Queueing Syst (2015) 79:53–85
DOI 10.1007/s11134-014-9427-5

A Tandem Queueing model for an appointment-based service system

Jianzhe Luo · Vidyadhar G. Kulkarni · Serhan Ziya

Received: 3 April 2013 / Revised: 9 October 2014 / Published online: 24 October 2014

Abstract We develop a queueing model for an appointment-based service system that consists of two queues in tandem: the appointment queue followed by the service queue. Customers join the appointment queue when they call for appointments, stay there (not physically) until the appointment time comes, and then leave the appointment queue and may physically join the service queue, and wait there until served, or become a no-show. The probability of becoming a no-show depends on the time spent in the appointment queue. We build a *Tandem Queue* model that explicitly captures the dependence between these two queues. We then derive several performance measures of interest, such as service utilization and customer long-run average waiting times in both queues. We also develop several approximation methods to compute the long-run average waiting time in the service queue, including one which is particularly useful when service times in that queue are not exponential.

Keywords Queueing model · Appointment scheduling · Service operations

Mathematics Subject Classification 60K20 · 60K25 · 90B36 · 90B22 · 60J05

1 Introduction

Customers encounter two types of delays when accessing a health care service via an appointment system, namely appointment delay (indirect delay) and service delay

J. Luo (✉) · V. G. Kulkarni · S. Ziya
Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
e-mail: jzluo@email.unc.edu

V. G. Kulkarni
e-mail: vkulkarn@email.unc.edu

S. Ziya
e-mail: ziya@email.unc.edu

(direct delay) (see [15]). Appointment delay is the time elapsed from the moment a customer requests an appointment until the actual appointment time scheduled for that customer. Service delay is the time elapsed from the time a customer arrives at the service facility (the appointment time if he is punctual) until the time when he is actually served. Service providers prefer to keep appointment intervals short in order to minimize the server idle time. However, short appointment intervals tend to cause congestion in the waiting room, which results in long direct waiting times. On the other hand, if appointments are scheduled sparsely, customers may encounter long indirect delays. One of the main problems associated with long indirect delays is that they typically lead to high no-show rates. Studies on appointments from a number of different clinical settings including primary care [14], family medicine [25], mental health [10], OB/GYN [8], and health care referral services [2] found that patients are more likely to not show-up for their appointments when they have longer appointment delays. Customer no-show behavior results in the wastage of service resources while other customers encounter long waits in getting appointments. For instance, the time slot assigned to a customer who becomes a no-show may not be reassigned to another customer but may be wasted. As a result, while long direct delay might lead to customer dissatisfaction about the service, long appointment delays may not only cause dissatisfaction, but may also lead to appointment cancelations, no-shows, or customers choosing to be served elsewhere, which results in the loss of revenue for the practice (see [12]).

Despite the significant impact that both direct and indirect delays have on the performance of appointment systems, [15] point out that the majority of the literature on appointment scheduling has concentrated on the problem of balancing customer direct delay and server utilization over a service session. The typical decision variables in this stream of research work include the number of appointment slots, the length of each slot, the number of customers assigned to each slot, and so on (for example [1,6,7,9,16,17,19,26,27,31,34,35]).

Multi-type appointment scheduling has also been studied by a number of research articles. Green et al. [13] and Patrick et al. [30] formulate the appointment scheduling of outpatients and inpatients who share the use of a diagnostic medical facility as a discrete Markov decision process. Gupta and Denton [15] develop a dynamic programming model for scheduling regular and same-day patients by taking patient preferences into consideration. Kortbeek et al. [21] develop a model that considers both scheduled and walk-in patients and balances their respective access time and waiting time. Luo et al. [26] investigate the appointment scheduling problem in which scheduled service can be preemptively interrupted by emergencies.

Overbooking is also a commonly used strategy when scheduling appointments with no-shows and/or cancelations. LaGanga and Lawrence [23], Muthuraman and Lawley [28], Chakraborty et al. [4], and Robinson and Chen [32] develop appointment scheduling policies that use overbooking to compensate for patient no-shows with different assumptions on service demand and service time distribution.

Indirect delay has recently drawn more attention (for example [5,12,24,25]). Clearly, appointment scheduling should ideally take into account indirect and direct waiting times simultaneously because both of them affect customers' service experiences significantly. Thus, the objective of our paper is to investigate the appoint-

ment system design problem that aims to balance the utilization of expensive service resources and the delays (both direct and indirect) encountered by customers.

Specifically, we develop a *Tandem Queueing* model of an appointment system. The first queue, which we call the *appointment queue*, captures the waiting process of customers whose scheduled appointment times have not come yet. The appointment queue does not physically exist. It is more like a list where customers with their scheduled appointments are recorded. The time that a customer spends in this queue is exactly his indirect delay. The second queue, which we call the *service queue*, is the queue of customers who show up at the service facility at their appointment times. When a customer is assigned an appointment time, he joins the appointment queue. Then once his appointment time is due, he either shows up at the service facility punctually (joins the service queue) or becomes a no-show with a probability affected by his indirect waiting time. If he shows up, the time that he spends in the service queue is his direct waiting time.

To the best of our knowledge, our Tandem Queueing formulation is the first one that unifies the appointment and service processes in the same queueing model and explicitly captures their dependence. (This might, however, be a distinction shared by a paper by Zacharias and Armony [36], which uses a similar queueing formulation but focuses on questions that are completely different from ours.) In addition, the majority of the appointment-scheduling literature that considers customer no-show behavior commonly assumes a fixed no-show probability. However, that assumption implicitly implies that customers are insensitive to their appointment delays, which may not reflect the reality with sufficient accuracy. Some recent empirical studies indicate that the longer the appointment delays, the more likely the patients are to cancel their appointments or become no-shows. In our formulation, a customer's no-show probability can be affected by the appointment delay he encounters during the appointment process. As we demonstrate by a simulation study in this paper, ignoring this dependence when in fact it is present, might lead to significant errors in the computation of expected service delays.

Following Green and Savin [12], Robinson and Chen [33], and Liu and Ziya [24], we carry out a steady-state analysis of a queueing system over an infinite-horizon but unlike these earlier papers, which consider single-stage queues, our analysis is for a two-stage tandem queue as we consider both appointment and service queues. We obtain expressions for several performance measures such as the server utilization, the long-run average appointment delay and service delay, and the long-run average probability that an arrival finds the service queue busy. We are particularly interested in appointment systems with high traffic intensity because this is the situation in which a service system may benefit most from the appointment mechanism. It is important to note that computation of these performance measures is not possible unless one assumes finite capacity for the appointment queue. Thus, computations for the infinite capacity case will have to be approximated by either setting the queue capacity arbitrarily large (finite) or using another method as we explain in Sect. 6.

These performance measures are very useful in the proper design and operation of an appointment system. For example, one may design an appointment system to maximize the server utilization subject to service level constraints such as upper bounds of both types of delays. Even though we do not solve such an optimization problem in this

paper, our Tandem Queueing model provides useful analytical results and numerical procedures that can be utilized for further investigation.

The remainder of this paper is organized as follows. In Sect. 2 we give a detailed problem description. Sections 3 and 4 investigate the infinite appointment queue and the service queue, respectively, and obtain the performance measures of interest. In Sect. 5, we study the system with finite appointment queue, which makes computation of various performance measures possible. In Sect. 6, we propose an interpolation approximation method of the direct delay and compare it with an existing tandem queueing approximation method. A numerical study is presented in Sect. 7. Finally, Sect. 8 summarizes our findings and conclusions.

2 Problem description

We consider an appointment scheduling system that consists of two stages, schematically shown in Fig. 1. Customers first call to request appointments and are given specific appointment times in a first-come, first-served (FCFS) manner. In other words, each customer is scheduled at the first available appointment time. The queue formed in this appointment process is referred to as *appointment queue*. Then customers show up at the service facility at their individual scheduled appointment times with a certain probability and are served in a FCFS manner. The queue formed in this service process is referred to as *service queue*. The appointment queue may have finite or infinite capacity. We assume that appointment requests arrive according to a Poisson process with rate λ and appointments can only be scheduled at equidistant time epochs with distance d, which is controlled by the system designer. For instance, suppose the service provider decides to schedule appointments every half an hour, such as 8:00, 8:30, 9:00, etc, and a customer calls at 10:20 to request an appointment, then he will be scheduled at 10:30 if there is no scheduled appointment in the system, or 30 min later than the last appointment time that has already been assigned. When a customer's appointment time is due, that customer is removed from the appointment list. The total waiting time that customer spends in the appointment queue (indirect waiting time) affects his show-up probability at the service facility. If he shows up, he joins the service queue. Thus, customers either arrive on time at their scheduled appointment times or become no-shows. No walk-in customers are allowed. In the service queue, we assume that there is a single server and customers are served in the order of their arrivals. The service times for all customers are independent and exponentially distributed with rate μ, which are assumed to be independent of everything else.

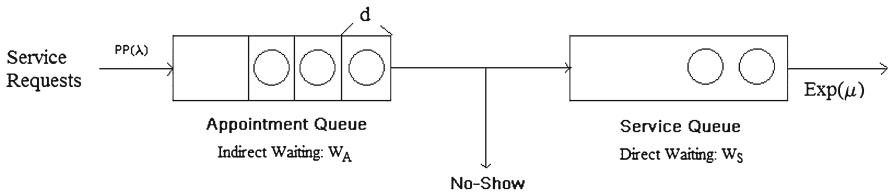

Fig. 1 Tandem Queueing model

3 Infinite capacity appointment queue

Given the model description in Sect. 2, the infinite capacity appointment queue can be seen as a "modified" $M/G/1$ queue as we describe in detail in this section. It is important to highlight the fact that the appointment queue is *not* a standard $M/G/1$ queue or more specifically not an $M/D/1$ queue because of the fact that the service time of a customer who enters an empty appointment queue is shorter than d while the service time of a customer who enters a non-empty appointment queue is always d. (Note that for the appointment queue, service time does not correspond to time spent for an actual service.)

Let $X(t)$ be the number of customers waiting in the appointment queue at time t. Define

$$X_n = X(nd^-), \quad n = 0, 1, 2, \ldots.$$

Note that $\{X_n, n \geq 0\}$ is *not* the usual embedded Markov chain obtained by observing an $M/D/1$ queue after each departure, although we shall see that it has the same transition probability matrix. Define (assuming the limits exist)

$$\pi_i^A = \lim_{n\to\infty} P(X_n = i), \quad i = 0, 1, 2, \ldots$$

and

$$p_i^A = \lim_{t\to\infty} P(X(t) = i), \quad i = 0, 1, 2, \ldots.$$

Also define the generating function

$$\phi(z) = \sum_{j=0}^{\infty} \pi_j^A z^j.$$

Let A_n be the number of arrivals during $[(n-1)d, nd)$, $(n \geq 1)$, and define

$$\rho = \lambda d, \quad a_i = P(A_n = i) = e^{-\rho}\frac{\rho^i}{i!}, \quad i = 0, 1, 2, \ldots,$$
$$\psi(z) = \sum_{j=0}^{\infty} a_j z^j = e^{\rho(z-1)}. \tag{1}$$

With this notation we have the following:

Theorem 1 *$\{X_n,\ n \geq 0\}$ is an aperiodic DTMC with the state space $S = \{0, 1, 2, \ldots\}$ and the transition probability matrix T given by:*

$$T = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & \dots \\ a_0 & a_1 & a_2 & a_3 & \dots \\ & a_0 & a_1 & a_2 & \dots \\ & & a_0 & a_1 & \dots \\ & & & a_0 & \dots \\ & & & & \ddots \end{bmatrix}. \tag{2}$$

It is positive recurrent if and only if

$$\rho < 1. \tag{3}$$

Assuming positive recurrence, we have

$$\phi(z) = (1-\rho)\frac{(1-z)\psi(z)}{\psi(z)-z}. \tag{4}$$

Furthermore,

$$p_i^A = \frac{\pi_{i+1}^A}{\rho}, \quad i = 0, 1, 2, \dots. \tag{5}$$

Proof We have $X_{n+1} = \max(X_n - 1, 0) + A_{n+1}$. The theorem follows from this and the fact that $\{A_n, n \geq 1\}$ are iid. The derivation of the limiting generating function is standard (see [22, Thm. 7.10]).

