Stationary Birth-and-Death Processes Fit to Queues with Periodic Arrival Rate Functions

James Dong and Ward Whitt

School of Operations Research and Information Engineering,
Cornell University, Ithaca, NY 14850 jd748@cornell.edu

Industrial Engineering and Operations Research
Columbia University, New York, NY, 10027 ww2040@columbia.edu

June 17, 2015

Abstract

To better understand what queueing models are appropriate for complex service systems such as hospital emergency departments, we suggest fitting a general state-dependent birth-and-death (BD) process to system data recording the number in system over a time interval To facilitate interpretation of the fitted BD rate functions, we investigate the consequences of fitting a BD process to a multi-server $M_t/GI/s$ queue with a nonhomogeneous Poisson arrival process having a periodic time-varying rate function. The fitted death rates consistently have the same piecewise-linear structure previously found for the $GI/GI/s$ model, independent of the service-time distribution, but the fitted birth rates have a very different structure, with a similar linear structure around the average occupancy, but constant limits at large and small arguments. Under minor regularity conditions, the fitted BD process has the same steady-state distribution as the original queue length process as the sample size increases. The steady-state distribution can be estimated efficiently by fitting a parametric function to the observed birth and death rates.

Keywords: birth-and-death processes; grey-box stochastic models; fitting stochastic models to data; queues with time-varying arrival rate; speed ratio; transient behavior.

1 Introduction

The purpose of this paper is to help fit appropriate queueing models to data from complex queueing systems, such as hospital emergency departments. This paper is a sequel to [12] in which we investigated fitting a general state-dependent birth-and-death (BD) stochastic process to an observation of the number of customers in a complex queueing system over some time interval, assuming that state changes occur one at a time. The fitted BD process, or some modification of it (e.g., the time-scaled version in (1.2) of [12]), may serve as a useful model, but we primarily regard the fitted BD as a diagnostic tool to learn what queueing models are appropriate.

Following common practice [47], we estimate the birth rate in state k from data over an interval $[0, t]$ by $\bar{\lambda}_k \equiv \bar{\lambda}_k(t)$, the number of arrivals observed in that state, divided by the total time spent in that state, while the death rate in state k is estimated by $\bar{\mu}_k \equiv \bar{\mu}_k(t)$, the number of departures observed in that state, divided by the total time spent in that state. For a BD process, those are the maximum likelihood estimators of the actual birth and death rates.

Actual service systems may have complex time-dependence and stochastic dependence that may be difficult to assess directly. Fitting a BD process may be a useful way to probe into system data. In [12] we referred to this as *grey-box stochastic modeling*. In [12] we applied this analysis to various conventional $GI/GI/s$ queueing models. We saw how the fitted rates $\{\bar{\lambda}_k, \bar{\mu}_k\}$ differ from the corresponding $M/M/s$ model, for given overall arrival rate λ and individual service rate μ. We saw that they differ in systematic ways that enabled us to see a *signature* of the $GI/GI/s$ model.

Here we consider many-server $M_t/GI/s$ queueing models with sinusoidal periodic arrival rate functions. We find that the fitted death rates have the same simple linear structure as seen for $GI/GI/s$ models, but we find significant differences in the fitted birth rates. Overall, we see a signature of the $M_t/GI/s$ model with sinusoidal arrival rates.

1.1 An Emergency Room Example

To illustrate how the results here can be applied, we show the fitted BD rates for an Israeli emergency department studied in [46]. Figure 1 shows the estimated birth rate (left), death rate (center) and death rate divided by the state (right) for the ED over a 25-week period. The ED is the same as studied in §3 of [2]. The data used in [46] included about 25,000 patient visits to the internal unit of the ED over a 25 week period from December 2004 to May 2005.

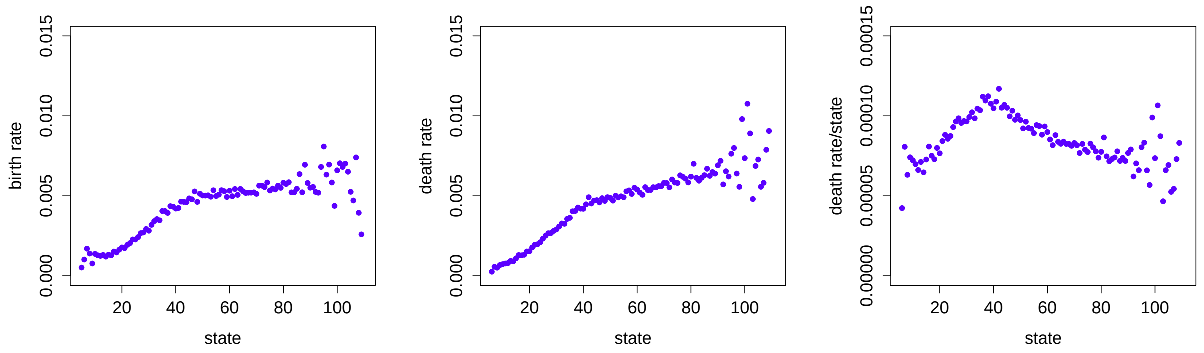

Figure 1: The fitted state-dependent birth rate $\bar{\lambda}_k$ (left), death rate $\bar{\mu}_k$ (center) and death rate divided by the state, $\bar{\mu}_k/k$ (right) obtained from arrival and departure data in an Israeli emergency department over 25 weeks, taken from [46]

It is well known that the arrivals to an ED vary strongly over time, just as in most service systems; see Figure 9 of [21]. Thus, a natural candidate rough aggregate model for an ED is the $M_t/GI/s$ queue for $s = \infty$ or relatively large s, which has a nonhomogeneous Poisson process (NHPP) as its arrival process, independent and identically distributed (i.i.d.) service (length-of-stay, LoS) times with some general (non-exponential, perhaps lognormal) distribution, s servers, unlimited waiting space and service in order of arrival.

The present paper helps interpret Figure 1. First, Figures 3, 5 and 12 in §2.2 and §2.7 provide strong support for two conclusions (which otherwise would not be evident): First, the fitted birth rates are roughly consistent with an NHPP (M_t) arrival process having a periodic arrival rate function. Second, the fitted death rates are inconsistent with i.i.d. service times. This second negative conclusion may be easier to see by looking at the state-dependent death rate divided by the state, so that is why we display that as the third plot in Figure 1. Extensive simulations show that the fitted death rates are approximately proportional to the state k in an $M_t/GI/\infty$ model with a periodic arrival rate function, and approximately piecewise-linear with finitely many servers.

These tentative conclusions about the ED based on the analysis of $M_t/GI/s$ queues in this paper are strongly supported by further data analysis in [46]. The data analysis in [46] supports an $M_t/G_t/\infty$, where there is strong time-dependence in the service-time distribution as well as the arrival rate function. That conclusion in turn is consistent with other observations, e.g., see [2, 40] and references there. The fitted BD is convenient because it quickly exposes the difficulty with the service times.

1.2 Fitting the Erlang-A Model to Data

The fitted BD process may also provide a useful statistical test of the classical $M/M/s+M$ Erlang-A model in settings where it may be applied. The Erlang-A model is a stationary birth-and-death (BD) process with four parameters: the arrival rate λ, the service rate μ, the number of servers s and the individual customer abandonment rate from queue θ; see [18, 22, 34] and references therein. The familiar $M/M/s/0$ Erlang B (loss) and $M/M/s \equiv M/M/s/\infty$ Erlang C (delay) models are the special cases in which $\theta = \infty$ and $\theta = 0$.

These models are convenient because there are so few parameters, but these parameters are typically fit to data in a different way. The arrival rate λ and service rate μ are typically estimated as the reciprocals of the average interarrival time and service time, respectively, but the abandonment rate is more complicated because of censoring; it is often better to estimate the hazard rate of the customer patience distribution; see [3].

For successful applications, it is important to investigate to what extent the model is consistent with service system data. That is most often done by estimating the distributions of the interarrival times and service times to see if they are nearly exponential, but there are many other ways the system can differ from the model. Service systems typically have time-varying arrival rates and there may be significant dependence among interarrival times and service times. The number of servers may vary over time as well and the servers are often actually heterogeneous [17]. Indeed, careful statistical analysis of service system data can be quite complicated, e.g., see [2, 3, 25, 26, 27].

In this paper we investigate an alternative way to fit the Erlang-A model to data, which we propose doing in addition to the standard fitting procedure, to check consistency. Given that the data are from the Erlang-A model, we will see simple linear structure in the estimated birth and death rates. With enough data, we will see that

$$\bar{\lambda}_k = \lambda, \quad k \geq 0, \quad \text{and} \quad \bar{\mu}_k = (k \wedge s)\mu + (k-s)^+\theta, \quad k \geq 1, \tag{1}$$

where $a \wedge b \equiv \min\{a, b\}$ and $(a)^+ \equiv \max\{a, 0\}$. By this procedure, we can estimate all four parameters and test if the model is appropriate. A direct BD fit of the form (1) may indicate that the model should be effective even though some other tests fail. For example, experience indicates that a good model fit can occur by this BD rate fit even though the servers are heterogeneous and the service-time distribution is not exponential. Moreover, in those cases we may find that the Erlang-A model works well in setting staffing levels.

1.3 When the Erlang-A Model Does Not Fit

However, what do we conclude if the BD fit does not yield the birth and death rate functions in (1)? Some insights are relatively obvious. For example, if we do not see death rates with two linear pieces joined at some level s, then we can judge that the number of servers probably was not constant during the measurement period. But it remains to carefully evaluate how to interpret departures from the simple Erlang structure in (1).

We may also consider directly applying the fitted BD process even if we do not see the Erlang-A structure in (1), because BD processes are remarkably tractable. If we happen to find piecewise-linear fits, then we may find diffusion approximations with large scale, as in [4], which is not limited to the classical Erlang models in [18, 23]. It is well known that we can calculate the steady-state distribution of a general BD process by solving local balance equations. We also can efficiently calculate first-passage-time distributions in general BD processes [1].

1.4 With Enough Data, The Steady-State Distribution Is Always Matched

A remarkable property of the fitted BD process is that (asymptotically, as the sample size increases) the steady-state distribution of the system is always matched by the steady-state distribution of the fitted BD process. Thus, this fitted BD process is one stochastic model that accurately describes the distribution of the steady-state number in system *based on the data used*.

To elaborate, the steady-state distribution of the fitted BD model, denoted by $\bar{\alpha}^e_k \equiv \bar{\alpha}^e_k(t)$ (with superscript e indicating the estimated rates), is well defined (under regularity conditions [43]) and characterized as the unique probability vector satisfying the local balance equations,

$$\bar{\alpha}^e_k \bar{\lambda}_k = \bar{\alpha}^e_{k+1} \bar{\mu}_{k+1}, \quad k \geq 0. \tag{2}$$

To obtain reasonable rate estimates for which $\bar{\alpha}^e_k$ is indeed well defined and unique, we truncate the state space to a region of states that are visited relatively frequently. Throughout this paper, we assume that the limiting values of the rates as $t \to \infty$ exist so we omit the t. We use large sample sizes in our simulations to justify this assumption.

In [43] we cautioned against drawing unwarranted positive conclusions if the fitted BD steady-state distribution $\{\bar{\alpha}^e_k : k \geq 0\}$ in (2) closely matches the empirical steady-state distribution, $\{\bar{\alpha}_k : k \geq 0\}$, where $\bar{\alpha}_k \equiv \bar{\alpha}_k(t)$ is the proportion of total time spent in each state, because these two distribution are automatically closely related. Indeed, as has been known for some time (e.g., see Chapter 4 of [15]), under regularity conditions, these two distributions coincide asymptotically as t (and thus the sample size) increases, even if the actual system evolves in a very different way from the fitted BD process. For example, the actual process $\{Q(t) : t \geq 0\}$ might be non-Markovian (as in [12]) or have a time-varying arrival rate (as here). Stochastic comparisons between the two distributions, depending on the beginning and ending states, were also derived in [43]. If the ending state coincides with the initial state, then these two empirical distributions are identical for any sample size!

Even though a close match between the empirical steady-state distribution, $\{\bar{\alpha}_k\}$, and the steady-state distribution of the fitted BD model, $\{\bar{\alpha}^e_k\}$, does not nearly imply that the actual system evolves as a BD process, we think that the fitted BD model has the potential to become a useful modeling and analysis tool, providing insight into the actual system. Of course, if the actual system can be well modeled by a standard BD model, such as one of the classical Erlang models, then we will see a good fit to that model with enough data. Of primary interest here is to be able to see deviations from classical models through the fitted birth and death rates.

