# Appointment-driven queueing systems with non-punctual customers

Oualid Jouini[1] • Saif Benjaafar[2]• Bingnan Lu[2] • Siqiao Li[3,4]• Benjamin Legros[5]

[1] *Université Paris-Saclay, CentraleSupélec, Laboratoire Genie Industriel, 3 rue Joliot-Curie 91190 Gif-sur-Yvette, France*

[2] *Department of Industrial and Systems Engineering, University of Minnesota, 100 Union Street SE, Minneapolis, MN 55455, USA*

[3] *Industrial Engineering Department, Shanghai Jiaotong University, Dongchun Road 800, Shanghai, China*

[4] *Mathematics Department, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1183 HV Amsterdam, Netherlands*

[5] *EM Normandie, Laboratoire Métis, 64 Rue du Ranelagh, 75016 Paris, France*

oualid.jouini@centralesupelec.fr • saif@umn.edu • luxx0389@umn.edu • l.s.q.li@vu.nl •

benjamin.legros@centraliens.net

**Abstract**

We consider a single server queueing system where a finite number of customers arrive over time to receive service. Arrivals are driven by appointments, with a scheduled appointment time associated with each customer. However, customers are not necessarily punctual and may arrive either earlier or later than their scheduled appointment times or may not show up at all. Arrival times relative to scheduled appointments are random. Customers are not homogeneous in their punctuality and show up behavior. The time between consecutive appointments is allowed to vary from customer to customer. Moreover, service times are assumed to be random with a $\gamma-$Cox distribution, a class of phase-type distributions known to be dense in the field of positive distributions. We develop both exact and approximate approaches for characterizing the distribution of the number of customers seen by each arrival. We show how this can be used to obtain the distribution of waiting time for each customer. We prove that the approximation provides an upper bound for the expected customer waiting time when non-punctuality is uniformly-distributed. We also examine the impact of non-punctuality on system performance. In particular, we prove that non-punctuality deteriorates waiting time performance regardless of the distribution of non-punctuality. In addition, we illustrate how our approach can be used to support individualised appointment scheduling.

**Keywords:** appointment-driven arrivals, finite arrivals, customer punctuality, customer no-shows

**Mathematics Subject Classification (2010):** 90B22, 68M20, 60K25

# 1 Introduction

There are numerous service systems where the arrivals of customers are driven by scheduled appointments. Examples include arrivals to healthcare facilities, government agencies (e.g., social services), the offices of tax and financial service providers, academic advising offices at universities, restaurants and wellness centers, just to name a few[1]. Despite this prevalence, analytical tools for the performance evaluation of these systems are relatively limited. Existing approaches from queueing theory mainly rely on steady state analysis of queueing systems with exogenous arrival processes. There are several important differences between queueing systems with exogenous arrival processes and systems with appointment-driven arrivals (ADA). In particular, systems with ADA are characterised by (1) a finite number of customers (e.g., the set of customers who have been scheduled at a clinic in a given day), so that steady-state analysis cannot be applied[2], (2) arrivals that are in part determined by known scheduled appointment times, (3) appointment times that may not be equally spaced, and (4) the possibility of customer non-punctuality and no-shows. These differences can be further compounded in settings in which customers are heterogeneous in their behaviors in terms of punctuality and likelihood of show up.

In this paper, we consider such a system. In particular, we consider a system with a finite number of customers and a single server, where each customer has a scheduled appointment but customers are not necessarily punctual and may arrive earlier or later than their appointment times. We allow for appointments to be arbitrarily spaced so that they are not necessarily equally spaced. To achieve a better understanding of how non-punctuality and no-shows affect the system, we allow for non-punctuality, a random variable with a general distribution, and the probability of show up to be customer-specific. The service times are assumed to be random with a $\gamma-$Cox distribution, a class of phase-type distributions known to be dense in the field of positive distributions [23]. (For ease of presentation, we first consider the case of the exponential distribution, a special case of a $\gamma-$Cox distribution.)

We develop an exact analytical approach that allows us to compute the distribution of waiting time for the customer with the $n$-th appointment from which various moments can be readily computed. The approach hinges on a recursive relationship between the conditional probability $p_{n,i}$ of customer $n$ finding, upon arrival (if the customer were to show up), $i$ customers already in

[1] With the advent of social distancing, it is likely that appointment driven arrivals will become even more prevalent.
[2] Steady state analysis can of course provide useful insights in some cases; see for example [12, 20, 36].

the system and the vector of conditional probabilities $p_{n-1,j}$ for $j = 1, ..., n-2$ similarly defined for customer $n-1$.

We illustrate the usefulness of our approach by describing numerical results that examine the impact of not accounting for non-punctuality and no-shows. We provide analytical support by proving that non-punctuality always deteriorates waiting time performance regardless of the distribution of non-punctuality. We also illustrate how our approach can be used to support *online* appointment scheduling[3] where the objective is to minimize completion time subject to a service level constraint on waiting time (scheduling that takes into account the punctuality and no-show behavior of each customer) and compare the performance of such a scheduling scheme to a scheme where all appointment times are equally spaced.

A difficulty in carrying out the exact approach is the computation involving the distribution of customer inter-arrival times, which relies on a convolution of the probability distribution functions of the arrival times of customers $n$ and $n-1$, conditional on customer $n-1$ finding $j$ customers in the system. Hence, while the exact approach is feasible for small to moderately sized problems, it is computing intensive for large problems. To address this limitation, we describe an approximate approach that is computationally efficient. The approximation retains all the steps of the exact approach, except for the one involving computing the distribution of customer inter-arrival times where conditional probabilities are replaced by their unconditional counterparts. We show that the approximation provides significant savings in computational effort with a relatively modest sacrifice in accuracy. We prove, using the theory of majorization, that the approximation provides an upper bound for the exact expected customer waiting time when the non-punctuality is uniformly-distributed.

The rest of the paper is organized as follows. In Section 2, we discuss related literature. In Section 3, we describe the problem and the analytical approach. In Section 4, we examine the impact of non-punctuality on performance. In Section 5, we illustrate how our performance evaluation approach can be embedded in an optimization problem to obtain optimal schedules. In Section 6, we describe the approximate approach and assess its performance. In Section 7, we provide concluding comments.

[3]We use the term online to refer to the realistic setting where customers are assigned an appointment time at the time they request one, taking into account previous appointments and the characteristics of the associated customers.

## 2 Related literature

There is an extensive literature on systems where arrivals are determined by appointments times. The typical application is appointment scheduling in healthcare. We refer the reader to Cayirli and Veral [3] and Gupta and Denton [10] for surveys of early contributions in this area and Ahmad et al. [1] and Zacharias and Yunes [44] for a review of more recent literature. Zacharias and Yunes [44] provide a useful classification of the literature based on features such as service time distribution, no-shows, non-punctuality, emergency demand and customer heterogeneity. In much of this literature, the focus is on determining schedules that can effectively balance the tradeoff between resource utilization (e.g., that of medical staff) and customer (patient) delay. In particular, a typical formulation is one that minimizes the sum of overtime cost and customer delay cost.

Two main streams of literature can be distinguished: a stream that focuses on *inter-day* dynamics and a stream that focuses on *intra-day* dynamics. The first stream accounts for the time between when a customer makes a request for an appointment and the date of the appointment, which may be days later. This time is often referred to as *indirect* waiting time since the customer may be able to carry on with other activities prior the day of the appointment. The second stream focuses on the waiting time, typically measured in minutes, experienced by the customer when the customer shows up to the appointment. This waiting time is often referred to as *direct* waiting time. Our paper belongs to this second stream. Therefore, in the remainder of this section, we limit our discussion to this stream and refer the reader to Green and Savin [9], Zacharias and Armony [42], and the references therein for a discussion of the first stream. Exploring the relationship between the inter-day and intra-day dynamics is of course important for many applications and is an area that merits further study; see for example Feldman et al. [8].

The literature on intra-day dynamics can be divided into two sub-streams: one assumes customers are punctual and one allows for customer non-punctuality. We review relevant papers below, focusing on those that are most closely related.

**Punctual customers.** Wang [37] considers a problem similar to ours, except that customers are always punctual and always show up. He considers the problem of selecting appointment times to minimize total cost (measured as the sum of delay cost and completion cost). He develops a recursive procedure for computing expected waiting times. This procedure is integrated into a

non-linear optimization algorithm for computing appointment times. This work is extended in [38] for the case of exponentially-distributed service times with heterogeneous rates and to a setting where there is flexibility in how patients are sequenced. The joint sequencing and appointment scheduling is also studied in [25] and in [14] by allowing for no-shows.

Hassin and Mendel [11] consider the problem where customers may not always show up and study the impact for no-shows. They do so for the case where the no-show probability is the same for all customers and service times are independent and have identical exponential distributions. They develop a procedure for determining optimal appointment times for the case where the objective is to minimize the cost of customer waiting time and completion cost. In doing so, they rely on the fact that the objective function is convex. Other related literature include [22, 24, 19] and the references therein. Our paper complements this literature by considering a more general setting by allowing for no-show probabilities to be heterogeneous, service time distributions to be Cox-distributed and by considering non-punctuality. Moreover, we consider an alternative approach to generating appointment times (minimizing completion time subject to a constraint on customer waiting time).

Millhiser and Valenti [28] develop a numerical approach for computing the probability of customer waiting time and completion (among other performance measures) in a setting similar to ours and by considering heterogeneous service times and no-show probabilities. Using this approach, Millhiser et al. [30] propose a framework for optimizing appointment times so as to satisfy a constraint on the waiting time that each customer experiences. This approach for appointment scheduling is similar to the online optimization approach we discuss in Section 5. Chen et al. [6] use the theory of majorization to study properties of the optimal schedule for a similar problem with punctual customers.

The literature we have discussed so far considers, as we do in this paper, a continuous time setting. There is also literature that treats time as being discrete and where appointments may only be scheduled at discrete points in time. The resulting appointment scheduling problem is typically formulated as a non-linear discrete optimization problem. Depending on its features, the problem is solved either exactly or using an approximation. Examples include [16, 45, 34, 21, 43] and the references therein.

**Non-punctual customers.** Although non-punctuality is quite prevalent in practice [17] academic literature that accounts for it is relatively limited. Deceuninck et al. [7] provide an excellent review of this literature. Among the papers that consider non-punctuality, some rely on simulation; see for example [4, 18, 41, 46]. Papers that provide analytical results include [7, 13, 44]. Jiang et al. [13] use a scenario-based stochastic programming approach to formulate an appointment scheduling problem with non-punctuality and no-shows. They generate scenarios for possible realizations of the various random variables involved. Each set of realizations leads to a deterministic problem. Combining these problems leads to an approximation of the original stochastic problem. Deceuninck et al. [7] consider a discrete time system and develop an approach for sequencing customers and assigning appointment times while accounting for non-punctuality and no-shows. Their approach for evaluating performance relies on a modified Lindley recursion. The appointment schedule optimization relies on a local search algorithm. Zacharias and Yunes [44] also consider a discrete time setting with non-punctuality, no-shows, and walk-in customers. They formulate the problem as a non-linear integer program. They show that the objective function, the sum of customer delay cost and overtime cost, for the case they consider is super-modular and component-wise convex, which they leverage to construct an efficient solution algorithm. Mercer [27] considers a system with equally spaced appointment times, identical show up probabilities, and identical distributions for service times and lateness. Under the assumption of an infinite number of arrivals, he derives the equilibrium distribution of the queue length.

Our paper is also related to literature that considers queueing systems with a finite number of arrivals, though not driven by appointments. Examples include [33] and [39]. Parlar and Sharafali [33] consider a model with a finite number of arrivals motivated by the arrival process of customers at airport check-in counters. Customers arrive independently of each other, with arrivals modeled as a "death process" from a finite population of travelers. Wang et al. [39] study a queueing model with a finite number of arrivals where inter-arrival times and service times are independent and heterogeneous. Both single and multi-server settings are considered. They examine the effect of heterogeneity in inter-arrival and service times on waiting times. In this paper, we build on the approach in Wang et al. [39] to study a system with a finite number of arrivals that are driven by appointments and incorporate both non-punctuality and no-shows.

In summary, our paper complements the existing literature by considering a system in con-

tinuous time where the arrival of customers is driven by appointments and where customers are non-punctual and may not always show up. Our treatment is general in the sense that we allow for no-show probabilities and non-punctuality to be heterogeneous across customers. We also consider a fairly general class of service times. We provide methodologies for both exact and approximate analysis. In the spirit of Millhiser et al. [29, 30], we illustrate the usefulness of our approach to online scheduling where the objective is to minimize completion time subject to a constraint on the waiting time for each customer. We enrich the study of non-punctuality by providing additional insights into its impact on schedules and performance. A preliminary conference version of this paper is Jouini and Benjaafar [15]. To our knowledge, that paper is among the first in the literature to study an analytical model that incorporates customer non-punctuality.

# 3 Problem description and analysis

## 3.1 Problem description

We consider a queueing system with a single server and a finite number of customers who arrive over time. There are $M$ customers who are scheduled to arrive. We denote by $d_n$, for $n = 1, \cdots, M$, the appointment time of the $n$-th customer. We index customers by their appointment times and we assume that $d_n \leq d_m$ if $n < m$. Customer $n$ has a probability $\alpha_n$ of showing up, independently of all other events. If a customer shows up, she may do so earlier or later than her appointment time. More specifically, the customer may show up at a random time between $d_n - \tau_n^l$ and $d_n + \tau_n^u$. In other words, the arrival time of customer $n$ can be described by a random variable with finite support $d_n - \tau_n^l$ and $d_n + \tau_n^u$. We refer to this random variable as $D_n$ and allow it to have a general distribution with probability distribution function (pdf) denoted by $f_n$ and cumulative distribution function (cdf) denoted by $F_n$. We use interchangeably the terms arrival distribution and non-punctuality distribution. Note that we allow for this distribution function to be customer specific[4]. We further assume that customers arrive in the order of their appointment times, so that $D_{n-1} < D_n$ or equivalently $d_n - d_{n-1} \geq \tau_{n-1}^u + \tau_n^l$. Therefore, customer arrival times are non-overlapping. This assumption is made for tractability and also to avoid the thorny issue around whether or not to proceed with the service of a customer who shows up earlier than the customer

[4]This is an important feature in applications, such as healthcare, where data may be available, or can be collected, on the punctuality of different customers and where punctuality of different customers can vary significantly.

scheduled before her and whether or not to preempt her service once that the customer with the earlier appointment shows up. This assumption is reasonable in settings where the times between appointments are large relative to customer lateness (for example, two successive appointment times are 50 minutes apart but customers are at most 25 minutes early or late) or when the service provider does not allow lateness to exceed a specified threshold. We place no other assumptions on the distribution of customer arrival times.

Upon arrival, a customer goes immediately into service if the server is available. If not, the customer joins the queue where she waits for service. Service times are independent and exponentially-distributed with mean service rate $\mu$ (in Section 3.4, we extend the analysis to the case where service times follow a homogeneous Cox-distributed). We assume that the server is available to start work exactly at $d_1$ (the scheduled time of the first customer). The server remains available until the last customer has completed service. The server has no prior knowledge of whether or not a particular customer will show up. Therefore if customer $M$ shows up, the server shuts down as soon as customer $M$ completes service. If not, the server shuts down after the latest possible arrival time of customer $d_M + \tau_M^u$ and as soon as the last customer present in system completes service. We assume that customers are processed in the order of their appointment times. We also assume that the system is work-conserving with the server never idling when there are customers in the queue.

## 3.2 Analysis

Our approach consists of first deriving the stationary probability of the system state seen by a new arrival, conditioned on system states seen by the previous arrival. Then, from the schedule, we compute the distribution of inter-arrival time between customer $n$ and $n-1$, conditioned on the system state when the customer $n-1$ arrives. Finally, we compute the conditional waiting time by combining results from the first two steps and characterize the unconditional waiting time by averaging over all possibilities.

### 3.2.1 Distribution of the number of customers at arrival instants

A full schedule is defined by the vector $\delta = (d_1, ..., d_M)$. Without loss of generality, we choose $d_1 = \tau_1^l$ such that the origin of time is at $d_1 - \tau_1^l = 0$. We denote by $R_n$ the random variable that describes the number of customers found (would have been found) in the system by customer $n$ if

she shows up (does not show up). This means that the total number of customers in the system at time $D_n$ is $R_n + 1$ ($R_n$) if customer $n$ shows up (does not show up). We let $p_{n,i} = \Pr\{R_n = i\}$ refer to the probability that the $n$-th customer finds (would have found), upon arrival if she shows up (she does not), $i$ customers already in the system (i.e., in the queue or in service), for $0 \leq i \leq n-1$ and $1 \leq n \leq M$.

Let $X_n$ be the random variable describing the inter-arrival time between customers $n$ and $n+1$ (note that we associate an arrival time with a customer regardless of whether or not she actually shows up; if the customer does not show up, we refer to this as a *virtual* arrival), where $X_n = D_{n+1} - D_n$ for $1 \leq n \leq M-1$. We denote by $h_n(.)$ the pdf of the random variable $X_n$. We have $d_{n+1} - d_n - \tau_{n+1}^l - \tau_n^u \leq X_n \leq d_{n+1} - d_n + \tau_{n+1}^u + \tau_n^l$. Note that the random variables $D_n$ and $D_{n+1}$ do not overlap; thus, $X_n \geq 0$. Figure 1 graphically illustrates the customer arrival dynamics.

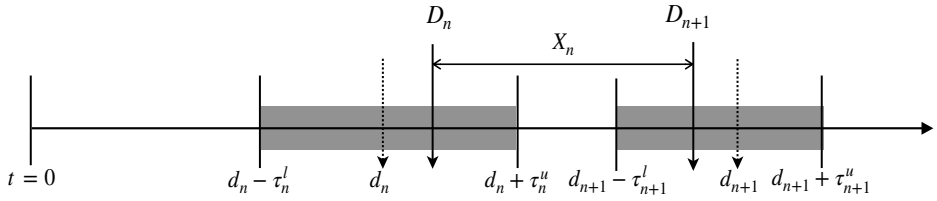



Figure 1: The appointment-driven arrival process

**Initialization: the case of $n = 1$ and $n = 2$.** For $n = 1$, $p_{1,0} = 1$ and $p_{1,i} = 0$ for $i \neq 0$ since the first customer always finds the system empty if she shows up. However she may have to wait to start service because the server starts work exactly at $d_1$. We shall discuss this matter later in this section.

For $n = 2$, we have $p_{2,0} = 1 - p_{2,1}$. In what follows, we compute $p_{2,1}$. To do so, we separate the cases of whether customer 1 arrives early ($D_1 < d_1$), or she arrives late ($D_1 \geq d_1$). Recall that $d_1 = \tau_1^l$. Then the probability $p_{2,1}$ may be written as

$$\begin{aligned} p_{2,1} &= \alpha_1 \Pr\{D_1 < d_1\} p_{2,1|D_1<d_1} + \alpha_1 \Pr\{D_1 \geq d_1\} p_{2,1|D_1 \geq d_1} \\ &= \alpha_1 \left( \int_0^{d_1} f_1(x)\, dx \right) p_{2,1|D_1<d_1} + \alpha_1 \left( \int_{d_1}^{d_1+\tau_1^u} f_1(x)\, dx \right) p_{2,1|D_1 \geq d_1}, \end{aligned} \tag{1}$$

where $p_{2,1|D_1<d_1}$ ($p_{2,1|D_1 \geq d_1}$) is the conditional probability that customer 2 sees customer 1 in the system upon arrival, given that customer 1 arrives early (late).

Customer 1 arrives early. Given that customer 1 arrives early, we have

$$p_{2,1|D_1<d_1} = \Pr\{D_2 - d_1 < \varepsilon_\mu\}, \tag{2}$$

where $\varepsilon_\mu$ is an exponential random variable with rate $\mu$. Then

$$p_{2,1|D_1<d_1} = \int_{d_2-d_1-\tau_2^l}^{d_2-d_1+\tau_2^u} \Pr\{x < \varepsilon_\mu\} f_2(x+d_1)\, dx = \int_{d_2-d_1-\tau_2^l}^{d_2-d_1+\tau_2^u} e^{-\mu x} f_2(x+d_1)\, dx. \tag{3}$$

Customer 1 arrives late. Given that customer 1 arrives late, we have

$$p_{2,1|D_1\geq d_1} = \Pr\{D_2 - D_1 < \varepsilon_\mu \mid D_1 \geq d_1\} = \int_{d_2-d_1-\tau_2^l-\tau_1^u}^{d_2-d_1+\tau_2^u} e^{-\mu x} h_{1|D_1\geq d_1}(x)\, dx, \tag{4}$$

where $h_{1|D_1\geq d_1}(x)$ defined on $d_2 - d_1 - \tau_2^l - \tau_1^u \leq x \leq d_2 - d_1 + \tau_2^u$ is the pdf of the random variable $(D_2 - D_1)|D_1 \geq d_1$; the conditional inter-arrival time given that customer 1 arrives late. This pdf can be obtained as follows:

$$h_{1|D_1\geq d_1}(x) = \int_{-\infty}^{\infty} f_2(u) f_{1|D_1\geq d_1}(u-x)\, du, \tag{5}$$

where

$$f_{1|D_1\geq d_1}(x) = \begin{cases} \frac{f_1(x)}{1-F_1(d_1)} & \text{if } d_1 \leq x \leq d_1 + \tau_1^u \text{ and} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

In the convolution in (5), the intersection of the supports of the two functions is

$$\Omega_{2,x} = [\max(d_2 - \tau_2^l, d_1 + x), \min(d_2 + \tau_2^u, d_1 + \tau_1^u + x)].$$

Note that $\min(d_2 + \tau_2^u, d_1 + \tau_1^u + x) \geq \max(d_2 - \tau_2^l, d_1 + x)$ always holds for $d_2 - d_1 - \tau_2^l - \tau_1^u \leq x \leq d_2 - d_1 + \tau_2^u$. Thus, (5) can be rewritten as

$$h_{1|D_1\geq d_1}(x) = \int_{\Omega_{2,x}} f_2(u) \frac{f_1(u-x)}{1-F_1(d_1)}\, du. \tag{7}$$

Substituting (3) and (4) into (1) leads to $p_{2,1}$ and $p_{2,0} = 1 - p_{2,1}$.