Next we prove Eq. 5. Since the appointment queue length process has jumps of size ± 1, the number of appointments as seen by an arrival to the appointment queue and the number of appointments left behind by a departure from the appointment queue have the same limiting distribution. The limiting probability that a departure leaves i customers behind is the limiting probability that there are $i+1$ customers in the appointment queue at nd^-, $n \to \infty$. The result then follows from using PASTA and $\pi_0^A = 1 - \rho$. □

As the reader might have noticed, T, the transition probability matrix for the DTMC $\{X_n,\ n \geq 0\}$ is structurally the same as the transition probability matrix of the DTMC one would get by looking at departure points in a standard $M/G/1$ queue. But, it is important to note that in our case the DTMC is constructed differently by looking at the system state right before every potential appointment time (integer multiples of d). Hence, even though the limiting distribution for the two DTMCs would be structurally identical, the way one would obtain the limiting distribution for the corresponding queue length processes (when described as continuous-time processes) is different.

In particular, using Theorem 1, we can readily obtain the following results for a stable appointment queue. First, using $\pi^A = \pi^A T$, we get

$$\pi_0^A = 1 - \rho, \quad \pi_1^A = (1-\rho)(e^\rho - 1).$$

Then using Eq. 5, we get

$$p_0^A = \frac{\pi_1^A}{\rho} = \frac{(1-\rho)(e^\rho - 1)}{\rho} > 1 - \rho.$$

We know that $1 - \rho$ is the probability that an $M/D/1$ system is empty. Thus the probability that the appointment queue is empty is higher than the probability that the $M/D/1$ system is empty, again emphasizing the fact that $\{X(t), t \geq 0\}$ is not an $M/D/1$ queue. The limiting expected number of appointments is given by

$$L_A = \lim_{t\to\infty} E(X(t)) = \frac{\rho}{2(1-\rho)},$$

and the limiting expected indirect delay is given by

$$W_A = \frac{d}{2(1-\rho)}. \tag{6}$$

4 Service queue with no-shows

Next we study the service queue. Since the customers departing the appointment queue either join service queue or become no-shows, it is clear that the appointment queue and the service queue are dependent. Hence we study their joint behavior in this section.

The no-show model needs to capture the commonly observed phenomenon that the no-show probability increases with indirect waiting time. Here, we consider a no-show behavior model described by a single parameter $\alpha \in (0, 1]$, which indirectly induces this characteristic. Specifically, we assume that if a customer departing the appointment queue leaves behind k customers in the appointment queue, he will join the service queue with probability α^k or become a no-show with probability $1 - \alpha^k$. This assumed structure on no-show probabilities leads to some convenient properties, which in the end make our analysis possible.

Note that if a customer departing the appointment queue leaves behind k customers there, then these k customers must have arrived during this customer's indirect waiting time. So k is positively correlated with the indirect waiting time. Thus the show-up probability decreases as k increases, or as the indirect waiting time increases (although the precise dependence on the indirect waiting time is much more complicated). In addition, $0 < \alpha \leq 1$ is the no-show parameter that reflects how sensitive customers are to their indirect waits, lower α implies more sensitivity. Green and Savin [12] use a similar idea for modeling customer no-show behavior.

Now let Y_n be the number of customers in the service queue at time nd^-, $n = 1, 2, \ldots$. The next theorem describes the bivariate stochastic process $\{(Y_n, X_n), n \geq 0\}$, where X_n is as in the previous section. We define $Q = \text{diag}(0, \alpha^0, \alpha^1, \alpha^2, \ldots)$, $M = (I - Q)T$, $N = QT$, where T is as in Eq. 2. We also let $A_k = b_{k-1}M + b_k N$, $k \geq 0$, where $b_i = e^{-\mu d}\frac{(\mu d)^i}{i!}$, $i = 0, 1, 2, \ldots, b_{-1} = 0$, and $B_k = \sum_{i=k}^{\infty} A_{i+1}$, $k = 0, 1, \ldots$. Then, we have:

Theorem 2 *$\{(Y_n, X_n), n = 1, 2, \ldots\}$ is a two-dimensional DTMC with the state space $\{(i, j) : i \geq 0, j \geq 0\}$ and transition probability matrix*

$$P = \begin{bmatrix} B_0 & A_0 & & \\ B_1 & A_1 & A_0 & \\ B_2 & A_2 & A_1 & A_0 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

It is positive recurrent if

$$\rho < 1, \tag{7}$$

$$\text{and } \lambda < \mu. \tag{8}$$

Proof We prove the theorem by using Foster's criterion. Let $C_n = 1$ if a customer joins the service queue at time nd, and zero otherwise. We have

$$P(C_n = 1 | X_n = j) = \begin{cases} \alpha^{j-1} & \text{if } j \geq 1 \\ 0 & \text{if } j = 0. \end{cases} \tag{9}$$

Also, let D_n be the number of customers who leave the service queue in $[(n-1)d, nd)$, $n \geq 1$. Then

$$(Y_{n+1}, X_{n+1}) = (Y_n + C_n - D_{n+1}, (X_n - 1)^+ + A_{n+1}), \quad n \geq 0.$$

Since D_{n+1} depends only on $Y_n + C_n$ (in fact, $D_{n+1} \sim \min(P(\mu d), Y_n + C_n)$, where $P(\mu d)$ is a Poisson random variable with parameter μd), and C_n depends only on X_n and $\{A_n, n \geq 1\}$ are iid, it follows that $\{(Y_n, X_n), n \geq 0\}$ is a DTMC. The transition probability matrix follows in a straightforward manner.

Now consider the potential function $v(i, j) = i + j$. Then the drift can be computed as

$$\begin{aligned} d(i, j) &= E(v(Y_{n+1}, X_{n+1}) - v(Y_n, X_n) | Y_n = i, X_n = j) \\ &= \begin{cases} \rho - 1 + \alpha^{j-1} - E(\min(P(\mu d), i + C_n | X_n = j)) & \text{if } j \geq 1 \\ \rho - 1 - E(\min(P(\mu d), i)) & \text{if } j = 0. \end{cases} \end{aligned}$$

Suppose the conditions in Eqs. 7 and 8 hold. Fix $0 < \epsilon < \min(1 - \rho, \mu d - \rho)$, and define

$$\begin{aligned} i^* &= \min\{i \geq 0 : \mu d - E(\min(P(\mu d), i + C_n | X_n = j)) < \epsilon\}, \\ j^* &= \min\{j \geq 1 : \alpha^{j-1} \leq \epsilon\}. \end{aligned}$$

Note that i^* is finite since $E(\min(P(\mu d), i + C_n | X_n = j))$ monotonically increases to μd as $i \to \infty$. Consider the finite set $H = \{(i, j) : i < i^*, j < j^*\}$. Now suppose

$(i, j) \notin H$. Then, either $i \geq i^*$ or $j \geq j^*$. If $i \geq i^*$, then $\rho < 1$ implies that we have

$$d(i, 0) = \rho - 1 - E(\min(P(\mu d), i)) < 0$$

and $\lambda < \mu$ implies that

$$\begin{aligned} d(i, j) = \rho - \mu d - (1 - \alpha^{j-1}) \\ + (\mu d - E(\min(P(\mu d), i + C_n | X_n = j))) < 0, \quad j \geq 1. \end{aligned}$$

Similarly, if $j \geq j^*$, then we get

$$d(i, j) = \rho - 1 + \alpha^{j-1} - E(\min(P(\mu d), i + C_n | X_n = j)) < 0, \quad i \geq 0.$$

Thus the drift is strictly negative for all $(i, j) \notin H$. Hence by Foster's criterion, the conditions in Eqs. 7 and 8 are sufficient for positive recurrence of $\{(Y_n, X_n), n \geq 0\}$. □

The next theorem gives a stronger sufficient condition for the stability of $\{Y_n, n \geq 0\}$.

Theorem 3 $\{Y_n, n \geq 0\}$ *is stable* (*that is, has a limiting distribution*) *if*

$$\rho < 1, \quad \textit{and} \tag{10}$$

$$\rho - \alpha\mu d + \phi(\alpha) < 1, \tag{11}$$

where ϕ *is as in Eq. 4.*

Proof Suppose $X_0 = 0$. Let $N_0 = 0$ and

$$N_{r+1} = \min\{n > N_r : X_n = 0\}, \quad r \geq 0.$$

Then $\tilde{Y}_r = Y_{N_r}$ is a Markov chain, and $\{Y_n, n \geq 0\}$ is a Markov regenerative process with the embedded Markov renewal sequence $\{(\tilde{Y}_r, N_r), r \geq 0\}$. See [22] for relevant definitions. We know that $\beta = E(N_{r+1} - N_r)$ is the expected value of a busy cycle in the appointment queue, and it is finite if the condition in Eq. 10 is satisfied. Hence, from Theorem 9.9 in [22], it suffices to prove that $\{\tilde{Y}_n, n \geq 0\}$ is positive recurrent. We use Foster's criterion to prove this. We have

$$\tilde{Y}_{r+1} = \tilde{Y}_r + \sum_{n=N_r+1}^{N_{r+1}} (C_n - D_{n+1}).$$

Using the potential function $v(i) = i$, we see that the drift is given by

$$d(i) = E(\tilde{Y}_{r+1} - \tilde{Y}_r | \tilde{Y}_r = i) = E\left(\sum_{n=N_r+1}^{N_{r+1}} C_n\right) - E\left(\sum_{n=N_r+1}^{N_{r+1}} D_{n+1} | \tilde{Y}_r = i\right).$$

Now $\{C_n, n \geq 0\}$ is a regenerative process that regenerates at time $\{N_r, r \geq 0\}$. We also know that $\gamma = \lim_{n\to\infty} E(C_n) = \lim_{n\to\infty} E(\alpha^{X_n - 1} 1_{\{X_n > 0\}}) = (\phi(\alpha) - (1-\rho))/\alpha$, where ϕ is the limiting generating function of X_n, as given in Eq. 4. Using the theory of regenerative processes, we see that

$$E\left(\sum_{n=N_r+1}^{N_{r+1}} C_n\right) = \beta\gamma, \text{ and } E\left(\sum_{n=N_r+1}^{N_{r+1}} D_{n+1} | \tilde{Y}_r = i\right) \leq \beta\mu d.$$

Furthermore, $E\left(\sum_{n=N_r+1}^{N_{r+1}} D_{n+1} | \tilde{Y}_r = i\right)$ increases to $\beta\mu d$ as $i \to \infty$. Now suppose

$$\beta\mu d - \beta\gamma > 0. \tag{12}$$

Then, for any $0 < \epsilon < \beta\mu d - \beta\gamma$ there exists an i^* such that for $i > i^*$

$$\begin{aligned} d(i) &= E\left(\sum_{n=N_r+1}^{N_{r+1}} C_n\right) - E\left(\sum_{n=N_r+1}^{N_{r+1}} D_{n+1} | \tilde{Y}_r = i\right) \\ &= \beta\gamma - \beta\mu d + \left(\beta\mu d - E\left(\sum_{n=N_r+1}^{N_{r+1}} D_{n+1} | \tilde{Y}_r = i\right)\right) \\ &< \beta\gamma - \beta\mu d + \epsilon < 0. \end{aligned}$$

Equation 12 can be rearranged to get Eq. 11. The result then follows from Foster's criterion. □

Now suppose the stability conditions in Eqs. 10 and 11 are satisfied. Define

$$\pi_{k,j} = \lim_{n\to\infty} P(Y_n = k, X_n = j), \quad k, j \geq 0.$$

Let

$$\pi_k = [\pi_{k,0}, \pi_{k,1}, \ldots], \quad k \geq 0$$

and e be a column vector with all coordinates equal to 1. The following theorem gives the structure of the limiting distribution of the DTMC $\{(Y_n, X_n), n = 1, 2, \ldots\}$.