1.5 Operational Analysis

The remarkable match between the steady-state distribution of the fitted BD process and the actual system is partially the basis for early work on *operational analysis.* In early performance analysis of computer systems, Buzen and Denning [5, 6, 11] advocated working with BD processes fit directly to data as part of a general operational analysis directly. The goal was to understand performance empirically, directly from data, without using customary stochastic models. Key support for this approach was provided by conservation laws that must hold among the statistics collected, as in Little's law.

The flaw of course is that *accurate description does not imply accurate prediction*; i.e., accurate description of the steady-state distribution over the period when the data are collected does not imply accurate prediction of system performance at a later time when the system has changed. Thus, we prefer to think of there actually being an underlying stochastic model, which can be used for prediction, provided that we can properly identify it. With that in mind, we think of the fitted BD process as a way to obtain partial information about the underlying model.

Problems with a direct application of operational analysis are discussed in §§4.6-4.7 in [15]. In that context, though, [12] and this paper provides the first comparison between an underlying stochastic process model and the operational analysis BD model fit to data. For either to be useful in prediction, the future system of interest should be like the current system being measured. To judge whether candidate models are appropriate, we think that it is appropriate to apply statistical analysis to analyze the measurements. Sound statistical analysis, as in [3, 26, 27], can strongly support an underlying stochastic model, which will behave differently from the fitted BD model if the data are inconsistent with the BD model, as we show here.

1.6 Periodic Queues

We started our investigation of fitted BD processes in [12] by looking carefully at BD fits to the number in system in $GI/GI/s$ queues. We continue here by looking carefully at BD fits to the number in system in $M_t/GI/s$ queues, having NHPP arrival processes with sinusoidal arrival rate functions and i.i.d. service times, paying especial attention to the case of $s = \infty$ servers. The sinusoidal arrival rate function is a stylized arrival rate function that is similar to actual arrival rate functions estimated from data.

Our goal in the present paper is to consider many-server queues with periodic arrival rates. These have been studied in [9, 13, 14, 16, 24, 31, 32, 33, 35, 37, 38, 44, 45] and references therein. As in [12], we want to understand how the fitted birth and death rates depend on the model structure. We find that the fitted birth and death rates provide very useful information about the structure of the actual model. In this paper we concentrate on $M_t/GI/s$ multi-server queues, where the arrival process is a an NHPP with a periodic arrival rate function, emphasizing the tractable limiting case of the infinite-server (IS) model [13, 14]. For these models, there is a proper steady-state distribution, which is the time average of the time-dependent distributions over each periodic cycle. For the special case of the $M_t/M/\infty$ model with a sinusoidal arrival rate function, the steady-state distribution is studied in [45].

There are very few available results for actually computing the steady-state distribution in periodic queues. For Markovian models, the steady-state distribution may be calculated by numerically solving ordinary differential equations, possibly simplified by closure approximations [39]. However, simulation seems to be the only available method for non-Markovian models. Thus, a significant contribution in this paper is to provide a new way to estimate the steady-state distribution from data, either from system measurments or simulation; see §2.8. We suggest fitting parametric functions to estimated birth and death rates and then solving the local balance equations in (2). This approach has potential because the fitted birth rates and death rates often have more elementary structure, such as linearity. There is efficiency in our proposed estimation procedure because there are much fewer parameters to estimate.

1.7 Organization

We start in §2 by reporting results of simulation experiments for $M_t/GI/s$ queueing models with sinusoidal arrival rates, which help us interpret Figure 1, and serve to motivate theoretical results that follow. In §3 and §4 we develop supporting theory. In §5 we summarize the notation and in §6 we draw conclusions.

2 Simulation Experiments

All the models considered in this paper will be $M_t/GI/s$ queueing models, having an NHPP (the M_t) as an arrival process, which is independent of i.i.d. service times distributed as a random variable S with mean $E[S] = 1/\mu = 1$ and a general distribution, s servers, $1 \leq s \leq \infty$, and unlimited waiting space. Moreover, we consider the stylized sinusoidal arrival rate function

$$\lambda(t) \equiv \bar{\lambda}\left(1 + \beta \sin\left(\gamma t\right)\right), \tag{3}$$

where the cycle is $c = 2\pi/\gamma$. There are three parameters: (i) the average arrival rate $\bar{\lambda}$, (ii) the relative amplitude β and (iii) the time scaling factor γ or, equivalently the cycle length $c = 2\pi/\gamma$. Our base model is the $M_t/M/\infty$ model, which is the special case of the $M_t/GI/s$ model in which $s = \infty$, S has an exponential distribution and $\beta = 10/35$.

2.1 Designing the Simulation Experiments

The simulation experiments were conducted much as in the prequel to this paper [12]. We generated the NHPP arrival process by thinning a Poisson process with rate equal to the maximum arrival rate over a sine cycle. Since we use relative amplitude $\beta = 10/35$, with $\bar{\lambda} = 35$ a proportion $10/(35+10) = 10/45 = 0.222$ of the potential arrivals were not actual arrivals. The fitted birth and death rates as well as the empirical mass function were estimated using 30 independent replications of 1.5 million potential arrivals before thinning. Overall, that means about $45 \times (35/45) = 35$ million arrivals in each experiment. Multiple i.i.d. repetitions were performed to confirm high accuracy within the regions shown. In order to compare the transient behavior of the fitted BD process to the original process, we simulated a separate version of the fitted BD process in a similar manner. To compute the first passage times starting from steady state (see §2.6), the process is initialized in steady state by choosing the initial state from the estimated steady-state distribution.

2.2 Comparing the Fitted Rates in the $M_t/M/\infty$ and $GI/M/\infty$ Models

Our main hypothesis is that the fitted birth and death rates can reveal features of the underlying model. To compare the impact of predictable deterministic variability in the arrival process, as manifested in a time-varying arrival rate function, to stochastic variability, we see how the fitted birth rates differ in the $M_t/M/\infty$ IS model with a sinusoidal arrival rate function and the stationary $GI/M/\infty$ model with a renewal process having an interarrival time more variable than the exponential distribution. (When the service-time distribution is exponential with mean 1, the fitted death rates coincide with the exact death rates in both cases, i.e., $\bar{\mu}_k = k$; see Theorem 3.1 of [12] and Theorem 3.3 here.) However, the fitted birth rates are revealing.

In [12] we found that, when the actual arrival rate is n (provided that n is not too small), with the service rate fixed at $\mu = 1$, the fitted birth rates in state k, denoted by $\lambda_{n,k}$, tended to have the form

$$\bar{\lambda}_{n,k} \approx (n + b(k-n)) \vee 0, \qquad (4)$$

where $b \approx 1 - 2/(1+c_a^2)$, a constant in the interval $[-1,1]$, with c_a^2 being the *squared coefficient of variation* (scv, variance divided by the square of the mean) of the interarrival-time distribution of the renewal arrival process. This is illustrated in Figure 2, which shows the fitted birth rates and death rates in five $GI/M/\infty$ models with arrival rate $\lambda = 39$ and service rate $\mu = 1$. The five interarrival-time distributions are Erlang E_4, E_2, M, and hyperexponential, H_2 with $c_a^2 = 2$ and $c_a^2 = 4$.

Figure 2 shows that the fitted birth rates tend to be approximately linear (over the region where the process visits relatively frequently, so that there are ample data for the estimation), with $\lambda_{n,n} = n$ and slope increasing as the variability increases. This is consistent with greater variability in the arrival process leading to a larger steady-state number in system. For $c_a^2 < 1$, the slope is negative; for $c_a^2 > 1$, the slope is positive. As c_a^2 increases to ∞, the slope approaches 1.

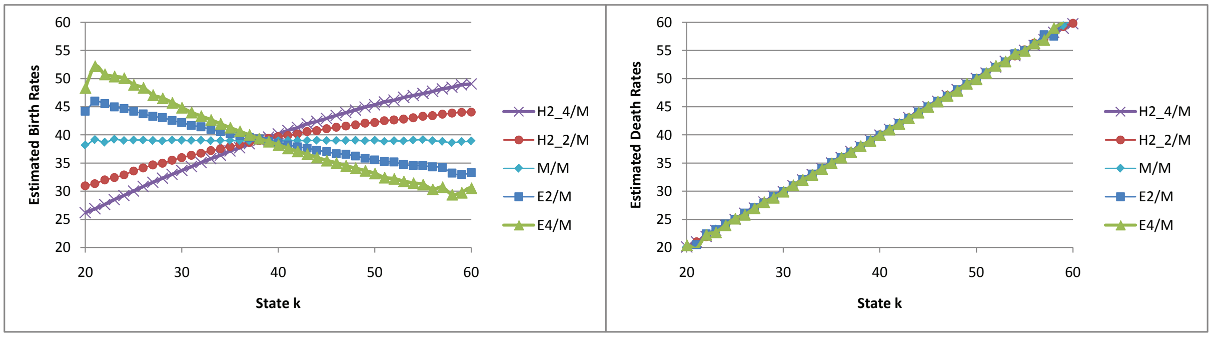

Figure 2: Fitted birth rates and death rates for five $G/M/\infty$ models with $\lambda = 39$ and $\mu = 1$.

We now consider the $M_t/M/\infty$ IS model with the sinusoidal arrival rate function in (3). Very roughly, we expect the predictable variability of a nonhomogeneous Poisson arrival process with a periodic arrival rate function to correspond approximately to a stationary model with a renewal arrival process having an interarrival-time distribution that is more variable than an exponential distribution [36]. That means we expect to see something like the fitted birth rates with increasing linear slopes in Figure 2. And indeed that is exactly what we do see, but restricted to a subinterval centered at the long-run average $\lambda_{n,n} = n$, as illustrated in Figure 3.

The evolution of a BD queue primarily depends on the birth and death rates λ_k and μ_k through their difference, the drift $\delta_k \equiv \lambda_k - \mu_k$, $k \geq 0$. Thus, we plot the drift functions associated with the $G/M/\infty$ and $M_t/M/\infty$ models in Figures 2 and 3 in Figure 4. These show that there is drift toward the overall mean in all cases, which is stronger when there is less variability.

Similar results hold for models with finitely many servers. We show the results paralleling Figure 3 for the case of 40 servers in Figure 5.

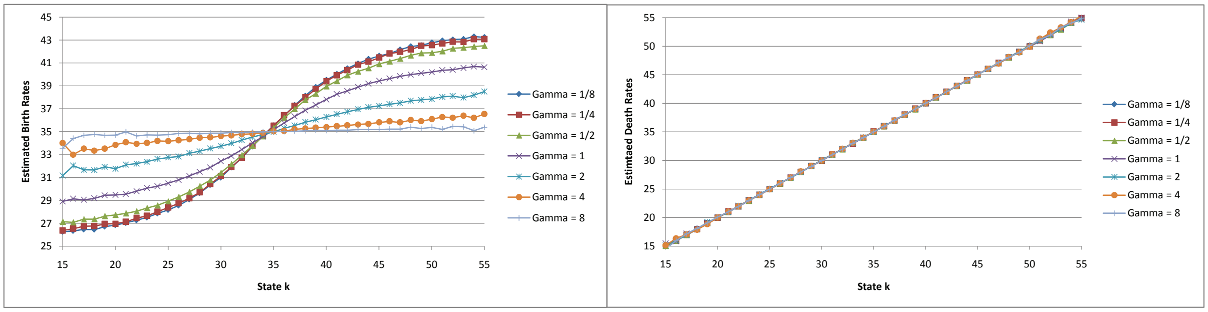

Figure 3: Fitted birth rates (left) and fitted death rates (right) for the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 35$ and $\beta\bar{\lambda} = 10$ and 7 values of γ ranging from 1/8 to 8.

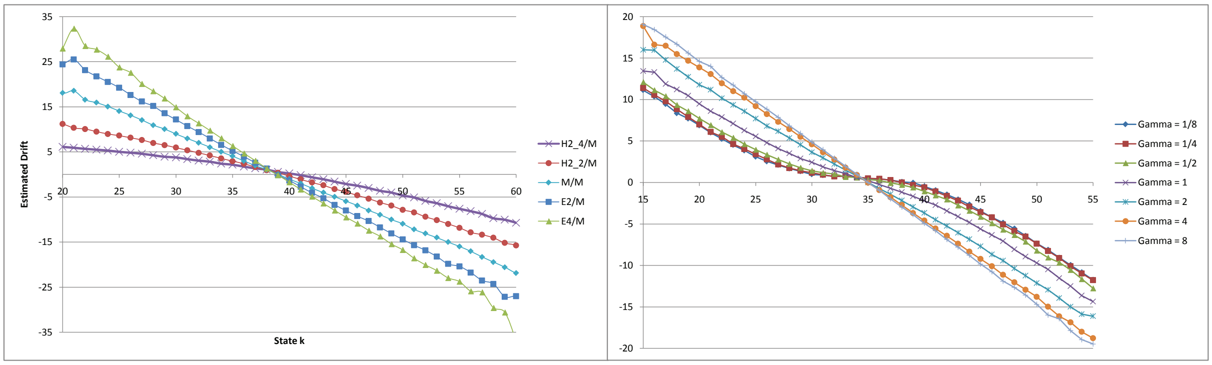

Figure 4: The estimated drift functions (birth rates minus death rates) for the $G/M/\infty$ model in Figure 2 (left) and the $M_t/M/\infty$ model in Figure 3 (right).