**Iteration: the case of $n \geq 3$.** For $3 \leq n \leq M$, we separate the cases $i = 0$ and $1 \leq i \leq n-1$ to compute $p_{n,i}$. Consider first $p_{n,i}$ for $3 \leq n \leq M$ and $1 \leq i \leq n-1$. Conditioning on the number of customers found upon arrival by customer $n-1$, we have

$$p_{n,i} = \sum_{j=i-1}^{n-2} p_{n-1,j} \Pr\{R_n = i \mid R_{n-1} = j\}. \tag{8}$$

We distinguish the cases of whether customer $n-1$ shows up or not. For customer $n$ to find $i$ customers upon arrival given that customer $n-1$ found (would have found) $j$ customers, there must be $j-i+1$ service completions ($j-i$ service completions) between the arrival times of customer $n-1$ and customer $n$. This also means that, once customer $n-1$ arrives but before customer $n$ does (the duration is $X_{n-1}$), exactly $j-i+1$ service completions ($j-i$ service completions) occur. Since the server has an exponential service time, the number of customers served during $X_{n-1}$ follows a Poisson process with rate $\mu$. Denoting by $h_{n-1,j}(.)$ the pdf of the conditional inter-arrival time between customers $n-1$ and $n$, given that the former finds (would have found) $j$ customers in the system if she shows up (she does not), for $0 \leq j \leq n-2$, we have

$$\begin{aligned} p_{n,i} = \alpha_{n-1} \sum_{j=i-1}^{n-2} p_{n-1,j} \int_{d_n - d_{n-1} - \tau_n^l - \tau_{n-1}^u}^{d_n - d_{n-1} + \tau_n^u + \tau_{n-1}^l} \frac{(\mu x)^{j+1-i}}{(j+1-i)!} e^{-\mu x} h_{n-1,j}(x)\, dx \\ + (1-\alpha_{n-1}) \sum_{j=i}^{n-2} p_{n-1,j} \int_{d_n - d_{n-1} - \tau_n^l - \tau_{n-1}^u}^{d_n - d_{n-1} + \tau_n^u + \tau_{n-1}^l} \frac{(\mu x)^{j-i}}{(j-i)!} e^{-\mu x} h_{n-1,j}(x)\, dx, \end{aligned} \tag{9}$$

for $3 \leq n \leq M$ and $1 \leq i \leq n-1$, with the convention that an empty sum is equal to 0.

A difficulty in the recursive approach is the characterization of the function $h_{n-1,j}(.)$. By definition, $h_{n-1,j}(.)$ is the pdf of the random variable $X_{n-1,j} = D_n - D_{n-1,j}$, where $D_n$ is the unconditional arrival time of customer $n$ and $D_{n-1,j}$ is the conditional arrival time of customer $n-1$, given that she finds (would have found) $j$ customers in the system if she shows up (she does not show up). Then, $h_{n-1,j}(x) = \int_{-\infty}^{\infty} f_n(u) f_{n-1,j}(u-x)\, du$. The support of $D_n$ is $[d_n - \tau_n^l, d_n + \tau_n^u]$ and the support of $D_{n-1,j}$ is the same as the support of $D_{n-1}$, i.e., $[d_{n-1} - \tau_{n-1}^l, d_{n-1} + \tau_{n-1}^u]$. Therefore, the intersection between the two supports is

$$\Omega_{n,x} = [\max(d_n - \tau_n^l, d_{n-1} - \tau_{n-1}^l + x), \min(d_n + \tau_n^u, d_{n-1} + \tau_{n-1}^u + x)],$$

for $d_n - d_{n-1} - \tau_n^l - \tau_{n-1}^u \leq x \leq d_n - d_{n-1} + \tau_n^u$. So,

$$h_{n-1,j}(x) = \int_{\Omega_{n,x}} f_n(u) f_{n-1,j}(u-x)\, du. \tag{10}$$

Applying Bayes' theorem, we may write

$$f_{n-1,j}(t) = \frac{p_{n-1,j|D_{n-1}=t} \times f_{n-1}(t)}{p_{n-1,j}}, \tag{11}$$

where $p_{n-1,j|D_{n-1}=t}$ is the probability that customer $n-1$ finds $j$ customers in the system given that she shows up at time $t$. It is a function of time $t$ that can be calculated using $p_{n-2,k}$ and $f_{n-2,k}(t')$. In particular,

$$\begin{aligned} p_{n-1,j|D_{n-1}=t} = \alpha_{n-2} \sum_{k=j-1}^{n-3} p_{n-2,k} \int_{d_{n-2}-\tau_{n-2}^l}^{d_{n-2}+\tau_{n-2}^u} \frac{(\mu(t-t'))^{k+1-j}}{(k+1-j)!} e^{-\mu(t-t')} f_{n-2,k}(t')\, dt' + \\ (1-\alpha_{n-2}) \sum_{k=j}^{n-3} p_{n-2,k} \int_{d_{n-2}-\tau_{n-2}^l}^{d_{n-2}+\tau_{n-2}^u} \frac{(\mu(t-t'))^{k-j}}{(k-j)!} e^{-\mu(t-t')} f_{n-2,k}(t')\, dt', \end{aligned} \tag{12}$$

for $1 \leq j \leq n-2$ and $n > 3$. The case $n = 3$ needs to be treated separately due to the fact that the first service starts exactly at $d_1$. In this case, we have

$$p_{2,1|D_2=t} = \alpha_1 \left( \int_{d_1}^{d_1+\tau_1^u} e^{-\mu(t-t')} f_{1,0}(t')\ dt' + \int_{d_1-\tau_1^l}^{d_1} e^{-\mu(t-d_1)} f_{1,0}(t')\ dt' \right). \tag{13}$$

Finally, the probabilities $p_{n,0}$ can be obtained using

$$p_{n,0} = 1 - \sum_{i=1}^{n-1} p_{n,i}, \tag{14}$$

for $3 \leq n \leq M$.

Using the above expressions, the probabilities $p_{n,i}$ for $1 \leq n \leq M$ and $0 \leq i \leq n-1$ can now be computed recursively starting with $n = 1$. The overall approach for computing the $p_{n,i}$s is summarized in Figure 2.

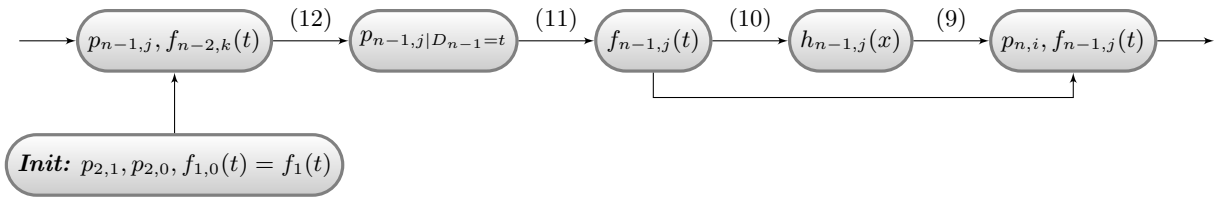



Figure 2: Steps for computing the probabilities $p_{n,i}$

### 3.2.2 Waiting time distributions

Having obtained the probability distribution of the number of customers in the system observed by an arriving customer, we can characterize the distribution of waiting time of each customer. Let $W_n$, a random variable, denote the waiting time in the queue of customer $n$, if she shows up, and let $\mathbb{E}[W_n^k]$ be the corresponding $k$-th moment for $k \geq 1$. (For the rest of the paper, $\mathbb{E}[Z]$ denotes the expected value of a given random variable $Z$ and $\mathbb{E}[Z^k]$ the $k$-th moment). Then

$$\mathbb{E}[W_n^k] = \sum_{i=1}^{n-1} p_{n,i} \mathbb{E}[W_{n,i}^k], \tag{15}$$

for $2 \leq n \leq M$, where $W_{n,i}$ is the random variable denoting the waiting time in queue of customer $n$, given that customer $n$ shows up and finds $i$ customers upon arrival. Since service times are independent and exponentially-distributed with parameter $\mu$, $W_{n,i}$ has an $i$-Erlang distribution with $i$ phases and rate $\mu$ per phase. Using Equation (15) and knowing that $\mathbb{E}[W_{n,i}] = \dfrac{i}{\mu}$ and $\mathbb{E}[W_{n,i}^2] = \dfrac{i(i+1)}{\mu^2}$, we obtain

$$\mathbb{E}[W_n] = \sum_{i=1}^{n-1} p_{n,i} \frac{i}{\mu} \text{ and } \mathbb{E}[W_n^2] = \sum_{i=1}^{n-1} p_{n,i} \frac{i(i+1)}{\mu^2}, \tag{16}$$

for $2 \leq n \leq M$.

Moreover, we have

$$\Pr\{W_{n,i} < t\} = 1 - \sum_{j=0}^{i-1} \frac{(\mu t)^j}{j!} e^{-\mu t}, \tag{17}$$

for $t \geq 0$. Consequently,

$$\begin{aligned} \Pr\{W_n < t\} &= p_{n,0} + \sum_{i=1}^{n-1} p_{n,i} \Pr\{W_{n,i} < t\} \\ &= 1 - \sum_{i=1}^{n-1} \sum_{j=0}^{i-1} p_{n,i} \frac{(\mu t)^j}{j!} \, e^{-\mu t}. \end{aligned} \tag{18}$$

The case $n = 1$ is treated separately. Assume customer 1 shows up. If she arrives before $d_1$, then she has to wait for the server's work starting time at $d_1$. If not, she immediately enters service with no waiting (recall that $d_1 - \tau_1^l = 0$). Then

$$\mathbb{E}[W_1^k] = \int_0^{d_1} (d_1 - x)^k f_1(x) \, dx. \tag{19}$$

Also,

$$\begin{aligned} \Pr\{W_1 < t\} &= \Pr\{d_1 - D_1 < t \mid D_1 < d_1\} \Pr\{D_1 < d_1\} + \Pr\{D_1 \geq d_1\} \\ &= (1 - \Pr\{D_1 \leq d_1 - t \mid D_1 < d_1\}) \Pr\{D_1 < d_1\} + \Pr\{D_1 \geq d_1\}. \end{aligned} \tag{20}$$

The pdf of the random variable $D_1 | D_1 < d_1$ is defined on $[0, d_1]$ and is given by $\dfrac{f_1(x)}{\Pr(D_1 < d_1)}$. After some algebra, Equation (20) leads to

$$\Pr\{W_1 < t\} = 1 - \int_0^{max(d_1 - t, 0)} f_1(x) \, dx. \tag{21}$$

### 3.3 An example: A symmetric system with uniformly-distributed non-punctuality

In this section, we illustrate the steps in characterizing the probability distributions for the special case of a symmetric system where customer non-punctuality is homogeneous and has the uniform distribution. In Appendix A, we also provide discussion for the case where customer non-punctuality has a symmetric triangular distribution. For brevity, we limit our discussion to the initialization steps (the iterative steps follow in a straightforward way as the general approach).

When the non-punctuality is symmetric and uniformly-distributed, we have, for $1 \leq n \leq M$, $\tau_n^u = \tau_n^l = \tau$, which leads to $f_n(t) = \dfrac{1}{2\tau}$ on $[d_n - \tau, d_n + \tau]$ and 0 otherwise. For $n = 2$, we have

$p_{2,0} = 1 - p_{2,1}$, so we complete the initialization step by computing $p_{2,1}$ as follows:

$$\begin{aligned} p_{2,1} &= \alpha_1 \Pr\{D_1 < d_1\} p_{2,1|D_1<d_1} + \alpha_1 \Pr\{D_1 \geq d_1\} p_{2,1|D_1 \geq d_1} \\ &= \alpha_1 \frac{1}{2} \int_{d_2-d_1-\tau}^{d_2-d_1+\tau} e^{-\mu x} f_2(x+d_1)\, dx + \alpha_1 \frac{1}{2} \int_{d_2-d_1-\tau_2^l-\tau}^{d_2-d_1+\tau} e^{-\mu x} h_{1|D_1 \geq d_1}(x)\, dx \\ &= \frac{\alpha_1}{2} \left( \frac{e^{\mu\tau} - e^{-\mu\tau}}{2\mu\tau} + \int_{d_2-d_1-2\tau}^{d_2-d_1+\tau} e^{-\mu x} h_{1|D_1 \geq d_1}(x)\, dx \right). \end{aligned} \tag{22}$$

Using (7), we can compute $h_{1|D_1 \geq d_1}(x)$ on $[d_2 - d_1 - \tau_2^l - \tau_1^u, d_2 - d_1 + \tau_2^u]$ as

$$\begin{aligned} h_{1|D_1 \geq d_1}(x) &= \int_{d_2-\tau}^{d_2+\tau} f_2(u) \frac{f_1(u-x)}{1-F(d_1)} \mathbb{1}_{\{d_1 \leq u-x \leq d_1+\tau\}}\, du \\ &= \int_{d_2-\tau}^{d_2+\tau} \frac{1}{2\tau} \frac{1}{\tau} \mathbb{1}_{\{d_1+x \leq u \leq d_1+\tau+x\}}\, du \\ &= \frac{1}{2\tau^2} \big( \min(d_2+\tau, d_1+\tau+x) - \max(d_2-\tau, d_1+x) \big) \\ &= \begin{cases} \frac{x+d_1-d_2+2\tau}{2\tau^2}, & \text{for } d_2-d_1-2\tau \leq x < d_2-d_1-\tau \\ \frac{1}{2\tau}, & \text{for } d_2-d_1-\tau \leq x \leq d_2-d_1 \\ \frac{d_2-d_1+\tau-x}{2\tau^2}, & \text{for } d_2-d_1 < x \leq d_2-d_1+\tau. \end{cases} \end{aligned} \tag{23}$$

Combining (22) and (23), we obtain

$$p_{2,1} = \alpha_1 \frac{e^{-\mu(d_2-d_1)}}{4\mu^2\tau^2} (e^{\mu\tau} - e^{-\mu\tau})(e^{\mu\tau} + \tau\mu - 1).$$

### 3.4 Systems with $\gamma-$Cox-distributed service times

In this section, we extend the analysis to the case where service times follow a homogeneous $\gamma-$Cox distribution (or simply $\gamma-$*Cox distribution*). This family of distributions is dense in the field of positive distributions as shown in [23] and has the advantage of being easier to manipulate than a general phase-type distribution. The $\gamma-$Cox distribution is a special case of a Cox distribution. Under the $\gamma-$Cox distribution, service time for each customer consists of a succession of independent and homogeneous exponential phases, each with rate $\gamma$. The maximal number of phases is $m$. This distribution is defined by the probabilities $q_0, q_1, \cdots, q_{m-1}$, with $0 \leq q_i \leq 1$, where $q_i$ is the probability to continue service after having reached phase $i$. With probability $1 - q_i$, a service in phase $i$ ends at the end of phase $i$. Instantaneous service is possible, with probability $1 - q_0$. At the

end of phase $m$, the service always ends, i.e., $q_m = 0$. The cdf of the $\gamma-$Cox distribution, $G_\gamma(t)$, and the pdf, $g_\gamma(t)$, are given respectively by

$$G_\gamma(t) = 1 - e^{-\gamma t}\sum_{k=0}^{m-1}\frac{(\gamma t)^k}{k!}\prod_{i=0}^{k} q_i \text{ and} \tag{24}$$

$$g_\gamma(t) = \gamma e^{-\gamma t}\sum_{k=0}^{m-1}\frac{(\gamma t)^k}{k!}(1-q_{k+1})\prod_{i=0}^{k} q_i, \tag{25}$$

for $t \geq 0$.

To characterize the waiting time distribution, we apply a similar recursive method to the one used for the case of exponentially-distributed service times. In that case, the system state (number of customers ahead in the system) seen by a new arrival leads to a deterministic number of exponential phases that represent the waiting time of this new arrival. Under the $\gamma-$Cox distribution, knowing just the system status at arrival instances (number of customers ahead in the system and the current phase of the customer in service) is not sufficient as we need to account for the random number of exponential phases for each customer ahead.

A direct method, considering all possible paths for the total number of service phases ahead, would have a high computational cost because of the large number of convolutions involved. To avoid this complication, we study an equivalent version of the system by allowing the number of phases associated with the service of a customer to be realized at the time of arrival instead of during service, and by defining the system state to be the *actual* number of phases of all customers ahead in the system who remain to be served. This version is equivalent to the original one because the service times of customers are independent. Let $R_n$ denote the random variable that describes the system state found (would have been found) by customer $n$ if she shows up (does not show up), and let $p_{n,r} = \Pr\{R_n = r\}$, for $0 \leq r \leq (n-1)m$ and $1 \leq n \leq M$. Let $\Theta_n$ denote the random variable that describes the number of exponential phases of which the service of customer $n$ consists, and $\theta_{n,i}$ denote the probability that $\Theta_n = i$, for $0 \leq i \leq m$ and $1 \leq n \leq M$. Because service times are homogeneous across customers, we have $\theta_{n,0} = \theta_0 = 1 - q_0$ and $\theta_{n,i} = \theta_i = (1-q_i)\prod_{k=0}^{i-1} q_i$, for $1 \leq i \leq m$.

We first characterize the probability $p_{n,r}$, i.e., the probability that customer $n$ finds upon her arrival that the remaining service of customers in the system consists of $r$ exponential phases, for $1 \leq n \leq M$ and $0 \leq r \leq (n-1)m$.

**Initialization: the case of $n = 1$ and $n = 2$.** For $n = 1$, $p_{1,0} = 1$ and $p_{1,r} = 0$ for $r \neq 0$, since the first customer always finds the system empty if she shows up. However she may have to wait to start service because the server starts work exactly at $d_1$. For $n = 2$, we have $p_{2,0} = 1 - \sum_{r=1}^{m} p_{2,r}$. Next, we compute $p_{2,r}$, for $1 \leq r \leq m$. Similar to the analysis in the exponential case, we separate the cases of whether customer 1 arrives early ($D_1 < d_1$) or late ($D_1 \geq d_1$). Recall that $d_1 = \tau_1^l$; then the probability $p_{2,r}$ can be written as

$$\begin{aligned} p_{2,r} &= \alpha_1 \Pr\{D_1 < d_1\} p_{2,r|D_1<d_1} + \alpha_1 \Pr\{D_1 \geq d_1\} p_{2,r|D_1 \geq d_1} \\ &= \alpha_1 \left( \int_0^{d_1} f_1(x)\, dx \right) p_{2,r|D_1<d_1} + \alpha_1 \left( \int_{d_1}^{d_1+\tau_1^u} f_1(x)\, dx \right) p_{2,r|D_1 \geq d_1}, \end{aligned} \tag{26}$$

for $1 \leq r \leq m$, where $p_{2,r|D_1<d_1}$ ($p_{2,r|D_1 \geq d_1}$) is the conditional probability that customer 2 sees $r$ remaining phases of customer 1 in the system upon arrival, given that customer 1 arrives early (late). In order for customer 2 to see $r$ phases, customer 1 must have at least $r$ phases and $\Theta_1 - r$ phases have been completed between the beginning of the service of customer 1 (who shows up) and the arrival of customer 2. To know the exact number of phases completed, we further condition on the number of phases of customer 1. Therefore, if customer 1 arrives early (work starts at $d_1$), we have

$$p_{2,r|D_1<d_1} = \sum_{i=r}^{m} \theta_i \int_{d_2-d_1-\tau_2^l}^{d_2-d_1+\tau_2^u} \frac{(\gamma x)^{i-r}}{(i-r)!} e^{-\gamma x} f_2(x + d_1)\, dx, \tag{27}$$

for $1 \leq r \leq m$. If customer 1 arrives late (work starts at $D_1 \geq d_1$), we have

$$p_{2,r|D_1 \geq d_1} = \sum_{i=r}^{m} \theta_i \int_{d_2-d_1-\tau_2^l-\tau_1^u}^{d_2-d_1+\tau_2^u} \frac{(\gamma x)^{i-r}}{(i-r)!} e^{-\gamma x} h_{1|D_1 \geq d_1}(x)\, dx, \tag{28}$$

for $1 \leq r \leq m$, where $h_{1|D_1 \geq d_1}(x)$ is defined similarly to the case with exponential service time, and can be obtained from (1). Using (27) and (28) and substituting in (26), we obtain $p_{2,r}$ for $1 \leq r \leq m$ and then $p_{2,0}$.