Theorem 4 *Let R be a matrix that satisfies*

$$R = Re^{\mu d(R-I)}M + e^{\mu d(R-I)}N. \tag{13}$$

Let η be a solution to

$$\begin{aligned} &\eta = \eta(I - R)^{-1}(T - R), \\ &\eta(I - R)^{-1}e = 1. \end{aligned} \tag{14}$$

Then the limiting distribution is given by

$$\pi_k = \eta R^k, \quad k \geq 0.$$

Proof From the standard analysis of $G/M/1$ type queue (see [29]), (although with infinite number of phases) we know that

$$\pi_k = \pi_0 R^k, \quad k = 1, 2, \ldots,$$

where

$$R = \sum_{k=0}^{\infty} R^k A_k, \quad \pi_0 \left(I - \sum_{k=0}^{\infty} R^k B_k \right) = 0, \quad \pi_0 (I - R)^{-1} e = 1.$$

We can also show with some algebraic manipulations that

$$\begin{aligned}
\sum_{k=0}^{\infty} R^k A_k &= R \sum_{k=0}^{\infty} R^k b_k M + \sum_{k=0}^{\infty} R^k b_k N = R e^{(R-I)\mu d} M + e^{(R-I)\mu d} N \\
\sum_{k=0}^{\infty} R^k B_k &= \sum_{k=0}^{\infty} R^k \sum_{i=k}^{\infty} (b_i M + b_{i+1} N) = \sum_{i=0}^{\infty} \sum_{k=0}^{i} R^k (b_i M + b_{i+1} N) \\
&= \sum_{i=0}^{\infty} (I - R)^{-1} (I - R^{i+1}) (b_i M + b_{i+1} N) \\
&= (I - R)^{-1} \left(T - R \sum_{i=0}^{\infty} R^i b_i M - \sum_{i=0}^{\infty} R^i b_i N \right) \\
&= (I - R)^{-1} (T - R).
\end{aligned}$$

Hence the result. □

Unfortunately, the above theorem does not yield a computational method to compute the joint distribution of (Y_n, X_n), since the R matrix is of infinite size. One way to circumvent this difficulty is to restrict the $\{X_n, n \geq 0\}$ process to a finite state-space. We study this model in the next section.

5 Finite appointment queue

Suppose that the maximum size of the appointment queue is $K < \infty$. Then, $\{X_n, n \geq 0\}$ is a DTMC with state-space $\{0, 1, \ldots, K\}$ and transition probability matrix:

$$\tilde{T} = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & \dots & a_{K-1} & 1 - \sum_{k=0}^{K-1} a_k \\ a_0 & a_1 & a_2 & a_3 & \dots & a_{K-1} & 1 - \sum_{k=0}^{K-1} a_k \\ & a_0 & a_1 & a_2 & \dots & a_{K-2} & 1 - \sum_{k=0}^{K-2} a_k \\ & & & \ddots & & & \vdots \\ & & & & & a_0 & 1 - a_0 \end{bmatrix}.$$

Let $\tilde{\phi}(z) = \lim_{n\to\infty} E(z^{X_n})$ be its limiting generating function. Then we can think of $\{(Y_n, X_n), n \geq 0\}$ as a $G/M/1$ type queue with X_n as the phase, and Y_n as the queue-length. One can use the ideas of Theorem 3 to show that this queue is stable if

$$\rho - \alpha\mu d + \tilde{\phi}(\alpha) < 1. \tag{15}$$

Let

$$\tilde{\pi}_{i,j} = \lim_{n\to\infty} P(Y_n = i, X_n = j), \quad i \geq 0, \quad 0 \leq j \leq K, \text{ and } \tilde{\pi}_i = [\tilde{\pi}_{i,0}, \tilde{\pi}_{i,1}, \dots, \tilde{\pi}_{i,K}].$$

Theorem 5 *Suppose the condition in Eq. 15 is satisfied. Let $\tilde{R}$ be the minimal matrix R that satisfies*

$$R = Re^{\mu d(R-I)}\tilde{M} + e^{\mu d(R-I)}\tilde{N}, \tag{16}$$

where $\tilde{Q} = \text{diag}(0, \alpha^0, \alpha^1, \alpha^2, \dots, \alpha^{K-1})$, $\tilde{M} = (I - \tilde{Q})\tilde{T}$, $\tilde{N} = \tilde{Q}\tilde{T}$. Let $\tilde{\eta}$ be a solution to

$$\tilde{\eta} = \tilde{\eta}(I - \tilde{R})^{-1}(\tilde{T} - \tilde{R}), \quad \tilde{\eta}(I - \tilde{R})^{-1}e = 1.$$

Then the limiting distribution is given by

$$\tilde{\pi}_i = \tilde{\eta}\tilde{R}^i, \quad i \geq 0.$$

Proof The proof is very similar to that of Theorem 4 and is omitted. The existence of $\tilde{R}$ follows from the general theory of $G/M/1$ type queues. □

There exist efficient algorithms to compute the $(K + 1) \times (K + 1)$ matrix $\tilde{R}$, and hence the above theorem provides an efficient method of computing the limiting distribution of $\{(Y_n, X_n), n \geq 0\}$.

5.1 Performance measures

Next we obtain three performance measures of interest for the service queue, namely, the long-run average direct waiting time, the long-run fraction of time that the server is idle, and the probability that a customer arriving to the server queue is delayed assuming a finite appointment queue of size k. Let $Z(t) = (Y(t), \tilde{X}(t))$, $t \geq 0$, where $\tilde{X}(t) = X_n$, $nd \leq t < (n + 1)d$, and $Y(t)$ is the number of customers in the service queue at time t. Then $\{Z(t), t \geq 0\}$ is a Markov regenerative process with an embedded Markov renewal sequence $\{((Y_n, X_n), nd), n = 1, 2, \dots\}$. Hence the

steady-state distribution of DTMC $\{(Y_n, X_n), n = 1, 2, \ldots\}$, namely $\tilde{\pi}_i, i = 0, 1, \ldots$, can be used to derive the steady-state joint distribution of $Z(t)$.

Define $\alpha_{k,j} = E(\text{Time spent by } Y(t) \text{ in state } j \text{ during } [0, d) \mid Y(0) = k)$. Then for $j = 1, 2, \ldots, k = j, j + 1, \ldots,$

$$\alpha_{k,j} = \int_0^d \frac{(\mu t)^{k-j} e^{-\mu t}}{(k-j)!}\, \mathrm{d}t = \frac{1}{\mu}\left(1 - \sum_{r=0}^{k-j} b_r\right).$$

Theorem 6 *Let* $p_j = \lim_{t\to\infty} P(Y(t) = j), \quad j = 0, 1, 2, \ldots,$ *and* $e = [1, 1, \ldots, 1]$, *then*

$$p_j = \frac{1}{\mu d}\tilde{\pi}_{j-1}\tilde{Q}e, \quad j = 1, 2, \ldots, \tag{17}$$
$$p_0 = 1 - \frac{\tilde{\pi}_0}{\mu d}(I - \tilde{R})^{-1}\tilde{Q}e,$$

where $\tilde{\pi}_0 = \tilde{\eta}$ (*as shown in Theorem 5*). *The steady-state service queue length is*

$$L_S = \frac{\tilde{\pi}_0}{\mu d}(I - \tilde{R})^{-2}\tilde{Q}e.$$

The long-run average direct delay is

$$W_S = \frac{\tilde{\pi}_0}{\lambda_S \mu d}(I - \tilde{R})^{-2}\tilde{Q}e, \tag{18}$$

where

$$\lambda_S = \frac{\tilde{\phi}(\alpha) - \tilde{\phi}(0)}{\alpha d},$$

and the probability of delay in the service queue is

$$p_{\text{delay}} = 1 - \frac{\tilde{\pi}_0\tilde{Q}e}{\tilde{\pi}_0(I - \tilde{R})^{-1}\tilde{Q}e}.$$

Proof First we have the following equalities

$$\sum_{k=j-1}^{\infty}\sum_{r=0}^{k+1-j} b_r\tilde{\pi}_k = \sum_{i=0}^{\infty}\sum_{r=0}^{i} b_r\tilde{\pi}_{i+j-1} = \sum_{r=0}^{\infty}\sum_{i=r}^{\infty} b_r\tilde{\pi}_{i+j-1} = \sum_{k=j}^{\infty}\sum_{r=0}^{\infty} b_r\tilde{\pi}_{k-1+r}, \tag{19}$$

$$\sum_{k=j}^{\infty}\sum_{r=0}^{k-j} b_r\tilde{\pi}_k = \sum_{k=0}^{\infty}\sum_{r=0}^{k} b_r\tilde{\pi}_{k+j} = \sum_{r=0}^{\infty}\sum_{k=r}^{\infty} b_r\tilde{\pi}_{k+j} = \sum_{k=j}^{\infty}\sum_{r=0}^{\infty} b_r\tilde{\pi}_{k+r}. \tag{20}$$

In addition, from P (as defined in Theorem 2 whose block matrices now have finite dimension $K+1$) and the fact that $Pe = e$ we have

$$\tilde{\pi}_k e = \sum_{r=0}^{\infty} \tilde{\pi}_{k+r} b_r (I - \tilde{Q}) e + \sum_{r=0}^{\infty} \tilde{\pi}_{k+r-1} b_r \tilde{Q} e. \tag{21}$$

Thus, from the standard analysis of Markov regenerative processes (see [22, Thm. 9.30]), with some algebraic manipulations, we can show that

$$p_j = \frac{1}{\mu d} \left[\sum_{k=j-1}^{\infty} \tilde{\pi}_k \left(1 - \sum_{r=0}^{k+1-j} b_r \right) \tilde{Q} e + \sum_{k=j}^{\infty} \tilde{\pi}_k \left(1 - \sum_{r=0}^{k-j} b_r \right) (I - \tilde{Q}) e \right]$$

(Using (19), (20), and (21))

$$= \frac{1}{\mu d} \left[\tilde{\pi}_{j-1} \tilde{Q} e + \sum_{k=j}^{\infty} \left(\tilde{\pi}_k e - \sum_{r=0}^{\infty} b_r \tilde{\pi}_{k-1+r} \tilde{Q} e - \sum_{r=0}^{\infty} b_r \tilde{\pi}_{k+r} (I - \tilde{Q}) e \right) \right]$$

$$= \frac{\tilde{\pi}_{j-1}}{\mu d} \tilde{Q} e, \quad j = 1, 2, \ldots.$$

Hence the steady-state service queue length is

$$L_S = \sum_{j=1}^{\infty} j p_j = \frac{\tilde{\pi}_0}{\mu d} (I - \tilde{R})^{-2} \tilde{Q} e.$$