Figure 5 shows the piecewise-linear death rates, with two linear components, joined at the number of servers, that are characteristic of multi-server queues. Figure 2 of [12] displays similar plots for $GI/GI/s$ queues. However, the estimated birth rates in Figures 3 and 5 are unlike those of any $GI/GI/s$ queue. Theorems 4.3 and 4.4 establish finite bounds and heavy-traffic limits for the fitted birth rates, consistent with these figures.

2.3 The Steady-State Distribution of the $M_t/M/\infty$ Model

The estimated BD rates yield corresponding estimates of the steady-state distribution by solving the local balance equation (2). The estimated steady-state distributions for the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 35$ and $\beta = 10/35$ for different ranges of γ are shown in Figure 6. On the left (right) is shown different cases varying in a power of 10 (2). Many of the plots on the left coincide, so that we see convergence as $\gamma \uparrow \infty$ and as $\gamma \downarrow 0$. Indeed, the relevant ranges for intermediate behavior can be said to be $1/8 \leq \gamma \leq 8$ for these parameters $\bar{\lambda} = 35$ and $\beta = 10/35$, with the limits serving as effective approximations outside this interval.

Explicit formulas and asymptotic expressions for the steady-state distribution of the number in system in the $M_t/M/\infty$ IS model with the sinusoidal arrival rate function in (3) were established in [45] by applying [13]. Since these results are relevant here, we review some of this material. By §5 of [13], the number of customers in the system (or the number of busy servers), $Q(t)$, starting

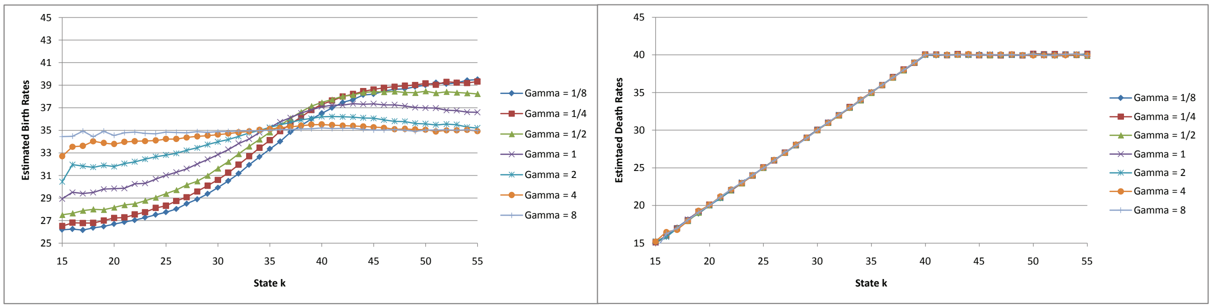

Figure 5: Fitted birth rates (left) and fitted death rates (right) for the $M_t/M/40$ queue with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 35$ and $\beta\bar{\lambda} = 10$ and 7 values of γ ranging from 1/8 to 8.

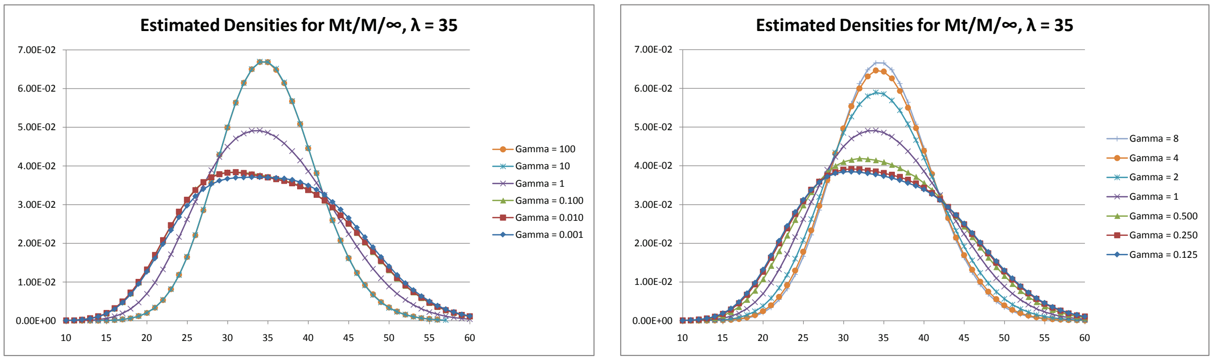

Figure 6: the estimated steady state number in the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 35$ and $\beta = 10/35$ for different ranges of γ.

empty in the distant past, has a Poisson distribution at each time t with mean

$$m(t) \equiv E[Q(t)] = \bar{\lambda}(1 + s(t)), \quad s(t) = \frac{\beta}{1+\gamma^2}\left(\sin(\gamma t) - \gamma\cos(\gamma t)\right). \tag{5}$$

Moreover,

$$s^U \equiv \sup_{t\geq 0} s(t) = \frac{\beta}{\sqrt{1+\gamma^2}} \tag{6}$$

and

$$s(t_0^m) = 0 \quad \text{and} \quad \dot{s}(t_0^m) > 0 \quad \text{for} \quad t_0^m = \frac{\cot^{-1}(1/\gamma)}{\gamma}. \tag{7}$$

The function $s(t)$ increases from 0 at time t_0^m to its maximum value $s^U = \beta/\sqrt{1+\gamma^2}$ at time $t_0^m + \pi/(2\gamma)$. The interval $[t_0^m, t_0^m + \pi/(2\gamma)]$ corresponds to its first quarter cycle.

Let Z be a random variable with the steady-state probability mass function (pmf) of $Q(t)$; its pmf is a mixture of Poisson pmf's. In particular,

$$P(Z = k) = \frac{\gamma}{2\pi}\int_0^{2\pi/\gamma} P(Q(t) = k)\,dt, \quad k \geq 0, \tag{8}$$

The moments of Z are given by the corresponding mixture

$$E[Z^k] = \frac{\gamma}{2\pi}\int_0^{2\pi/\gamma} E[Q(t)^k]\,dt, \quad k \geq 1,$$

so that $E[Z] = \bar{\lambda}$. For more details, see [45].

2.4 Transient Behavior

It should be evident that the transient behavior of the fitted BD process and the original process have significant differences. In particular, there is no periodicity in the fitted BD process. The differences are particularly striking with small γ, i.e., for long cycles $c(\gamma) = 2\pi/\gamma$. That is dramatically illustrated in Figure 7, which compares the sample paths of the number in system of the two processes for the $M_t/M/\infty$ queue with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 100$ and $\beta = 10/35$ and $\gamma = 0.01$. Since $\gamma = 0.01$, the cycle length is 628. Hence in the time interval $[0, 4000]$ we see a bit more than six cycles, but there is no periodic behavior in the fitted BD process.

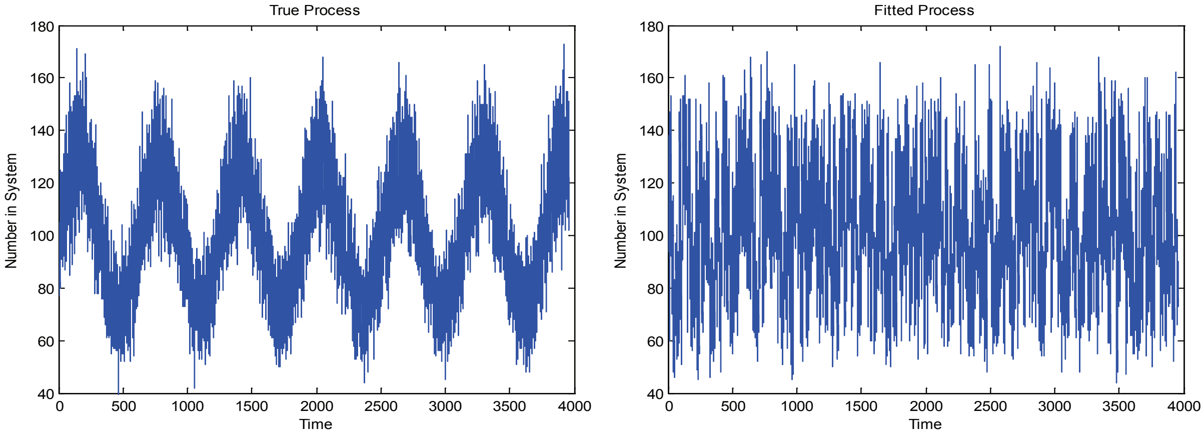

Figure 7: sample paths of the number in system for the original process (left) and the fitted BD process (right) for the $M_t/M/\infty$ queue with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 100$ and $\beta = 10/35$ and $\gamma = 0.01$.

However, the sample paths are not always so strikingly different. Indeed, the sample paths get less different as γ increases. Figures 8 and 9 illustrate by showing the sample paths for $\gamma = 1$ and $\gamma = 10$ over the interval $[0, 40]$. For $\gamma = 1$, there are again 6.28 sine cycles, but for $\gamma = 10$, there are 62.8 cycles. In these cases, the sample paths look much more similar. From Figures 8 and 9, we conclude that we might well use the fitted BD process to describe the transient behavior as well as the steady-state behavior for $\gamma \geq 1$, i.e., for relatively short cycles. Periodic arrival rates with short cycles often arise in practice in appointment-generated arrivals, where the actual arrivals are randomly distributed about the scheduled appointment times; see [28, 29] and references therein.

2.5 Limits for Small and Large γ

The behavior of the fitted BD process can be better understood by limits for the steady-state distribution of the $M_t/M/\infty$ model as $\gamma \uparrow \infty$ and as $\gamma \downarrow 0$. First, as $\gamma \uparrow \infty$, even though the arrival rate function oscillates more and more rapidly, the cumulative arrival rate function $\Lambda(t) \equiv \int_0^t \lambda(s)\, ds$ converges to the linear function $\bar{\lambda}t$. Consequently, the arrival process converges to a stationary Poisson process (M) with the average arrival rate $\bar{\lambda}$ and the steady-state number in system converges to the Poisson steady state distribution in associated the stationary $M/M/\infty$ model with mean $\bar{\lambda}$. That follows from Theorem 1 of [41] and references therein. As a consequence,

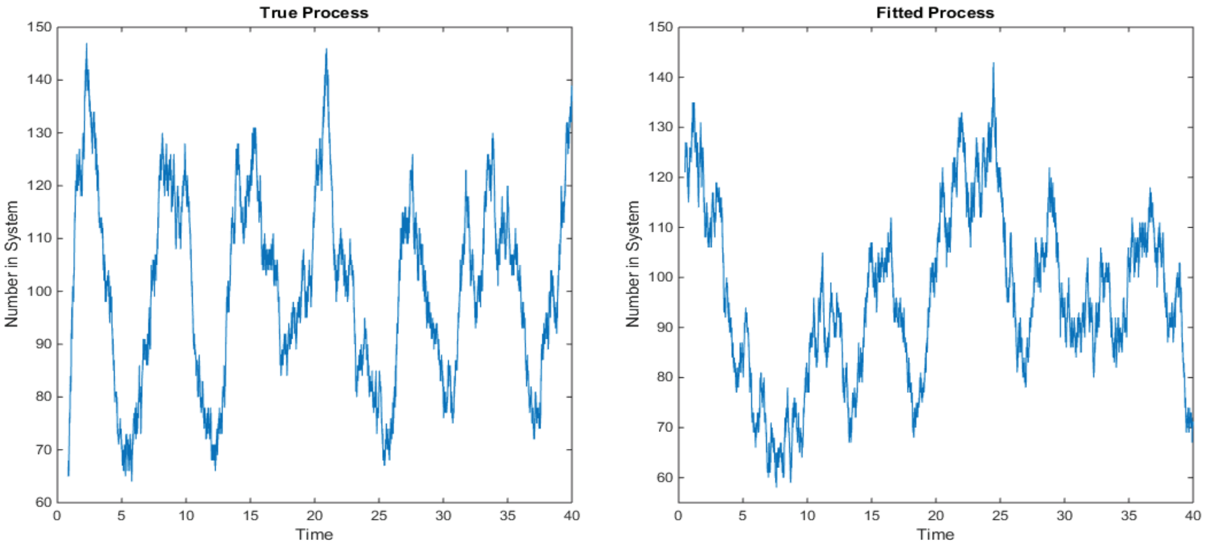

Figure 8: sample paths of the number in system for the original process (left) and the fitted BD process (right) for the $M_t/M/\infty$ queue with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 100$ and $\beta = 10/35$ and $\gamma = 1.0$.