**Iteration: the case of $n \geq 3$.** For $3 \leq n \leq M$, we separate the cases $r = 0$ and $1 \leq r \leq (n-1)m$. Consider first $p_{n,r}$ for $3 \leq n \leq M$ and $1 \leq r \leq (n-1)m$. Our approach is based on conditioning on the number of service phases of customer $n-1$, the number of phases found ahead (would have been found) upon arrival by customer $n-1$, and whether customer $n-1$ shows up or not. Notice that these conditions are independent. Denoting by $h_{n-1,j}(.)$ the pdf of the conditional inter-arrival

time between customers $n-1$ and $n$, given that the former finds (would have found) $j$ phases in the system if she shows up (she does not), we have

$$\begin{aligned}
p_{n,r} &= \alpha_{n-1}\sum_{i=0}^{m}\theta_i \sum_{j=\max(r-i,0)}^{(n-2)m} p_{n-1,j}\Pr\{R_n = r \mid R_{n-1}=j, \Theta_{n-1}=i, \text{ customer } n-1 \text{ shows up}\} \\
&\quad + (1-\alpha_{n-1})\sum_{j=r}^{(n-2)m} p_{n-1,j}\Pr\{R_n = r \mid R_{n-1}=j, \text{ customer } n-1 \text{ does not show up}\} \\
&= \alpha_{n-1}\sum_{i=0}^{m}\theta_i \sum_{j=\max(r-i,0)}^{(n-2)m} p_{n-1,j}\int_{d_n-d_{n-1}-\tau_n^l-\tau_{n-1}^u}^{d_n-d_{n-1}+\tau_n^u+\tau_{n-1}^l} \frac{(\gamma x)^{j+i-r}}{(j+i-r)!}e^{-\gamma x}h_{n-1,j}(x)\,dx \\
&\quad + (1-\alpha_{n-1})\sum_{j=r}^{(n-2)m} p_{n-1,j}\int_{d_n-d_{n-1}-\tau_n^l-\tau_{n-1}^u}^{d_n-d_{n-1}+\tau_n^u+\tau_{n-1}^l} \frac{(\gamma x)^{j-r}}{(j-r)!}e^{-\gamma x}h_{n-1,j}(x)\,dx, \qquad (29)
\end{aligned}$$

for $3 \le n \le M$ and $1 \le r \le (n-1)m$, with the convention that an empty sum is equal to 0. By redefining $f_{n-1,j}(.)$ and $p_{n-1,j|D_n=t}$ based on the number of phases ahead (instead of the number of customers ahead as in the case with exponential service time) observed by customer $n-1$, we can obtain $h_{n-1,j}(x)$ and $f_{n-1,j}(t)$ using (10) and (11). The only missing term in (11) is $p_{n-1,j|D_n=t}$, which can be obtained as

$$\begin{aligned}
p_{n-1,j|D_{n-1}=t} &= \alpha_{n-2}\sum_{i=0}^{m}\theta_i \sum_{k=max(j-i,0)}^{(n-3)m} p_{n-2,k}\int_{d_{n-2}-\tau_{n-2}^l}^{d_{n-2}+\tau_{n-2}^u} \frac{(\gamma(t-t'))^{k+i-j}}{(k+i-j)!}e^{-\mu(t-t')}f_{n-2,k}(t')\,dt' \\
&\quad +(1-\alpha_{n-2})\sum_{k=j}^{(n-3)m} p_{n-2,k}\int_{d_{n-2}-\tau_{n-2}^l}^{d_{n-2}+\tau_{n-2}^u} \frac{(\gamma(t-t'))^{k-j}}{(k-j)!}e^{-\mu(t-t')}f_{n-2,k}(t')\,dt', \qquad (30)
\end{aligned}$$

for $1 \le j \le (n-2)m$ and $n > 3$, and

$$p_{2,j|D_2=t} = \alpha_1\sum_{i=j}^{m}\theta_i\left(\int_{d_1}^{d_1+\tau_1^u}\frac{(\gamma x)^{i-r}}{(i-r)!}e^{-\gamma(t-t')}f_{1,0}(t')\;dt' + \int_{d_1-\tau_1^l}^{d_1}\frac{(\gamma x)^{i-r}}{(i-r)!}e^{-\gamma(t-d_1)}f_{1,0}(t')\;dt'\right), \qquad (31)$$

for $1 \le j \le m$. Using (10), (11), (30) and (31), we can obtain $p_{n,r}$ for $1 \le r \le (n-1)m$ using (29). Finally, the probabilities $p_{n,0}$ can be computed as $p_{n,0} = 1-\sum_{i=1}^{(n-1)m} p_{n,r}$ for $3 \le n \le M$. With the same procedure as shown in Figure 2, the probabilities $p_{n,r}$ for $1 \le n \le M$ and $0 \le r \le (n-1)m$ can now be computed recursively starting with $n = 1$.

The above procedure allows us to determine the distribution of $W_{n,r}$, $P(W_{n,r} < t)$, for $n > 1$ and the moments of the waiting time by applying the same computation as in Section 3.2.2. The

details are provided in Appendix B.

**Special cases.** As mentioned at the beginning of this section, $\gamma-$Cox distributions are dense in the field of positive distributions. Below, we specify the parameters of the $\gamma$-Cox distribution for several common distributions. We refer the reader to Section 3 in [23] for additional details.

- Exponential with rate $\mu$: Set $m = 1$, $q_0 = 1$, and $\gamma = \mu$.
- Erlang with $m$ phases and rate $m\mu$ per phase: Set $q_0 = q_1 = \cdots = q_{m-1} = 1$, and $\gamma = m\mu$.
- Deterministic with duration $\tau$: Set $\frac{m}{\mu} = \tau$ in the Erlang case and let $m$ and $\mu$ go to infinity.
- Hyperexponential with parameters $(\mu_n, p_n)$ with $\mu_n > 0$ and $p_n \in [0, 1]$ for $n = 1, 2, \cdots, N$ (i.e., with probability $p_n$ the service time follows an exponential distribution with rate $\mu_n$, for $n = 1, 2, \cdots, N$): Set $q_0 = 1$, $q_i = \frac{\sum_{n=1}^{N} p_n \left(\frac{\gamma}{\gamma+\mu_i}\right)^i}{\sum_{n=1}^{N} p_n \left(\frac{\gamma}{\gamma+\mu_i}\right)^{i-1}}$, for $0 < i \leq m-1$ and let first $m$ and then $\gamma$ go to infinity.
- Hypoexponential with rates $\mu_n$, with $\mu_n > 0$, $\mu_n \neq \mu_m$, for $1 \leq n, m \leq N$ (i.e., the service is a succession of $N$ exponential phases with rate $\mu_n$, for $n = 1, 2, \cdots, N$, where the rates may be different): Set $q_0 = 1$ and $q_i = \frac{\sum_{n=1}^{N} p_n \left(\frac{\gamma}{\gamma+\mu_i}\right)^i \prod_{n \neq m} \frac{\mu_m}{\mu_m - \mu_n}}{\sum_{n=1}^{N} p_n \left(\frac{\gamma}{\gamma+\mu_i}\right)^{i-1} \prod_{n \neq m} \frac{\mu_m}{\mu_m - \mu_n}}$, for $0 < i \leq m-1$ and let first $m$ and then $\gamma$ go to infinity.

# 4 Numerical results: the impact of non-punctuality

In this section, we present numerical results to assess the impact of customer non-punctuality on expected waiting time, which also allows us to assess the error in evaluating waiting times that would be introduced if customers were assumed to be always punctual. To measure the impact of non-punctuality, we evaluate the percentage difference in expected waiting time between a system where customers are non-punctual, $\mathbb{E}[W_{non-punctual}]$ and one where customers are always punctual (that is, customers show up on exactly their appointment times), $\mathbb{E}[W_{punctual}]$, where the percentage is computed as $\frac{\mathbb{E}[W_{non-punctual}] - \mathbb{E}[W_{punctual}]}{\mathbb{E}[W_{punctual}]} \times 100\%$. In both the punctual and non-punctual systems, we allow for no-shows, where the probability of no-shows is the same in both systems (in the punctual system, the customers that do show up, they do it on time). Appointment times are

constructed as follows. Appointment time for customer $n$ is $d_n = d_{n-1} + \frac{1}{\mu_{n-1}}$, for $2 \leq n \leq M$. For customer 1, appointment time is $d_1 = \tau_1$. This appointment scheme is perhaps consistent with some that are observed in practice that seek to limit the idleness of the servers (introducing additional slack time between appointments would reduce customer waiting time but increase server idleness). In Section 5, we discuss additional numerical results where appointment times are optimized and may not be equally spaced as chosen here.

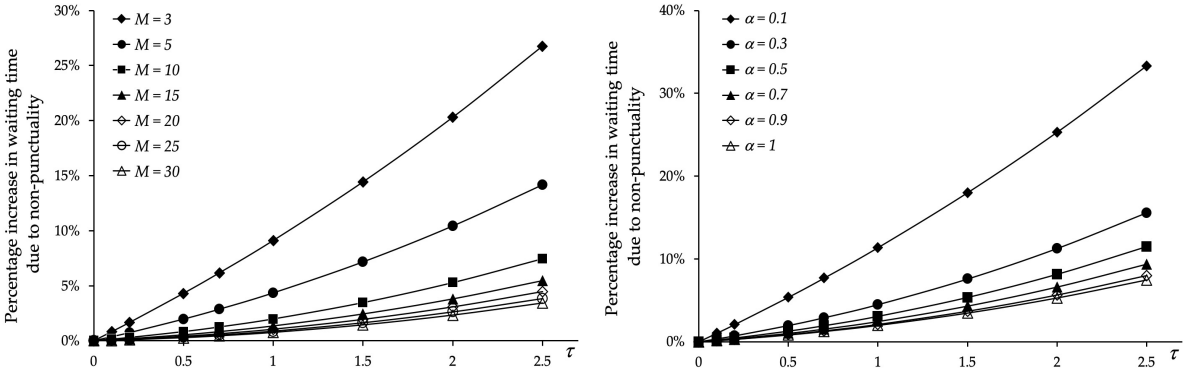
The results, representative of a much larger set, are shown in Figures 3 and 4. The results shown in Figure 3 are for systems where service times are exponentially-distributed. To examine the impact of service time variability, we choose to generate results for non-exponential distribution. The results in Figure 4 are for systems where the service times are $m$-Erlang-distributed ($m$ exponential phases with rate $m\mu$ per phase) with the same mean but different coefficients of variation ($CV$). Moreover, for the sake of brevity, the results presented are only for systems where non-punctuality has the uniform distribution. Results obtained (but not shown here) for other distributions of punctuality yield similar insights. The following observations can be made.

- As shown in Figure 3, non-punctuality can significantly increase customers' waiting time (or, equivalently, ignoring non-punctuality can lead to significant errors in waiting time estimation).

- The effect of non-punctuality is more significant when the number of customers is small. It is also more significant when the show up probability is low. This is perhaps surprising, as one might expect the impact of non-punctuality to be greater when there are more customers in the system (either because of a larger $M$ or a higher $\alpha$). The effect appears to be due to the fact that, when either $M$ or $\alpha$ are large, expected waiting time is relatively large even when the customers are punctual. Introducing non-punctuality does increase waiting time, but the effect is relatively small.

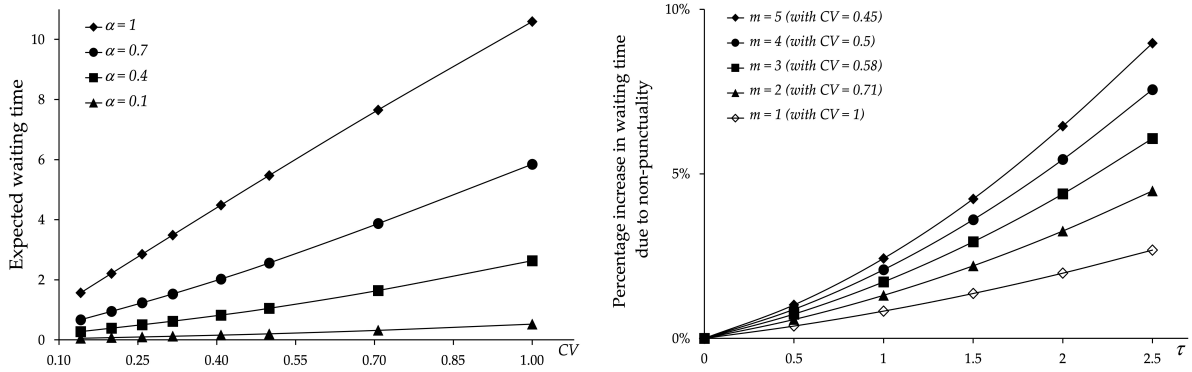
- As shown in Figure 4, the impact of non-punctuality is more significant for systems with low service time variability (the results in Figure 4) are for a system where service times are $m$-Erlang-distributed. Coefficient of variation ($CV$) is varied by varying the number of phases $m$ for fixed $\mu$. Again, this is perhaps surprising and appears due to the fact that when service time variability is high, so is congestion. Therefore, the introduction of non-punctuality,

which introduces variability in the arrival process, has a relatively smaller effect.




(a) $\alpha = 1$ (b) $M = 10$

Figure 3: The impact of non-punctuality ($\tau_n^l = \tau_n^u = \tau$, $\alpha_n = \alpha$, $\mu = 0.2$)




(a) Impact of the service time variability ($\tau = 0$) (b) Impact of non-punctuality ($\alpha = 1$)

Figure 4: The impact of non-punctuality with $m$-Erlang service times ($M = 10$, $\mu = 0.1$, $\tau_n^l = \tau_n^u = \tau$, $\alpha_n = \alpha$)

We also observe that in all cases, the expected waiting time when customers are punctual is lower than the expected waiting time in the case when customers are non-punctual (i.e., the punctual case produces a lower bound on the expected waiting time for the non-punctual case). In what follows, we provide analytical support for this observation and show that it is true regardless of the distribution of punctuality.

**Proposition 1.** *Consider two systems denoted by System 1 and System 2. They are identical except for the arrival pattern of customer $n$. In System 1 customer $n$ is punctual (arrives at exactly $d_n$). In System 2 she is not punctual (arrives at a random moment $D_n$). If $\mathbb{E}[D_n] = d_n$, then the expected waiting time of customer $m$ ($m \geq n$) is higher under System 2 than that under System 1 (i.e., $E(W_n|System\ 1) \leq E(W_n|System\ 2)$).*

The proof of Proposition 1 can be found in Appendix C. It is based on a sample path argument showing that the completion time of customer $m$ is convex with respect to the arrival time of customer $n$.

# 5 Application: Appointment scheduling subject to a service level constraint

In this section, we illustrate how our approach can be used as a basis for generating optimal appointment times. We consider a setting where appointments are generated *online*, with requests for appointments arriving over time and appointments provided at the time of arrival of such requests. Appointments are generated taking into account the characteristics of the customer requesting the appointment (punctuality, likelihood of show up and distribution of service time), the set of previously scheduled appointments and the characteristics of the corresponding customers.

Similar to Millhiser et al. [29, 30], we consider a problem formulation that guarantees a minimum service level for each individual customer. Specifically, we consider a setting where the objective is to minimize the expected service completion time for each customer $n$ assuming she shows up, denoted by $\mathbb{E}[C_n]$, while meeting a service level requirement on the waiting time of this customer $n$. Service levels could be specified in a variety of ways, including a requirement that expected waiting time for each customer does not exceed a certain threshold or that the probability of waiting time exceeding a certain threshold is less than a specified level. Note that such an approach obviates the need, observed in typical appointment scheduling formulations, for determining parameter values for the cost of waiting for customers and the cost of overtime for servers (an overtime cost is incurred when the service completion time of the last customer exceeds a specified threshold[5]). This approach is also more equitable, in the sense that it treats all customers equally and avoids that certain customers experience excessive waiting times. More importantly, this approach is more practical since it allows us to provide a customer with an appointment time when that request is made. That is, we do not need to wait for all the requests to become known before we assign appointment times. This perhaps corresponds to the practical case where customers call an appointment line and are provided with an appointment on the spot (i.e., without knowledge about future appointment

[5]Typical formulations from the appointment scheduling literature adopt the objective of minimizing the sum of the cost of expected waiting time for customers and the cost of expected overtime for the server.

requests).

The optimal appointment for the $n$-th customer, for $n = 2, ..., M$ can be obtained by solving the following optimization problem (without loss of generality, we let $d_1^* = \tau_1$):

$$\min_{d_n} \quad \mathbb{E}[C_n] \tag{32}$$

$$\text{subject to} \quad \mathbb{E}[W_n | d_1 = d_1^*, \cdots, d_{n-1} = d_{n-1}^*] \leq SL, \text{ and} \tag{33}$$

$$\mathbb{E}[C_n] = \mathbb{E}[D_n] + \mathbb{E}[W_n] + \frac{1}{\mu}, \tag{34}$$

where the decision variable is the appointment time $d_n$ such that $d_n \geq d_{n-1}^*$ (with $d_j^*$ denoting the optimal solution to the above problem at $n = j$). The online problem is solved for each customer $n$ (for $n = 2, ..., M$), so $M - 1$ times. The obtained schedule guarantees the earliest expected time for individual customers to exit the system while maintaining reasonable waiting time. Note that the optimal appointment time of a customer, say customer $j$, is generated based on the optimal appointment times of all the customers that preceded her. These are known upon the arrival of customer $j$. In other words, $d_1^*, ..., d_{j-1}^*$ are inputs for the problem at rank $j$.

In Proposition 2, we describe a stochastic ordering result for $W_n$ that allows us to characterize the optimal appointment times. We first define the concept of stochastic ordering we use.

**Definition 1.**

1. *A real random variable $Z_1$ is stochastically larger than a real random variable $Z_2$ if*

$$Pr\{Z_1 > t\} \geq Pr\{Z_2 > t\}, \textit{for all } t \in \mathbb{R}.$$

   *It is equivalent to say that $Z_1$ first-order stochastically (FOS) dominates $Z_2$.*

2. *Let $Z_\beta$ be a real random variable with parameter $\beta \in \mathbb{R}$. We say that $Z_\beta$ stochastically decreases in $\beta$ if $Z_\beta$ FOS dominates $Z_{\hat{\beta}}$ whenever $\beta \leq \hat{\beta}$, for $\beta, \hat{\beta} \in \mathbb{R}$.*

**Proposition 2.** *For $2 \leq n \leq M$, the random variable $W_n$ stochastically decreases in $d_n$.*

The proof of Proposition 2 is given in Appendix C.2. The results below immediately follow.

**Corollary 1.** *The following holds for $2 \leq n \leq M$,*

1. *The k-th moment $\mathbb{E}[W_n^k]$, $k \geq 1$, decreases in $d_n$ and so does $Pr\{W_n \geq t\}$ for all $t \in \mathbb{R}^+$,*

2. $\mathbb{E}[C_n]$ *increases in* $d_n$, *and*

3. *Server idle time prior to the arrival of customer* $n$, $\mathbb{E}[(D_n - C_{n-1})^+]$, *increases in* $d_n$ *(with* $Z^+$ *defined as* $\max(Z, 0)$ *for a given real random variable* $Z$*).*

An implication of Corollary 1 is that the optimal appointment time for customer $n$ is the smallest $d_n$ that meets the service level constraint (e.g., $\mathbb{E}[W_n] \leq SL$). Since $\mathbb{E}[W_n^k]$ (and also $\Pr\{W_n \geq t\}$) is continuous and strictly decreasing in $d_n$, $d_n^*$ can be computed efficiently.

To illustrate the appointment schedules generated using the above approach, we consider a setting with 12 customers and nine scenarios that correspond to different combinations of parameter values for show up and non-punctuality as specified in Table 1 (other parameter values are specified in Table 2). The scenarios cover cases with homogeneous/heterogeneous and symmetric/asymmetric non-punctuality; and cases with homogeneous/heterogeneous no-shows with no-show probabilities ranging from 5% to 30% as suggested in [3]. In practice, information about customers' no-show and non-punctuality may be available for returning customers based on their visit history. It may also be possible to estimate this information from customer characteristics, such as age, type of service being provided, cell phone ownership, and travel distance from home. Data analytics approaches could be employed for this purpose (Cheong et al. [5], Mohammadi et al. [32]).

In Table 2, we provide numerical results for the optimal time between consecutive appointments (i.e., the difference between $d_n^*$ and $d_{n-1}^*$). We also provide corresponding results for expected waiting time averaged over customers 2 to $M$ (in the optimization problem, $d_1^* = \tau_1$ is exogenously specified) and expected total completion time, $\mathbb{E}[C_M]$. The results in Table 2 illustrate how the time between appointments is affected by the show up and non-punctuality characteristics of different customers. The results from Scenarios 1-5 (customers having similar non-punctuality but different no-show characteristics) show how different no-show profiles lead to different schedules, illustrating the importance of accounting for no-shows. Scenarios 1, 2, and 5 illustrate, perhaps as expected, that lower show up probabilities lead to shorter times between appointments. The results for Scenarios 3 and 4 illustrate how time between appointments are affected by patterns of customer show up behavior, with higher (lower) show up probabilities associated with higher (lower) time between appointments.

Table 1: Scenarios for the experiments

| Scenarios | Varying Parameters | |
|---|---|---|
| 0 | $\tau_n = \tau = 0$, and $\alpha_n = \alpha = 1$ | |
| 1 | $\tau_n = \tau = 2$ | $\alpha_1 = \alpha_2 = \cdots = \alpha_{12} = 0.95$ |
| 2 | | $\alpha_1 = \alpha_2 = \cdots = \alpha_{12} = 0.75$ |
| 3 | | $\alpha_1 = 0.95$; $\alpha_2 = 0.75$; $\alpha_3 = 0.95$; ...; and $\alpha_{12} = 0.75$ |
| 4 | | $\alpha_1 = \alpha_2 = \cdots = \alpha_6 = 0.95$; and $\alpha_6 = \alpha_7 = \cdots = \alpha_{12} = 0.75$ |
| 5 | $\alpha_n = \alpha = 1$ | $\tau_1 = \tau_2 = \cdots = \tau_{12} = 2$ |
| 6 | | $\tau_1 = \tau_2 = \cdots = \tau_{12} = 4$ |
| 7 | | $\tau_1 = 6$; $\tau_2 = 2$; $\tau_3 = 6$; ...; and $\tau_{12} = 2$ |
| 8 | | $\tau_1^u = 6$; $\tau_2^u = 2$; $\tau_3^u = 6$; $\tau_4^u = 2$; ...; and $\tau_{12}^u = 2$<br>$\tau_1^l = 2$; $\tau_2^l = 6$; $\tau_3^l = 2$; $\tau_4^l = 6$; ...; and $\tau_{12}^l = 6$ |

Table 2: Time between appointments, $x_{n-1}^*$ (exponential service times, $\mu = 0.1$, $M = 12$, $SL^* = 5$, and parameters as in Table 1)

| $n-1$ | Scenarios | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 6.93 | 7.01 | 4.64 | 7.01 | 7.01 | 7.52 | 8.29 | 8.70 | 5.46 |
| 2 | 15.06 | 14.58 | 12.35 | 12.34 | 14.58 | 15.09 | 15.21 | 15.45 | 19.21 |
| 3 | 15.80 | 15.34 | 13.10 | 15.33 | 15.34 | 15.85 | 16.00 | 15.90 | 11.99 |
| 4 | 16.10 | 15.64 | 13.40 | 13.40 | 15.64 | 16.14 | 16.24 | 16.47 | 20.29 |
| 5 | 16.24 | 15.78 | 13.54 | 15.77 | 15.78 | 16.28 | 16.42 | 16.30 | 12.39 |
| 6 | 16.32 | 15.86 | 13.63 | 13.62 | 15.86 | 16.37 | 16.53 | 16.69 | 20.50 |
| 7 | 16.37 | 15.90 | 13.67 | 15.90 | 13.68 | 16.41 | 16.55 | 16.43 | 12.56 |
| 8 | 16.40 | 15.94 | 13.70 | 13.71 | 13.71 | 16.44 | 16.55 | 16.75 | 20.57 |
| 9 | 16.42 | 15.96 | 13.73 | 15.95 | 13.73 | 16.47 | 16.59 | 16.51 | 12.58 |
| 10 | 16.44 | 15.97 | 13.74 | 13.74 | 13.74 | 16.48 | 16.59 | 16.78 | 20.62 |
| 11 | 16.45 | 15.98 | 13.76 | 15.98 | 13.76 | 16.49 | 16.64 | 16.55 | 12.59 |
| $\frac{1}{M-1}\sum_{n=2}^{M}\mathbb{E}[W_n]$ | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 |
| $\mathbb{E}[C_M]$ | 183.54 | 180.94 | 156.25 | 169.77 | 169.81 | 186.54 | 190.60 | 193.52 | 189.77 |

The results also illustrate the importance of accounting for non-punctuality. The results in Scenarios 0, 5 and 6 show that higher non-punctuality leads to longer times between appointments and, therefore, longer expected completion times. The results in Scenarios 6, 7, and 8 illustrate the impact of heterogeneity in non-punctuality (i.e., differences in non-punctuality among customers). In particular, the results show that higher heterogeneity leads to longer times between appointments and longer completion times (e.g., contrast the results for Scenario 6 with those for Scenario 7). This effect can perhaps be explained by the fact that higher heterogeneity in non-punctuality translates into higher variability in customer inter-arrival times. The results for Scenario 8 show that asymmetric non-punctuality can have a significant impact on the time between appointments, with customers more likely to be late (early) scheduled later (earlier).