Note that the actual arrive rate to the service queue is $\lambda_S = \frac{1}{d} \sum_{i=1}^{K} \tilde{\pi}_i^A \alpha^{i-1} = \frac{\tilde{\phi}(\alpha) - \tilde{\phi}(0)}{\alpha d}$, where $\tilde{\pi}_i^A$ is the finite capacity version of π_i^A as defined in Sect. 3. Applying Little's Law, we obtain the long-run average direct waiting time as follows:

$$W_S = \frac{\tilde{\pi}_0}{\lambda_S \mu d} (I - \tilde{R})^{-2} \tilde{Q} e.$$

Finally, the expression for p_{delay} can easily be obtained by computing the probability that the server is busy at an appointment time conditional on there is an arrival at that time, i.e.,

$$p_{\text{delay}} = 1 - \frac{\sum_{k=1}^{K} \tilde{\pi}_{0,k} \alpha^{k-1}}{\sum_{i=0}^{\infty} \sum_{k=1}^{K} \tilde{\pi}_{i,k} \alpha^{k-1}} = 1 - \frac{\tilde{\pi}_0 \tilde{Q} e}{\tilde{\pi}_0 (I - \tilde{R})^{-1} \tilde{Q} e}.$$

□

6 Approximating the expected direct waiting time

If the appointment queue has a finite capacity, i.e., the service provider does not schedule new appointments when the number of scheduled appointments exceeds some fixed finite level, then our results in Sect. 5 can be used to compute the expected waiting time in the service queue. However, if there is no such queueing capacity restriction, computation of the expected direct waiting time is not possible as it requires complete knowledge of the infinite size matrix $\tilde{R}$, which cannot be obtained. Therefore, it is of interest to develop methods that help determine the expected direct waiting time approximately. One way to obtain an approximation is by setting the appointment queue capacity to some arbitrarily large number and using the corresponding expected waiting time. Again, our results in Sect. 5 can be used to that end. Nevertheless, it must be noted that the larger the assumed queue capacity, the better will be the approximation but also the longer it would take to compute the matrix $\tilde{R}$ and the corresponding expectations. Thus, developing approximations that are computationally more efficient would be important.

It is also important to note that while our model and results are based on the assumption that service times are exponentially distributed, it is known that in practice exponential distribution may not be a good choice for service times. Many studies have in fact found lognormal distribution to be a better fit (see for example [3,20]). Thus, approximations are also needed to investigate systems where service times have some general distributions.

6.1 Approximating the service queue by a $G/M/1$ queue

The main difficulty in analyzing the service queue arises from the fact that its arrival process is not a renewal process. However, by assuming that it is a renewal process and appropriately choosing the probability distribution for the interarrival times, we can use the known results for the $G/M/1$ queue to approximate the expected waiting time in the service queue. In this section, we develop three different methods of approximating the service queue as a $G/M/1$ queue. To our knowledge, while the first two (Sects. 6.1.1 and 6.1.2) have not been proposed and studied before, the third one (Sect. 6.1.3) is due to [11].

6.1.1 Iid arrivals

Recall the definition of C_n from the proof of Theorem 2. Its dependence on X_n is given by Eq. 9. Thus, if X_0 has the stationary distribution, then $\{C_n, n \geq 0\}$ is a stationary process with marginal distribution given by

$$P(C_n = 1) = \sum_{i=1}^{\infty} \alpha^{i-1} P(X_n = i) = (\phi(\alpha) - \phi(0))/\alpha, \tag{22}$$

$$P(C_n = 0) = 1 - (\phi(\alpha) - \phi(0))/\alpha, \tag{23}$$

where the generating function ϕ is as given in Eq. 4. Clearly, $\{C_n, n \geq 0\}$ is not a Markov chain, even though $\{X_n, n \geq 0\}$ is. However, we can approximate it to generate tractable approximations to the service queue.

The simplest approximation is to *assume* that $\{C_n, n \geq 0\}$ is a sequence of iid random variables with common distribution given by Eqs. 22 and 23. Let $\beta = (\phi(\alpha) - \phi(0))/\alpha$. Then it is clear that the inter-arrival times to the service queue are iid with $P(\text{Inter-arrival time } = kd) = (1-\beta)^{k-1}\beta, k \geq 1$. Thus, under this assumption, the service queue becomes a standard $G/M/1$ queue with the Laplace–Stieltjes transform (LST) of the inter-arrival times given by

$$\tilde{G^I}(s) = \frac{\exp(-sd)\beta}{1 - \exp(-sd)\beta}.$$

Using standard results from $G/M/1$ queues (see [22]) we see that the service queue is stable if $\beta < \mu d$. Assuming stability, there exists a unique solution in (0, 1) to the equation

$$\theta^I = \tilde{G^I}(\mu(1-\theta^I)).$$

We use the superscript I to indicate the iid assumption. This equation can be solved numerically very easily by iterative techniques. The expected waiting time in the service system in steady state is given by

$$W_S^I = \frac{1}{\mu(1-\theta^I)}.$$

6.1.2 Markovian arrivals

The next simplest method to approximate the process $\{C_n, n \geq 0\}$ (see Sect. 6.1.1) is to *assume* that $\{C_n, n \geq 0\}$ is a stationary Markovian sequence. Define

$$\Theta(i, j) = P(C_{n+1} = j | C_n = i), \quad i, j = 0, 1.$$

After some tedious calculations, we can show that

$$\Theta(1,1) = \frac{\psi(\alpha)}{\alpha^2} \cdot \frac{\phi(\alpha^2) - \phi(0)}{\phi(\alpha) - \phi(0)} - \frac{\psi(0)\pi_1^A}{\phi(\alpha) - \phi(0)},$$
$$\Theta(0,1) = \frac{\alpha^2(\psi(\alpha) - \psi(0)) + \alpha\psi(\alpha)(\phi(\alpha) - \phi(0)) - \psi(0)(\phi(\alpha^2) - \phi(0))}{\alpha^2(\alpha - (\phi(\alpha) - \phi(0)))},$$

where ψ is given by Eq. 1, and

$$\phi(0) = 1 - \rho, \quad \psi(0) = \exp(-\rho), \quad \pi_1^A = (\exp(\rho) - 1)(1 - \rho).$$

Note also that $\Theta(1,0) = 1 - \Theta(1,1), \quad \Theta(0,0) = 1 - \Theta(0,1)$.

Due to the assumed Markovian nature of the $\{C_n, n \geq 0\}$ process, the inter-arrival times to the service are again iid with common distribution

$$P(\text{Inter-arrival time} = kd) = \begin{cases} \Theta(1,1) & \text{if } k = 1, \\ \Theta(1,0)\Theta(0,0)^{k-2}\Theta(0,1) & \text{if } k \geq 2. \end{cases}$$

Thus the service queue is again a standard $G/M/1$ queue, with the LST of the inter-arrival time

$$\tilde{G}^M(s) = \frac{\exp(-sd)\Theta(1,1) - \exp(-2sd)\det(\Theta)}{1 - \exp(-sd)\Theta(0,0)},$$

where $\det(\Theta)$ is the determinant of Θ. Assuming stability, there exists a unique solution in $(0, 1)$ to the equation

$$\theta^M = \tilde{G}^M(\mu(1 - \theta^M)).$$

This equation can be solved numerically very easily by iterative techniques. The expected waiting time in the service system in steady state is given by

$$W_S^M = \frac{1}{\mu(1 - \theta^M)}.$$

Clearly the Markovian approximation is expected to work better than the iid approximation. One can further improve the approximation by assuming that $\{C_n, n \geq 0\}$ is a Markov process of order $k \geq 2$, that is, by assuming C_{n+1} depends on the history only via $C_n, C_{n-1}, \cdots, C_{n-k+1}$. However this further improvement is obtained at a considerable cost. First, computing the transition probabilities becomes more formidable, and second, the service queue is no longer a $G/M/1$ queue, and has to be analyzed using phase type methods. Hence we shall not pursue this avenue any further.

6.1.3 Tandem queueing network approximation

We are aware of one existing method that can be used to approximate the expected direct waiting time. It is called the tandem queueing network (TQN) approximation and was developed by Girish and Hu [11]. In this section, we apply this general approximation method to our model. The idea behind TQN is to ignore the dependence between the appointment queue and the service queue and approximate the arrival process to the service queue by a renewal process with interrenewal times defined by a phase type distribution (specifically a mixed Erlang distribution) whose parameters are chosen so that the first three moments of the distribution match with the first three non-central moments of the interdeparture times from the appointment queue.

Let D denote the time between two arbitrary consecutive departures from the appointment queue and let $M_i = E(D^i)$ for $i = 1, 2, 3$. Then, we can prove the following theorem.

Theorem 7 *Suppose that the show-up probability parameter $\alpha = 1$. Then first three non-central moments of the departure process from the appointment queue are given by*

$$M_1 = \frac{1}{\lambda},\ M_2 = d^2\left[1 - p_0^A + p_0^A \frac{1+a_0}{(1-a_0)^2}\right],\ M_3 = d^3\left[1 - p_0^A + p_0^A \frac{1+4a_0+a_0^2}{(1-a_0)^3}\right],$$

where $p_0^A = \frac{1-\lambda d}{\lambda d}(e^{\lambda d} - 1)$ and $a_0 = e^{-\lambda d}$ (see Sect. 3 for the definition of p_0^A and a_0).

Proof Consider an arbitrary departure from the appointment queue. Theorem 1 together with its proof shows that this arbitrary departure leaves an empty appointment queue behind with probability p_0^A. So with probability $1 - p_0^A$, the time until the next departure is d and with probability p_0^A it is Nd where N is a geometric random variable with parameter $1 - a_0$. Hence

$$M_1 = (1 - p_0^A)d + p_0^A \sum_{i=1}^{\infty} a_0^{i-1}(1-a_0)id = d\left[1 - p_0^A + \frac{p_0^A}{1-a_0}\right] = \frac{1}{\lambda}.$$

M_2 and M_3 can be obtained similarly. □

The calculations required to determine M_2 and M_3 are too involved when $\alpha < 1$. Therefore, to make a comparison between TQN approximation and other approximations, we only consider scenarios with $\alpha = 1$.

Using a procedure provided by Johnson and Taaffe [18], we construct a probability distribution whose first three moments are given as above. Specifically, we first define a random variable $\tilde{D} = \tilde{p}\text{Erlang}(n, \lambda_1) + (1 - \tilde{p})\text{Erlang}(n, \lambda_2)$ and choose $\tilde{p}$, n, λ_1, and λ_2 so that the moments of the probability distribution for $\tilde{D}$ are the same as those for the distribution of D. This gives us

$$n = \min\left\{k \in \mathbb{Z} \,|\, k \geq \frac{M_1^2}{M_2 - M_1^2},\ k \geq \frac{2(M_2 - M_1^2)^2 + M_1^2 M_2 - M_1 M_3}{M_1 M_3 - (M_2 - M_1^2)(M_2 - 2M_1^2)}\right\},$$
$$\lambda_1 = \frac{2D_1}{D_2 + \sqrt{D^2 - 4D_1 D_3}},\ \lambda_2 = \frac{2D_1}{D_2 - \sqrt{D^2 - 4D_1 D_3}},\ \tilde{p} = \lambda_1 \frac{1 - \lambda_2 A}{\lambda_1 - \lambda_2}.$$

where

$$A = \frac{M_1}{n},\ B = \frac{M_2}{n(n+1)},\ C = \frac{M_3}{n(n+1)(n+2)},$$
$$D_1 = A^2 - B,\ D_2 = AB - C,\ D_3 = B^2 - AC.$$

Let $\tilde{G}^T(s)$ be the LST of $\tilde{D}$. It is given by

$$\tilde{G}^T(s) = \tilde{p}\left(\frac{\lambda_1}{s+\lambda_1}\right)^n + (1-\tilde{p})\left(\frac{\lambda_2}{s+\lambda_2}\right)^n.$$

Then we approximate the service queue by a $G/M/1$ queue (or more precisely, a $PH/M/1$ queue) with iid inter-arrival times with LST given above. Here the superscript T stands for the TQN approximation.