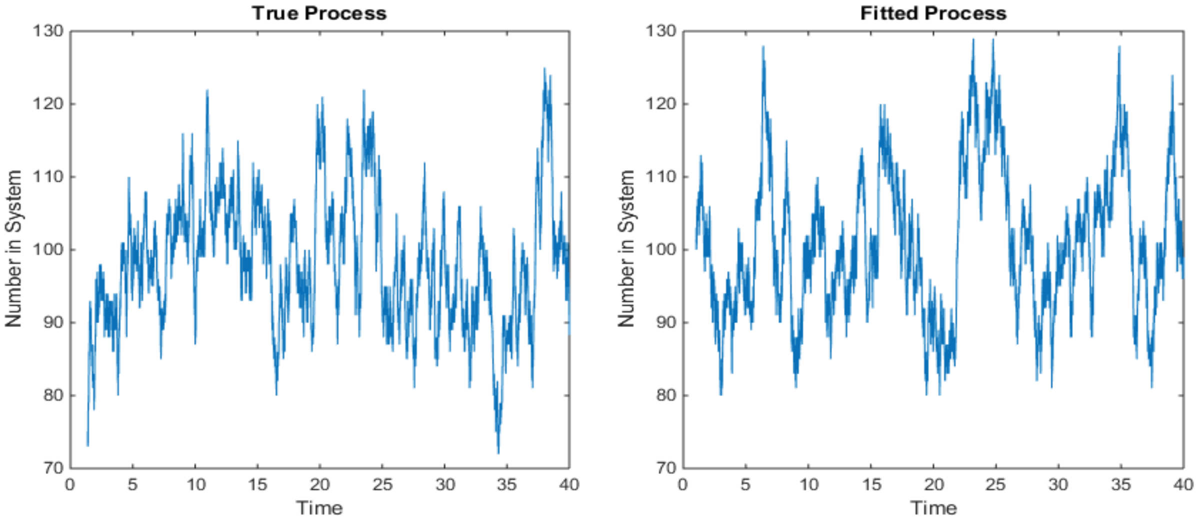

Figure 9: sample paths of the number in system for the original process (left) and the fitted BD process (right) for the $M_t/M/\infty$ queue with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 100$ and $\beta = 10/35$ and $\gamma = 10$.

as $\gamma \uparrow \infty$ we must have the fitted birth rates in the fitted BD process converge to the constant birth rates of a Poisson process, and that is precisely what we see as γ increases in Figure 3.

Second, as $\gamma \downarrow 0$, the cycles get longer and longer, so that the system behaves at each time t as a stationary model with the instantaneous arrival rate at that particular time t. That is the perspective of the pointwise stationary approximation (PSA) for queues with time-varying arrival rates [20], which is asymptotically correct for the $M_t/M/\infty$ model as $\gamma \downarrow 0$. That follows from Theorem 1 of [42]. As a consequence, as $\gamma \downarrow 0$ we must have the fitted birth rates in the fitted BD process converge to a proper limit, corresponding to an appropriate average of the birth rates seen at each time t for t in a sinusoidal cycle, and that is precisely what we see as γ increases in Figure 3. In particular, the limit Z_0 of the steady-state variable $Z \equiv Z_\gamma$ as $\gamma \downarrow 0$ is the mixture of the steady-state distributions. That is, by combining the PSA limit with (8), we see that

$$P(Z_0 = k) = \frac{1}{2\pi} \int_0^{2\pi} P(Q_0(t) = k)\, dt, \quad k \geq 0, \tag{9}$$

where $Q_0(t)$ has a Poisson distribution with mean $m_0(t) = \lambda_1(t)$, where we let $\gamma = 1$. In particular, this limit as $\gamma \downarrow 0$ becomes independent of γ.

These two limits as $\gamma \uparrow \infty$ and as $\gamma \downarrow 0$ can be seen by comparing the sample paths of the fitted BD processes for different γ. This is especially interesting for the long-cycle case. Figure 10 illustrates by showing the sample paths of the number in system for the fitted BD process in the $M_t/M/\infty$ queue with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 100$ and $\beta = 10/35$ and $\gamma = 0.1$ (left) and $\gamma = 0.01$ (right). The plots of different interval lengths show that the fitted BD processes are very similar.

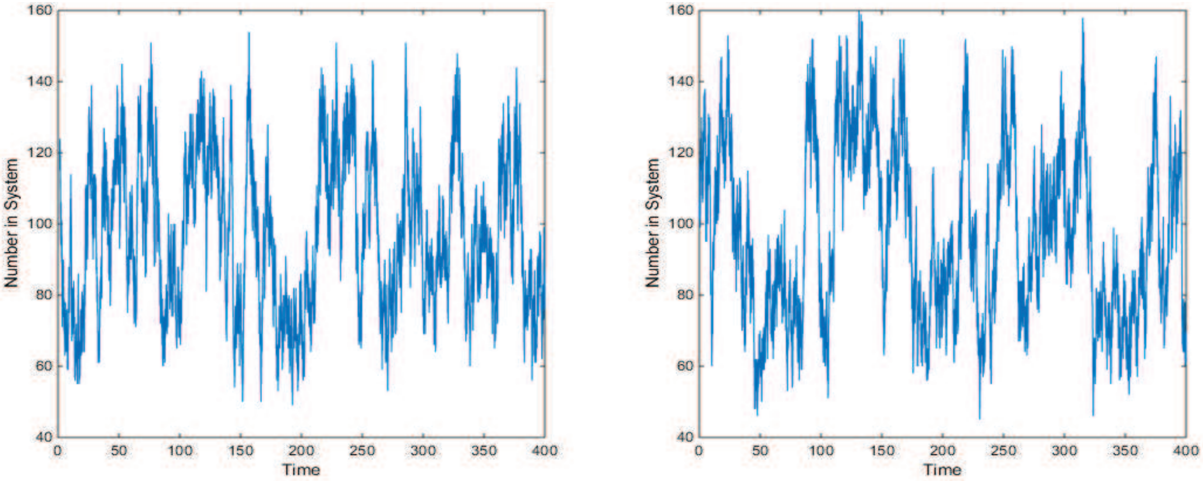

Figure 10: sample paths of the number in system for the fitted BD process in the $M_t/M/\infty$ queue with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 100$ and $\beta = 10/35$ and $\gamma = 0.1$ (left) and $\gamma = 0.01$ (right).

2.6 Speed Ratios: Very Different Limits for the Finite-Server Models

The stationary Poisson limit as $\gamma \uparrow \infty$ is the same in $M_t/GI/s$ models with s servers and general service times, but the limit as $\gamma \downarrow 0$ can be very different. Indeed, the limiting behavior will be very different if the finite-server model is overloaded with instantaneous traffic intensity $\rho(t) \equiv \lambda(t)/s\mu > 1$ at some time within its periodic cycle. If $\rho(t) > 1$ for some values of t and if we make γ very small, then these overload periods extend for longer and longer times, so that there can be a significant queue buildup. Indeed, proper limits as $\gamma \downarrow 0$ can only be obtained by adding additional scaling. This interesting phenomenon is discussed in [8].

The great difference between finite-server and infinite-server models as $\gamma \downarrow 0$ is illustrated by the speed ratios, introduced in [12] to partially characterize the transient behavior. To define the speed ratio, let $T(p, q)$ be the first passage time from the p^{th} percentile of the steady-state distribution to the q^{th} percentile of the steady-state distribution in the original process, and let $T_f(p, q)$ be the first passage time from the p^{th} percentile of the steady-state distribution to the q^{th} percentile of the steady-state distribution in the fitted BD process. These first passage times are fully specified for the fitted BD process because it is a Markov process, but they are not completely specified in the original model, because the stochastic process $\{Q(t) : t \geq 0\}$ is in general not Markov. Thus we need to specify the initial conditions. We understand the system to be in steady-state, so the initial condition is the steady-state distribution of the process conditional on starting at percentile p.

We in fact estimate the expected first passage times for the original process from simulations, by considering successive alternating visits to the p^{th} and q^{th} percentiles of the steady-state distribution. As an approximation, which we regard as reasonable as long as p is not too close to q, we will assume that these successive first passage times are i.i.d. We estimate the expected values of these first passage times by sample averages and estimate 95% confidence intervals under the i.i.d. assumption. The rate at which these transitions occur can be defined by

$$r(p,q) \equiv \frac{1}{E[T(p,q)]} \quad \text{and} \quad r_f(p,q) \equiv \frac{1}{E[T_f(p,q)]}.$$

The associated (p,q)-*speed ratio* can be defined by

$$\omega(p,q) \equiv \frac{r(p,q)}{r_f(p,q)} = \frac{E[T_f(p,q)]}{E[T(p,q)]}.$$

To obtain further simplification, we assume that $q = 1-p$ with $0 < p < 1/2$ and consider round trips, so that

$$T(p) = T(p, 1-p) + T(1-p, p) \quad \text{and} \quad T_f(p) = T_f(p, 1-p) + T_f(1-p, p),$$

$$r(p) \equiv \frac{1}{E[T(p)]} \quad \text{and} \quad r_f(p) \equiv \frac{1}{E[T_f(p)]}$$

and the p-*speed ratio* can be defined by

$$\omega(p) \equiv \frac{r(p)}{r_f(p)} = \frac{E[T_f(p)]}{E[T(p)]}. \tag{10}$$

Figure 11, plots the speed ratios in (10) for the case $p = 0.1$ for the $M_t/M/\infty$ (left) and $M_t/M/40$ models with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 35$ and $\beta = 10/35$ as a function of the parameter γ. Consistent with our previous discussion, Figure 11 shows that the speed ratios approach 1 as γ increases, but we see very different behavior as $\gamma \downarrow 0$. The finite limit for the $M_t/M/\infty$ model confirms the limit of the steady-state distributions, whereas the divergence for the $M_t/M/40$ model shows the divergence of the 40-server models, due to the persistent overload over long time intervals.

2.7 Different Service Distributions: Near Insensitivity

We have also conducted corresponding simulation experiments for the $M_t/GI/\infty$ model with non-exponential service-time distributions. Figure 12 shows the fitted rates for the H_2 service distributions with scv $c^2 = 2$ just as in §2 of [12]. The corresponding plots for the E_2 distribution are in an appendix (maintained by the authors on their web pages); they look similar. Figure 13 shows the associated steady-state mass functions for H_2 and E_2 service times, which also look similar.

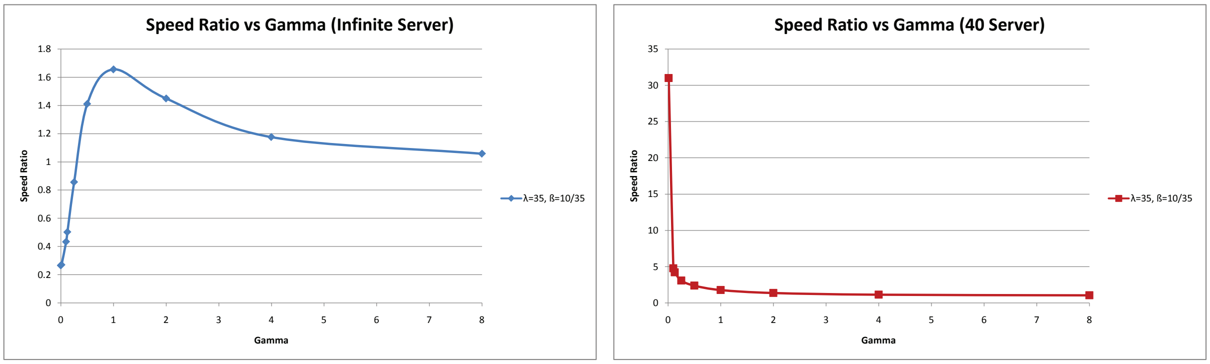

Figure 11: plots of the speed ratios in the $M_t/M/\infty$ (left) and $M_t/M/40$ models with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 35$ and $\beta = 10/35$ as a function of the parameter γ.