Although the numerical results shown in Table 2 are for scenarios where the time between appointments is mainly either increasing or alternating between increasing and then decreasing, it is easy to construct scenarios where this is not the case. In particular, depending on the combination of no-show and non-punctuality parameters of the different customers, it is possible to observe other patterns, including a pattern of increasing and then decreasing time between appointments (dome-shaped), deceasing and then increasing (valley-shaped), or indeterminate. For brevity, results for such scenarios are omitted.

Next, we contrast the individualized appointment schedules generated using our online approach (as illustrated in Table 2) to those generated using an approach in which appointments are equally spaced (as illustrated in Table 3). Equally spaced schedules are not uncommon in practice and have been extensively considered in the literature. For the results shown in Table 3, the time between appointments is generated using the same objective of minimizing the overall expected completion time. Specifically, to allow for a fair comparison, the time between appointments is chosen to be the smallest value such that either $\mathbb{E}[W_n] \leq SL^*$, for $2 \leq n \leq M$ (Case 1 in Table 3) or $\frac{1}{M-1}\sum_{n=2}^{M}\mathbb{E}[W_n] \leq SL^*$ (Case 2 in Table 3). Case 1 corresponds to the situation where we require that each customer experience an expected waiting time less than the threshold while Case 2 corresponds to the case where we require the expected waiting time averaged over all customers is less than the threshold. Note that, under Case 2, some customers may experience an expected waiting time that is larger than the threshold.

Note that to generate the equally spaced schedules, we assume that we wait for all customer

Table 3: Customer's expected waiting time, $\mathbb{E}[W_n]$, with equally spaced appointment schedule (exponential service times, $\mu = 0.1$, $M = 12$, $SL^* = 5$, and parameters as in Table 1)

| Case 1 | Scenarios | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 0.00 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 1.00 | 1.50 | 2.25 |
| 2 | 1.96 | 2.07 | 2.04 | 2.26 | 2.19 | 2.07 | 2.20 | 2.29 | 1.49 |
| 3 | 2.97 | 3.04 | 3.01 | 2.77 | 3.26 | 3.04 | 3.13 | 3.23 | 3.60 |
| 4 | 3.60 | 3.64 | 3.61 | 3.75 | 3.94 | 3.64 | 3.71 | 3.74 | 2.62 |
| 5 | 4.01 | 4.04 | 4.02 | 3.68 | 4.41 | 4.04 | 4.10 | 4.17 | 4.41 |
| 6 | 4.30 | 4.32 | 4.30 | 4.39 | 4.75 | 4.32 | 4.37 | 4.37 | 3.07 |
| 7 | 4.51 | 4.53 | 4.51 | 4.12 | 5.00 | 4.53 | 4.56 | 4.62 | 4.75 |
| 8 | 4.67 | 4.68 | 4.67 | 4.71 | 4.46 | 4.68 | 4.70 | 4.70 | 3.28 |
| 9 | 4.79 | 4.79 | 4.78 | 4.36 | 4.17 | 4.79 | 4.81 | 4.86 | 4.91 |
| 10 | 4.88 | 4.88 | 4.87 | 4.89 | 4.01 | 4.88 | 4.89 | 4.88 | 3.39 |
| 11 | 4.95 | 4.95 | 4.94 | 4.49 | 3.90 | 4.95 | 4.95 | 5.00 | 5.00 |
| 12 | 5.00 | 5.00 | 5.00 | 5.00 | 3.84 | 5.00 | 5.00 | 4.99 | 3.45 |
| $x^*$ | 16.29 | 15.83 | 13.59 | 14.96 | 15.25 | 16.33 | 16.47 | 16.50 | 17.56 |
| $\frac{1}{M-1}\sum_{n=2}^{M}\mathbb{E}[W_n]$ | 4.15 | 4.18 | 4.16 | 4.04 | 3.99 | 4.18 | 4.22 | 4.26 | 3.63 |
| $\mathbb{E}[C_M]$ | 194.15 | 191.08 | 166.48 | 181.60 | 183.58 | 196.66 | 200.19 | 202.51 | 212.59 |

| Case 2 | Scenarios | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 0.00 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 1.00 | 2.25 | 1.50 |
| 2 | 2.19 | 2.30 | 2.27 | 2.55 | 2.49 | 2.30 | 2.43 | 1.81 | 2.52 |
| 3 | 3.39 | 3.45 | 3.42 | 3.23 | 3.81 | 3.45 | 3.54 | 4.49 | 3.62 |
| 4 | 4.17 | 4.20 | 4.18 | 4.44 | 4.71 | 4.20 | 4.26 | 3.47 | 4.26 |
| 5 | 4.72 | 4.74 | 4.72 | 4.47 | 5.37 | 4.74 | 4.76 | 5.77 | 4.80 |
| 6 | 5.13 | 5.13 | 5.12 | 5.35 | 5.87 | 5.13 | 5.13 | 4.28 | 5.10 |
| 7 | 5.44 | 5.42 | 5.43 | 5.13 | 6.27 | 5.42 | 5.41 | 6.43 | 5.43 |
| 8 | 5.68 | 5.66 | 5.66 | 5.88 | 5.71 | 5.65 | 5.62 | 4.73 | 5.57 |
| 9 | 5.87 | 5.84 | 5.85 | 5.53 | 5.41 | 5.84 | 5.79 | 6.81 | 5.79 |
| 10 | 6.02 | 5.98 | 6.00 | 6.20 | 5.23 | 5.98 | 5.92 | 5.00 | 5.86 |
| 11 | 6.14 | 6.10 | 6.13 | 5.79 | 5.10 | 6.10 | 6.03 | 7.04 | 6.02 |
| 12 | 6.24 | 6.19 | 6.23 | 6.41 | 5.02 | 6.19 | 6.11 | 5.17 | 6.04 |
| $x^*$ | 15.21 | 14.78 | 12.54 | 13.73 | 13.97 | 15.29 | 15.48 | 15.63 | 15.56 |
| $\frac{1}{M-1}\sum_{n=2}^{M}\mathbb{E}[W_n]$ | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 |
| $\mathbb{E}[C_M]$ | 183.54 | 180.80 | 156.13 | 169.39 | 170.66 | 186.41 | 190.38 | 193.07 | 193.16 |

requests for appointments to be realized before appointment times are determined. In other words, appointment times are not generated online at the time customers request one. We do so to allow for a fair comparison between the two approaches and to generate insights into the performance of each.

Considering first the results of Case 1 (and contrasting them with the results in Table 2), we can see that requiring appointments to be equally spaced leads to higher expected completion times. The longer expected completion times are due to the fact that, in Case 1, additional slack time must be inserted to ensure that all the expected waiting times for all customers are within the threshold. Moreover, we can see that requiring appointments to be equally spaced leads to greater differences (than those observed in the results of Table 2) among the completion times under the different scenarios. This is because an approach that allows for appointments to be unequal is more effective at adjusting the schedule in accordance to the characteristics of various customers (e.g., their punctuality).

Considering the results of Case 2 (and contrasting them again with the results in Table 2), we can see that the expected completion times are much more comparable, especially in settings without too much heterogeneity in no-shows and non-punctuality. This is of course achieved at the expense of having some customers (more than half in the cases shown) experience a longer expected waiting time than the threshold $SL^*$.

We conclude this section with the following three extended remarks.

**Remark 1: Stopping criteria for customer scheduling.** We have so far assumed that the number of customers $M$ is exogenously specified. In many applications, the number of customers who can be scheduled is constrained by the total time the service facility is available. In the context of a healthcare facility, this would correspond to the facility's number of working hours in a day. In that case, customers are provided an appointment in a given day as long as the quoted appointment time falls within the facility's working hours. In other words, the last customer, denoted by $M^*$, to be given an appointment in a given day is given by $M^* = \arg\max_{n\geq 2}(d_n \leq d_{max})$, where $d_{max}$ refers to the facility's closing time (or some other threshold for the latest allowed appointment time). Alternatively, $M^*$ could be chosen so that the expected completion time is before the closing time (or some other threshold). That is, $M^* = \arg\max_{n\geq 2}(\mathbb{E}[C_n] \leq d_{max})$. Customers who cannot be scheduled in a given day are scheduled at the next feasible future day.

**Remark 2: The traditional formulation of the appointment scheduling problem.** In contrast to the online scheduling approach we describe in this paper, the approach studied in much of the existing literature is offline and assumes that appointment times are generated once the set of customers requesting appointments is known. That is, appointments are generated for all the customers at once. Such an approach is appropriate when customers do not expect an appointment immediately upon making a request and the service facility has flexibility in informing them later of their appointment times. More significantly, much of the existing literature adopts a cost-based formulation for the objective function, with appointments generated so as to minimize a weighted sum of customer waiting cost and facility service cost that is increasing in the completion time of the last customer. In other words, the objective is to generate appointment times that minimize a function of the form $G(x_1, \ldots, x_{M-1}) = G_w(W_1, \ldots, W_M) + G_c(C_M)$, where $G_w(.)$ and $G_c(.)$ are, respectively, cost functions associated with customer waiting and the utilization of the service facility. Instantiations of these functions can be found in, among others, [2, 7, 13, 44, 40], with linear functions being the most common.

Whether the online or the offline approach, with a service level constraint or a cost-based objective function, is appropriate would depend on the requirements of the application. Note that both the service level-based and the cost-based approach do address the trade-off, though differently, between customer waiting time and service facility utilization. In fact, the cost-based approach has the features of a Lagrangian relaxed version of the service level-based approach (with the constraint elevated into the objective function). Moreover, if the cost function associated with customer waiting is convex (instead of being linear), then the optimal solution generated by the cost-based approach would tend to limit the differences in waiting time across customers.

**Remark 3: Patterns in time between appointments.** It has been shown in the literature that time between appointments produced by an approach that minimizes the sum of linear costs of waiting time and service facility utilization tends to be first increasing and then decreasing (or dome-shaped). This is the case when customers are homogeneous in their no-show and non-punctuality parameters. Under the online and service level-based approach, the time between appointments is increasing when customers are homogeneous. For both formulations, customer heterogeneity breaks down the consistency in the time between appointment patterns.

# 6 An approximate approach

The exact approach described in Section 3 can be computing-intensive when the number of customers is large. Much of this computational effort is exerted in characterizing the functions $h_{n,j}$ for every combination of $n$ and $j$ (recall that $h_{n,j}$ is the pdf of the inter-arrival time between customers $n$ and $n+1$, given that customer $n$ encounters (would have encountered) $j$ customers in the system if she shows up (she does not)). In this section, we examine the extent to which approximating the function $h_{n,j}$ by the function $h_n$, the unconditional inter-arrival time between customer $n$ and $n+1$, can be effective. With this change, the state probability $p_{n,i}$ would be approximated by $\tilde{p}_{n,i}$, where $\tilde{p}_{1,0} = p_{1,0} = 1$, $\tilde{p}_{2,1} = p_{2,1}$, $\tilde{p}_{2,0} = p_{2,0}$, and

$$\begin{aligned}\tilde{p}_{n,i} = \alpha_{n-1} \sum_{j=i-1}^{n-2} \tilde{p}_{n-1,j} \int_{d_n - d_{n-1} - \tau_n^l - \tau_{n-1}^u}^{d_n - d_{n-1} + \tau_n^u + \tau_{n-1}^l} \frac{(\mu x)^{j+1-i}}{(j+1-i)!} e^{-\mu x} h_{n-1}(x)\, dx \\ + (1-\alpha_{n-1}) \sum_{j=i}^{n-2} \tilde{p}_{n-1,j} \int_{d_n - d_{n-1} - \tau_n^l - \tau_{n-1}^u}^{d_n - d_{n-1} + \tau_n^u + \tau_{n-1}^l} \frac{(\mu x)^{j-i}}{(j-i)!} e^{-\mu x} h_{n-1}(x)\, dx,\end{aligned} \tag{35}$$

with $\tilde{p}_{n,0} = 1 - \sum_{i=1}^{n-1} \tilde{p}_{n,i}$, for $3 \le n \le M$ and $1 \le i \le n-1$. Note that Equation (35) gives the exact value for $p_{n,i}$ under the case of punctual customers.

Similarly to Section 3.2.2, we can compute the waiting time distribution for each customer. In what follows we denote by $\mathbb{E}[\widetilde{W}_n]$ the approximated expected value of the waiting time for customer $n$.

To evaluate the effectiveness of the approximation, we carried out extensive numerical experiments involving a wide range of combinations of problem parameter values, including the number of customers, $M$, the supports of the punctuality distribution, $\tau^l$ and $\tau^u$, the probabilities of no-shows $\alpha_n$, the time between appointments, and the degree of parameter symmetry among customers. We also considered various distributions for punctuality and service times. Representative numerical examples are shown (additional results are available upon request) in Tables 5-8 in Appendix D. The scenarios are chosen such that we vary the value of one parameter at a time, keeping all other parameters fixed. In each case, we obtain the values $\mathbb{E}[W]$ and $\mathbb{E}[\widetilde{W}]$, the exact and approximated expected values for waiting time averaged over all the customers, where $\mathbb{E}[W] = \frac{1}{M}\sum_{n=1}^{M} \alpha_n \mathbb{E}[W_n]$ and $\mathbb{E}[\widetilde{W}] = \frac{1}{M}\sum_{n=1}^{M} \alpha_n \mathbb{E}[\widetilde{W}_n]$. We also report on the percentage difference in these values, $\Delta_W = \frac{\mathbb{E}[\widetilde{W}] - \mathbb{E}[W]}{\mathbb{E}[W]} \times 100\%$.

The following observations can be made.

- In most of the cases shown, the difference between the approximated and exact values are a few percentages (the average across all cases is less than 6%).
- The percentage difference increases with the number of customers and with the width of the range of non-punctuality, $\tau^l + \tau^u$. However, the increase in the differences appears to increase in all cases at a decreasing rate, suggesting that there may be a cap on the maximum amount of discrepancy between the approximated and exact values.
- The effect of increasing the no-show probability has a non-monotonic effect on the percentage difference between the approximated and exact value, first increasing and then decreasing, suggesting that the percentage difference increases with the uncertainty about no-shows.
- The gains in computational effort from using the approximation can be substantial when the number of customers is large. For example, for a system with 10 customers, the approximate method computes the waiting time distribution within 10 seconds while the exact method takes approximately 2.5 minutes. Within 2.5 minutes, the approximate method can evaluate the performance of a system with 30 customers. This would take approximately 4 hours for the exact method (see Figure 5 in Appendix D for an illustration).

We also observe that the approximation consistently yields an upper bound on the exact value of expected waiting time. In what follows, we provide analytical support for this observation. In particular, we show that, under the assumption that non-punctuality has the uniform distribution, that the expected waiting time for each customer is bounded above by the expected waiting produced by the approximation. We first define the following concepts.

**Definition 2.**

1. *For $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, let $x_{(1)} \leq \cdots \leq x_{(n)}$ denote the components of $x$ in increasing order. Then, $x_\uparrow = (x_{(1)}, \ldots, x_{(n)})$ is called the* ***increasing arrangement*** *of $x$.*
2. *For $x, y \in \mathbb{R}^n$, $x$ is said to be* ***weakly supermajorized*** *by $y$ (denoted as $x \prec^w y$), if*

$$\sum_1^k x_{(i)} \geq \sum_1^k y_{(i)}$$

*for* $k = 1, \ldots, n$.

**Proposition 3.** *If customer non-punctuality has the uniform distribution, i.e.,*

$$f_n(x) = \begin{cases} \frac{1}{\tau_n^l + \tau_n^u} & \text{if } x \in [d_n - \tau_n^l, d_n + \tau_n^h] \text{ and} \\ 0 & \text{otherwise,} \end{cases}$$

*then the expected waiting time of customer* $n$ *is bounded above by expected waiting time obtained from the approximation, i.e.,*

$$\mathbb{E}[W_n] \;\leq\; \mathbb{E}[\widetilde{W}_n], \tag{36}$$

*for* $1 \leq n \leq M$. *Moreover,*

$$(\tilde{p}_{n,i\geq 0}, \tilde{p}_{n,i\geq 1}, \cdots, \tilde{p}_{n,i\geq n-1}) \prec^w (p_{n,i\geq 0}, p_{n,i\geq 1}, \cdots, p_{n,i\geq n-1}), \tag{37}$$

*for* $1 \leq n \leq M$, *where* $\tilde{p}_{n,i\geq l} = \sum_{i=l}^{n-1} \tilde{p}_{n,i}$ *and* $p_{n,i\geq l} = \sum_{i=l}^{n-1} p_{n,i}$.

A sketch of the proof is as follows (a detailed proof is given in Appendix C.3). We use induction to prove (37), which by definition leads to (36) using the following inequality

$$\mathbb{E}[W_n] = \sum_{l=1}^{n-1} p_{n,i\geq l} \;\leq\; \sum_{l=1}^{n-1} \tilde{p}_{n,i\geq l} = \mathbb{E}[\widetilde{W}_n]\,. \tag{38}$$

Under the exact and approximate methods, we recursively compute $p_{n,i}$ and $\tilde{p}_{n,i}$ by first conditioning on the system states observed by customer $n$ upon her arrival (as in (8)) and further by conditioning on the inter-arrival time (as in (9)). In contrast to the exact method, where the latter conditioning is taken based on the former one, the approximate method treats the two conditionings independently. When computing $p_{n,i}$ and $\tilde{p}_{n,i}$ in the proof, we switch the order of conditionings and extend the inequality generated from the induction assumption of (37) at $n-1$ across the conditionings, which implies that (37) holds at $n$.

Finally note that the approximate method can be used instead of the exact one for the appointment scheduling problem in Section 5. In the case of uniform-distributed non-punctuality, Proposition 3 ensures that the appointment times obtained from the approximation lead to waiting times that are below the service level thresholds. In other words, the obtained solution using the

approximation is always feasible. We have repeated the experiments in Table 2 using the approximate method in order to assess its quality for the online optimization problem. The results shown in Table 4 illustrate the quality of the approximation. For each value in this table, we provide between parentheses the corresponding optimal value from Table 2. We observe that the obtained times between appointments, while being longer, are quite close to the optimal ones. The average absolute difference (excluding the punctual case of Scenario 0) is 0.099 and the average relative difference is 0.634%. This is useful given the computational efficiency of the approximation.

Table 4: Time between appointments obtained using the approximate method (exponential service times, $\mu = 0.1$, $M = 12$, $SL^* = 5$, and parameters as in Table 1)

| $n-1$ | Scenarios | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 6.93 (6.93) | 7.01 (7.01) | 4.64 (4.64) | 7.01 (7.01) | 7.01 (7.01) | 7.52 (7.52) | 8.29 (8.29) | 8.70 (8.70) | 5.46 (5.46) |
| 2 | 15.06 (15.06) | 14.64 (14.58) | 12.40 (12.35) | 12.40 (12.34) | 14.64 (14.58) | 15.15 (15.09) | 15.43 (15.21) | 15.53 (15.45) | 19.43 (19.21) |
| 3 | 15.80 (15.80) | 15.39 (15.34) | 13.15 (13.10) | 15.38 (15.33) | 15.39 (15.34) | 15.90 (15.85) | 16.18 (16.00) | 16.27 (15.90) | 12.18 (11.99) |
| 4 | 16.10 (16.10) | 15.68 (15.64) | 13.45 (13.40) | 13.45 (13.40) | 15.68 (15.64) | 16.19 (16.14) | 16.46 (16.24) | 16.56 (16.47) | 20.47 (20.29) |
| 5 | 16.24 (16.24) | 15.82 (15.78) | 13.59 (13.54) | 15.82 (15.77) | 15.82 (15.78) | 16.33 (16.28) | 16.61 (16.42) | 16.69 (16.30) | 12.61 (12.39) |
| 6 | 16.32 (16.32) | 15.91 (15.86) | 13.68 (13.63) | 13.68 (13.62) | 15.91 (15.86) | 16.42 (16.37) | 16.68 (16.53) | 16.77 (16.69) | 20.68 (20.50) |
| 7 | 16.37 (16.37) | 15.95 (15.90) | 13.72 (13.67) | 15.95 (15.90) | 13.73 (13.68) | 16.46 (16.41) | 16.73 (16.55) | 16.82 (16.43) | 12.73 (12.56) |
| 8 | 16.40 (16.40) | 15.99 (15.94) | 13.75 (13.70) | 13.75 (13.71) | 13.75 (13.71) | 16.49 (16.44) | 16.76 (16.55) | 16.85 (16.75) | 20.76 (20.57) |
| 9 | 16.42 (16.42) | 16.00 (15.96) | 13.78 (13.73) | 16.00 (15.95) | 13.78 (13.73) | 16.51 (16.47) | 16.78 (16.59) | 16.87 (16.51) | 12.78 (12.58) |
| 10 | 16.44 (16.44) | 16.02 (15.97) | 13.79 (13.74) | 13.79 (13.74) | 13.79 (13.74) | 16.53 (16.48) | 16.80 (16.59) | 16.89 (16.78) | 20.80 (20.62) |
| 11 | 16.45 (16.45) | 16.03 (15.98) | 13.80 (13.76) | 16.03 (15.98) | 13.80 (13.76) | 16.54 (16.49) | 16.80 (16.64) | 16.90 (16.55) | 12.80 (12.59) |
| $\frac{1}{M-1}\sum_{n=2}^{M}\mathbb{E}[W_n]$ | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 |
| $\mathbb{E}[C_M]$ | 183.54 | 181.44 | 156.76 | 170.27 | 170.31 | 187.04 | 192.53 | 195.84 | 191.70 |

# 7 Concluding remarks

In this paper, we studied a queueing system where the arrivals of customers are driven by appointments. We considered a continuous time setting where customers are not necessarily punctual and may not show up at all, with both punctuality and no-shows being heterogeneous across customers. We developed both exact and approximate approaches for characterizing the distribution of waiting for each customer and showed that the approximation provides an upper bound for the expected customer waiting time when non-punctuality is uniformly-distributed. We illustrated how our approach can be used to support online appointment scheduling that guarantees a specified service level for each customer. We also examined the impact of non-punctuality on system performance and provided a proof that non-punctuality deteriorates waiting time performance regardless of the distribution of inter-arrival times.

