Article Title: Appointment Systems under Service Level Constraints

February 12, 2017

Abstract

In this paper, we tackle the classical appointment scheduling problem from a completely new angle. We study an appointment system where a finite number of customers are scheduled to arrive in such a way that (1) the expected waiting time of each individual customer cannot exceed a given threshold, and (2) the appointment times are set as early as possible (without breaking the waiting time constraint). Using a transient queueing analysis approach, we analytically characterize the structure of the optimal appointment schedule and prove the limiting behavior of our system. Compared with the literature, our paper brings unique features in both modeling perspectives and analysis methods. We discuss in detail these new features in the following subsections.

1. Modeling Perspective

The fundamental principle of appointment scheduling is on the balance between servers’ idling (when appointments are scheduled far from each other) and customers’ waiting (when appointments are scheduled close to each other). For decades, appointment scheduling has drawn significant attention in the queueing, optimization, operations management, and health care research communities. As pointed out by Cayirli and Veral (2003), the overwhelming majority of the studies assign unit costs (weights) to servers’ idling and customers’ waiting and then search appointment schedules that minimize the expected total system cost which is a linear combination of servers’ idling time and customers’ waiting time.

Despite the fruitful results available in academia, the implementation or guidance of appointment scheduling in practice is still very limited. Many service firms are still using simple rules of thumb. The authors have discussed with practitioners in different service industries and found out, among others, four main concerns that obstruct the application of results from academic literature to industry.

First, as the optimal appointment schedules are found through minimizing the sum of servers’ idling cost and customers’ waiting cost, it is obviously true that the resulting schedules depend critically on the relative costs of servers’ idling and customers’ waiting. Therefore, obtaining accurate cost parameters becomes a crucial issue in the application of theoretical results. However, it is difficult to estimate consumers’ waiting cost. Fries and Marathe (1981) relate the difficulty in estimating waiting cost to the connection between customers’ waiting and the issues of goodwill as well as the cost of society. We also notice that long waiting times would lead to reneging and negative word of mouth, which further complicates the estimation of the cost.

Second, most of the literature models customers’ waiting cost as a linear function of waiting time. However, in reality, the magnitude of customers’ annoyance from waiting may not be proportional to the length of waiting time. From recent empirical studies (see, e.g., Baron et al. 2016), in various service encounters, customers’ perception of waiting reveals a threshold type behavior in time. In many service industries, this acceptance threshold can be obtained from customer satisfaction or complaint surveys (see, e.g., Baron et al. 2016). The firms usually consider the acceptance thresholds as their performance targets.

Third, and very importantly, a schedule that minimizes the total system cost may not lead to equal waiting experiences for each individual customer. Hassin and Mendel (2008)
provide numerical results showing that for both the dome-shape system (appointment intervals initially increase and then decrease) and the equal-space system (appointment intervals stay fixed), customers who are scheduled to come later wait longer than those who are scheduled to come earlier. Cayirli and Veral (2003) highlight that the increasing waiting trend is observed under most commonly studied appointment systems. The inequity in waiting time among customers certainly leads to fairness issues, which would clearly create problems in practice.

Fourth, besides the usual concept of waiting time which describes the duration from the time when a customer arrives and joins the queue to the time when she starts her service, there is another important measure which captures the duration from the time a customer requests service to the time when she arrives (i.e., her appointment time). This can be viewed as the indirect waiting time. However indirect waiting time is often ignored in the literature.

Unlike a traditional appointment system that minimizes the sum of servers’ idling cost and customers’ waiting cost, in this paper, we study an appointment system under a specific service level constraint, that is, the expected waiting time of each individual customer must be less than a certain value. Customers are then given the earliest possible appointment times without breaking the service level constraint. Our model resolves the above four concerns simultaneously. (1) Our model only deals with time, and cost is never involved. Thus, our results can be applied in practice without any cost estimation. This resolves the first concern. (2) A unique feature of our model is the service level constraint which gives the upper limit of the expected waiting time of each individual customer. As a result, our appointment schedule ensures fairness among customers. None of the customers wait longer than the acceptance threshold in expectation. This resolves the second and third concerns. (3) Since each individual customer is scheduled to arrive as early as possible, her indirect waiting time is minimized. This resolves the fourth concern. In addition to these, our model has many other advantages. (4) When customers are given the earliest possible appointment times, the servers’ idling time and overtime are automatically minimized (without breaking the service level constraint). (5) Our model can be viewed as both prospective scheduling (while the appointment times of all the customers are decided together at once) and sequential scheduling (while the appointment time of each customer is set one after another at the time when service is requested). The interpretation of our problem in the prospective scheduling setting is to find the earliest possible appointment times for all the customers such that the service level constraint is fulfilled, while the interpretation in the sequential scheduling setting is, given that all the previous inter-appointment times are minimized while keeping the service level constraint valid, we need to find the shortest inter-appointment time for the next customer such that the service level constraint is still valid.

(2) Analysis Method In the past few years, there has been a growing body of literature on appointment scheduling from the optimization community. The studies there mainly focus on applying optimization techniques (e.g. robust optimization) to develop computationally tractable programming models (or approximations) for searching the optimal appointment schedules. On the other hand, appointment scheduling has received relatively less recent attention in the queueing community. This is, in part, due to the nature of appointment systems that (1) there is only a finite number of arrivals; and (2) the inter-arrival times between customers may not be equal. These features create difficulties in applying standard queueing methodology which relies on steady state analysis (and therefore assumes infinite arrivals) and requires homogeneous inter-arrival times. As a matter of fact, very few analytical results exist on the structural properties of optimal schedules.

In this paper, we take the queueing approach to explore the structure of optimal appointment schedules. We study a system with a single server and a finite number of customers to schedule. Customer service times are i.i.d. and follow an exponential distribution. Taking the advantage of the Markov property of the exponential distribution, each consumer’s expected waiting time can be represented as a linear function of the probabilities for him to see different numbers of patients in the system. We apply the theory of majorization to explore the majorization relationship between two probability vectors of seeing at least n consumers in the system for each two consecutive consumers m and m + 1. Based on the majorization relationship, we prove the optimal schedule intervals in our problem set up
show a nondecreasing pattern. That is, the inter-appointment time between the \( m^{th} \) and \((m + 1)^{th}\) arrivals is no less than the inter-appointment time between the \((m - 1)^{th}\) and \(m^{th}\) customers. We further show our system converges asymptotically to a D/M/1 queueing system as the number of arrivals approaches infinity. To the best of our knowledge, this paper is the first one to provide the analytical proof of the optimal schedule pattern in the appointment scheduling literature. Although this increasing pattern is observed numerically in some literatures from this modelling perspective, sequential scheduling with service level constraints, there is no existing literature on the theoretical proof of this pattern.

We further extend our method to prove the same pattern holds in the service system with single server but i.i.d Erlang distributed consumer service time. Worth to mention that Erlang service time distribution, which still holds the Markov property in someway that helps mathematical tractability, is a tremendously relaxed assumption. We also discuss the extension of our results to systems with multiple servers.

References