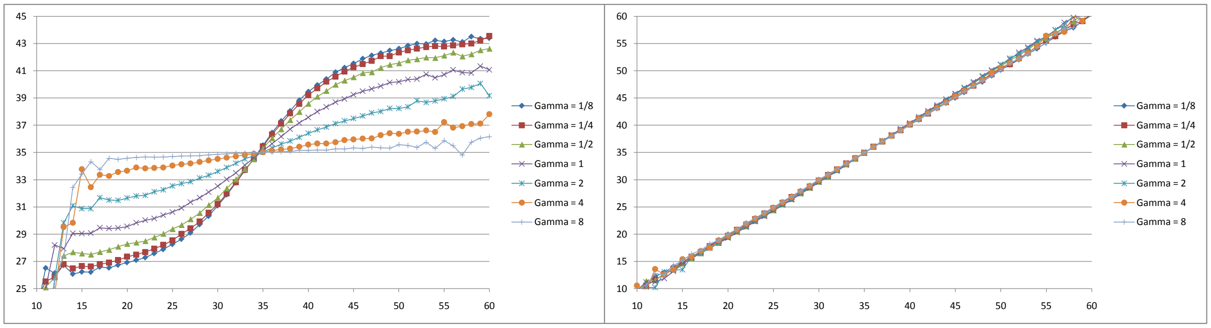

Figure 12: Fitted birth rates (left) and fitted death rates (right) for the $M_t/H_2/\infty$ model with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 35$ and $\beta\bar{\lambda} = 10$ and 7 values of γ ranging from 1/8 to 8. (The service scv is $c^2 = 2$.)

Indeed, the agreement is so good that it is natural to wonder whether the fitted birth rate, fitted death rate and steady-state pmf have an insensitivity property, i.e., depend on the service-time distribution only through its mean. However, closer examination show that it is not so. Plots for the D, E_2, M and H_2 service distributions having mean 1 with the same arrival rate function having $\gamma = 2$ exhibit significant differences, thus providing a concrete counterexample. Thus, we conclude that the model possesses a near-insensitivity property with regard to the service-time distribution.

We do see that the insensitivity property does hold asymptotically as $\gamma \downarrow 0$, which is to be expected. In that limit the PSA approximation is valid, so that at time t the model has a time-varying distribution equal to the steady-state distribution of the stationary $M/GI/\infty$ model with constant arrival rate equal to $\lambda(t)$.

Finally, consistent with [10], the transient behavior does not possess an insensitivity property. That is demonstrated by Figure 14, which shows that there are discernible differences among the speed ratios for the three service distributions, but the differences are not great.

2.8 Estimating the Steady-State Distribution

In this section we investigate how we can efficiently estimate the steady-state distribution by fitting parametric functions to the estimated birth and death rates and then solve the local balance

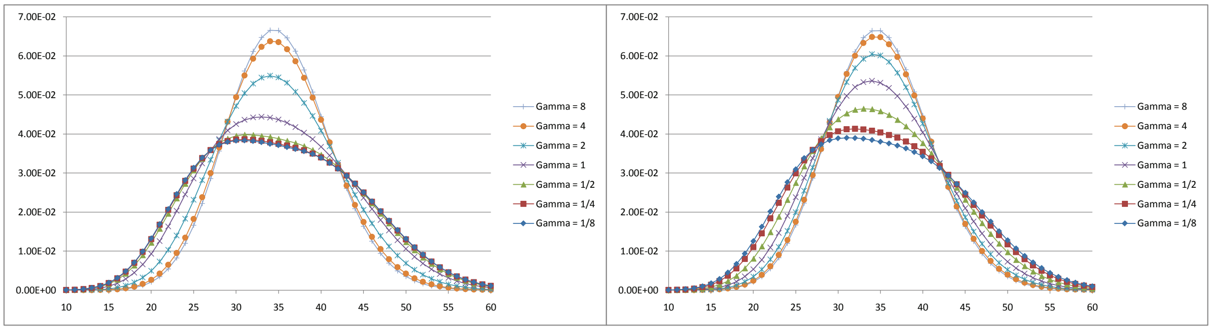

Figure 13: Fitted steady-state mass functions for the $M_t/H_2/\infty$ model (left) and the $M_t/E_2/\infty$ model (right) for with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 35$ and $\beta\bar{\lambda} = 10$ and 7 values of γ ranging from 1/8 to 8.

equation (2). First, for IS model with $E[S] = 1$, we do not need to consider the death rates, because we have $\bar{\mu}_k \approx k$ throughout. Hence, we concentrate on the birth rates. For larger values of γ, a linear function works well, but not for smaller values of γ. As our parametric function, we choose

$$\lambda_k^p = a \arctan b(k - c) + d, \tag{11}$$

which is nondecreasing in k with finite limits as k increases and decreases, and has the parameter four-tuple (a, b, c, d). We let $c = d = \bar{\lambda}$, so that leaves only the two parameters a and b.

Figures 15, 16 and 17 show the fitted mass function and birth rates for the three gamma values: $\gamma = 1/8$, $1/2$ and 2, respectively. These were constructed using the Matlab curve fitting toolbox, which fits by least squares. The figures show that the special arctangent function in (11) does much better than a linear fit for small γ, but a simple linear fit works well for large γ. The parameter pairs in the three cases were $(a, b) = (7.541, 0.125)$, $(6.682, 0.1253)$ and $(3.577, 0.0744)$, respectively. The main point is that a parametric fit based on only two parameters yields an accurate fit to a mass function that can be quite complicated.

3 Supporting Theory: The Periodic $M_t/GI/s$ Queueing Model

We now develop supporting theory, aiming to explain what we saw in the imulation plots. Let $A(t)$ count the number of arrivals in the interval $[0, t]$. We assume that the arrival rate function $\lambda(t)$ is a periodic continuous function with periodic cycle of length c. Let $\bar{\lambda}$ be the long-run average arrival rate, with

$$\bar{\lambda} \equiv \frac{1}{c}\int_0^c \lambda(s)\, ds = \lim_{t\to\infty} \frac{A(t)}{t}. \tag{12}$$

(That is consistent with (3) in the special case of a sinusoidal arrival rate function.) Let the service times be distributed as a random variable S with cumulative distribution function (cdf) G and mean $E[S] \equiv 1/\mu < \infty$. Let the (long-run) traffic intensity be defined by $\bar{\rho} \equiv \bar{\lambda}E[S]/s = \bar{\lambda}/s\mu$.

Let $Q(t)$ denote the number of customers in the system at time t and let $P(Q(t) = k)$, $k \geq 0$, be its time-dependent pmf. As indicated in [24], because of the NHPP arrival process, the stochastic process $\{Q(t) : t \geq 0\}$ is a regenerative processes, with the events $\{Q(nc + t) = 0\}$, $n \geq 1$, for any fixed t, $0 \leq t < c$, being regenerative events. As a consequence, we have a well defined periodic steady-state distribution when $\bar{\rho} < 1$.

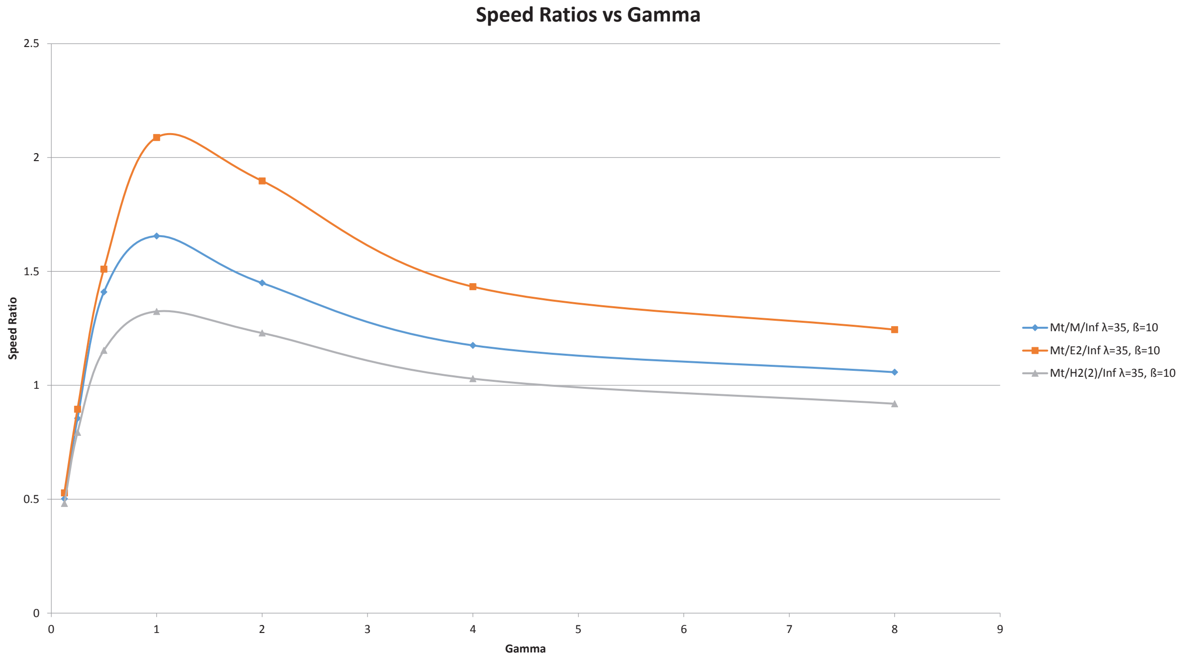

Figure 14: Speed ratios for the $M_t/GI/\infty$ model with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 35$ and $\beta\bar{\lambda} = 10$ as a function of γ for three different service distributions.

Theorem 3.1 (*periodic steady-state distribution*) *If $\bar{\rho} < 1$ in the periodic $M_t/GI/s$ queueing model, then a dynamic steady-state pmf $\alpha(t)$, $0 \le t < c$, and an overall steady-state pmf α^c are well defined probability vectors with*

$$\begin{aligned} \alpha_k(t) &\equiv \lim_{n\to\infty} P(Q(nc+t)=k) = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} 1_{\{Q(jc+t)=k\}}, \quad 0 \le t < c, \quad \textit{and} \\ \alpha_k^c &\equiv \frac{1}{c}\int_0^c \alpha_k(t)\,dt = \lim_{t\to\infty}\frac{1}{t}\int_0^t 1_{\{Q(s)=k\}}\,ds, \quad k \ge 0. \end{aligned} \tag{13}$$

Let $\bar{\lambda}_k(t)$ and $\bar{\mu}_k(t)$ be the fitted birth rate and death rate in state k from data over $[0,t]$, obtained as indicated in §1. Our theoretical results will be for the limits $\bar{\lambda}_k(\infty)$ and $\bar{\mu}_k(\infty)$ obtained by letting $t \to \infty$. In the $M_t/GI/s$ model, the arrival rate actually depends only on time, not the state. Hence, we can obtain the following explicit expressions for the fitted rates with ample data.

Theorem 3.2 (*fitted birth and death rates with ample data*) *In the periodic $M_t/GI/s$ queueing model with $\bar{\rho} < 1$,*

$$\bar{\lambda}_k(\infty) = \frac{\int_0^c \alpha_k(t)\lambda(t)\,dt}{\int_0^c \alpha_k(t)\,dt} = \frac{\int_0^c \alpha_k(t)\lambda(t)\,dt}{c\alpha_k^c} \tag{14}$$

and

$$\bar{\mu}_{k+1}(\infty) = \frac{\alpha_k^c \bar{\lambda}_k(\infty)}{\alpha_{k+1}^c} = \frac{\int_0^c \alpha_k(t)\lambda(t)\,dt}{c\alpha_{k+1}^c}. \tag{15}$$

for $\alpha_k(t)$ and α_k^c in (13).