There are several avenues for future research. For example, it would be of interest to extend the analysis to systems with multiple servers, systems with walk-in customers, and systems with differing customer priorities. Such extensions would complement existing results; see for example Zacharias and Yunes [44] and Zacharias and Pinedo [43] for systems with multiple servers and walk-ins in discrete time and Wang et al. [40] for systems with walk-ins. It would also be of interest to extend the analysis by relaxing the non-overlapping assumption regarding the non-punctuality of consecutive customers. This is particularly important when the variability of non-punctuality is high (Cayirli and Veral [3]). A good starting point is Samorani and Ganguly [35], which considers the "Wait-Preempt Dilemma" arising when customers do not arrive according to their scheduled order. Finally, it would be useful to carry out comparative studies grounded in specific applications from practice to assess the real world benefit from using an appointment scheduling approach that accounts for non-punctuality and no-shows. In doing so, it would be useful to compare outcomes under different appointment scheduling approaches (e.g., online scheduling vs offline and cost-based vs. service level-based) and the sensitivity of each to non-punctuality and no-shows.

## References


[1] Amir Ahmadi-Javid, Zahra Jalali, and Kenneth Klassen. Outpatient appointment systems in healthcare: A review of optimization studies. European Journal of Operational Research, 258(1):3–34, 2017.

[2] T. Cayirli and E. Veral. Outpatient Scheduling in Health care: A Review of Literature. Production and Operations Management, 12:519–549, 2003.

[3] Tugba Cayirli and Emre Veral. Outpatient scheduling in health care: a review of literature. Production and operations management, 12(4):519–549, 2003.

[4] Tugba Cayirli, Emre Veral, and Harry Rosen. Designing appointment scheduling systems for ambulatory care services. Health care management science, 9(1):47–58, 2006.

[5] Seunggyun Cheong, Robert Bitmead, and John Fontanesi. Modeling scheduled patient punctuality in an infusion center. Lecture Notes in Management Science, 5:46–56, 2013.

[6] Chen David, Wang Rowan, Yan Zhenzhen, and Saif Benjaafar. Appointment scheduling under a service level constraint. 2021. Working paper, The Chinese University of Hong Kong, Shenzhen.

[7] Matthias Deceuninck, Dieter Fiems, and Stijn De Vuyst. Outpatient scheduling with unpunctual patients and no-shows. European Journal of Operational Research, 265(1):195–207, 2018.

[8] Jacob Feldman, Nan Liu, Huseyin Topaloglu, and Serhan Ziya. Appointment scheduling under patient preference and no-show behavior. Operations Research, 62(4):794–811, 2014.

[9] Linda V Green and Sergei Savin. Reducing delays for medical appointments: A queueing approach. Operations Research, 56(6):1526–1538, 2008.

[10] Diwakar Gupta and Brian Denton. Appointment scheduling in health care: Challenges and opportunities. IIE transactions, 40(9):800–819, 2008.

[11] Refael Hassin and Sharon Mendel. Scheduling arrivals to queues: A single-server model with no-shows. Management Science, 54(3):565–572, 2008.

[12] Birger Jansson. Choosing a good appointment system-a study of queues of the type D/M/1. Operations Research, 14(2):292–312, 1966.

[13] Bowen Jiang, Jiafu Tang, and Chongjun Yan. A stochastic programming model for outpatient appointment scheduling considering unpunctuality. Omega, 82:70–82, 2019.

[14] Ruiwei Jiang, Siqian Shen, and Yiling Zhang. Integer programming approaches for appointment scheduling with random no-shows and service durations. Operations Research, 65(6):1638–1656, 2017.

[15] Oualid Jouini and Saif Benjaafar. Appointment scheduling with non-punctual arrivals. IFAC Proceedings Volumes, 42(4):235–239, 2009.

[16] Guido Kaandorp and Ger Koole. Optimal outpatient appointment scheduling. Health Care Management Science, 10(3):217–229, 2007.

[17] Song-Hee Kim, Ward Whitt, and Won Chul Cha. A data-driven model of an appointment-generated arrival process at an outpatient clinic. INFORMS Journal on Computing, 30(1):181–199, 2018.

[18] Kenneth Klassen and Reena Yoogalingam. Appointment system design with interruptions and physician lateness. International Journal of Operations and Production Management, 33(3-4):394–414, 2013.

[19] Alex Kuiper, Benjamin Kemper, and Michel Mandjes. A computational approach to optimized appointment scheduling. Queueing Systems, 79(1):5–36, 2015.

[20] Alex Kuiper, Michel Mandjes, and Jeroen de Mast. Optimal stationary appointment schedules. Operations Research Letters, 45(6):549–555, 2017.

[21] Linda R LaGanga and Stephen R Lawrence. Appointment overbooking in health care clinics to improve patient service and clinic performance. Production and Operations Management, 21(5):874–888, 2012.

[22] Ho-Shiang Lau and Amy Hing-Ling Lau. A fast procedure for computing the total system cost of an appointment schedule for medical and kindred facilities. IIE Transactions, 32(9):833–839, 2000.

[23] Benjamin Legros, Oualid Jouini, and Ger Koole. A uniformization approach for the dynamic control of queueing systems with abandonments. Operations Research, 66(1):200–209, 2018.

[24] Jianzhe Luo, Vidyadhar G Kulkarni, and Serhan Ziya. Appointment scheduling under patient no-shows and service interruptions. Manufacturing & Service Operations Management, 14(4):670–684, 2012.

[25] Ho-Yin Mak, Ying Rong, and Jiawei Zhang. Appointment scheduling with limited distributional information. Management Science, 61(2):316–334, 2015.

[26] Albert W Marshall, Ingram Olkin, and Barry C Arnold. Inequalities: theory of majorization and its applications, volume 143. Springer, 1979.

[27] A Mercer. Queues with scheduled arrivals: A correction, simplification and extension. Journal of the Royal Statistical Society: Series B (Methodological), 35(1):104–116, 1973.

[28] William P Millhiser and B.C. Valenti. Delay distributions in appointment systems with generally and non-identically distributed service times and no-shows. 2012. Available on SSRN: http://ssrn.com/abstract=2045074.

[29] William P Millhiser and Emre A Veral. Designing appointment system templates with operational performance targets. IIE Transactions on Healthcare Systems Engineering, 5(3):125–146, 2015.

[30] William P Millhiser, Emre A Veral, and Benedetto C Valenti. Assessing appointment systems' operational performance with policy targets. IIE Transactions on Healthcare Systems Engineering, 2(4):274–289, 2012.

[31] Dragoslav S Mitrinovic and Petar M Vasic. Analytic inequalities, volume 61. Springer, 1970.

[32] Iman Mohammadi, Huanmei Wu, Ayten Turkcan, Tammy Toscos, and Bradley N Doebbeling. Data analytics and modeling for appointment no-show in community health centers. Journal of primary care & community health, 9:2150132718811692, 2018.

[33] Mahmut Parlar and Moosa Sharafali. Dynamic allocation of airline check-in counters: a queueing optimization approach. Management Science, 54(8):1410–1424, 2008.

[34] Lawrence W Robinson and Rachel R Chen. A comparison of traditional and open-access policies for appointment scheduling. Manufacturing & Service Operations Management, 12(2):330–346, 2010.

[35] Michele Samorani and Subhamoy Ganguly. Optimal sequencing of unpunctual patients in high-service-level clinics. Production and Operations Management, 25(2):330–346, 2016.

[36] Alfonso Soriano. Comparison of two scheduling systems. Operations Research, 14(3):388–397, 1966.

[37] Patrick Wang. Optimally scheduling n customer arrival times for a single-server system. Computers & Operations Research, 24(8):703–716, 1997.

[38] Patrick Wang. Sequencing and scheduling n customers for a stochastic server. European journal of operational research, 119(3):729–738, 1999.

[39] Rowan Wang, Oualid Jouini, and Saif Benjaafar. Service systems with finite and heterogeneous customer arrivals. Manufacturing & Service Operations Management, 16(3):365–380, 2014.

[40] Shan Wang, Nan Liu, and Guohua Wan. Managing appointment-based services in the presence of walk-in customers. Management Science, 66(2):667–686, 2020.

[41] Fan Yue, Qiying Hu, Fan Yue, Qiying Hu, Fan Yue, Qiying Hu, Fan Yue, and Qiying Hu. Minimizing total cost in outpatient scheduling with unpunctual arrivals. In International Conference on Service Systems and Service Management, 2016.

[42] Christos Zacharias and Mor Armony. Joint panel sizing and appointment scheduling in outpatient care. Management Science, 63(11):3978–3997, 2017.

[43] Christos Zacharias and Michael Pinedo. Managing customer arrivals in service systems with multiple identical servers. Manufacturing & Service Operations Management, 19(4):639–656, 2017.

[44] Christos Zacharias and Tallys Yunes. Multimodularity in the stochastic appointment scheduling problem with discrete arrival epochs. Management Science, 66(2):744–763, 2020.

[45] Bo Zeng, Ayten Turkcan, Ji Lin, and Mark Lawley. Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. Annals of Operations Research, 178(1):121–144, 2010.

[46] Han Zhu, Youhua Chen, Eman Leung, and Xing Liu. Outpatient appointment scheduling with unpunctual patients. International Journal of Production Research, 56(5):1982–2002, 2018.

## Acknowledgement

The first and last authors were funded by NPRPC Grant No. NPRP11C-1229-170007 from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the authors.

## Appendix A Triangular distribution for punctuality

Consider the case where customer non-punctuality is homogeneous and follows a symmetric triangular distribution. For $1 \leq n \leq M$, we have

$$f_n(x) = \begin{cases} \frac{(x-d_n+\tau)}{\tau^2} & \text{if } d_n - \tau \leq x < d_n, \\ \frac{(d_n+\tau-x)}{\tau^2} & \text{if } d_n \leq x \leq d_n + \tau, \text{ and} \\ 0 & \text{otherwise.} \end{cases} \tag{39}$$

Similarly to the uniform distribution case, we calculate $p_{2,1}$ to complete the initialization step, since we have $f_1(t)$ from the definition of $f_n(t)$ and $p_{2,0} = 1 - p_{2,1}$. Here, we have $\Pr\{D_1 < d_1\} = \Pr\{D_1 \geq d_1\} = \frac{1}{2}$. When customer 1 arrives before $d_1$, we have

$$\begin{aligned} p_{2,1|D_1<d_1} &= \int_{d_2-d_1-\tau}^{d_2-d_1+\tau} e^{-\mu x} f_2(x+d_1)\, dx \\ &= \int_{d_2-d_1-\tau}^{d_2-d_1} e^{-\mu x} \frac{(x+d_1-d_2+\tau)}{\tau^2}\, dx + \int_{d_2-d_1}^{d_2-d_1+\tau} e^{-\mu x} \frac{(d_2+\tau-x-d_1)}{\tau^2}\, dx \\ &= \frac{(e^{\mu\tau}-1)^2\, e^{-\mu(-d_1+d_2+\tau)}}{\mu^2\tau^2}. \end{aligned} \tag{40}$$

To derive $p_{2,1|D_1>d_1}$, we first compute $h_{2,1|D_1\geq d_1}(x)$, the pdf of the random variable $D_2-D_1|D_1\geq d_1$. Using (7), we can compute $h_{1|D_1\geq d_1}(x)$ on $[d_2-d_1-\tau_2^l-\tau_1^u, d_2-d_1+\tau_2^u]$ as

$$\begin{aligned}
h_{2,1|D_1\geq d_1}(x) &= \int_{\max\{d_2-\tau,\, d_1+x\}}^{\min\{d_2+\tau,\, d_1+\tau+x\}} f_2(u) f_{1|D_1\geq d_1}(u-x)\, du \\
&= \int_{\max\{d_2-\tau,\, d_1+x\}}^{\min\{d_2+\tau,\, d_1+\tau+x\}} f_2(u) \frac{f_1(u-x)}{\bar{F}_1(d_1)}\, du \\
&= \frac{2}{\tau^4}\Bigg( \mathbb{1}_{\{d_1+x\leq d_2\}} \int_{\max\{d_2-\tau,\, d_1+x\}}^{d_2} (u-d_2+\tau)(d_1+\tau-u+x)\, du \\
&\qquad + \mathbb{1}_{\{d_2\leq d_1+x+\tau\}} \int_{d_2}^{\min\{d_2+\tau,\, d_1+\tau+x\}} (d_2+\tau-u)(d_1+\tau-u+x)\, du \Bigg) \\
&= \frac{2}{\tau^4}\Bigg[ \mathbb{1}_{\{d_1+x\leq d_2\}} C_{min} \left( \tau + \frac{(C_{min})^2}{6\tau} - \frac{C_{min}}{2} - \frac{1}{2}(d_2-d_1-x) \right) \\
&\qquad + \mathbb{1}_{\{d_2\leq d_1+x+\tau\}} D_{min} \left( \frac{1}{2}(x+d_1-d_2+\tau) - \frac{1}{6\tau} {D_{min}}^2 \right) \Bigg],
\end{aligned} \tag{41}$$

where $C_{min} = \min\{\tau, d_2-d_1-x\}$ and $D_{min} = \min\{\tau, d_1+\tau+x-d_2\}$. By substituting $h_{1|D_1\geq d_1}(x)$ in Equation (4) by (41), we obtain

$$\begin{aligned}
p_{2,1|D_1\geq d_1} &= \int_{d_2-d_1-\tau_2^l-\tau_1^u}^{d_2-d_1+\tau_2^u} e^{-\mu x} h_{1|D_1\geq d_1}(x)\, dx \\
&= \int_{d_2-d_1-\tau_2^l-\tau_1^u}^{d_2-d_1} e^{-\mu x} C_{min} \left( \tau_1^u + \frac{(C_{min})^2}{6\tau_2^l} - \frac{\tau_1^u C_{min}}{2\tau_2^l} - \frac{1}{2}(d_2-d_1-x) \right) dx \\
&\quad + \int_{d_2-d_1-\tau_1^u}^{d_2-d_1+\tau_2^u} e^{-\mu x} D_{min} \left( \frac{1}{2}(x+d_1-d_2+\tau_1^u) - \frac{1}{6\tau_2^u} {D_{min}}^2 \right) dx \\
&= -\frac{\left(\mu^2\tau^2 e^{3\mu\tau}(2\mu\tau-3) + \mu^2\tau^2(5\mu\tau+3) + 12e^{2\mu\tau} - 6e^{\mu\tau}(\mu\tau(\mu\tau+2)+2)\right) e^{-\mu(-d_1+d_2+\tau)}}{3\mu^4\tau^4}.
\end{aligned} \tag{42}$$

With (40) and (42), we complete the initialization step by computing $p_{2,1}$ as

$$\begin{aligned}
p_{2,1} &= \alpha_1 \Pr\{D_1 < d_1\} p_{2,1|D_1<d_1} + \alpha_1 \Pr\{D_1 \geq d_1\} p_{2,1|D_1\geq d_1} \\
&= \frac{\alpha_1 \left(-5\mu^3\tau^3 - \mu^2\tau^2 e^{3\mu\tau}(2\mu\tau-3) + 3e^{2\mu\tau}\left(\mu^2\tau^2-4\right) + 12e^{\mu\tau}(\mu\tau+1)\right) e^{-\mu(-d_1+d_2+\tau)}}{6\mu^4\tau^4}.
\end{aligned}$$

# Appendix B Details for the analysis of $\gamma-$Cox-distributed service times

The moments of the waiting time can be obtained similarly to the exponential service time case. We have

$$\mathbb{E}[W_n^k] = \sum_{r=1}^{(n-1)m} p_{n,r}\mathbb{E}[W_{n,r}^k], \tag{43}$$

for $2 \leq n \leq M$, where $W_{n,r}$ is the random variable denoting the waiting time in the queue of customer $n$, given that customer $n$ shows up and finds $r$ actual phases of service in system remain to be serviced (i.e., the system state $R_n$ is $r$). Since service times follows independent $\gamma$-Cox distribution with $m$ phases, the completion time of each phase is independently and exponentially-distributed with rate $\gamma$. Therefore, $W_{n,r}$ has an $r$-Erlang distribution with $r$ phases and rate $\gamma$ per phase. Using Equation (43) and knowing that $\mathbb{E}[W_{n,r}] = \dfrac{r}{\gamma}$ and $\mathbb{E}[W_{n,r}^2] = \dfrac{r(r+1)}{\gamma^2}$, we obtain

$$\mathbb{E}[W_n] = \sum_{r=1}^{(n-1)m} p_{n,r}\frac{r}{\gamma} \text{ and } \mathbb{E}[W_n^2] = \sum_{r=1}^{(n-1)m} p_{n,r}\frac{r(r+1)}{\gamma^2}, \tag{44}$$

for $2 \leq n \leq M$. Moreover, we have

$$\Pr\{W_{n,r} < t\} = 1 - \sum_{j=0}^{r-1} \frac{(\gamma t)^j}{j!}\, e^{-\gamma t}, \tag{45}$$

for $t \geq 0$. Consequently,

$$\begin{aligned} \Pr\{W_n < t\} &= p_{n,0} + \sum_{r=1}^{(n-1)m} p_{n,r}\Pr\{W_{n,r} < t\} \\ &= 1 - \sum_{r=1}^{(n-1)m}\sum_{j=0}^{r-1} p_{n,r}\frac{(\gamma t)^j}{j!}\, e^{-\gamma t}. \end{aligned} \tag{46}$$

The case $n = 1$ is treated separately. The moments and distribution of the first customer's waiting time can be obtained exactly the same as for the exponential case, using Equations (19)-(21).

# Appendix C Proof of Propositions

## C.1 Proof of Proposition 1

We first state several definitions and lemmas that will be used in the proof. We denote by $S_n$ the random variable of the service time of customer $n$, and by $A_n$ the random variable for the arrival time of customer $n$. For a given schedule $\delta = (d_1, d_2, \cdots, d_M)$, we have $A_n = D_n$ if customer $n$ is not punctual, and $A_n = \mathbb{E}[D_n]$ if customer $n$ is punctual.

We use $\gamma_n \in \{0, 1\}$ to denote the type of punctuality of customer $n$, where $\gamma_n = 0$ if customer $n$ arrives with non-punctuality at time $D_n \in [d_n - \tau_n^l, d_n + \tau_n^u]$, and $\gamma_n = 1$ if customer $n$ arrives with punctuality at time $\mathbb{E}[D_n]$. Let us denote by $\Gamma = (\gamma_1, \cdots, \gamma_M)$ the customer's punctuality profile and use $A_n(\Gamma)$, $W_n(\Gamma)$, and $C_n(\Gamma)$ to represent the arrival, waiting, and completion time of customer $n$ under the profile $\Gamma$, where $A_n(\Gamma) = D_n$ if $\gamma_n = 0$, and $A_n(\Gamma) = \mathbb{E}[D_n]$ if $\gamma_n = 1$. For $0 \leq k \leq M$, let $\Gamma_k$ denote the profile where the first $k$ customers are punctual and the last $M - k$ customers are non-punctual.

