_6B_1-x_6^\star)$, and note

$$\begin{aligned}\limsup_{\epsilon\downarrow0}\psi_\epsilon(\sigma_6)&\leqslant\limsup_{\epsilon\downarrow0}\mathbb{E}\ell_\epsilon(W_i+\sigma_6B_1-x_6^\star)\\&=\mathbb{E}\ell(W_i+\sigma_6B_1-x_6^\star)=\psi(\sigma_6).\end{aligned} \tag{28}$$

Together, (27) and (28) prove that $\psi(\cdot)$ is non-decreasing.

References

Araman, V., & Glynn, P. (2012). Fractional Brownian motion with $H<1/2$ as a limit of scheduled traffic. *Journal of Applied Probability, 49*, 710–718.

Asmussen, S. (2003). *Applied probability and queues. Applications of mathematics* (2nd ed., Vol. 51). New York, NY, USA: Springer-Verlag. Stochastic modelling and applied probability.

Bailey, N. T. J. (1952). A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society. Series B (Methodological), 14*(2), 185–199.

Bailey, N. T. J. (1954). Queueing for medical care. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 3*(3), 137–145.

Begen, M., & Queyranne, M. (2011). Appointment scheduling with discrete random durations. *Mathematics of Operations Research, 36*, 240–257.

Bickel, P. J., & Doksum, K. A. (2001). *Mathematical statistics* (Vol. I). Upper Saddle River, NJ, USA: Prentice Hall.

Brahimi, M., & Worthington, D. J. (1991). Queueing models for out-patient appointment systems – A case study. *The Journal of the Operational Research Society, 42*(9), 733–746.

Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: a review of literature. *Production and Operations Management, 12*(4), 519–549.

Denton, B., & Gupta, D. (2003). A sequential bounding approach for optimal appointment scheduling. *IIE Transactions, 35*, 1003–1016.

Feller, W. (1971). *An introduction to probability theory and its applications* (2nd ed., Vol. II). New York, NY, USA: John Wiley & Sons.

Ferguson, T. S. (1967). *Mathematical statistics: A decision theoretic approach*. New York, NY, USA: Academic Press.

Fries, B. E., & Marathe, V. P. (1981). Determination of optimal variable-sized multiple-block appointment systems. *Operations Research, 29*(2), 324–345.

Hardy, G. H., Littlewood, J. E., & Pólya, G. (1934). *Inequalities* (2nd ed.). Cambridge, UK: Cambridge University Press.

Hassin, R., & Mendel, S. (2008). Scheduling arrivals to queues: A single-server model with no-shows. *Management Science, 54*(3), 565–572.

Ho, C.-J., & Lau, H.-S. (1992). Minimizing total cost in scheduling outpatient appointments. *Management Science, 38*(12), 1750–1764.

Hopp, W. J., & Spearman, M. L. (1995). *Factory physics* (3rd ed.). New York, NY, USA: McGraw-Hill/Irwin.

Jogdeo, K. (1977). Association and probability inequalities. *The Annals of Statistics, 5*(3), 495–504.

Kaandorp, G. C., & Koole, G. (2007). Optimal outpatient appointment scheduling. *Health Care Management Science, 10*(3), 217–229.

Klassen, K. J., & Rohleder, T. R. (1996). Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management, 14*(2), 83–101.

Klassen, K. J., & Rohleder, T. R. (2004). Outpatient appointment scheduling with urgent clients in a dynamic, multi-period environment. *International Journal of Service Industry Management, 15*(2), 167–186.

Kuiper, A., Kemper, B., & Mandjes, M. (2014). A computational approach to optimized appointment scheduling. *Queueing Systems*, http://dx.doi.org/10.1007/s11134-014-9398-6.

Lehaney, B., Clarke, S. A., & Paul, R. J. (1999). A case of an intervention in an outpatients department. *The Journal of the Operational Research Society, 50*(9), 877–891.

Lindley, D. V. (1952). The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society, 48*(2), 277–289.

Mitrinović, D. S., & Vasić, P. M. (1974). History, variations and generalisations of the Čebyšev inequality and the question of some priorities. *Publications of the Faculty of Electrical Engineering of the University of Belgrade, Series Mathematics and Physics, 461–497*, 1–30.

Mondschein, S. V., & Weintraub, G. Y. (2003). Appointment policies in service operations: A critical analysis of the economic framework. *Production and Operations Managements, 12*(2), 266–286.

Patrick, J., Puterman, M. L., & Queyranne, M. (2008). Dynamic multipriority patient scheduling for a diagnostic resource. *Operations Research, 56*(6), 1507–1525.

Pegden, C. D., & Rosenshine, M. (1990). Scheduling arrivals to queues. *Computers & Operations Research, 17*(4), 343–348.

Pinedo, M. L. (2001). *Scheduling: Theory, algorithms, and systems*. New York, NY: Prentice Hall.

Rising, E. J., Baron, R., & Averill, B. (1973). A systems analysis of a university-health-service outpatient clinic. *Operations Research, 21*(5), 1030–1047.

Roberts, J., Mocci, U., & Virtamo, J. (Eds.). (1996). *Broadband network teletraffic – Performance evaluation and design of broadband multiservice networks: Final report of action COST 242. Lecture notes in computer science* (Vol. 1155). Springer.

Robinson, L., & Chen, R. (2003). Scheduling doctors' appointments: Optimal and empirically-based heuristic policies. *IEE Transactions, 35*, 295–307.

Rohleder, T. R., & Klassen, K. J. (2000). Using client-variance information to improve dynamic appointment scheduling performance. *Omega, 28*(3), 293–302.

Swisher, J. R., Jacobsen, S. H., Jun, J. B., & Balci, O. (2001). Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. *Computers & Operations Research, 28*(2), 105–125.

Tijms, H. (1986). *Stochastic modelling and analysis – A computational approach. Wiley series in probability and mathematical statistics: Applied probability and statistics*. Chichester, UK: John Wiley & Sons.

Wang, P. P. (1999). Sequencing and scheduling n customers for a stochastic server. *European Journal of Operational Research, 119*(3), 729–738.
Welch, J. D. (1964). Appointment systems in hospital outpatient departments. *Operational Research Quarterly, 15*(3), 224–232.
Welch, J. D., & Bailey, N. T. J. (1952). Appointment systems in hospital outpatient departments. *The Lancet, 259*(6718), 1105–1108.
Whitt, W. (1972). Complements to heavy-traffic limit theorems for the GI/G/1 queue. *Journal of Applied Probability, 9*, 185–191.