The service queue is stable if $\lambda < \mu$ (since $\alpha = 1$). Assuming stability, there exists a unique solution in $(0, 1)$ to the equation

$$\theta^T = \tilde{G}^T(\mu(1 - \theta^T)).$$

and the expected waiting time in the service system in steady state is given by

$$W_S^T = \frac{1}{\mu(1 - \theta^T)}.$$

The extension of this method to non-exponential service times is discussed in Sect. 6.3.

6.2 Interpolation approximation for the service queue

The stability of the appointment queue requires that $d \in [0, \frac{1}{\lambda})$. We can obtain simple exact expressions for the expected direct waiting time when d approaches one of the either two end points of this interval. Let W_0 and $W_{1/\lambda}$ denote the limit of the expected direct waiting time as d approaches 0 and as d approaches $1/\lambda$, respectively.

When $d = 0$, the appointment queue disappears, the service queue works exactly like an $M/M/1$ queue and thus

$$W_0 = \frac{1}{\mu - \lambda}.$$

When $d = 1/\lambda$, the appointment queue is null-recurrent and there is a departure from the appointment queue (consequently an arrival to the service queue) every d units of time with probability 1. Hence the service queue can be seen as a $D/M/1$ queue with inter-arrival time $d = \frac{1}{\lambda}$. Let $\rho = \frac{\lambda}{\mu}$ be the traffic intensity. Then, from the standard $G/M/1$ queue analysis, we have

$$W_{1/\lambda} = \frac{1}{\mu(1 - \beta)},$$

where $\beta \in (0, 1)$ is the unique root of the following equation:

$$\beta = e^{-\frac{1-\beta}{\rho}}.$$

Our approximation is based on the idea that W_0 and $W_{1/\lambda}$ are, respectively, upper and lower bounds on the expected waiting time for any $d \in [0, 1/\lambda)$ and multiplying them with appropriate weights, i.e., using interpolation, should provide a reasonable approximation. More specifically, we propose

$$W_S^{\mathrm{INT}} = \gamma W_0 + (1-\gamma) W_{1/\lambda},$$

where γ is given by

$$\gamma = \rho^{\frac{\lambda d}{2(1-\lambda d)}}.$$

The expression we use for γ is a result of a number of analytical and numerical observations we made regarding the properties γ "should" satisfy. First, γ should ideally converge to 1 as d approaches 0 and converge to 0 as d approaches $1/\lambda$ so that $W_S^{\mathrm{INT}} = W_S$ at the two end points of the possible value interval for d. This is true for our choice of γ above. Second, our numerical study strongly suggested that W_S, the actual expected waiting time, is decreasing in d with a rate that is dependent on the traffic intensity (see Sect. 7 for details). Specifically, when the traffic intensity is high, as d increases from 0 to $1/\lambda$, the rate of decline in W_S is initially flat but increases exponentially after a certain point. On the other hand, when the traffic intensity is low, the rate of decline is flatter. This led us to numerically test several functions that are close relatives of the specific form of γ given above and we eventually picked this particular form as the best.

It is also important to note that the exponent in the expression for γ, i.e., $\frac{\lambda d}{2(1-\lambda d)}$, is the expected appointment queue length in steady state. Large expected appointment queue lengths indicate that the arrival process to the service queue is closer to a deterministic process and shorter expected appointment queue lengths indicate a more sporadic arrival process (closer to Poisson) for the service queue. This means that the approximation should be close to $W_{1/\lambda}$ when the expected appointment queue length is large and close to W_0 when the expected appointment queue length is small. This is also achieved by our choice of γ.

6.3 Approximation with general service time distribution

In this section, we explain how the TQN and the interpolation approximations can be generalized for systems where service time has general distribution. Denote the mean and variance of the service time by $\frac{1}{\mu}$ and σ_S^2, respectively. Recall that the arrival process to the appointment queue is Poisson with rate λ. Denote $\rho = \frac{\lambda}{\mu}$. We first consider the interpolation approximation method we propose in Sect. 6.2. When $d = 0$, the service queue operates as an $M/G/1$ queue, which means W_0 is given by the Pollaczek–Khintchine formula

$$W_0 = \frac{\rho^2 + \lambda^2 \sigma_S^2}{2\lambda(1-\rho)} + \frac{1}{\mu}.$$

When $d = \frac{1}{\lambda}$, the service queue operates as a $D/G/1$ queue. While we do not have an exact expression for the expected waiting time in this case, there are known approximations. In particular, we use the following formula to obtain the approximate expected direct delay (see $G/G/1$ queue approximation in [29, p. 341]):

$$W_{1/\lambda} = \frac{\lambda\sigma_S^2}{2(1-\rho)}\frac{\rho^2+\lambda^2\sigma_S^2}{1+\lambda^2\sigma_S^2} + \frac{1}{\mu}$$

with a slight abuse of notation since the above expression is not exact.

Then, letting $\gamma = \rho^{\frac{\lambda d}{2(1-\lambda d)}}$, as in Sect. 6.2, we propose

$$W_S^{\mathrm{INT}} = \gamma W_0 + (1-\gamma) W_{1/\lambda}$$

as an approximation for the expected direct delay when $d \in [0, 1/\lambda]$.

Next we generalize the TQN approximation method introduced in Sect. 6.1.3 to systems where service times have a general distribution. Recall that this approximation assumes that the arrival process to the service queue is a renewal process and the first three moments of the interarrival time distribution match with those for the interdeparture time for the appointment queue. Thus, in this case, the service queue is assumed to be a $G/G/1$ queue with mean and variance for interarrival times given by $M_1 = \frac{1}{\lambda}$ and $\sigma_A^2 = M_2 - M_1^2$, respectively. Hence, using the same $G/G/1$ queue approximation above we get

$$W_S^T = \frac{\lambda(\sigma_A^2+\sigma_S^2)}{2(1-\rho)}\frac{\rho^2+\lambda^2\sigma_S^2}{1+\lambda^2\sigma_S^2} + \frac{1}{\mu}.$$

Finally, note that using the same $G/G/1$ approximation for the expected waiting time, one can also generalize the iid and Markovian approximations to the setting where service times are not exponentially distributed.

7 Numerical study

In this section, we have three goals. First, we investigate the relationship between d, the time between two consecutive appointments, and the expected waiting times in both queues. This study demonstrates how our analysis of the Tandem Queueing model can be used to provide insights into appointment scheduling decisions and the trade-off between direct and indirect waiting times. Second, we study the effects of assuming a fixed no-show probability when in fact the true no-show probabilities depend on customers' appointment delays. Finally, we investigate how our approximations for the expected direct waiting time perform and how they compare with the existing alternative, the TQN approximation method.

One issue we needed to deal with for our computational study was to decide the length at which the appointment queue should be truncated. This is not an issue for the approximations but is needed to compute the expected waiting time in the service queue using the exact expressions for our Tandem Queueing model. To determine how large k, the capacity of the appointment queue, should be in order for our computations to be close enough to the actual numbers, we simulated the model under the infinite capacity assumption with various parameter settings, and compared the expected direct waiting time obtained through simulation with that obtained numerically under the assumption

that the appointment queue size was $K < \infty$. We gradually increased k and stopped when the simulation results were close enough to numerical results.

More specifically, we assumed exponential service times with rate μ, Poisson appointment request arrivals with rate λ. We fixed $\lambda = 1$ and constructed 114 scenarios by choosing μ from the set $\{1.1, 1.5, 2\}$, α from the set $\{0.95, 1\}$, varying d from 0.05 to 0.95 with increments of 0.05. For each scenario, we started with $K = 20$, simulated the system for 500 independent runs, and constructed 95 % confidence intervals for the difference between the mean direct delay obtained from simulation and that obtained numerically using our analysis. We repeated this procedure by gradually increasing K and stopped the first time the constructed confidence interval contained 0, which would mean that the numerical computation assuming finite capacity K would not be statistically different from the computation through simulation of the infinite capacity system at 95 % significance level. As a result, we found that for all the scenarios tested, $K = 60$ is large enough. In fact, in most cases where d is not close to 1, $K = 20$ is sufficient but in the following we set $K = 60$ regardless of the values of the other model parameters.

7.1 The effect of changing time between appointments on the expected waiting times

By choosing d, the time between two consecutive appointments, the service provider can influence the expected direct and indirect waiting times. The two waiting times are closely related to each other. Specifically, as one increases d, because of the appointment times that are more spread out, the expected indirect waiting time increases. On the other hand, more spread out appointments slow down the arrival process to the service queue and as a result the expected direct waiting time decreases. One can argue that direct waiting times are typically more "costly" than indirect waiting times since customers (or patients) are physically waiting in the service queue while waiting for an appointment does not prevent them from being engaged in other activities while waiting. One can easily develop an objective function that penalizes these two types of waiting appropriately and using our analytical results determine the optimal value of d that minimizes this function. Alternatively, instead of quantifying waiting costs, one can look at how the expected waiting times change with d, observe the trade-off between two types of waiting, and balance the two types of waiting appropriately.

Figures 2, 3, 4, and 5 demonstrate the effect of changes in d visually. Figure 2 assumes high traffic intensity and all customers showing-up for their appointments. As we can observe from the figure, changes in d do not have a significant impact on either type of waiting as long as it is less than 0.7. It appears that to reduce the expected direct waiting time in any significant way, d needs to be set to close to 1 but one should note that this comes with a significant increase in the expected indirect waiting time. Differing from Fig. 2, Fig. 3 assumes a light load on the system. Here, we can see that as in the heavy load case, only when d is sufficiently large (starting around 0.6), changes in d have significant effect on the waiting times. However, in the light load case, even when d is close to 1, we do not observe a significant decline in the expected direct waiting time while the expected indirect waiting time increases rapidly. This is because when the load on the system is light, even when customers

