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