For $a_k, s_k \in \mathbb{R}$, let $\vec{a}_k = (a_1, \cdots, a_k)$, and $\vec{s}_k = (s_1, \cdots, s_k)$. For $k = 0, \cdots, M$, we define the function $h_k(\vec{a}_k, \vec{s}_k) : \mathbb{R}^{2k} \to \mathbb{R}$ as follows:

$$
\begin{aligned}
&h_0(\vec{a}_0, \vec{s}_0) = h_0 = d_1, \\
&h_k(\vec{a}_k, \vec{s}_k) = \max(h_{k-1}(\vec{a}_{k-1}, \vec{s}_{k-1}), a_k) + s_k \quad \text{for } k = 1, \cdots, M.
\end{aligned}
$$

**Proposition 4.** *For $k \in \{1, \cdots, M\}$, $h_n(\vec{a}_n, \vec{s}_n)$ is convex with respect to $a_k$ for $k \leq n \leq M$.*

*Proof.* First, note that the function $h_{k-1}(\vec{a}_{k-1}, \vec{s}_{k-1})$ only relies on the first $k-1$ elements of $\vec{a}_n$ and $\vec{s}_n$, and is constant with respect to $a_l$ and $s_l$, for $l \geq k$. From standard results, we know that $\max(C, f(x))$ is convex in $x$ if $f(x)$ is convex in $x$ and $C$ is constant with respect to $x$. It is then easy to see that $h_k(\vec{a}_k, \vec{s}_k) = \max(h_{k-1}(\vec{a}_{k-1}, \vec{s}_{k-1}), a_k) + s_k$ is convex in $a_k$. Let us consider $n > k$ and assume that $h_{n-1}(\vec{a}_{n-1}, \vec{s}_{n-1})$ is convex in $a_k$. Again, we can see that $h_n(\vec{a}_n, \vec{s}_n) = \max(h_{n-1}(\vec{a}_{n-1}, \vec{s}_{n-1}), a_n) + s_{m+1}$ is convex in $a_k$. Therefore, by induction, we have shown that $h_n(\vec{a}_n, \vec{s}_n)$ is convex in $a_k$ for all $n \geq k$, which finishes the proof of the proposition. □

For $1 \leq k \leq n \leq M$, we define the functions $g_{n,k}$ and $\hat{g}_{n,k} : \mathbb{R}^{2n-2} \to \mathbb{R}$ as follows. If $n = k$,

$$g_{k,k}((a_1, \cdots, a_{k-1}), \vec{s}_{k-1}) = \mathbb{E}[(h_{k-1}((a_1, \cdots, a_{k-1}), \vec{s}_{k-1}) - D_k)^+],$$
$$\hat{g}_{k,k}((a_1, \cdots, a_{k-1}), \vec{s}_{k-1}) = (h_{k-1}((a_1, \cdots, a_{k-1}), \vec{s}_{k-1}) - \mathbb{E}[D_k])^+.$$

Otherwise, for $n > k$,

$$g_{n,k}((a_1, \cdots, a_{k-1}, a_{k+1}, \cdots, a_n), \vec{s}_{n-1}) = \mathbb{E}[(h_{n-1}((a_1, \cdots, a_{k-1}, D_k, a_{k+1}, \cdots, a_{n-1}), \vec{s}_{n-1}) - a_n)^+],$$
$$\hat{g}_{n,k}((a_1, \cdots, a_{k-1}, a_{k+1}, \cdots, a_n), \vec{s}_{n-1}) = (h_{n-1}((a_1, \cdots, a_{k-1}, \mathbb{E}[D_k], a_{k+1}, \cdots, a_{n-1}), \vec{s}_{n-1}) - a_n)^+.$$

By applying Jensen's inequality and Proposition 4, we may write

$$g_{n,k} \geq \hat{g}_{n,k}, \tag{47}$$

uniformly on $\mathbb{R}^{2n-2}$, for $1 \leq k \leq n \leq M$.

Next, we show that for a fixed schedule $\delta = (d_1, d_2, \cdots, d_M)$, the expected waiting time of all customers decreases as we have more punctual customers at the beginning of the schedule. In particular, we want to show

$$\mathbb{E}[W_n(\Gamma_{k-1})] \geq \mathbb{E}[W_n(\Gamma_k)] \quad \text{for } n = 1, 2, \cdots, M, \tag{48}$$

for every $k = 1, 2, \cdots, M$. This means the customer's expected waiting time under the case where all customers are non-punctual (i.e., $\mathbb{E}[W_n(\Gamma_0)]$) is higher than that under the case where all customers are punctual (i.e., $\mathbb{E}[W_n(\Gamma_M)]$).

Let $C_0 = d_1$. Therefore, the waiting time and completion time of each customer can be characterized by the following equations:

$$W_n = (C_{n-1} - A_n)^+,$$
$$C_n = \max(C_{n-1}, A_n) + S_n\,.$$

Let $\vec{A}_n(\Gamma) = (A_1(\Gamma), A_2(\Gamma), \cdots, A_n(\Gamma))$ and $\vec{S}_n = (S_1, S_2, \cdots, S_n)$ denote respectively the random vectors of the arrival times and service times of the first $n$ customers with the punctuality profile

$\Gamma$. It follows that

$$W_n(\Gamma) = (C_{n-1}(\Gamma) - A_n(\Gamma))^+ \quad \text{for } n = 1, \cdots, M, \text{ and}$$

$$C_n(\Gamma) = \begin{cases} h_0 = d_1 & \text{for } n = 0 \\ h_n(\vec{A}_n(\Gamma), \vec{S}_n) & \text{for } n = 1, \cdots, M. \end{cases}$$

Consider a fixed $k = 1, \cdots, M$. The first $k-1$ customers are punctual under both profiles $\Gamma_{k-1}$ and $\Gamma_k$. Hence, the expected waiting for customer $n < k$ are the same (i.e., $\mathbb{E}[W_n(\Gamma_{k-1})] = \mathbb{E}[W_n(\Gamma_k)]$ for $n < k$).

For customer $n = k$, there are two possible cases. If $n = k = 1$, we have $\mathbb{E}[W_1(\Gamma_0)] = \mathbb{E}[(C_0(\Gamma_0) - A_1(\Gamma_0))^+] = \mathbb{E}[(d_1 - D_1)^+] \geq (d_1 - \mathbb{E}[D_1])^+ = \mathbb{E}[(d_1 - \mathbb{E}[D_1])^+] = \mathbb{E}[(C_0(\Gamma_1) - A_1(\Gamma_1))^+] = \mathbb{E}[W_1(\Gamma_1)]$, where the inequality is obtained by applying Jensen's inequality. Otherwise, for $n = k > 1$, and we have

$$\begin{aligned}
&\mathbb{E}[W_k(\Gamma_{k-1})] \\
=&\mathbb{E}[(C_{k-1}(\Gamma_{k-1}) - A_k(\Gamma_{k-1}))^+] \\
=&\mathbb{E}[(h_{k-1}(\vec{A}_{k-1}(\Gamma_{k-1}), \vec{S}_{k-1}) - D_k)^+] \\
=&\mathbb{E}[\mathbb{E}[(h_{k-1}(\vec{A}_{k-1}(\Gamma_{k-1}), \vec{S}_{k-1}) - D_k)^+ \mid A_1(\Gamma_{k-1}), \cdots, A_{k-1}(\Gamma_{k-1}), \vec{S}_{k-1}]] \\
=&\mathbb{E}[g_{k,k}(\vec{A}_{k-1}(\Gamma_{k-1}), \vec{S}_{k-1})] \\
=&\mathbb{E}[g_{k,k}(\vec{A}_{k-1}(\Gamma_k), \vec{S}_{k-1})] \\
\geq&\mathbb{E}[\hat{g}_{k,k}(\vec{A}_{k-1}(\Gamma_k), \vec{S}_{k-1})] \\
=&\mathbb{E}[(h_{k-1}(\vec{A}_{k-1}(\Gamma_k), \vec{S}_{k-1}) - \mathbb{E}[D_k])^+] \\
=&\mathbb{E}[(C_{k-1}(\Gamma_k) - A_k(\Gamma_k))^+] \\
=&\mathbb{E}[W_k(\Gamma_k)].
\end{aligned}$$

The inequality is due to (47) and also the fact that the functions $g_{n,k}$ and $\hat{g}_{n,k}$ are integrable given

the finite range of customer's non-punctuality. Similarly, for customer $n > k$, we obtain

$$
\begin{aligned}
&\mathbb{E}[W_n(\Gamma_{k-1})] \\
=&\mathbb{E}[(C_{n-1}(\Gamma_{k-1}) - A_n(\Gamma_{k-1}))^+] \\
=&\mathbb{E}[(h_{n-1}(\vec{A}_{n-1}(\Gamma_{k-1}), \vec{S}_{n-1}) - A_n(\Gamma_{k-1}))^+] \\
=&\mathbb{E}[\mathbb{E}[(h_{n-1}(\vec{A}_{n-1}(\Gamma_{k-1}), \vec{S}_{n-1}) - A_n(\Gamma_{k-1}))^+ \mid A_1(\Gamma_{k-1}), \cdots, A_{k-1}(\Gamma_{k-1}), A_{k+1}(\Gamma_{k-1}), \cdots, A_n(\Gamma_{k-1}), \vec{S}_{n-1}]] \\
=&\mathbb{E}[g_{n,k}(A_1(\Gamma_{k-1}), \cdots, A_{k-1}(\Gamma_{k-1}), A_{k+1}(\Gamma_{k-1}), \cdots, A_n(\Gamma_{k-1}), \vec{S}_{n-1})] \\
=&\mathbb{E}[g_{n,k}(A_1(\Gamma_k), \cdots, A_{k-1}(\Gamma_k), A_{k+1}(\Gamma_k), \cdots, A_n(\Gamma_k), \vec{S}_{n-1})] \\
\geq&\mathbb{E}[\hat{g}_{n,k}(A_1(\Gamma_k), \cdots, A_{k-1}(\Gamma_k), A_{k+1}(\Gamma_k), \cdots, A_n(\Gamma_k), \vec{S}_{n-1})] \\
=&\mathbb{E}[(h_{n-1}(A_1(\Gamma_k), \cdots, A_{k-1}(\Gamma_k), \mathbb{E}[D_k], A_{k+1}(\Gamma_k), \cdots, A_{n-1}(\Gamma_k), \vec{S}_{n-1}) - A_n(\Gamma_k))^+] \\
=&\mathbb{E}[(h_{n-1}(\vec{A}_{n-1}(\Gamma_k), \vec{S}_{n-1}) - A_n(\Gamma_k))^+] \\
=&\mathbb{E}[(C_{n-1}(\Gamma_k) - A_n(\Gamma_k))^+] \\
=&\mathbb{E}[W_n(\Gamma_k)].
\end{aligned}
$$

In conclusion, we have proved that (47) holds and that $\mathbb{E}[W_n(\Gamma_0)] \geq \mathbb{E}[W_n(\Gamma_M)]$, which finishes the proof of the proposition. □

## C.2 Proof of Proposition 2

*Proof.* Consider customer $n$ with the two possible appointment times $\hat{d}_n$ and $d_n$, such that $\hat{d}_n > d_n$. For these two appointment times, the random variables $\widehat{D}_n$ and $D_n$ correspond to the arrival times, $\widehat{V}_n = C_{n-1} - \widehat{D}_n$ and $V_n = C_{n-1} - D_n$ correspond to the difference between the completion time of customer $n-1$ and the arrival time of customer $n$, and $\widehat{W}_n = \max(0, \widehat{V}_n)$ and $W_n = \max(0, V_n)$ correspond to the waiting times, respectively. In what follows, we prove that $W_n$ FOS dominates $\widehat{W}_n$.

For $t \in [\hat{d}_n - \tau_n^l, \hat{d}_n + \tau_n^u]$, we have $f_{\widehat{V}_n}(t) = f_{V_n}(t - (\hat{d}_n - d_n))$. Since $\hat{d}_n - d_n > 0$, $\widehat{D}_n$ is stochastically larger than $D_n$. In other words, $\widehat{D}_n$ FOS dominates $D_n$. Thus, $-\widehat{D}_n$ FOS dominates $-D_n$, which implies given the independence of $C_{n-1}$, $\widehat{D}_n$ and $D_n$ that $C_{n-1} - \widehat{D}_n$ FOS dominates $C_{n-1} - D_n$. Since $\max(\cdot, 0)$ is a non-decreasing function, we have $\widehat{W}_n$ FOS dominates $W_n$, which completes the proof of the proposition. □

## C.3 Proof of Proposition 3

When arrival times are uniformly-distributed, we would like to prove that the expected waiting time of customer $n$ computed using the exact method, $\mathbb{E}[W_n]$, is bounded above by the one computed using the approximate method, $\mathbb{E}[\widetilde{W}_n]$. In other words, we want to show that

$$\mathbb{E}[W_n] \;\leq\; \mathbb{E}[\widetilde{W}_n] \quad \text{, for} \quad 1 \leq n \leq M, \tag{49}$$

with $f_n(x) = \dfrac{1}{\tau_n^l + \tau_n^u}$ on $[d_n - \tau_n^l, d_n + \tau_n^u]$ and $f_n(x) = 0$ otherwise.

Moreover, we want to prove

$$(\tilde{p}_{n,i\geq 0}, \tilde{p}_{n,i\geq 1}, \cdots, \tilde{p}_{n,i\geq n-1}) \prec^w (p_{n,i\geq 0}, p_{n,i\geq 1}, \cdots, p_{n,i\geq n-1}),$$

for $1 \leq n \leq M$, where $\tilde{p}_{n,i\geq l} = \sum_{i=l}^{n-1} \tilde{p}_{n,i}$ and $p_{n,i\geq l} = \sum_{i=l}^{n-1} p_{n,i}$.

First, we state several definitions and results that will be used throughout the proof. For any $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, let $x_{[1]} \geq \cdots \geq x_{[n]}$ denote the components of $x$ in decreasing order, and let $x_\downarrow = (x_{[1]}, \ldots, x_{[n]})$ denote the *decreasing rearrangement* of $x$. Similarly, let $x_{(1)} \leq \cdots \leq x_{(n)}$ denote the components of x in increasing order, and let $x_\uparrow = (x_{(1)}, \ldots, x_{(n)})$ denote the *increasing arrangement* of $x$. Let $\mathbb{D}$ denote the subspace of descending vectors in $\mathbb{R}^n$, in particular $\mathbb{D} = \{(x_1, \ldots, x_n) : x_1 \geq \cdots \geq x_n\}$. Similarly, we have $\mathbb{D}^+ = \{(x_1, \ldots, x_n) : x_1 \geq \cdots \geq x_n \geq 0\}$.

**Definition 3.** *For $x, y \in \mathbb{R}^n$,*

$$x \prec_w y \qquad if \quad \sum_1^k x_{[i]} \leq \sum_1^k y_{[i]}, \quad k = 1, \ldots, n,$$

*and*

$$x \prec^w y \qquad if \quad \sum_1^k x_{(i)} \geq \sum_1^k y_{(i)}, \quad k = 1, \ldots, n.$$

*$x$ is said to be **weakly submajorized** by $y$, if $x \prec_w y$, and $x$ is said to be **weakly supermajorized** by $y$, if $x \prec^w y$. In either case, $x$ is said to be **weakly majorized** by $y$ ($y$ **weakly majorizes** $x$). Moreover, $x$ is said to be **majorized** by $y$ ($y$ **majorizes** $x$), denoted by $x \prec y$ if both cases hold.*

It is easy to see that

$$x \prec y \Leftrightarrow -x \prec -y, \tag{50}$$

$$x \prec_w y \Leftrightarrow -x \prec^w -y. \tag{51}$$

**Theorem 1.** *(Theorem A.7, p.86 in [26])*

*Let $\phi$ be a real-valued function, defined and continuous on $\mathbb{D}$, and continuously differentiable on the interior of $\mathbb{D}$. Denote the partial derivative of $\phi$ with respect to its kth argument by $\phi_{(k)}$: $\phi_{(k)}(z) = \partial\phi(z)/\partial z_k$. Then*

$$\phi(x) \leq \phi(y) \quad \textit{whenever} \quad x \prec_w y \textit{ on } \mathbb{D},$$

*if and only if,*

$$\phi_{(1)}(z) \geq \phi_{(2)}(z) \geq \cdots \geq \phi_{(n)}(z) \geq 0,$$

*i.e., the gradient $\nabla\phi(z) \in \mathbb{D}$, for all $z$ in the interior of $\mathbb{D}$.*

**Lemma 1.** *(Theorem H.3.b, p.136 in [26])*

*If $x, y \in \mathbb{D}$ and $x \prec_w y$, then*

$$\sum x_i u_i \leq \sum y_i u_i \qquad \textit{for all} \quad u \in \mathbb{D}^+.$$

**Proposition 5.** *If $x, y \in \mathbb{D}$ and $y \prec^w x$, then for each $k \in \{1, \ldots, n\}$, we have*

$$\sum_{i=k}^{n} x_i u_{i-k+1} \leq \sum_{i=k}^{n} y_i u_{i-k+1} \qquad \textit{, for all} \quad u \in \mathbb{I}^+,$$

*where $\mathbb{I}^+ = \{(x_1, \ldots, x_n) : 0 \leq x_1 \leq \cdots \leq x_n\}$.*

*Proof.* Take $x, y \in \mathbb{D}$ with $y \prec^w x$. Let $\hat{x}$ be the reverse arrangement of $x$, in particular

$$\hat{x}_i = x_{n+1-i} \quad \text{, for } i \in \{1, \ldots, n\},$$

and by definition we have $\hat{y} \prec^w \hat{x}$. Using Equation (51), we have $-\hat{y} \prec_w -\hat{x}$ with $-\hat{x} \in \mathbb{D}$ and $-\hat{y} \in \mathbb{D}$. Take $u \in \mathbb{I}^+$ and let $\hat{u}$ be the reverse arrangement of $u$. We have $\hat{u} \in \mathbb{D}^+$. Moreover, for

any $k \in \{1, \ldots, n\}$, we have

$$(-\hat{y}_1, \cdots, -\hat{y}_{n-k+1}) \prec_w (-\hat{x}_1, \cdots, -\hat{x}_{n-k+1}),$$

$$(-\hat{y}_1, \cdots, -\hat{y}_{n-k+1}) \in \mathbb{D}, \quad (-\hat{x}_1, \cdots, -\hat{x}_{n-k+1}) \in \mathbb{D},$$

$$(\hat{u}_k, \cdots, \hat{u}_n) \in \mathbb{D}^+.$$

It follows from Lemma 1 that

$$\sum_{i=1}^{n-k+1} -\hat{y}_i \hat{u}_{i+k-1} \leq \sum_{i=1}^{n-k+1} -\hat{x}_i \hat{u}_{i+k-1},$$

and therefore we obtain

$$\begin{aligned}
\sum_{i=k}^{n} x_i u_{i-k+1} &= \sum_{i=k}^{n} \hat{x}_{n+1-i} \hat{u}_{n+k-i} \\
&= \sum_{i=1}^{n-k+1} \hat{x}_i \hat{u}_{i+k-1} \\
&\leq \sum_{i=1}^{n-k+1} \hat{y}_i \hat{u}_{i+k-1} \\
&= \sum_{i=k}^{n} \hat{y}_{n+1-i} \hat{u}_{n+k-i} = \sum_{i=k}^{n} y_i u_{i-k+1}.
\end{aligned}$$

Note that this proposition could be also proven by applying Theorem 1 with $\phi(z) = \sum_{i=1}^{k} -z_{n+1-i} u_{k+1-i}$, for $k \in \{1, \ldots, n\}$. □

**Theorem 2** (Chebyshev Integral Inequality)**.** *(Theorem 9, p.39 in [31])*
*Let $f$ and $g$ be real and integrable functions on $[a, b]$ and let them both be either increasing or decreasing. Then*

$$\frac{1}{b-a} \int_a^b f(x) g(x) dx \geq \frac{1}{b-a} \int_a^b f(x) dx \, \frac{1}{b-a} \int_a^b g(x) dx.$$

*If one function is increasing and the other is decreasing, the reverse inequality holds.*

We use the tilde sign to denote variables computed using the approximation method, such as

$\widetilde{W}_n$ for customer's waiting time. Let us define the following notations:

$$p_{n,i\geq l} = \sum_{i=l}^{n-1} p_{n,i},$$

and

$$p_{n,i\leq l} = \sum_{i=0}^{l} p_{n,i},$$

for $0 \leq l \leq n-1$, and similarly for $\tilde{p}_{n,i\geq l}$ and $\tilde{p}_{n,i\leq l}$. We use $g(n,\lambda)$ and $G(n,\lambda)$ to denote the pdf and cdf of a Poisson distribution with rate $\lambda$. We have $g(n,\lambda) = \frac{\lambda^n}{n!}e^{-\lambda}$ and $G(n,\lambda) = \sum_{i=0}^{n} \frac{\lambda^i}{i!}e^{-\lambda}$ if $n \geq 0$, and $g(n,\lambda) = G(n,\lambda) = 0$ otherwise.

Let us recall the differences between the exact and approximate methods we developed in the main paper. Instead of the conditional distribution of inter-arrival time $h_{n,j}(\cdot)$ used in the exact method (as shown in Equation (9)), we use the unconditional inter-arrival time distribution $h_n(\cdot)$ as the approximation for $3 \leq n \leq M$ (as shown in Equation (35)). Since no approximation is involved in the computation of $\tilde{p}_{n,i}$ for $n = 1, 2$, the expected waiting time for customers 1 and 2 are the same from both methods. Therefore, to prove Equation (49), we only need to show

$$\mathbb{E}[W_n] \;\leq\; \mathbb{E}[\widetilde{W}_n] \quad \text{, for} \quad 3 \leq n \leq M,$$

which is equivalent to

$$\sum_{l=1}^{n-1} p_{n,i\geq l} \;\leq\; \sum_{l=1}^{n-1} \tilde{p}_{n,i\geq l}, \tag{52}$$

for $3 \leq n \leq M$.

In the following, we use induction to prove that

$$(\tilde{p}_{n,i\geq 0}, \tilde{p}_{n,i\geq 1}, \cdots, \tilde{p}_{n,i\geq n-1}) \prec^{w} (p_{n,i\geq 0}, p_{n,i\geq 1}, \cdots, p_{n,i\geq n-1}), \tag{53}$$

for $3 \leq n \leq M$. By definition, Equation (53) leads to

$$\sum_{l=0}^{n-1} p_{n,i\geq l} \;\leq\; \sum_{l=0}^{n-1} \tilde{p}_{n,i\geq l}, \tag{54}$$

which is equivalent to Equation (52), since $p_{n,i\geq 0} = \tilde{p}_{n,i\geq 0} = 1$.

**Initialization:** For $n = 2$, we have $p_{2,i\geq l} = \tilde{p}_{2,i\geq l}$ for $l = 0, 1$, as no approximation is involved

when computing $\tilde{p}_{n,i}$. By definition, we have

$$(\tilde{p}_{2,i\geq 0}, \tilde{p}_{2,i\geq 1}) \prec^w (p_{2,i\geq 0}, p_{2,i\geq 1}).$$

**Induction:** Assume Equation (53) holds for $n-1$, which gives

$$(\tilde{p}_{n-1,i\geq 0}, \tilde{p}_{n-1,i\geq 1}, \cdots, \tilde{p}_{n-1,i\geq n-2}) \prec^w (p_{n-1,i\geq 0}, p_{n-1,i\geq 1}, \cdots, p_{n-1,i\geq n-2}). \tag{55}$$

Let us prove that

$$\sum_{l=k}^{n-1} p_{n,i\geq l} \;\leq\; \sum_{l=k}^{n-1} \tilde{p}_{n,i\geq l} \quad \text{for} \quad 0 \leq k \leq n-1. \tag{56}$$

This reduces to prove that

$$\sum_{l=k}^{n-1} p_{n,i\geq l} \;\leq\; \sum_{l=k}^{n-1} \tilde{p}_{n,i\geq l} \quad \text{for} \quad 1 \leq k \leq n-1, \tag{57}$$

since $p_{n,i\geq 0} = \tilde{p}_{n,i\geq 0} = 0$.