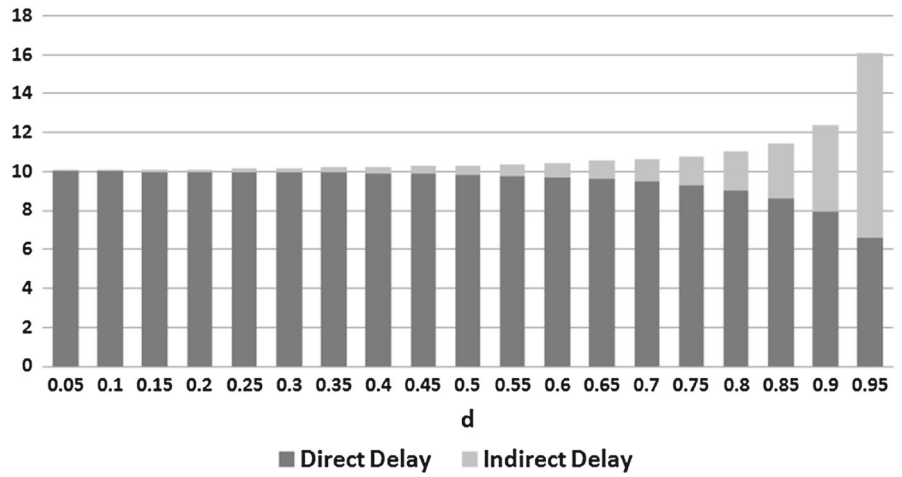

Fig. 2 $\lambda = 1, \mu = 1.1, \alpha = 1$

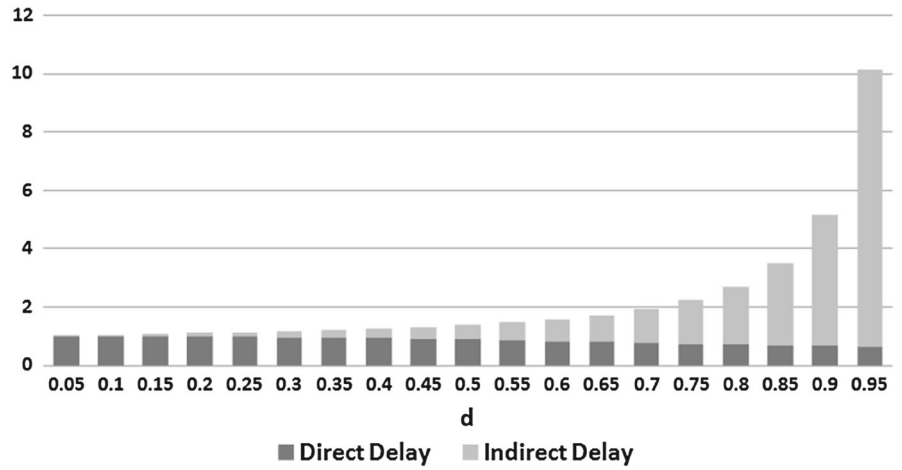

Fig. 3 $\lambda = 1, \mu = 2, \alpha = 1$

are not given appointments, the service queue will not tend to be long. As a result, giving appointments and increasing the time between appointments will simply add to the indirect waiting times without providing any help in reducing the direct waiting times. In short, perhaps not surprisingly, this observation suggests that service systems that are heavily loaded will benefit more from providing appointments and choosing the time between the appointments carefully. Figures 4 and 5 are for the heavy and light traffic scenarios, respectively, but differing from the previous figures in that they consider the possibility that the customers may not show-up for their appointments with probabilities that depend on their indirect delay. Specifically, both scenarios assume that $\alpha = 0.95$. These two figures make it even clearer that choosing the time between consecutive appointments is very important for heavily loaded systems. In this case,

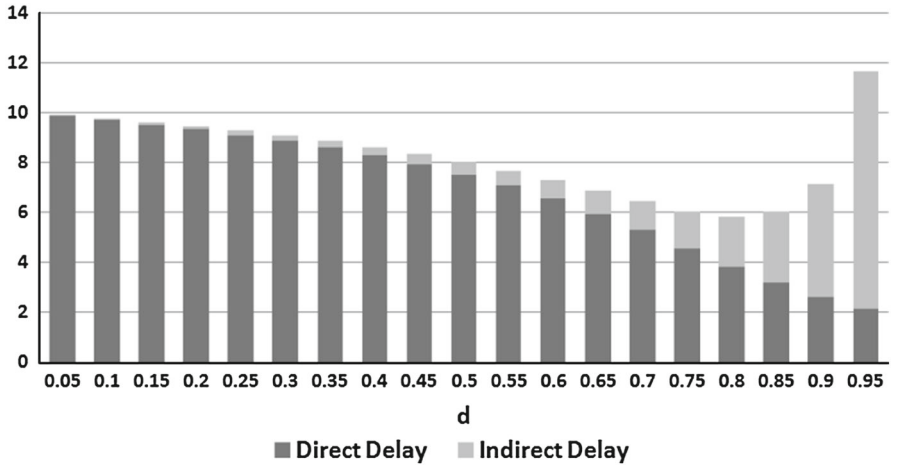

Fig. 4 $\lambda = 1, \mu = 1.1, \alpha = 0.95$

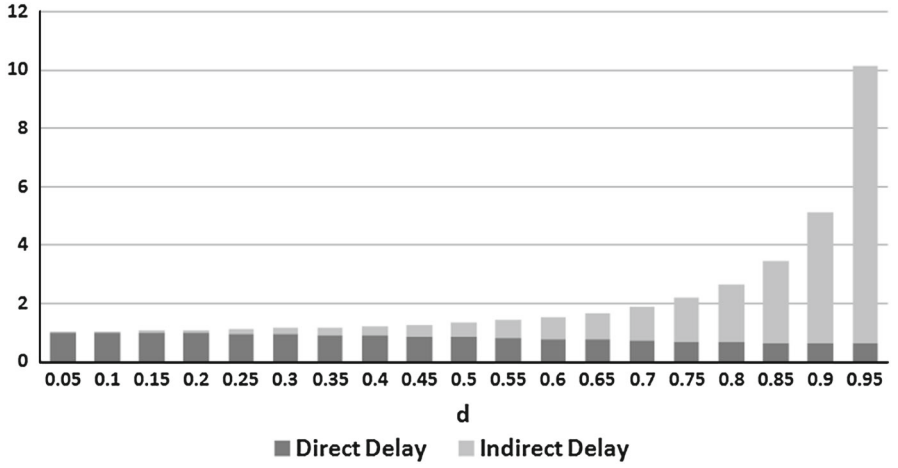

Fig. 5 $\lambda = 1, \mu = 2, \alpha = 0.95$

because of the fact that long appointment delays lead to high no-show probabilities, by setting appropriate inter-appointment time the service provider can decrease not only the expected direct waiting time but also the total expected waiting time. One should note however that the fact that the expected direct delay decreases faster in this case is due to some customers not showing up for their appointments. The service provider should also factor this in when choosing d.

7.2 The impact of capturing no-show dependency on direct delay

To investigate the importance of capturing the dependence of no-show probabilities on the appointment delays, we first ran a simulation of our queueing model with

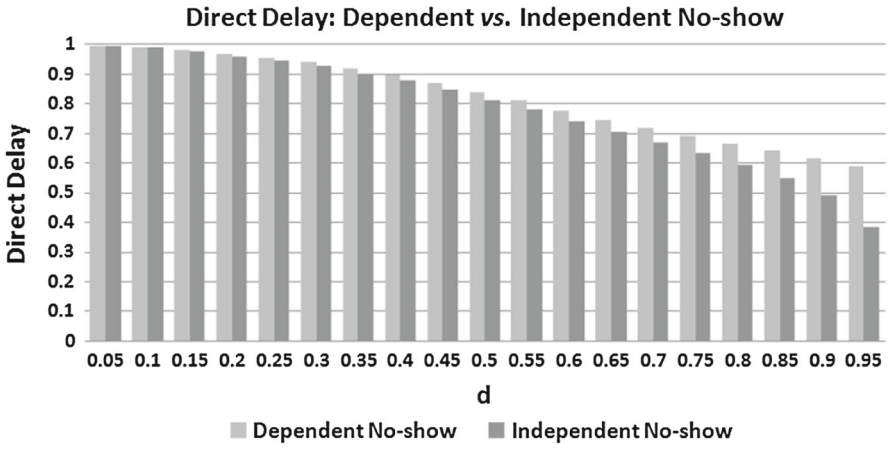

Fig. 6 $\lambda = 1, \mu = 2, \alpha = 0.95$

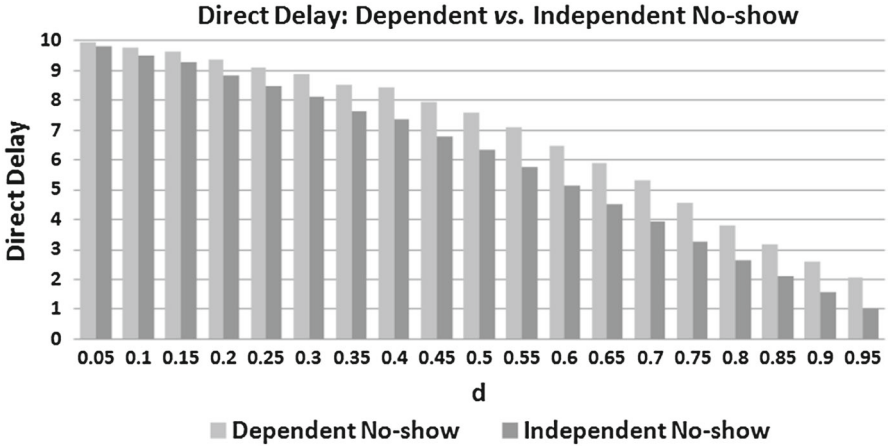

Fig. 7 $\lambda = 1, \mu = 1.1, \alpha = 0.95$

no-show probabilities assumed to be dependent on appointment delays, and obtained the fraction of customers who did not show-up for their appointments. Then, we ran another simulation using that same fraction as the independent fixed no-show probability. We computed the long-run average direct delay in both settings and under low and high traffic intensities separately.

As shown in Figs. 6 and 7, under both low and high traffic intensity scenarios, the independent no-show model clearly underestimates the direct delay if the actual underlying no-show probability is appointment-delay dependent. The difference may seem small for small d, which is expected since small values of d imply short appointment delays. However, we can see that as d increases and gets closer to the mean customer interarrival time, the underestimation increases up to 34.8 % under the low traffic intensity scenario and 51.7 % under the high traffic intensity scenario.

7.3 Testing the performance of the approximation methods

7.3.1 Exponential setting

We first compare the performance (in terms of approximating the actual mean direct delay) of the TQN approximation and our proposed Markovian, iid, and interpolation approximations under the exponential service time assumption. We fixed $\lambda = 1$ and $\alpha = 1$, and considered four scenarios with $\mu = 2, 1.5, 1.3$, and 1.1, respectively. These scenarios represent appointment systems with traffic intensity from low to high. For each scenario, we obtained the mean direct delay for different values of d ranging from 0.05 to 0.95 with increments of 0.05, using our exact analysis of the Tandem Queueing model (with queue truncated at $K = 60$), the interpolation approximation, the Markovian and iid approximations, and the TQN approximation. The results are shown in Fig. 8.

We can observe that when the traffic intensity is low, all approximation methods work well with the exception of the iid approximation. However, as the traffic intensity gets larger, while our interpolation approximation continues to work surprisingly well (the difference with the exact calculation being within 3 % in most cases), the performances of the other approximations decline visibly, particularly when d is away from zero and one. Thus, our results suggest that when service times are exponentially distributed, our interpolation approximation works consistently well across various traffic intensity levels but the TQN approximation and the Markovian arrival approximation work well only at relatively low traffic intensity levels. The iid approximation performs badly at all levels of traffic intensity.

7.3.2 Lognormal setting

Empirical studies demonstrate that lognormal distribution can be a good fit for service times under certain circumstances (see for example [3,20]). In this section, we investigate the performance of the generalized versions of the interpolation approximation and the TQN approximation as described in Sect. 6.3. When service times have a lognormal distribution we resort to simulation to find the expected direct delay and compare it with the results obtained from the approximation methods. In addition to our interpolation and TQN approximations, we also use our analysis for the exponential case to come up with another approximation we call *exponential approximation*. This approximation simply ignores the fact that service times are lognormal and computes the expected direct waiting time using our analytical results, which assume exponential service time distribution.

Again, we fix $\lambda = 1$ and $\alpha = 1$. In addition, we choose lognormal service time distributions with different combinations of service rate and coefficient of variation ($\text{CV} = \sqrt{Var(X)}/E[X]$). Specifically, we choose service rate $\mu \in \{1.1, 1.5, 2\}$, $\text{CV} \in \{0.5, 1, 1.5, 2\}$, and carry out performance comparisons for the above 12 scenarios. The following figures show the mean direct delay obtained from the simulation, the interpolation approximation, the TQN approximation, and the exponential approximation in each scenario.