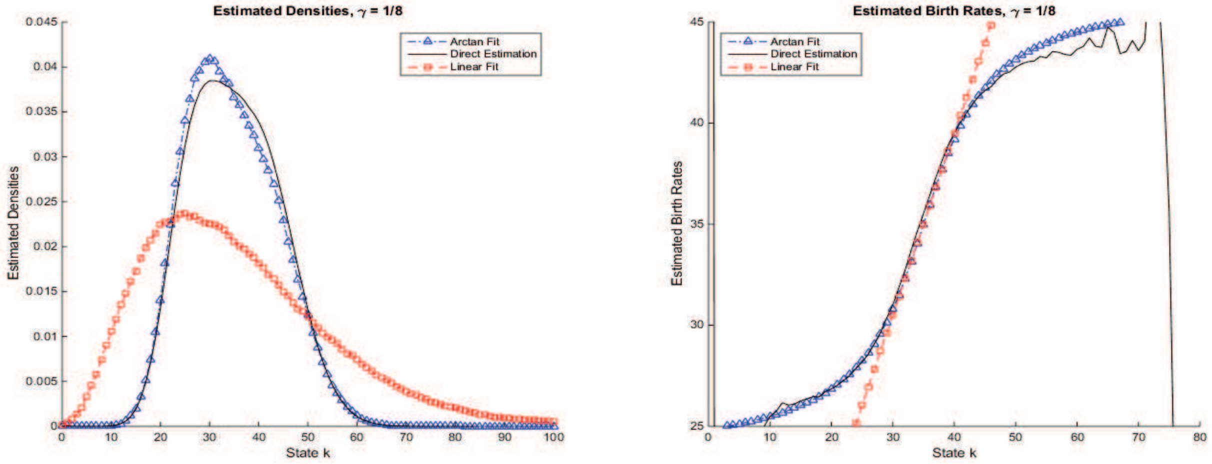

Figure 15: Fitted mass function (left) and birth rates (right) for the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 35$, $\beta\bar{\lambda} = 10$ and $\gamma = 0.125$

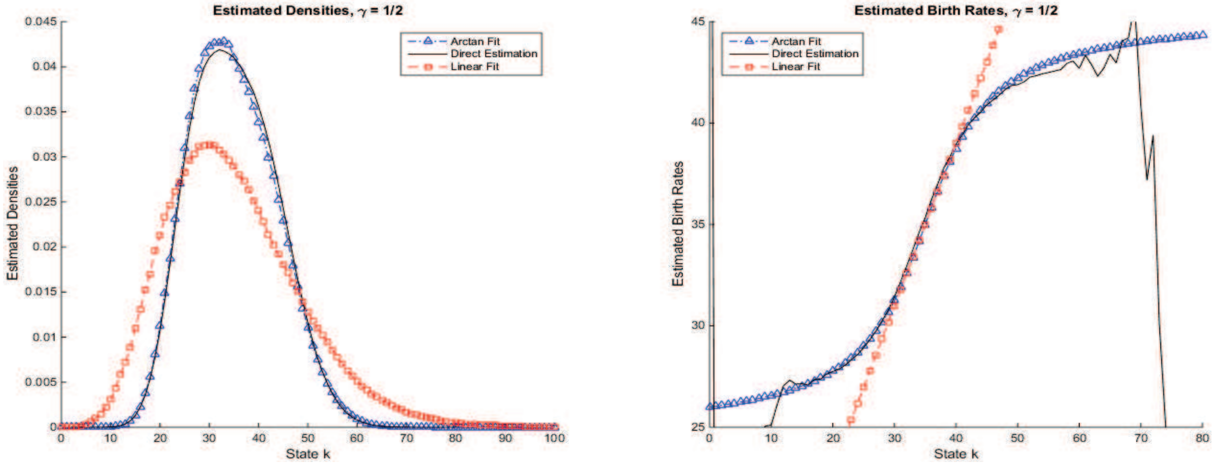

Figure 16: Fitted mass function (left) and birth rates (right) for the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 35$, $\beta\bar{\lambda} = 10$ and $\gamma = 0.5$

Proof. We use the regenerative structure to focus on (i) the expected number of arrivals in state k per cycle divided by the expected length of a cycle and (ii) the expected time spent in state k per cycle divided by the expected length of a cycle. We get (14) by looking at the ratio. Since the arrival rate depends only on time, we have (14). We then can apply the detailed balance equation in (2) to get (15). ▪

Theorems 3.1 and 3.2 can be applied in two ways. First, we can apply these theorems to learn about the fitted birth and death rates. They pose a strong constraint on the fitted birth and death rates because the detailed balance equation in (2) must hold. As a consequence, if we know either the fitted birth rates or the fitted death rates, then the others are determined as well. We will illustrate in our specific results below.

Second, we can apply the estimated birth and death rates to estimate the steady-state probability vector α^c in Theorem 3.1. Let $\bar{\alpha}^e(\infty)$ be the steady-state probability vector of the fitted BD process obtained from (2). Since $\bar{\alpha}^e$ coincides with α^c in (13), we can use the fitted BD model to calculate the steady-state distribution α^c in (13). To do so, we estimate the birth and death rates

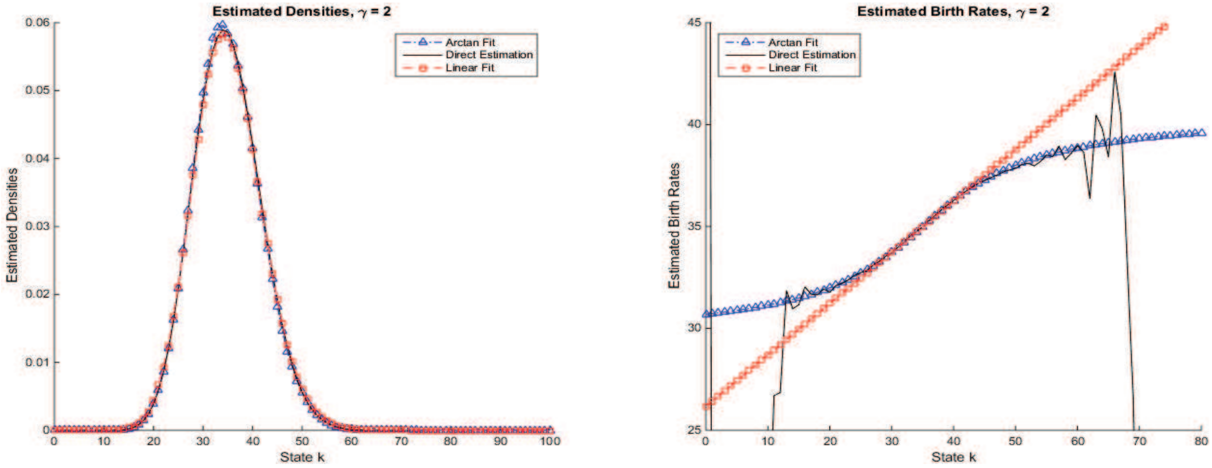

Figure 17: Fitted mass function (left) and fitted birth rates (right) for the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (3) having parameters $\bar{\lambda} = 35$, $\beta\bar{\lambda} = 10$ and $\gamma = 2.0$

and then apply the detailed balance equation in (2). Moreover, by developing analytical approximations for the fitted birth and death rates, we succeed in developing an analytical approximation for α^c.

We can immediately apply Theorem 3.2 to obtain bounds on the fitted birth rates. Since formula (14) expresses $\bar{\lambda}_k(\infty)$ as an average of the arrival rate function over one cycle, we can immediately deduce

Corollary 3.1 (*bounds on the fitted birth rates*) *In the periodic $M_t/GI/s$ queueing model starting empty in the distant past,*

$$\lambda_L \equiv \inf_{0 \le t < c} \lambda(t) \le \bar{\lambda}_k(\infty) \le \sup_{0 \le t < c} \lambda(t) \equiv \lambda_U. \tag{16}$$

3.1 The Periodic $M_t/M/s$ Model

For the special case of an exponential service-time distribution, i.e., for the $M_t/M/s$ model, the stochastic process $\{Q(t) : t \ge 0\}$ is Markov and more convenient explicit formulas are available.

We first observe that an analog of Theorem 3.1 of [12] also holds for the fitted death rates in the present time-varying case.

Theorem 3.3 (*explicit death rates*) *For the periodic $M_t/M/s$ model with $\bar{\rho} < 1$,*

$$\bar{\mu}_k(\infty) = \min\{k, s\}\mu, \quad k \ge 0, \tag{17}$$

so that

$$\bar{\lambda}_k(\infty) = \frac{\alpha^c_{k+1} \min\{k+1, s\}\mu}{\alpha^c_k}, \quad k \ge 0, \tag{18}$$

for α^c_k in (13).

Proof. As for Theorem 3.1 of [12], (17) follows from the lack of memory property of the exponential distribution. We then apply (2) to get (18). However, we now show that it is also possible to directly apply Theorem 3.1 of [12] here. We use the fact that the $M_t/M/s$ model has a proper dynamic periodic steady-state distribution with a period equal to the period of the arrival process, cf. [24]. For that model we can convert the arrival process to a stationary point process by simply randomizing where we start in the first cycle. If the period is of length d, then we start the arrival process at time t, where t is uniformly distributed over the interval $[0, d]$. That randomization converts the arrival process to a stationary point process, so that we can apply Theorem 3.1 of [12] (a). But then we observe that the randomization does not alter the limit (17). ■

We next observe that a geometric tail holds for the $M_t/M/s$ model with the same decay rate as for the associated stationary $M/M/s$ model with arrival rate $\bar{\lambda}$. Recall that a probability vector α has a geometric tail with decay rate σ if

$$\alpha_k \sim \zeta\sigma^k \quad \text{as} \quad k \to \infty, \tag{19}$$

for positive constants σ and ζ, i.e., if the ratio of the two sides in (19) converges to 1 as $k \to \infty$; see §3.3 of [12].

Theorem 3.4 *(geometric tail) For the $M_t/M/s$ model with $s < \infty$ and $\bar{\lambda} < s\mu$, the periodic steady-state pmf's $\alpha_k(t)$ and α^c_k in (13) possess a geometric tail as in (19) with the same decay rate as in the associated stationary $M/M/s$ model with arrival rate $\bar{\lambda}$; i.e.,*

$$\alpha_k(t) \sim \zeta_t\sigma_t^k \quad \textit{as} \quad k \to \infty \quad \textit{for each} \quad t, \quad 0 \le t < c, \tag{20}$$

and

$$\alpha^c_k \sim \zeta^c\sigma_c^k \quad \textit{as} \quad k \to \infty, \tag{21}$$

where

$$\sigma_c = \sigma_t = \sigma = \rho \equiv \frac{\bar{\lambda}}{s\mu}, \quad \zeta_t \ge \zeta \ge (1-\rho) \quad \textit{and} \quad \zeta^c \ge \zeta \ge (1-\rho) \tag{22}$$

with (ζ, σ), (ζ_t, σ_t) and (ζ^c, σ_c) denoting the asymptotic parameter pairs for $M/M/s$, $\alpha(t)$ and α^c. As a consequence,

$$\bar{\lambda}_k(\infty) \to \bar{\lambda} \quad \textit{as} \quad k \to \infty. \tag{23}$$

Proof. For each t in a cycle $[0, c]$, the tail behavior can be deduced by considering bounding discrete-time processes, looking at the system at times $t + kc$, $k \ge 0$. Both systems are bounded below by the discrete-time model that has all arrivals in each interval at the end of the interval and all departures at the beginning of the interval, while both systems are bounded above by the discrete-time model that has all arrivals in each interval at the beginning of the interval and all departures at the end of the interval. These two-discrete time systems are random walks with steady-state distributions satisfying (19) with common decay factor $\sigma = \rho$. A step in the random walk is the difference of two Poisson random variables $U - D$, where $EU = \bar{\lambda}c$ and $ED = s\mu c$, which have ratio $EU/ED = \bar{\lambda}/s\mu$, which in turn determines the decay rate. A stochastic comparison [7] then implies that $\beta_t \ge \beta$. For the final inequality in (22), we can compare the $M/M/s$ system to the corresponding $M/M/1$ model with a fast server, working at rate $s\mu$. The two systems have the same birth rate, while the $M/M/1$ system has death rates that are greater than or equal to those in the $M/M/s$ model. Hence, the steady-state distributions are ordered stochastically. Finally, the final limit in (23) follows from Theorem 3.3 and (21), where here $s\mu\sigma = s\mu\rho = \bar{\lambda}$. ■

We remark in closing this section that the periodic $M_t/M/\infty$ has different tail behavior; hence the assumption that $s < \infty$. We next start considering the IS model.

3.2 The Periodic Infinite-Server Model

We now consider the special case of the periodic $M_t/GI/\infty$ IS model, because it admits many explicit formulas, as shown in [13, 14, 35]. We let the model start in the indefinite past, so that it can be regarded as in periodic steady-state at time 0. This is achieved by assuming an explicit form for the arrival rate function, as in (3), and then assuming that the system started empty in the indefinite past.

By Theorem 1 of [14], the number in system has a Poisson distribution for each t with periodic mean function $m(t)$, with the same period c, where

$$m(t) = E[\lambda(t - S_e)]E[S] = E[S] \int_0^\infty \lambda(t - s) dG_e(s), \quad t \geq 0, \tag{24}$$

and S_e is a random variable with the stationary-excess cdf G_e associated with the service-time cdf G, i.e.,

$$G_e(t) \equiv P(S_e \leq t) \equiv \frac{1}{E[S]} \int_0^t (1 - G(s))\, ds, \quad t \geq 0. \tag{25}$$

Moreover, the departure process in the $M_t/GI/\infty$ model is a Poisson process with periodic rate function $\delta(t)$, with the same period c, where

$$\delta(t) = E[\lambda(t - S)] = \int_0^\infty \lambda(t - s) dG(s), \quad t \geq 0. \tag{26}$$

For the special case of a sinusoidal arrival rate function, an explicit expression for $m(t)$ is given in Theorem 4.1 of [13].

As a consequence, we have the following corollary to Theorem 3.1.