We start by providing an equivalent formation of $\Pr\{R_n = i \mid R_{n-1} = j\}$ as compared to what is used in Equations (8) and (9). Instead of conditioning on the inter-arrival time between customers $n-1$ and $n$, we can compute $\Pr\{R_n = i \mid R_{n-1} = j\}$ by conditioning on the arrival time of customer $n-1$, where we have

$$\begin{aligned}
&\Pr\{R_n = i \mid R_{n-1} = j\} \\
&= \int_{d_n-\tau_n^l}^{d_n+\tau_n^u} \int_{d_{n-1}-\tau_{n-1}^l}^{d_{n-1}+\tau_{n-1}^u} \Pr\{R_n = i \mid R_{n-1} = j, D_{n-1} = v, D_n = u\} f_{n-1,j}(v) f_n(u)\, dv\, du \\
&= \int_{d_n-\tau_n^l}^{d_n+\tau_n^u} \int_{d_{n-1}-\tau_{n-1}^l}^{d_{n-1}+\tau_{n-1}^u} [\alpha_{n-1} g(j+1-i, (u-v)\mu) + (1-\alpha_{n-1}) g(j-i, (u-v)\mu)] f_{n-1,j}(v) f_n(u)\, dv\, du,
\end{aligned} \tag{58}$$

for $0 \leq j \leq n-2$ and $0 \leq i \leq j+1$.

Thus

$$
\begin{aligned}
\sum_{l=k}^{n-1} p_{n,i\geq l} &= \sum_{l=k}^{n-1}\sum_{i=l}^{n-1} p_{n,i} \\
&= \sum_{l=k}^{n-1}\sum_{i=l}^{n-1}\sum_{j=i-1}^{n-2} p_{n-1,j}\Pr\{R_n = i \mid R_{n-1} = j\} \\
&= \sum_{l=k}^{n-1}\sum_{i=l}^{n-1}\sum_{j=i-1}^{n-2} p_{n-1,j}\int_{d_n-\tau_n^l}^{d_n+\tau_n^u}\int_{d_{n-1}-\tau_{n-1}^l}^{d_{n-1}+\tau_{n-1}^u} \big[\alpha_{n-1}g(j+1-i,(u-v)\mu) \\
&\qquad + (1-\alpha_{n-1})g(j-i,(u-v)\mu)\big] f_{n-1,j}(v) f_n(u)\,dv\,du \\
&= \int_{d_n-\tau_n^l}^{d_n+\tau_n^u} f_n(u)\int_{d_{n-1}-\tau_{n-1}^l}^{d_{n-1}+\tau_{n-1}^u} \sum_{l=k}^{n-1}\sum_{i=l}^{n-1}\sum_{j=i-1}^{n-2} \big[\alpha_{n-1}g(j+1-i,(u-v)\mu) \\
&\qquad + (1-\alpha_{n-1})g(j-i,(u-v)\mu)\big] p_{n-1,j} f_{n-1,j}(v)\,dv\,du \\
&= \int_{d_n-\tau_n^l}^{d_n+\tau_n^u} f_n(u)\int_{d_{n-1}-\tau_{n-1}^l}^{d_{n-1}+\tau_{n-1}^u} f_{n-1}(v) \sum_{l=k}^{n-1}\sum_{i=l}^{n-1}\sum_{j=i-1}^{n-2} \big[\alpha_{n-1}g(j+1-i,(u-v)\mu) \\
&\qquad + (1-\alpha_{n-1})g(j-i,(u-v)\mu)\big] p_{n-1,j|D_{n-1}=v}\,dv\,du \\
&= \int_{d_n-\tau_n^l}^{d_n+\tau_n^u} f_n(u)\int_{d_{n-1}-\tau_{n-1}^l}^{d_{n-1}+\tau_{n-1}^u} f_{n-1}(v)\Big[\alpha_{n-1}\sum_{l=k-1}^{n-2} G(l-k+1,(u-v)\mu) p_{n-1,i\geq l|D_{n-1}=v} \\
&\qquad + (1-\alpha_{n-1})\sum_{l=k}^{n-2} G(l-k,(u-v)\mu) p_{n-1,i\geq l|D_{n-1}=v}\Big]\,dv\,du \\
&= \int_{d_n-\tau_n^l}^{d_n+\tau_n^u} f_n(u)\Big[\alpha_{n-1}\sum_{l=k-1}^{n-2}\int_{d_{n-1}-\tau_{n-1}^l}^{d_{n-1}+\tau_{n-1}^u} f_{n-1}(v) G(l-k+1,(u-v)\mu) p_{n-1,i\geq l|D_{n-1}=v}\,dv \\
&\qquad + (1-\alpha_{n-1})\sum_{l=k}^{n-2}\int_{d_{n-1}-\tau_{n-1}^l}^{d_{n-1}+\tau_{n-1}^u} f_{n-1}(v) G(l-k,(u-v)\mu) p_{n-1,i\geq l|D_{n-1}=v}\,dv\Big]\,du.
\end{aligned}
\tag{59}
$$

Since $f_{n-1}(v)$ is constant on $[d_{n-1}-\tau_{n-1}^l, d_{n-1}+\tau_{n-1}^u]$, together with Theorem 2, we obtain

$$\begin{aligned}
&\int_{d_{n-1}-\tau_{n-1}^l}^{d_{n-1}+\tau_{n-1}^u} f_{n-1}(v)G(l-k+1,(u-v)\mu)p_{n-1,i\geq l|D_{n-1}=v}\,dv \\
\leq &\int_{d_{n-1}-\tau_{n-1}^l}^{d_{n-1}+\tau_{n-1}^u} f_{n-1}(v)G(l-k+1,(u-v)\mu)\,dv \int_{d_{n-1}-\tau_{n-1}^l}^{d_{n-1}+\tau_{n-1}^u} f_{n-1}(v)p_{n-1,i\geq l|D_{n-1}=v}\,dv \\
=&\mathbb{E}_{D_{n-1}}\big[G(l-k+1,(u-D_{n-1})\mu)\big]\int_{d_{n-1}-\tau_{n-1}^l}^{d_{n-1}+\tau_{n-1}^u} f_{n-1,i\geq l}(v)p_{n-1,i\geq l}\,dv \\
=&\mathbb{E}_{D_{n-1}}\big[G(l-k+1,(u-D_{n-1})\mu)\big]p_{n-1,i\geq l}\int_{d_{n-1}-\tau_{n-1}^l}^{d_{n-1}+\tau_{n-1}^u} f_{n-1,i\geq l}(v)\,dv \\
=&\mathbb{E}_{D_{n-1}}\big[G(l-k+1,(u-D_{n-1})\mu)\big]p_{n-1,i\geq l}\,,
\end{aligned} \tag{60}$$

where $f_{n-1,i\geq l}(v)$ is the pdf of the conditional arrival time of customer $n-1$, given she finds equal or more than $l$ customers in system upon her arrival. The first equality in Equation (60) is derived by applying Bayes' theorem. In particular

$$f_{n-1,i\geq l}(t) = \frac{p_{n-1,i\geq l|D_{n-1}=t}\times f_{n-1}(t)}{p_{n-1,i\geq l}}\,.$$

Similarly, we can write

$$\begin{aligned}
&\int_{d_{n-1}-\tau_{n-1}^l}^{d_{n-1}+\tau_{n-1}^u} f_{n-1}(v)G(l-k,(u-v)\mu)p_{n-1,i\geq l|D_{n-1}=v}\,dv \\
\leq &\mathbb{E}_{D_{n-1}}\big[G(l-k,(u-D_{n-1})\mu)\big]p_{n-1,i\geq l}\,.
\end{aligned} \tag{61}$$

Substituting Equation (59) with Equations (60) and (61) gives

$$\begin{aligned}
&\sum_{l=k}^{n-1} p_{n,i\geq l} \\
\leq &\int_{d_n-\tau_n^l}^{d_n+\tau_n^u} f_n(u)\Big[\alpha_{n-1}\sum_{l=k-1}^{n-2}\mathbb{E}_{D_{n-1}}\big[G(l-k+1,(u-D_{n-1})\mu)\big]p_{n-1,i\geq l} \\
&\qquad + (1-\alpha_{n-1})\sum_{l=k}^{n-2}\mathbb{E}_{D_{n-1}}\big[G(l-k,(u-D_{n-1})\mu)\big]p_{n-1,i\geq l}\Big]\,du.
\end{aligned} \tag{62}$$

Next, we compute $\sum_{l=k}^{n-1}\tilde{p}_{n,i\geq l}$ with a characterization of $\tilde{p}_{n,i}$ that is equivalent to what we used in our approximation. In the approximation method, we approximate $h_{n-1,j}$ in Equation (9) by

$h_{n-1}$, which leads to

$$\begin{aligned}\tilde{p}_{n,i} = \alpha_{n-1} \sum_{j=i-1}^{n-2} \tilde{p}_{n-1,j} \int_{d_n - d_{n-1} - \tau_n^l - \tau_{n-1}^u}^{d_n - d_{n-1} + \tau_n^u + \tau_{n-1}^l} g(j+1-i, x\mu) h_{n-1}(x)\, dx \\ + (1-\alpha_{n-1}) \sum_{j=i}^{n-2} \tilde{p}_{n-1,j} \int_{d_n - d_{n-1} - \tau_n^l - \tau_{n-1}^u}^{d_n - d_{n-1} + \tau_n^u + \tau_{n-1}^l} g(j-i, x\mu) h_{n-1}(x)\, dx,\end{aligned} \tag{63}$$

with

$$\begin{aligned}h_{n-1}(x) = \int_{\max(d_n - \tau_n^l, d_{n-1} - \tau_{n-1}^l + x)}^{\min(d_n + \tau_n^u, d_{n-1} + \tau_{n-1}^u + x)} f_n(u) f_{n-1}(u-x)\, du \\ \text{for } x \in [d_n - d_{n-1} + \tau_n^u + \tau_{n-1}^l, d_n - d_{n-1} - \tau_n^l - \tau_{n-1}^u].\end{aligned} \tag{64}$$

We can substitute $h_{n-1}(x)$ and reformulate $\tilde{p}_{n,i}$ by changing the integration variables and ranges, which gives

$$\begin{aligned}\tilde{p}_{n,i} = \sum_{j=i-1}^{n-2} \tilde{p}_{n-1,j} \int_{d_n - \tau_n^l}^{d_n + \tau_n^u} f_n(u) \int_{d_{n-1} - \tau_{n-1}^l}^{d_{n-1} + \tau_{n-1}^u} f_{n-1}(v) \big[\alpha_{n-1} g(j+1-i, (u-v)\mu) \\ + (1-\alpha_{n-1}) g(j-i, (u-v)\mu)\big]\, dv\, du\,.\end{aligned} \tag{65}$$

Then, we can compute $\sum_{l=k}^{n-1} \tilde{p}_{n,i\geq l}$ by using Equation (65). This implies

$$\begin{aligned}&\sum_{l=k}^{n-1} \tilde{p}_{n,i\geq l} \\ &= \sum_{l=k}^{n-1} \sum_{i=l}^{n-1} \sum_{j=i-1}^{n-2} \tilde{p}_{n-1,j} \int_{d_n - \tau_n^l}^{d_n + \tau_n^u} \int_{d_{n-1} - \tau_{n-1}^l}^{d_{n-1} + \tau_{n-1}^u} \big[\alpha_{n-1} g(j+1-i, (u-v)\mu) \\ &\qquad + (1-\alpha_{n-1}) g(j-i, (u-v)\mu)\big] f_{n-1}(v) f_n(u)\, dv\, du \\ &= \int_{d_n - \tau_n^l}^{d_n + \tau_n^u} f_n(u) \int_{d_{n-1} - \tau_{n-1}^l}^{d_{n-1} + \tau_{n-1}^u} f_{n-1}(v) \Big[\alpha_{n-1} \sum_{l=k-1}^{n-2} G(l-k+1, (u-v)\mu) \tilde{p}_{n-1,i\geq l}\, dv \\ &\qquad + (1-\alpha_{n-1}) \sum_{l=k}^{n-2} G(l-k, (u-v)\mu) \tilde{p}_{n-1,i\geq l}\Big]\, dv\, du \\ &= \int_{d_n - \tau_n^l}^{d_n + \tau_n^u} f_n(u) \Big[\alpha_{n-1} \sum_{l=k-1}^{n-2} \mathbb{E}_{D_{n-1}}\big[G(l-k+1, (u-D_{n-1})\mu)\big] \tilde{p}_{n-1,i\geq l} \\ &\qquad + (1-\alpha_{n-1}) \sum_{l=k}^{n-2} \mathbb{E}_{D_{n-1}}\big[G(l-k, (u-D_{n-1})\mu)\big] \tilde{p}_{n-1,i\geq l}\Big]\, du.\end{aligned} \tag{66}$$

Finally, we are ready to show that Equation (57) holds. By assumption, it follows from Equation (55) that

$$(\tilde{p}_{n-1,i\geq k}, \tilde{p}_{n-1,i\geq k+1}, \cdots, \tilde{p}_{n-1,i\geq n-2}) \prec^w (p_{n-1,i\geq k}, p_{n-1,i\geq k+1}, \cdots, p_{n-1,i\geq n-2}),$$

for all $k = 1, \cdots, n-2$. For fixed $u \in [d_n - \tau_n^l, d_n + \tau_n^u]$, it is obvious that $\Big(\mathbb{E}_{D_{n-1}}\big[G(l, (u - D_{n-1})\mu\big]\Big)_{l=1}^{n-2}$ and $\Big(\mathbb{E}_{D_{n-1}}\big[G(l-1, (u - D_{n-1})\mu\big]\Big)_{l=1}^{n-2}$ are positive and increasing in $l$. Therefore, according to Proposition 5, we have

$$\sum_{l=k}^{n-2} \mathbb{E}_{D_{n-1}}\big[G(l-k+1, (u - D_{n-1})\mu)\big] p_{n-1,i\geq l} \leq \sum_{l=k}^{n-2} \mathbb{E}_{D_{n-1}}\big[G(l-k+1, (u - D_{n-1})\mu)\big] \tilde{p}_{n-1,i\geq l},$$

which leads to

$$\sum_{l=k-1}^{n-2} \mathbb{E}_{D_{n-1}}\big[G(l-k+1, (u - D_{n-1})\mu)\big] p_{n-1,i\geq l} \leq \sum_{l=k-1}^{n-2} \mathbb{E}_{D_{n-1}}\big[G(l-k+1, (u - D_{n-1})\mu)\big] \tilde{p}_{n-1,i\geq l}, \tag{67}$$

since $p_{n-1,i\geq 0} = \tilde{p}_{n-1,i\geq 0} = 1$. Also, according to Proposition 5, we have

$$\sum_{l=k}^{n-2} \mathbb{E}_{D_{n-1}}\big[G(l-k, (u - D_{n-1})\mu)\big] p_{n-1,i\geq l} \leq \sum_{l=k}^{n-2} \mathbb{E}_{D_{n-1}}\big[G(l-k, (u - D_{n-1})\mu)\big] \tilde{p}_{n-1,i\geq l}. \tag{68}$$

Since Equations (67) and (68) hold for all $u \in [d_n - \tau_n^l, d_n + \tau_n^u]$, together with Equation (62) and (66), we deduce that

$$\sum_{l=k}^{n-1} p_{n,i\geq l} \;\leq\; \sum_{l=k}^{n-1} \tilde{p}_{n,i\geq l} \quad \text{for} \quad 1 \leq k \leq n-1,$$

which completes the proof of the proposition. □

# Appendix D Experiments related to Section 6

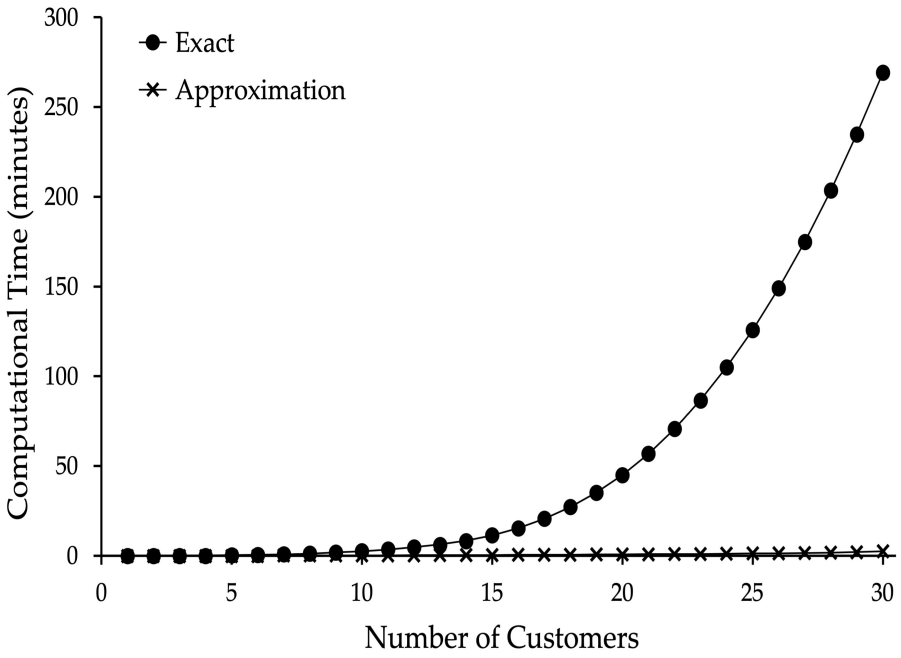



Figure 5: Exact v.s. Approximation ($\tau_n^l = \tau_n^u = \tau = 10$, $\alpha_n = \alpha = 1$, $x_n = x = 20$, $\mu = 0.1$)[6]

[6]Numerical results are generated using Wolfram Mathematica 12.1.1 on Mac OS 10.15.4 with an 18 cores Intel Xeon W processor and 64 GB Ram.

Table 5: Comparison between exact and approximate methods: Experiments 1, $\mu = 0.05$

| Scenarios | Parameters | | | | Exact, $\mathbb{E}[W]$ | | | | Approx., $\mathbb{E}[\widetilde{W}]$ | | | | $\Delta_W = \frac{\mathbb{E}[\widetilde{W}] - \mathbb{E}[W]}{\mathbb{E}[W]} \times 100\%$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Nber of customers | | | | Nber of customers | | | | Nber of customers | | | |
| | $\tau^l$ | $\tau^u$ | $\alpha$ | $d_n - d_{n-1}$ | 10 | 20 | 30 | 40 | 10 | 20 | 30 | 40 | 10 | 20 | 30 | 40 |
| Varying $\tau$ | 2.5 | 2.5 | 1 | 20 | 21.4 | 35.0 | 45.6 | 54.5 | 21.5 | 35.2 | 45.8 | 54.7 | 0.2 % | 0.3 % | 0.3 % | 0.4 % |
| | 5 | 5 | 1 | 20 | 21.8 | 35.4 | 45.9 | 54.9 | 22.0 | 35.8 | 46.6 | 55.7 | 1.0 % | 1.2 % | 1.4 % | 1.5 % |
| | 7.5 | 7.5 | 1 | 20 | 22.2 | 35.8 | 46.4 | 55.3 | 22.7 | 36.8 | 47.8 | 57.1 | 2.1 % | 2.8 % | 3.1 % | 3.3 % |
| | 10 | 10 | 1 | 20 | 22.8 | 36.4 | 47.0 | 55.9 | 23.6 | 38.2 | 49.5 | 59.1 | 3.7 % | 4.8 % | 5.4 % | 5.7 % |
| Varying $\tau^l - \tau^u$ | 1 | 9 | 1 | 20 | 21.4 | 35.2 | 45.8 | 54.7 | 21.6 | 35.6 | 46.4 | 55.5 | 1.0 % | 1.3 % | 1.4 % | 1.5 % |
| | 2 | 8 | 1 | 20 | 21.5 | 35.2 | 45.8 | 54.7 | 21.7 | 35.6 | 46.4 | 55.5 | 1.0 % | 1.3 % | 1.4 % | 1.5 % |
| | 3 | 7 | 1 | 20 | 21.5 | 35.2 | 45.8 | 54.8 | 21.7 | 35.7 | 46.4 | 55.6 | 1.0 % | 1.3 % | 1.4 % | 1.5 % |
| | 4 | 6 | 1 | 20 | 21.6 | 35.3 | 45.9 | 54.8 | 21.8 | 35.7 | 46.5 | 55.6 | 1.0 % | 1.2 % | 1.4 % | 1.5 % |
| | 5 | 5 | 1 | 20 | 21.8 | 35.4 | 45.9 | 54.9 | 22.0 | 35.8 | 46.6 | 55.7 | 1.0 % | 1.2 % | 1.4 % | 1.5 % |
| | 6 | 4 | 1 | 20 | 21.9 | 35.5 | 46.0 | 54.9 | 22.1 | 35.9 | 46.7 | 55.7 | 1.0 % | 1.3 % | 1.4 % | 1.5 % |
| | 7 | 3 | 1 | 20 | 22.1 | 35.6 | 46.1 | 55.0 | 22.3 | 36.1 | 46.8 | 55.8 | 1.0 % | 1.3 % | 1.4 % | 1.5 % |
| | 8 | 2 | 1 | 20 | 22.3 | 35.8 | 46.3 | 55.1 | 22.5 | 36.2 | 46.9 | 56.0 | 1.0 % | 1.2 % | 1.4 % | 1.5 % |
| | 9 | 1 | 1 | 20 | 22.6 | 36.0 | 46.4 | 55.3 | 22.8 | 36.4 | 47.0 | 56.1 | 1.0 % | 1.2 % | 1.4 % | 1.5 % |
| Varying $\alpha$ | 5 | 5 | 0.2 | 20 | 2.4 | 2.7 | 2.8 | 2.8 | 2.5 | 2.7 | 2.8 | 2.8 | 1.0 % | 1.3 % | 1.4 % | 1.4 % |
| | 5 | 5 | 0.4 | 20 | 5.5 | 6.5 | 6.9 | 7.0 | 5.6 | 6.6 | 7.0 | 7.2 | 1.2 % | 1.5 % | 1.7 % | 1.8 % |
| | 5 | 5 | 0.6 | 20 | 9.6 | 12.3 | 13.6 | 14.3 | 9.7 | 12.5 | 13.8 | 14.6 | 1.2 % | 1.6 % | 1.9 % | 2.0 % |
| | 5 | 5 | 0.8 | 20 | 14.9 | 21.5 | 25.4 | 28.1 | 15.1 | 21.8 | 25.9 | 28.7 | 1.1 % | 1.5 % | 1.8 % | 2.0 % |
| | 5 | 5 | 1 | 20 | 21.8 | 35.4 | 45.9 | 54.9 | 22.0 | 35.8 | 46.6 | 55.7 | 1.0 % | 1.2 % | 1.4 % | 1.5 % |
| Varying $\delta$ | 5 | 5 | 1 | 20 | 21.8 | 35.4 | 45.9 | 54.9 | 22.0 | 35.8 | 46.6 | 55.7 | 1.0 % | 1.2 % | 1.4 % | 1.5 % |
| | 5 | 5 | 1 | 22.5 | 17.5 | 26.3 | 32.2 | 36.6 | 17.6 | 26.7 | 32.7 | 37.3 | 1.0 % | 1.4 % | 1.7 % | 1.8 % |
| | 5 | 5 | 1 | 25 | 14.0 | 19.6 | 22.8 | 24.9 | 14.2 | 19.9 | 23.2 | 25.4 | 1.1 % | 1.5 % | 1.8 % | 1.9 % |
| | 5 | 5 | 1 | 27.5 | 11.3 | 14.8 | 16.6 | 17.6 | 11.4 | 15.1 | 16.9 | 17.9 | 1.1 % | 1.5 % | 1.7 % | 1.9 % |
| | 5 | 5 | 1 | 30 | 9.2 | 11.4 | 12.4 | 12.9 | 9.3 | 11.6 | 12.6 | 13.1 | 1.0 % | 1.4 % | 1.6 % | 1.7 % |
| | 5 | 5 | 1 | 35 | 6.1 | 7.1 | 7.5 | 7.7 | 6.2 | 7.2 | 7.6 | 7.8 | 0.9 % | 1.2 % | 1.3 % | 1.4 % |
| | 5 | 5 | 1 | 40 | 4.3 | 4.7 | 4.9 | 5.0 | 4.3 | 4.8 | 4.9 | 5.0 | 0.8 % | 1.0 % | 1.0 % | 1.0 % |