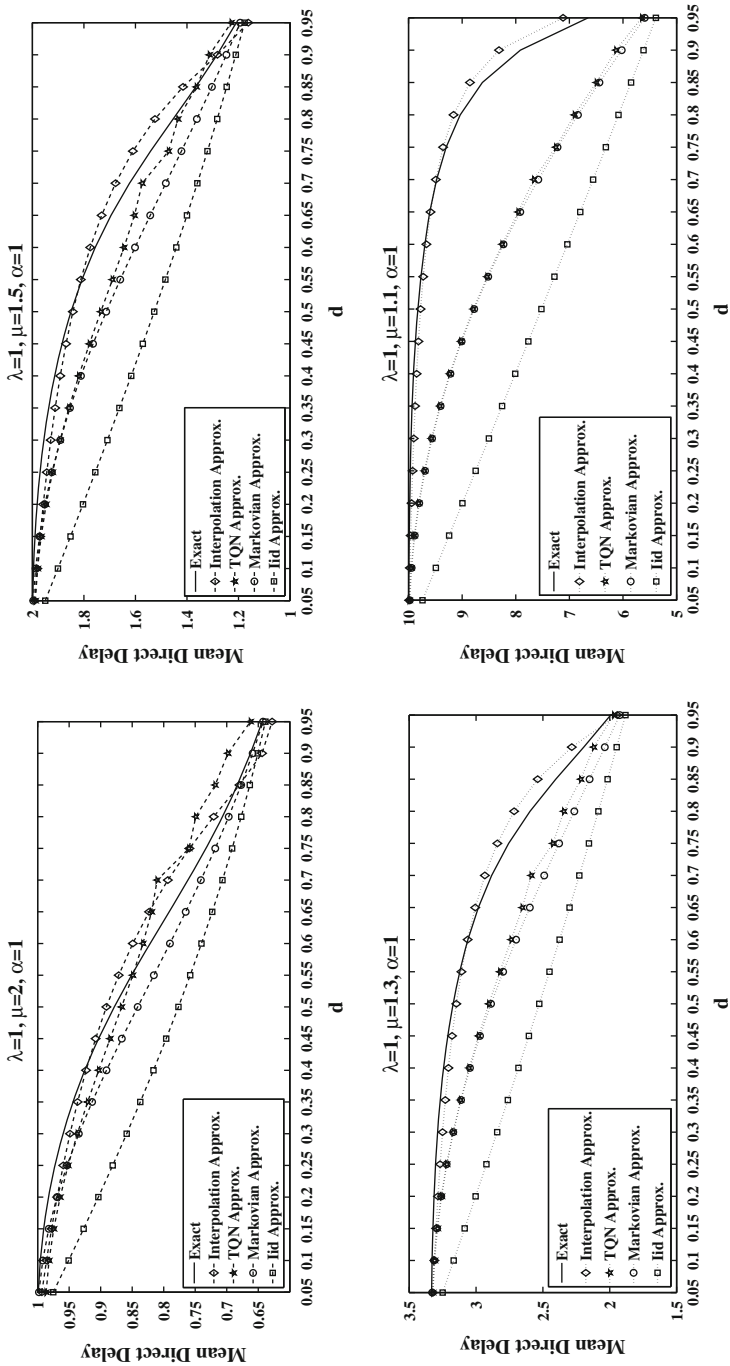

Fig. 8 Comparison of approximation methods under exponential setting

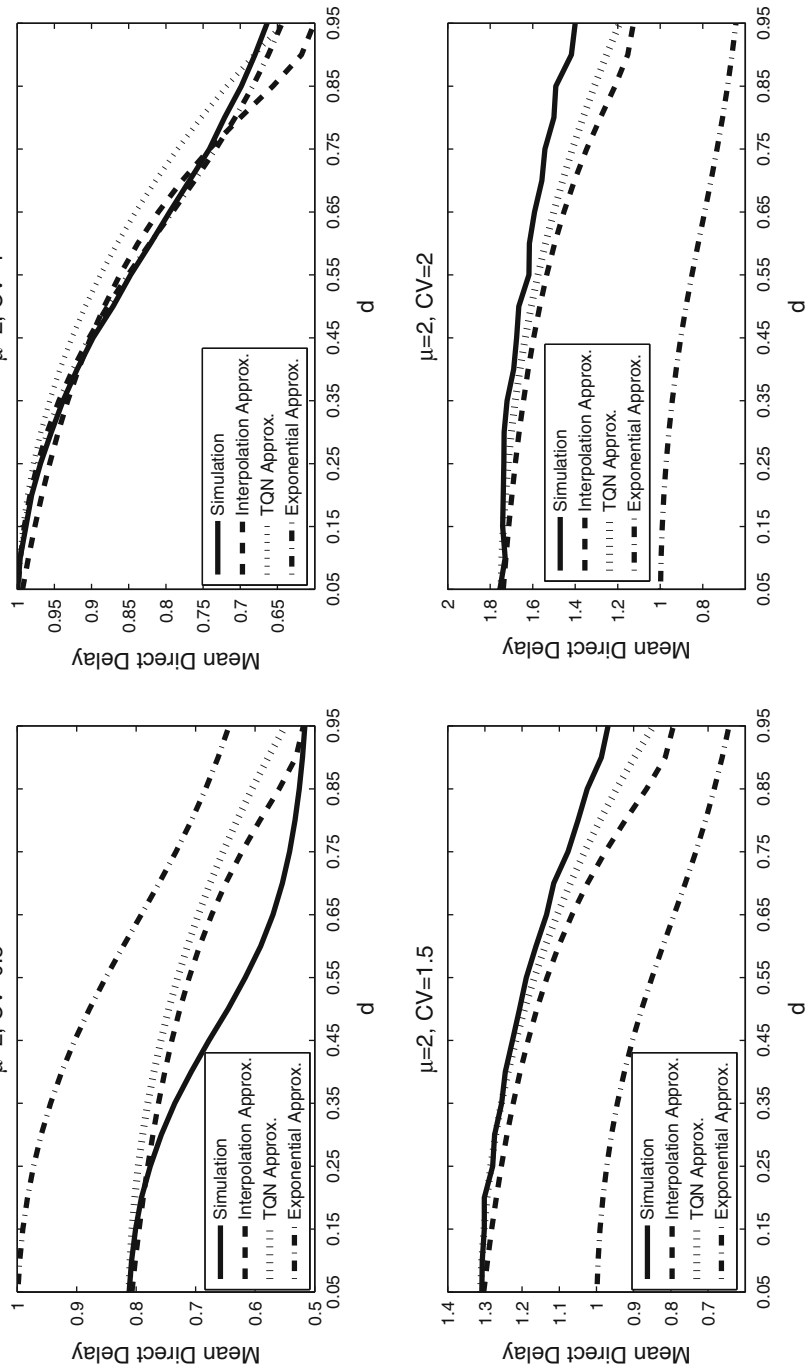

Fig. 9 Performance comparison under lognormal setting: light load

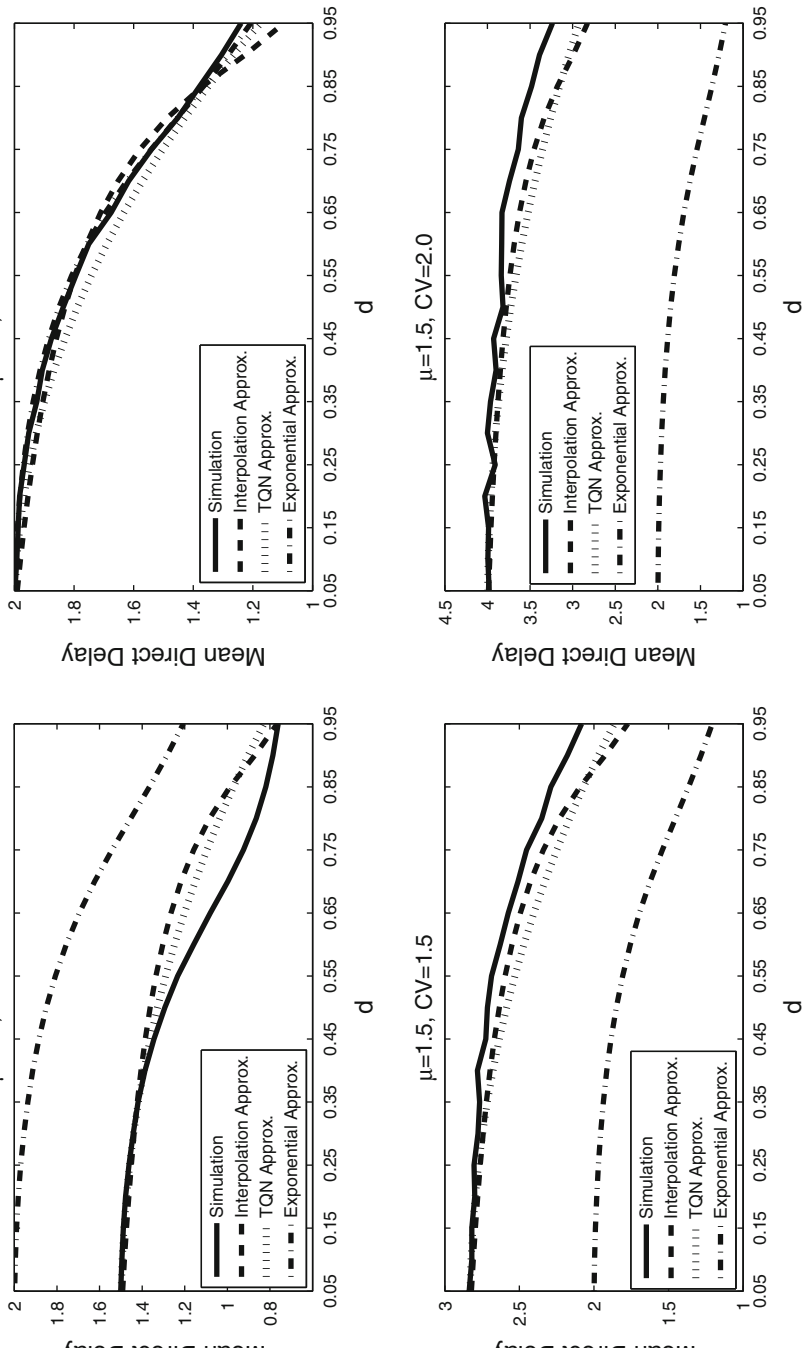

Fig. 10 Performance comparison under lognormal setting: medium load

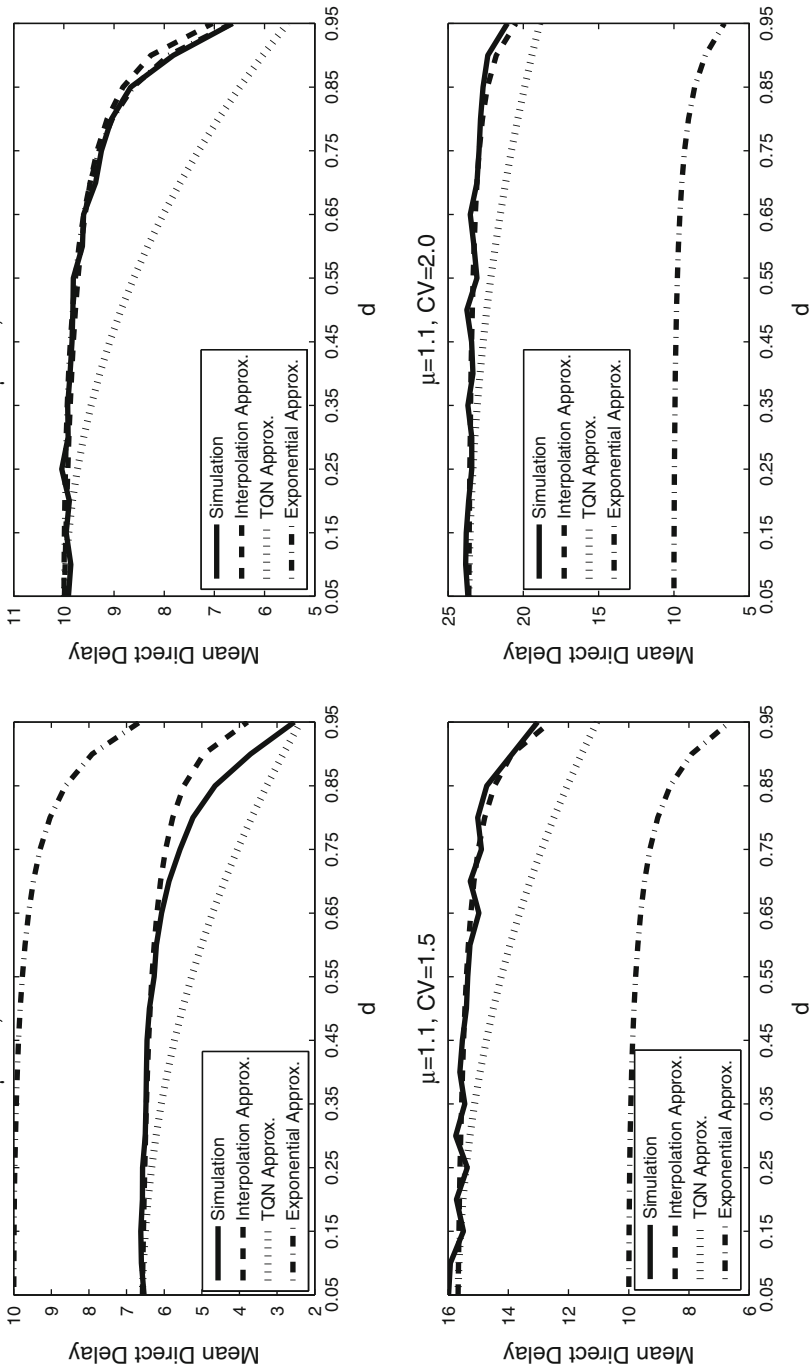

Fig. 11 Performance comparison under lognormal setting: heavy load

One can observe that when the traffic intensity is low or moderate (see Figs. 9, 10), the performances of our interpolation approximation and TQN approximation are comparable. When d is close to 0, both methods approximate the mean direct delay quite well regardless of the value of CV. However, when d is close to 1 in scenarios with large CV, the mean direct delay obtained by both approximation methods deviates from the simulation result somewhat visibly. This might be partially due to the $G/G/1$ approximation used in both methods. On the other hand, in the case where CV = 0.5, the approximations appear to be less accurate when d is in the middle of the two end points. It is also clear that across all scenarios, except of course when CV = 1, which corresponds to exponential distribution, the exponential approximation performs very poorly.

When the traffic intensity is high (see Fig. 11), our interpolation approximation method performs very well and clearly outperforms the TQN approximation method in all four scenarios. In addition, the exponential approximation performs very poorly across all scenarios except when CV is 1. These observations suggest that our proposed interpolation approximation method can be very useful for the estimation of the mean direct delay especially when service times have non-exponential distribution and system is heavily loaded.

8 Conclusions

The majority of past research on appointment scheduling has focused on the appointment-driven service process, and a lot of effort has been put on how to balance service utilization and customer direct delay. In this stream of research, the indirect delay has not been investigated in depth. There are two main reasons that customer indirect delay should also be considered when one aims to design an efficient appointment system. First, the indirect and direct delays are equally important factors that affect customers' experience in an appointment-based service system. A well-designed service system should balance both types of delays with service utilization. Second, some recent empirical studies have shown that customer no-show behavior is positively correlated to the indirect delay. This makes the consideration of indirect delay even more important as customer no-show behavior usually affects service utilization negatively.

With this motivation, in this paper we have studied an appointment system that consists of two queues in tandem. The first appointment queue captures the waiting process of customers whose scheduled appointment epochs have not come yet. The appointment queue is followed by the service queue that captures the waiting process of customers who have arrived at the service facility but whose services have not been completed yet. We develop a Tandem Queueing model to analyze the above system and obtain important performance measures of interest such as service utilization and mean direct and indirect delays.

In addition, to approximate the mean direct delay, we propose a simple interpolation approximation method as well as two other methods based on the idea of assuming that the service queue operates as a $G/M/1$ queue. All of these approximations bypass the need to solve the non-linear matrix equation derived in the Tandem Queueing model.

We also compare our approximation methods with the existing TQN approximation method under both exponential and lognormal service time settings.

We numerically verify that the mean direct delay monotonically decreases in the time between consecutive appointments while the mean indirect delay does the opposite. The mean total delay, depending on the no-show parameter, either monotonically increases in the time between appointments or is unimodal with a decreasing region followed by an increasing region. We further point out that even if the appointment mechanism may result in longer mean total delay, it may still be beneficial to customers because their reduced direct delay is usually more valuable than the corresponding increased indirect delay. In addition, we show that assuming a fixed no-show probability when in fact the true no-show probabilities depend on customers' appointment delays might lead to significant errors in the computation of expected service delays.

Finally, our numerical study shows that our proposed interpolation approximation method performs consistently well in all scenarios we have tested under the exponential setting with traffic intensity varying from high to low. It also performs quite well under the lognormal setting especially when the traffic intensity is high. In particular, our interpolation approximation method outperforms the existing TQN approximation method in all scenarios in which the traffic intensity is high.

Acknowledgments The work of the third author is supported in part by the National Science Foundation grant CMMI 1234212.