Corollary 3.2 (*periodic steady-state distribution in the IS model*) *In the periodic $M_t/GI/\infty$ queueing model starting empty in the distant past, $\alpha(t)$, $0 \leq t < c$ and α^c are well defined probability vectors with*

$$\alpha_k(t) = \pi_k(m(t)), \quad 0 \leq t < c, \quad \textit{and} \quad \alpha_k^c = \frac{1}{c} \int_0^c \pi_k(m(t))\, dt, \tag{27}$$

for $m(t)$ in (24), where $\pi_k(m)$ be the Poisson distribution with mean m, i.e.,

$$\pi_k(m) \equiv \frac{e^{-m} m^k}{k!}, \quad k \geq 0.$$

We now consider the fitted death rates estimated with ample data, i.e., $\bar{\mu}_k(\infty)$. To obtain the departure rate conditional on the number of busy servers, we use the following consequence of Theorem 2.1 of [19], which characterizes the time-varying distributions of the remaining service times in an $M_t/GI/\infty$ model, conditional on the number of busy servers, extending the classical result for the $M/GI/\infty$ model.

Theorem 3.5 (*remaining service times conditional on the number*) *Consider the periodic $M_t/GI/\infty$ queueing model starting empty in the distant past, where the service-time cdf G has pdf g. Conditional on $Q(t) = k$, the remaining service times at time t are distributed as k i.i.d. random variables with pdf*

$$g_{k,t}(x) = \frac{\int_0^\infty \lambda(t - u) g(x + u)\, du}{\int_0^\infty \lambda(t - u) G^c(u)\, du}, \quad x \geq 0,$$

which is independent of k.

We now apply Theorem 3.5 to obtain the following general result about the fitted death rates.

Theorem 3.6 (*fitted death rates*) *Consider the $M_t/GI/\infty$ queue with a periodic arrival rate function in the setting of Theorem 3.5. Conditional on $Q(t) = k$, the departure rate at time t is*

$$\delta_k(t) = k\delta_1(t) = kg_{k,t}(0) = \frac{k\mu E[\lambda(t-S)]}{E[\lambda(t-S_e)]} = \frac{k\delta(t)}{m(t)}. \tag{28}$$

Hence, paralleling the fitted birth rate in (14),

$$\bar{\mu}_k(\infty) = \frac{\int_0^c \alpha_k(t)\delta_k(t)\, dt}{c\alpha_k^c} = \frac{k\int_0^c \alpha_k(t)(\delta(t)/m(t))\, dt}{c\alpha_k^c}, \quad k \ge 1, \tag{29}$$

where $\alpha_k(t)$, α_k^c, $m(t)$ and $\delta(t)$ are given in (27), (24) and (26).

Proof. First, we get (28) directly from Theorem 3.5 and formulas (24) and (26). The first term in (29) can be taken as a definition. Then we apply (28). ■

Paralleling Corollary 3.1, Theorem 3.6 implies bounds for the fitted death rates. We get equality of the upper and lower bounds, recovering (17) for $s = \infty$, if S is exponential, because then $\delta(t) = m(t)\mu$ since S_e is distributed the same as S.

Corollary 3.3 (*bounds on the fitted death rates*) *In the periodic $M_t/GI/\infty$ queueing model starting empty in the distant past,*

$$\mu_L \equiv \inf_{0\le t<c}\{\delta(t)/m(t)\} \le \frac{\bar{\mu}_k(\infty)}{k} \le \sup_{0\le t<c}\{\delta(t)/m(t)\} \equiv \mu_U. \tag{30}$$

for $m(t)$ in (24) and $\delta(t)$ in (26).

Proof. Theorem 3.6 expresses $\bar{\mu}_k(\infty)/k$ as an average of $\delta(t)/m(t)$ over one cycle. ■

We now apply Theorem 3.5 to deduce a rate conservation property for this $M_t/GI/\infty$ model in each state over a periodic cycle.

Theorem 3.7 (*arrival and departure rates over a cycle*) *For the periodic $M_t/GI/\infty$ queueing model starting empty in the distant past,*

$$\int_0^c \alpha_k(t)\lambda(t)\, dt = \int_0^c \alpha_k(t)\delta(t)\, dt \quad \textit{for each} \quad k \ge 0 \tag{31}$$

for $\alpha_k(t)$ in (27), so that

$$\int_0^c \lambda(t)\, dt = \int_0^c \delta(t)\, dt. \tag{32}$$

Proof. Since the arrival rate at time t is $\lambda(t)$, independent of the state k, we can apply first (2) and then (28) to obtain

$$\begin{aligned}\int_0^c \alpha_k(t)\lambda(t)\, dt &= c\alpha_k^c\bar{\lambda}_k(\infty) = c\alpha_{k+1}^c\bar{\mu}_{k+1}(\infty) = \int_0^c \alpha_{k+1}(t)\delta_{k+1}(t)\, dt \\ &= \int_0^c \alpha_{k+1}(t)(k+1)[\delta(t)/m(t)]\, dt = \int_0^c \alpha_k(t)\delta(t)\, dt, \end{aligned} \tag{33}$$

as in (31). We add over k to get (32). ■

4 The IS Model with a Sinusoidal Arrival-Rate Function

We now consider the special case of the periodic $M_t/GI/\infty$ IS model with the sinusoidal arrival rate function in (3), as in all our simulation experiments. For this model we draw on previous results established in [13].

4.1 A General Service-Time Distribution

We can apply Corollary 3.1 to obtain explicit bounds on the fitted birth rates.

Corollary 4.1 (*bounds on the fitted birth rates*) *For the $M_t/GI/\infty$ model with sinusoidal arrival rate function in* (3) *having $0 < \beta < 1$, starting empty in the distant past,*

$$0 < (1-\beta) \le \frac{\bar{\lambda}_k(\infty)}{\bar{\lambda}} \le (1+\beta) < 2 \quad \textit{for all} \quad k \ge 0. \tag{34}$$

We now establish asymptotic results for the extreme cases in which the cycles are very long ($\gamma \downarrow 0$) or are very short ($\gamma \uparrow \infty$). We directly show the dependence on γ; e.g., by writing $\bar{\lambda}_k(\infty;\gamma)$. The following result is consistent with the known results that the arrival process converges to a stationary Poisson process, and the steady-state distribution converges to a Poisson distribution with mean $\bar{\lambda}/\mu$ as $\gamma \downarrow 0$; see Theorem 1 of [41].

Theorem 4.1 (*short cycles*) *For the $M_t/GI/\infty$ model with sinusoidal arrival rate function in* (3)*,*

$$\bar{\lambda}_k(\infty;\gamma) \to \bar{\lambda} \quad \textit{and} \quad \bar{\mu}_{k+1}(\infty;\gamma) \to (k+1)\mu \quad \textit{as} \quad \gamma \uparrow \infty \quad \textit{for all} \quad k \ge 0. \tag{35}$$

Proof. First, it is helpful to rewrite (14) so that the integrals are over a fixed interval, independent of γ. By making a change of variables $s = \gamma t$, we obtain

$$\bar{\lambda}_k(\infty;\gamma) = \frac{\int_0^{2\pi/\gamma} \alpha_k(t)\lambda(t)\, dt}{\int_0^{2\pi/\gamma} \alpha_k(t)\, dt} = \frac{\int_0^{2\pi} \alpha_k(s/\gamma)\lambda(s/\gamma)\, ds}{\int_0^{2\pi} \alpha_k(s/\gamma)\, ds} \tag{36}$$

The conclusion follows in two steps. First, $\lambda(s;\gamma) \to \bar{\lambda}$ as $\gamma \uparrow \infty$, uniformly in s over $[0, 2\pi]$. (Recall that $\lambda(0;\gamma) = \bar{\lambda}$ because $\sin(0) = 0$ and that $\sin(t) \to 0$ as $t \downarrow 0$.) Second, by Theorem 4.5 of [13], $m(t;\gamma) \to \bar{\lambda}/\mu$ as $\gamma \uparrow \infty$, uniformly in t. Hence, $\alpha_k(t;\gamma) \to \alpha_k(t;\infty)$ as $\gamma \uparrow \infty$, uniformly in t, where $\alpha_k(t;\infty)$ is the Poisson pmf with mean $\bar{\lambda}/\mu$, independent of t. For the fitted death rates, we apply (2) to write

$$\bar{\mu}_{k+1}(\infty;\gamma) = \frac{\bar{\lambda}_k(\infty;\gamma)\alpha^c_{k;\gamma}}{\alpha^c_{k+1;\gamma}} \to \frac{\bar{\lambda}\alpha^c_{k;\infty}}{\alpha^c_{k+1;\infty}} = (k+1)\mu \quad \text{as} \quad \gamma \uparrow \infty, \tag{37}$$

because $\alpha_k(t;\infty)$ is the Poisson pmf with mean $\bar{\lambda}/\mu$ independent of t. ■

We now turn to the case of long cycles, where the PSA is appropriate. Thus, the steady-state pmf α^c is the average of the individual steady-state pmf's for each t in the cycle; see Theorem 1 of [42]

Theorem 4.2 (*long cycles*) *For the $M_t/GI/\infty$ model with sinusoidal arrival rate function in* (3)*,*

$$\bar{\lambda}_k(\infty;\gamma) \to \frac{(k+1)\mu\alpha^c_{k+1;0}}{\alpha^c_{k;0}} \quad \textit{and} \quad \bar{\mu}_{k+1}(\infty;\gamma) \to (k+1)\mu \quad \textit{as} \quad \gamma \downarrow 0 \tag{38}$$

for all $k \ge 0$, where $\alpha^c_{k;0}$ is the time average of $\alpha^c_k(t;0)$ which is the Poisson pmf with mean $\bar{\lambda}\lambda_1(t)/\mu$, where $\lambda_1(t) = 1 + \beta\sin(t)$, $0 \le t \le 2\pi$.

Proof. By Theorem 4.4 of [13], $m(t/\gamma) \to \lambda(t)/\mu$ as $\gamma \downarrow 0$ uniformly in t. Hence, $\alpha_k(t;\gamma) \to \alpha_k(t;0)$ uniformly in t. We then apply this starting from (36), getting

$$\begin{aligned}\bar{\lambda}_k(\infty) &= \frac{\int_0^{2\pi/\gamma} \alpha_k(t)\lambda(t)\,dt}{\int_0^{2\pi/\gamma} \alpha_k(t)\,dt} = \frac{\int_0^{2\pi} \alpha_k(s/\gamma)\lambda(s/\gamma)\,ds}{\int_0^{2\pi} \alpha_k(s/\gamma)\,ds} \\ &\to \frac{\int_0^{2\pi} \alpha_k(s;0)\lambda(s;0)\,ds}{\int_0^{2\pi} \alpha_k(s;0)\,ds} = \frac{\int_0^{2\pi} (k+1)\mu\alpha_{k+1}(s;0)\,ds}{\int_0^{2\pi} \alpha_k(s;0)\,ds} = \frac{(k+1)\mu\alpha^c_{k+1;0}}{\alpha^c_{k;0}},\end{aligned}$$

because $\alpha_k(s;0)$ is the Poisson pmf with mean $\lambda(s;0)/\mu$ at time s. ∎

4.2 The $M_t/M/\infty$ Model with Sinusoidal Arrival Rate

In this section we provide stronger theoretical support to explain the fitted birth rate functions for the $M_t/M/\infty$ model with sinusoidal arrival rate shown in Figure 3. In particular, we determine tight bounds, verified by showing that these bounds are approached in the heavy-traffic limit.

As shown in [13], the $M_t/M/\infty$ model with sinusoidal arrival rate function in (3) is especially tractable. From (15) of [13], the number in system, $Q(t)$, has a Poisson distribution for each t with mean in (5). We use this expression to improve the bounds in Corollary 4.1 and obtain a simple proof of Theorem 4.1 in this case.