Table 6: Comparison between exact and approximate methods: Experiments 2, $\mu = 0.05$

| Scenarios | Parameters | | | | Exact, $\mathbb{E}[W]$ | | | | Approx., $\mathbb{E}[\widetilde{W}]$ | | | | $\Delta_W = \frac{\mathbb{E}[\widetilde{W}] - \mathbb{E}[W]}{\mathbb{E}[W]} \times 100\%$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Nber of customers | | | | Nber of customers | | | | Nber of customers | | | |
| | $\tau^l$ | $\tau^u$ | $\alpha$ | $d_n - d_{n-1}$ | 10 | 20 | 30 | 40 | 10 | 20 | 30 | 40 | 10 | 20 | 30 | 40 |
| Varying $\tau$ | 2.5 | 2.5 | 1 | 20 | 21.4 | 35.0 | 45.6 | 54.5 | 21.5 | 35.2 | 45.8 | 54.7 | 0.2 % | 0.3 % | 0.3 % | 0.4 % |
| | 5 | 5 | 1 | 20 | 21.8 | 35.4 | 45.9 | 54.9 | 22.0 | 35.8 | 46.6 | 55.7 | 1.0 % | 1.2 % | 1.4 % | 1.5 % |
| | 7.5 | 7.5 | 1 | 20 | 22.2 | 35.8 | 46.4 | 55.3 | 22.7 | 36.8 | 47.8 | 57.1 | 2.1 % | 2.8 % | 3.1 % | 3.3 % |
| | 10 | 10 | 1 | 20 | 22.8 | 36.4 | 47.0 | 55.9 | 23.6 | 38.2 | 49.5 | 59.1 | 3.7 % | 4.8 % | 5.4 % | 5.7 % |
| Varying $\tau^l - \tau^u$ | 2 | 18 | 1 | 20 | 22.2 | 36.0 | 46.7 | 55.7 | 23.0 | 37.7 | 49.2 | 58.8 | 3.7 % | 4.8 % | 5.4 % | 5.7 % |
| | 4 | 16 | 1 | 20 | 22.2 | 36.1 | 46.7 | 55.7 | 23.0 | 37.8 | 49.2 | 58.9 | 3.7 % | 4.8 % | 5.4 % | 5.7 % |
| | 6 | 14 | 1 | 20 | 22.3 | 36.1 | 46.8 | 55.8 | 23.2 | 37.9 | 49.3 | 58.9 | 3.7 % | 4.8 % | 5.4 % | 5.7 % |
| | 8 | 12 | 1 | 20 | 22.5 | 36.3 | 46.9 | 55.8 | 23.4 | 38.0 | 49.4 | 59.0 | 3.7 % | 4.8 % | 5.4 % | 5.7 % |
| | 10 | 10 | 1 | 20 | 22.8 | 36.4 | 47.0 | 55.9 | 23.6 | 38.2 | 49.5 | 59.1 | 3.7 % | 4.8 % | 5.4 % | 5.7 % |
| | 12 | 8 | 1 | 20 | 23.1 | 36.6 | 47.2 | 56.1 | 23.9 | 38.4 | 49.7 | 59.3 | 3.7 % | 4.8 % | 5.4 % | 5.7 % |
| | 14 | 6 | 1 | 20 | 23.5 | 36.9 | 47.4 | 56.3 | 24.3 | 38.7 | 49.9 | 59.5 | 3.6 % | 4.8 % | 5.4 % | 5.7 % |
| | 16 | 4 | 1 | 20 | 24.0 | 37.3 | 47.7 | 56.5 | 24.9 | 39.1 | 50.2 | 59.8 | 3.6 % | 4.8 % | 5.4 % | 5.7 % |
| | 18 | 2 | 1 | 20 | 24.6 | 37.7 | 48.0 | 56.8 | 25.5 | 39.5 | 50.6 | 60.1 | 3.6 % | 4.8 % | 5.4 % | 5.7 % |
| Varying $\alpha$ | 10 | 10 | 0.2 | 20 | 2.7 | 2.9 | 3.0 | 3.0 | 2.8 | 3.0 | 3.1 | 3.2 | 4.2 % | 5.4 % | 5.8 % | 6.0 % |
| | 10 | 10 | 0.4 | 20 | 6.0 | 6.9 | 7.2 | 7.4 | 6.2 | 7.3 | 7.7 | 7.9 | 4.6 % | 6.2 % | 6.8 % | 7.1 % |
| | 10 | 10 | 0.6 | 20 | 10.2 | 12.9 | 14.2 | 14.9 | 10.7 | 13.8 | 15.2 | 16.1 | 4.5 % | 6.4 % | 7.4 % | 8.0 % |
| | 10 | 10 | 0.8 | 20 | 15.7 | 22.3 | 26.2 | 29.0 | 16.4 | 23.6 | 28.1 | 31.2 | 4.2 % | 6.0 % | 7.0 % | 7.7 % |
| | 10 | 10 | 1 | 20 | 22.8 | 36.4 | 47.0 | 55.9 | 23.6 | 38.2 | 49.5 | 59.1 | 3.7 % | 4.8 % | 5.4 % | 5.7 % |
| Varying $\delta$ | 10 | 10 | 1 | 20 | 22.8 | 36.4 | 47.0 | 55.9 | 23.6 | 38.2 | 49.5 | 59.1 | 3.7 % | 4.8 % | 5.4 % | 5.7 % |
| | 10 | 10 | 1 | 22.5 | 18.3 | 27.2 | 33.1 | 37.5 | 19.1 | 28.7 | 35.2 | 40.2 | 4.0 % | 5.6 % | 6.4 % | 7.1 % |
| | 10 | 10 | 1 | 25 | 14.8 | 20.4 | 23.6 | 25.7 | 15.4 | 21.6 | 25.2 | 27.6 | 4.2 % | 5.9 % | 6.9 % | 7.6 % |
| | 10 | 10 | 1 | 27.5 | 12.0 | 15.5 | 17.2 | 18.2 | 12.5 | 16.4 | 18.4 | 19.6 | 4.2 % | 6.0 % | 6.9 % | 7.5 % |
| | 10 | 10 | 1 | 30 | 9.7 | 12.0 | 12.9 | 13.4 | 10.1 | 12.7 | 13.8 | 14.4 | 4.1 % | 5.7 % | 6.5 % | 6.9 % |
| | 10 | 10 | 1 | 35 | 6.6 | 7.5 | 7.9 | 8.0 | 6.9 | 7.9 | 8.3 | 8.5 | 3.6 % | 4.8 % | 5.2 % | 5.5 % |
| | 10 | 10 | 1 | 40 | 4.6 | 5.0 | 5.2 | 5.2 | 4.8 | 5.2 | 5.4 | 5.5 | 3.1 % | 3.9 % | 4.1 % | 4.2 % |

Table 7: Comparison between exact and approximate methods: Experiments 3, $\mu = 0.05$

| Scenarios | Parameters | | | | Exact, $\mathbb{E}[W]$ | | | | Approx., $\mathbb{E}[\widetilde{W}]$ | | | | $\Delta_W = \frac{\mathbb{E}[\widetilde{W}] - \mathbb{E}[W]}{\mathbb{E}[W]} \times 100\%$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Nber of customers | | | | Nber of customers | | | | Nber of customers | | | |
| | $\tau^l$ | $\tau^u$ | $\alpha$ | $d_n - d_{n-1}$ | 10 | 20 | 30 | 40 | 10 | 20 | 30 | 40 | 10 | 20 | 30 | 40 |
| Varying $\tau$ | 2.5 | 2.5 | 1 | 30 | 9.0 | 11.3 | 12.2 | 12.7 | 9.0 | 11.3 | 12.3 | 12.8 | 0.3 % | 0.4 % | 0.4 % | 0.4 % |
| | 5 | 5 | 1 | 30 | 9.2 | 11.4 | 12.4 | 12.9 | 9.3 | 11.6 | 12.6 | 13.1 | 1.0 % | 1.4 % | 1.6 % | 1.7 % |
| | 7.5 | 7.5 | 1 | 30 | 9.4 | 11.7 | 12.6 | 13.1 | 9.6 | 12.0 | 13.1 | 13.6 | 2.3 % | 3.2 % | 3.6 % | 3.9 % |
| | 10 | 10 | 1 | 30 | 9.7 | 12.0 | 12.9 | 13.4 | 10.1 | 12.7 | 13.8 | 14.4 | 4.1 % | 5.7 % | 6.5 % | 6.9 % |
| Varying $\tau^l - \tau^u$ | 1 | 9 | 1 | 30 | 9.0 | 11.3 | 12.3 | 12.8 | 9.1 | 11.5 | 12.5 | 13.1 | 1.1 % | 1.4 % | 1.6 % | 1.7 % |
| | 2 | 8 | 1 | 30 | 9.0 | 11.3 | 12.3 | 12.9 | 9.1 | 11.5 | 12.5 | 13.1 | 1.1 % | 1.4 % | 1.6 % | 1.7 % |
| | 3 | 7 | 1 | 30 | 9.0 | 11.4 | 12.3 | 12.9 | 9.1 | 11.5 | 12.5 | 13.1 | 1.1 % | 1.4 % | 1.6 % | 1.7 % |
| | 4 | 6 | 1 | 30 | 9.1 | 11.4 | 12.4 | 12.9 | 9.2 | 11.5 | 12.6 | 13.1 | 1.1 % | 1.4 % | 1.6 % | 1.7 % |
| | 5 | 5 | 1 | 30 | 9.2 | 11.4 | 12.4 | 12.9 | 9.3 | 11.6 | 12.6 | 13.1 | 1.0 % | 1.4 % | 1.6 % | 1.7 % |
| | 6 | 4 | 1 | 30 | 9.2 | 11.5 | 12.4 | 12.9 | 9.3 | 11.6 | 12.6 | 13.1 | 1.0 % | 1.4 % | 1.6 % | 1.7 % |
| | 7 | 3 | 1 | 30 | 9.4 | 11.5 | 12.5 | 12.9 | 9.5 | 11.7 | 12.7 | 13.2 | 1.0 % | 1.4 % | 1.6 % | 1.7 % |
| | 8 | 2 | 1 | 30 | 9.5 | 11.6 | 12.5 | 13.0 | 9.6 | 11.8 | 12.7 | 13.2 | 1.0 % | 1.4 % | 1.6 % | 1.7 % |
| | 9 | 1 | 1 | 30 | 9.6 | 11.7 | 12.6 | 13.0 | 9.7 | 11.8 | 12.8 | 13.2 | 1.0 % | 1.4 % | 1.6 % | 1.7 % |
| Varying $\alpha$ | 5 | 5 | 0.2 | 30 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 0.6 % | 0.7 % | 0.7 % | 0.7 % |
| | 5 | 5 | 0.4 | 30 | 2.6 | 2.8 | 2.9 | 2.9 | 2.7 | 2.9 | 2.9 | 3.0 | 0.8 % | 0.9 % | 0.9 % | 0.9 % |
| | 5 | 5 | 0.6 | 30 | 4.3 | 4.8 | 5.0 | 5.1 | 4.4 | 4.9 | 5.1 | 5.2 | 0.9 % | 1.1 % | 1.1 % | 1.2 % |
| | 5 | 5 | 0.8 | 30 | 6.5 | 7.6 | 8.0 | 8.2 | 6.5 | 7.7 | 8.1 | 8.3 | 1.0 % | 1.3 % | 1.4 % | 1.4 % |
| | 5 | 5 | 1 | 30 | 9.2 | 11.4 | 12.4 | 12.9 | 9.3 | 11.6 | 12.6 | 13.1 | 1.0 % | 1.4 % | 1.6 % | 1.7 % |
| Varying $\delta$ | 10 | 10 | 1 | 20 | 22.8 | 36.4 | 47.0 | 55.9 | 23.6 | 38.2 | 49.5 | 59.1 | 3.7 % | 4.8 % | 5.4 % | 5.7 % |
| | 10 | 10 | 1 | 22.5 | 18.3 | 27.2 | 33.1 | 37.5 | 19.1 | 28.7 | 35.2 | 40.2 | 4.0 % | 5.6 % | 6.4 % | 7.1 % |
| | 10 | 10 | 1 | 25 | 14.8 | 20.4 | 23.6 | 25.7 | 15.4 | 21.6 | 25.2 | 27.6 | 4.2 % | 5.9 % | 6.9 % | 7.6 % |
| | 10 | 10 | 1 | 27.5 | 12.0 | 15.5 | 17.2 | 18.2 | 12.5 | 16.4 | 18.4 | 19.6 | 4.2 % | 6.0 % | 6.9 % | 7.5 % |
| | 10 | 10 | 1 | 30 | 9.7 | 12.0 | 12.9 | 13.4 | 10.1 | 12.7 | 13.8 | 14.4 | 4.1 % | 5.7 % | 6.5 % | 6.9 % |
| | 10 | 10 | 1 | 35 | 6.6 | 7.5 | 7.9 | 8.0 | 6.9 | 7.9 | 8.3 | 8.5 | 3.6 % | 4.8 % | 5.2 % | 5.5 % |
| | 10 | 10 | 1 | 40 | 4.6 | 5.0 | 5.2 | 5.2 | 4.8 | 5.2 | 5.4 | 5.5 | 3.1 % | 3.9 % | 4.1 % | 4.2 % |

Table 8: Comparison between exact and approximate methods: Experiments 4, $\mu = 0.05$

| Scenarios | Parameters | | | | Exact, $\mathbb{E}[W]$ | | | | Approx., $\mathbb{E}[\widetilde{W}]$ | | | | $\Delta_W = \frac{\mathbb{E}[\widetilde{W}] - \mathbb{E}[W]}{\mathbb{E}[W]} \times 100\%$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Nber of customers | | | | Nber of customers | | | | Nber of customers | | | |
| | $\tau^l$ | $\tau^u$ | $\alpha$ | $d_n - d_{n-1}$ | 10 | 20 | 30 | 40 | 10 | 20 | 30 | 40 | 10 | 20 | 30 | 40 |
| Varying $\tau$ | 2.5 | 2.5 | 1 | 30 | 9.0 | 11.3 | 12.2 | 12.7 | 9.0 | 11.3 | 12.3 | 12.8 | 0.3 % | 0.4 % | 0.4 % | 0.4 % |
| | 5 | 5 | 1 | 30 | 9.2 | 11.4 | 12.4 | 12.9 | 9.3 | 11.6 | 12.6 | 13.1 | 1.0 % | 1.4 % | 1.6 % | 1.7 % |
| | 7.5 | 7.5 | 1 | 30 | 9.4 | 11.7 | 12.6 | 13.1 | 9.6 | 12.0 | 13.1 | 13.6 | 2.3 % | 3.2 % | 3.6 % | 3.9 % |
| | 10 | 10 | 1 | 30 | 9.7 | 12.0 | 12.9 | 13.4 | 10.1 | 12.7 | 13.8 | 14.4 | 4.1 % | 5.7 % | 6.5 % | 6.9 % |
| Varying $\tau^l - \tau^u$ | 2 | 18 | 1 | 30 | 9.4 | 11.8 | 12.8 | 13.3 | 9.8 | 12.5 | 13.6 | 14.3 | 4.2 % | 5.8 % | 6.5 % | 6.9 % |
| | 4 | 16 | 1 | 30 | 9.4 | 11.8 | 12.8 | 13.4 | 9.8 | 12.5 | 13.7 | 14.3 | 4.2 % | 5.7 % | 6.5 % | 6.9 % |
| | 6 | 14 | 1 | 30 | 9.5 | 11.9 | 12.8 | 13.4 | 9.9 | 12.5 | 13.7 | 14.3 | 4.1 % | 5.7 % | 6.5 % | 6.9 % |
| | 8 | 12 | 1 | 30 | 9.6 | 11.9 | 12.9 | 13.4 | 10.0 | 12.6 | 13.7 | 14.3 | 4.1 % | 5.7 % | 6.5 % | 6.9 % |
| | 10 | 10 | 1 | 30 | 9.7 | 12.0 | 12.9 | 13.4 | 10.1 | 12.7 | 13.8 | 14.4 | 4.1 % | 5.7 % | 6.5 % | 6.9 % |
| | 12 | 8 | 1 | 30 | 9.9 | 12.1 | 13.0 | 13.5 | 10.3 | 12.8 | 13.8 | 14.4 | 4.0 % | 5.7 % | 6.5 % | 6.9 % |
| | 14 | 6 | 1 | 30 | 10.2 | 12.2 | 13.1 | 13.6 | 10.6 | 12.9 | 13.9 | 14.5 | 4.0 % | 5.7 % | 6.4 % | 6.9 % |
| | 16 | 4 | 1 | 30 | 10.4 | 12.3 | 13.2 | 13.6 | 10.8 | 13.0 | 14.0 | 14.6 | 4.0 % | 5.6 % | 6.4 % | 6.8 % |
| | 18 | 2 | 1 | 30 | 10.8 | 12.5 | 13.3 | 13.7 | 11.2 | 13.2 | 14.1 | 14.6 | 3.9 % | 5.6 % | 6.4 % | 6.8 % |
| Varying $\alpha$ | 10 | 10 | 0.2 | 30 | 1.5 | 1.4 | 1.4 | 1.4 | 1.5 | 1.5 | 1.5 | 1.5 | 2.2 % | 2.7 % | 2.9 % | 3.0 % |
| | 10 | 10 | 0.4 | 30 | 2.9 | 3.1 | 3.1 | 3.1 | 3.0 | 3.2 | 3.2 | 3.2 | 2.9 % | 3.5 % | 3.7 % | 3.8 % |
| | 10 | 10 | 0.6 | 30 | 4.7 | 5.2 | 5.3 | 5.4 | 4.9 | 5.4 | 5.6 | 5.6 | 3.4 % | 4.3 % | 4.6 % | 4.7 % |
| | 10 | 10 | 0.8 | 30 | 6.9 | 8.0 | 8.4 | 8.6 | 7.2 | 8.4 | 8.8 | 9.1 | 3.8 % | 5.1 % | 5.5 % | 5.8 % |
| | 10 | 10 | 1 | 30 | 9.7 | 12.0 | 12.9 | 13.4 | 10.1 | 12.7 | 13.8 | 14.4 | 4.1 % | 5.7 % | 6.5 % | 6.9 % |
| Varying $\delta$ | 10 | 10 | 1 | 20 | 22.8 | 36.4 | 47.0 | 55.9 | 23.6 | 38.2 | 49.5 | 59.1 | 3.7 % | 4.8 % | 5.4 % | 5.7 % |
| | 10 | 10 | 1 | 22.5 | 18.3 | 27.2 | 33.1 | 37.5 | 19.1 | 28.7 | 35.2 | 40.2 | 4.0 % | 5.6 % | 6.4 % | 7.1 % |
| | 10 | 10 | 1 | 25 | 14.8 | 20.4 | 23.6 | 25.7 | 15.4 | 21.6 | 25.2 | 27.6 | 4.2 % | 5.9 % | 6.9 % | 7.6 % |
| | 10 | 10 | 1 | 27.5 | 12.0 | 15.5 | 17.2 | 18.2 | 12.5 | 16.4 | 18.4 | 19.6 | 4.2 % | 6.0 % | 6.9 % | 7.5 % |
| | 10 | 10 | 1 | 30 | 9.7 | 12.0 | 12.9 | 13.4 | 10.1 | 12.7 | 13.8 | 14.4 | 4.1 % | 5.7 % | 6.5 % | 6.9 % |
| | 10 | 10 | 1 | 35 | 6.6 | 7.5 | 7.9 | 8.0 | 6.9 | 7.9 | 8.3 | 8.5 | 3.6 % | 4.8 % | 5.2 % | 5.5 % |
| | 10 | 10 | 1 | 40 | 4.6 | 5.0 | 5.2 | 5.2 | 4.8 | 5.2 | 5.4 | 5.5 | 3.1 % | 3.9 % | 4.1 % | 4.2 % |