References

1. Bailey, N.T.J.: A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. J. R. Stat. Soc. Ser. B **14**, 185–199 (1952)
2. Bean, A.G., Talaga, J.: Predicting appointment breaking. J. Health Care Mark. **15**(1), 29–34 (1995)
3. Cayirli, T., Veral, E., Rosen, H.: Designing appointment scheduling systems for ambulatory care services. Health Care Manag. Sci. **9**(1), 47–58 (2006)
4. Chakraborty, S., Muthuraman, K., Lawley, M.: Sequential clinical scheduling with patient no-shows and general service time distributions. IIE Trans. **42**(5), 354–366 (2010)
5. Creemers, S., Lambrecht, M.: Queueing models for appointment-driven systems. Ann. Oper. Res. **178**(1), 155–172 (2010)
6. Denton, B., Gupta, D.: A sequential bounding approach for optimal appointment scheduling. IIE Trans. **35**(11), 1003–1016 (2003)
7. Doi, M., Chen, Y.M., Ōsawa, H.: A queueing model in which arrival times are scheduled. Oper. Res. Lett. **21**(5), 249–252 (1997)
8. Dreiher, J., Froimovici, M., Bibi, Y., Vardy, D.A., Cicurel, A., Cohen, A.D.: Nonattendance in obstetrics and gynecology patients. Gynecol. Obstet. Investig. **66**(1), 40–43 (2008)
9. Fries, B.E., Marathe, V.P.: Determination of optimal variable-sized multiple-block appointment systems. Oper. Res. **29**(2), 324–345 (1981)
10. Gallucci, G., Swartz, W., Hackerman, F.: Brief reports: impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. Psychiatr. Serv. **56**(3), 344–346 (2005)
11. Girish, M.K., Hu, J.Q.: Higher order approximations for tandem queueing networks. Queueing Syst. **22**(3), 249–276 (1996)
12. Green, L.V., Savin, S.: Reducing delays for medical appointments: a queueing approach. Oper. Res. **56**(6), 1526–1538 (2008)
13. Green, L.V., Savin, S., Wang, B.: Managing patient service in a diagnostic medical facility. Oper. Res. **54**(1), 11–25 (2006)
14. Grunebaum, M., Luber, P., Callahan, M., Leon, A.C., Olfson, M., Portera, L.: Predictors of missed appointments for psychiatric consultations in a primary care clinic. Psychiatr. Serv. **47**(8), 848–852 (1996)

15. Gupta, D., Denton, B.: Appointment scheduling in health care: challenges and opportunities. IIE Trans. **40**(9), 800–819 (2008)
16. Hassin, R., Mendel, S.: Scheduling arrivals to queues: a single-server model with no-shows. Manag. Sci. **54**(3), 565–572 (2008)
17. Jansson, B.: Choosing a good appointment system—a study of queues of the type (d, m, 1). Oper. Res. **14**(2), 292–312 (1966)
18. Johnson, M.A., Taaffe, M.R.: Matching moments to phase distributions: mixtures of Erlang distributions of common order. Stoch. Models **5**(4), 711–743 (1989)
19. Kaandorp, G.C., Koole, G.: Optimal outpatient appointment scheduling. Health Care Manag. Sci. **10**(3), 217–229 (2007)
20. Klassen, K.J., Rohleder, T.R.: Scheduling outpatient appointments in a dynamic environment. J. Oper. Manag. **14**(2), 83–101 (1996)
21. Kortbeek, N., Zonderland, M.E., Boucherie, R.J., Litvak, N., Hans, E.W.: Designing cyclic appointment schedules for outpatient clinics with scheduled and unscheduled patient arrivals. Memorandum 1968, Department of Applied Mathematics, University of Twente, Enschede, The Netherlands (2011)
22. Kulkarni, V.G.: Modeling and Analysis of Stochastic Systems. Chapman & Hall, London (1995)
23. LaGanga, L.R., Lawrence, S.R.: Clinic overbooking to improve patient access and increase provider productivity. Decis. Sci. **38**(2), 251–276 (2007)
24. Liu, N., Ziya, S.: Panel size and overbooking decisions for appointment-based services under patient no-shows. Forthcoming in Prod. Oper. Manag. (2014)
25. Liu, N., Ziya, S., Kulkarni, V.G.: Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. Manuf. Serv. Oper. Manag. **12**(2), 347–364 (2010)
26. Luo, J., Kulkarni, V.G., Ziya, S.: Appointment scheduling under patient no-shows and service interruptions. Manuf. Serv. Oper. Manag. **14**(4), 670–684 (2012)
27. Mercer, A.: Queues with scheduled arrivals: a correction, simplification and extension. J. R. Stat. Soc. Ser. B **35**(1), 104–116 (1973)
28. Muthuraman, K., Lawley, M.: A stochastic overbooking model for outpatient clinical scheduling with no-shows. IIE Trans. **40**(9), 820–837 (2008)
29. Neuts, M.F.: Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach. Johns Hopkins University Press, Baltimore (1981)
30. Patrick, J., Puterman, M.L., Queyranne, M.: Dynamic multi-priority patient scheduling for a diagnostic resource. Oper. Res. **56**(6), 1507–1525 (2008)
31. Pegden, C.D., Rosenshine, M.: Scheduling arrivals to queues. Comput. Oper. Res. **17**(4), 343–348 (1990)
32. Robinson, L.W., Chen, R.R.: A comparison of traditional and open-access policies for appointment scheduling. Manuf. Serv. Oper. Manag. **12**(2), 330–346 (2010)
33. Robinson, L.W., Chen, R.R.: Estimating the implied value of the customer's waiting time. Manuf. Serv. Oper. Manag. **13**(1), 53–57 (2011)
34. Stein, W.E., Côté, M.J.: Scheduling arrivals to a queue. Comput. Oper. Res. **21**(6), 607–614 (1994)
35. Wang, P.P.: Optimally scheduling N customer arrival times for a single-server system. Comput. Oper. Res. **24**(8), 703–716 (1997)
36. Zacharias, C., Armony, M.: Panel Sizing and Appointment Scheduling in Outpatient Medical Care. Working paper, Stern School of Business, New York University, New York, NY (2013)