Theorem 4.3 (*bounds for the fitted birth rates for the $M_t/M/\infty$ model with sinusoidal arrival rate function*) *In the $M_t/M/\infty$ IS queueing model with the sinusoidal arrival rate function in (3), starting empty in the distant past,*

$$\bar{\lambda}\left(1 - \frac{\beta}{\sqrt{1+\gamma^2}}\right) \le \bar{\lambda}_k(\infty) \le \bar{\lambda}\left(1 + \frac{\beta}{\sqrt{1+\gamma^2}}\right) \quad \textit{for all} \quad k \ge 0. \tag{39}$$

and

$$\bar{\lambda}_k(\infty) \to \bar{\lambda} \quad \textit{as} \quad \gamma \to \infty \quad \textit{for all} \quad k \ge 0. \tag{40}$$

Proof. We apply (2) to obtain the expression

$$\frac{(k+1)\bar{\lambda}_k(\infty)}{\bar{\mu}_{k+1}(\infty)} = \frac{\alpha^c_{k+1}}{\alpha^c_k}, \quad k \ge 0. \tag{41}$$

Since we have M service, $\bar{\mu}_{k+1}(\infty) = (k+1)\mu$. Hence we can write

$$\bar{\lambda}_k(\infty) = \frac{(k+1)\mu\alpha^c_{k+1}}{\alpha^c_k}, \quad k \ge 0. \tag{42}$$

Since the integrand in the integral representation of α^c_{k+1} in (27) differs from the the integrand in the integral representation of α^c_k by an extra factor of $m(t)/(k+1)$, we can insert the bounds on $m(t)$ in (18) of [13] to obtain (39). Clearly, (40) follows from (39). ∎

Remark 4.1 (*an alternative approach*) We conclude by mentioning that another approach to Theorem 4.3, which we plan to discuss elsewhere, is to apply the known functional weak law of large numbers, e.g., in [30, 33] expressing convergence of the queueing model to a deterministic fluid model in the heavy-traffic limit as $\bar{\lambda} \to \infty$. As $\bar{\lambda} \to \infty$, the scaled arrival process $\bar{A}_{\bar{\lambda}}(t) \equiv A_{\bar{\lambda}}(t)/\bar{\lambda}$, where $A_{\bar{\lambda}}(t)$ is the NHPP with arrival rate function in (3) converges to $\lambda_f(t) \equiv 1 + \beta\sin(\gamma t)$ and the associated scaled queue-length process $\bar{Q}_{\bar{\lambda}}(t) \equiv Q_{\bar{\lambda}}(t)/\bar{\lambda}$ converges to the mean function $s(t)$ in (5). For the limiting deterministic fluid model, each state that is visited is visited exactly twice, except the two extreme states and the middle state. Hence, the fitted arrival rate becomes the average of two arrival rates during the cycle.

We conclude by deriving a heavy-traffic limit showing that the lower and upper bounds established in Theorem 4.3

are attained in the heavy-traffic limit. This limit involves $\bar{\lambda}$, which is both the long-run average arrival rate and the long-run average number of busy servers. In the limit $\bar{\lambda} \to \infty$, any fixed state k, independent of $\bar{\lambda}$, will thus be a small state in the limit, so we should expect to see the minimum value of the increasing fitted birth rate function, as shown in Figure 3, in the first limit in (43) below. To have a relatively large state compared to $\bar{\lambda}$ asymptotically in the limit $\bar{\lambda} \to \infty$, we let the state index be $\lfloor m\bar{\lambda} \rfloor + k$ for suitably large m in the second limit. That yields the upper bound.

Theorem 4.4 (*heavy-traffic limits*) *In the $M_t/M/\infty$ IS queueing model with periodic arrival rate function, starting empty in the distant past,*

$$\begin{aligned} \frac{\bar{\lambda}_k(\infty)}{\bar{\lambda}} &\to 1 - \frac{\beta}{\sqrt{1+\gamma^2}} \quad \textit{as} \quad \bar{\lambda} \to \infty \quad \textit{and} \\ \frac{\bar{\lambda}_{\lfloor m\bar{\lambda} \rfloor + k}(\infty)}{\bar{\lambda}} &\to 1 + \frac{\beta}{\sqrt{1+\gamma^2}} \quad \textit{as} \quad \bar{\lambda} \to \infty \quad \textit{for} \quad m > 1/\log_e 2 \approx 1.44. \end{aligned} \tag{43}$$

Proof. We expand (42), writing

$$\bar{\lambda}_k(\infty) = \frac{\alpha^c_{k+1}(k+1)\mu}{\alpha^c_k} = \frac{\mu \int_0^c e^{-m(t)} m(t)^{k+1}\, dt}{\int_0^c e^{-m(t)} m(t)^k\, dt} \tag{44}$$

In each case of (43), we apply Laplace's method to the numerator and denominator of (44), after pre-multiplying both by the same appropriate term (so this term cancels). Let $x \equiv \bar{\lambda}/\mu$ and consider the first expression. In particular, After multiplying the numerator and denominator by e^x/x^k, we can express the denominator as

$$\int_0^c e^{-xs(t)}(1+s(t))^k\, dt \sim \sqrt{\frac{2\pi}{x|s''(x_0)|}}(1+s(x_0))^k e^{xs(x_0)} \quad \text{as} \quad x \to \infty,$$

where $\sim$ means that the ratio of the two sides converges to 1, $s(t) \equiv m_1(t) - 1$ for $m_1(t)$ in (5), where $c = 2\pi/\gamma$ and $x_0 = c - cot^{-1}(1/\gamma))/\gamma$ and $m(x_0) = (\bar{\lambda}/\mu)(1 - \beta/(\sqrt{1+\gamma^2}))$, by virtue of (16) and (18) in [13]. (The minus sign in the exponent of $e^{-xs(t)}$ means that we look for the most negative value of $s(t)$.) We have used the fact that the integral is dominated by an appropriate modification of the integrand at a single point when x becomes large. The ratio in (44) thus approaches $1 + s(x_0)$.

For the second expression, after multiplying the numerator and denominator by e^x/x^{x+k}, we can express the denominator as

$$\begin{aligned} \int_0^c e^{-xs(t)}(1+s(t))^{mx+k}\, dt &= \int_0^c e^{+x[m\log_e\{1+s(t)\} - s(t)]}(1+s(t))^k\, dt \\ &\sim \sqrt{\frac{2\pi}{x|f''(x_0)|}}(1+s(x_0))^k e^{xf(x_0)} \quad \text{as} \quad x \to \infty, \end{aligned}$$

where $f(t) \equiv m \log_e \{1 + s(t)\} - s(t)$, so that $x_0 = (c/4) + cot^{-1}(1/\gamma))/\gamma$ and $m(x_0) = (\bar{\lambda}/\mu)(1 + \beta/(\sqrt{1+\gamma^2}))$, again by (16) and (18) in [13]. (The plus sign in the exponent of $e^{+x[m \log_e \{1+s(t)\}-s(t)]}$ with $m > 1/\log_2 2$ means that we look for the most positive value of $s(t)$.) The ratio in (44) again approaches $1 + s(x_0)$. ■

5 Notation

We summarize key notation in Table 1, giving the meaning and where it was first introduced.

Table 1: A summary of the notation, giving the meaning and where it was introduced.

symbol	meaning	where
$\bar{\lambda}_k \equiv \bar{\lambda}_k(t)$	fitted birth rate in state k from data over $[0,t]$	§1
$\bar{\mu}_k \equiv \bar{\mu}_k(t)$	fitted death rate in state k from data over $[0,t]$	§1
λ	overall arrival rate	§1.2
μ	individual (per server) service rate	§1.2
s	number of servers (allowing $s = \infty$)	§1.2
θ	individual (per customer in queue) abandonment rate	§1.2
$\bar{\alpha}^e \equiv \bar{\alpha}^e_k$	steady-state pmf from fitted rates using (2)	§1.4
$\bar{\alpha} \equiv \bar{\alpha}_k$	directly estimated steady-state pmf	§1.4
$\lambda(t)$	the time-varying arrival rate function, as in (3)	§2
$\bar{\lambda}$	the average of the time-varying arrival rate, as in (3)	§2
β	the relative amplitude of the arrival rate, as in (3)	§2
γ	the frequency of the sinusoidal arrival rate, as in (3)	§2
$c \equiv c(\gamma) \equiv 2\pi/\gamma$	the cycle length of the sinusoidal arrival rate, as in (3)	§2
$\bar{\lambda}_{n,k} \equiv \bar{\lambda}_{n,k}(t)$	estimated birth rate for model with arrival rate n in (4)	§2.2
c_a^2	squared coefficient of variation (scv)	§2.2
b	estimated slope for fitted birth rate in $GI/M/\infty$ in (4)	§2.2
$\bar{\delta}_k \equiv \bar{\lambda}_k - \bar{\mu}_k$	estimated drift in state k from data over $[0,t]$	§2.2
$m(t)$	the mean number of busy servers in $M_t/GI/\infty$, in (5)	§2.3
$s(t) \equiv m(t)/\bar{\lambda}$	the scaled mean number of busy servers, as in (5)	§2.3
$\rho(t) \equiv \lambda(t)/s\mu$	the instantaneous traffic intensity	§2.6
$T(p,q)$	first passage time from the p^{th} to the q^{th} percentile	§2.6
$r(p,q)$	the corresponding rate (reciprocal)	§2.6
$\omega(p,q)$	the speed ratio	§2.6
$T(p) \equiv T(p,p)$	the special case of $T(p,q)$ when $q = p$	§2.6
$r(p) \equiv r(p,p)$	the special case of $r(p,q)$ when $q = p$	§2.6
$\omega(p) \equiv \omega(p,p)$	the speed ratio $\omega(p,q)$ when $q = p$	§2.6
λ^p_k	parametric function for estimating steady-state dist.	§2.8
$\bar{\rho} \equiv \bar{\lambda}E[S]/s$	the long-run average traffic intensity	§3
$\alpha_k(t)$	the dynamic periodic steady-state pmf	§3
α^c_k	the overall steady-state pmf with a periodic arrival rate	§3
(ζ, σ)	parameters of asymptotic geometric decay as in (19)	§3.1
G_e	stationary-excess cdf associated with service-time cdf G	§3.2
$\delta(t)$	the departure rate in an infinite-server model	§3.2
$\delta_k(t)$	the conditional departure rate	§3.2

6 Conclusions

We have continued our study of state-dependent birth-and-death (BD) processes fit to complex queueing systems, begun in [12]. In §1.2 we observed that this provides an alternate fitting procedure for the classical Erlang-A model, which can usefully supplement the standard fitting procedure, and thus provide a statistical test. In §1.4 we reviewed the important property that the steady-state distribution always matches the steady-state distribution of the system from which the data come, discussed in [43]. Thus, the fitted BD model could be applied directly.

However, the main purpose of this paper is to develop diagnostic tools to help determine what stochastic model is appropriate for a complex queueing system. That goal is well illustrated by Figure 1 showing the BD fit to emergency department arrival and departure data in §1.1. The results in the main paper help show that the ED data are roughly consistent with the arrival process being a nonhomogeneous Poisson process (NHPP) with a periodic arrival rate function, but are inconsistent with i.i.d. service times. Extensive simulation show that the fitted death rates are approximately proportional to k in all $M_t/GI/\infty$ models, unlike Figure 1. The need for a time-dependent service-time distribution is consistent with observations in [2, 40, 46].

We conducted extensive simulation experiments to study the impact of fitting general state-dependent birth-and-death (BD) processes to the observed queue length (number in system) in $M_t/GI/s$ models. These models have the sinusoidal arrival rate function in (3) with relative amplitude $\beta = 10/35$. In the experiments we considered arrival rates $\bar{\lambda} = 35$ and 100 (moderately large scale) for a range of scaling factors γ, yielding a range of sine cycles of length $2\pi/\gamma$.

From these experiments, we see that the death rates have the same linear structure as for the many-server $GI/GI/s$ models studied in [12], but we see significantly different fitted birth rates, as can be seen by comparing Figures 2 and 3. Theorems 4.3 and 4.4 establish finite bounds and heavy-traffic limits for the fitted birth rates, consistent with these figures. The simulation results in §§2.3-2.8 indicate that (i) for larger γ (shorter cycles) such as $\gamma \geq 1$, the fitted BD process may serve as a useful direct approximation for the original queue-length process, but (ii) for smaller γ (longer cycles) such as $\gamma \leq 0.1$, the transient behavior of the fitted BD process is very different. However, consistent with the theory in [43], we see that the fitted BD process consistently describes the steady-state distribution. In §2.8 we showed that a relatively simple two-parameter parametric function can be fit to the estimated birth rates in order to efficiently estimate first the fitted birth rate and then the steady-state distribution of the original system. The results here for known stochastic models should help interpret similar fitting to data from complicated service systems, as in [46].

As for [12], it remains to derive explicit formulas and asymptotic approximations for the fitted rates in these models, but we have obtained some analytical results and we have shown how simulation can be used to expose the essential structure of the fitted birth and death rates in important classes of queueing models.

Acknowledgement

This research was begun while the first author was an undergraduate in the IEOR Department at Columbia University. The second author acknowledges support from NSF grants CMMI 1066372 and 1265070. The authors thank Avishai Mandelbaum for facilitating access to his SEE data repository at the Technion in Israel and Xiaopei Zhang for preparing Figure 1 using that data.

References

[1] Abate, J. and Whitt, W. (1999). Computing Laplace transforms for numerical inversion via continued fractions. *INFORMS Journal on Computing* 11(4):394–405.

[2] Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y., Tseytlin, Y. and Yom-Tov, G. (2014). Patient flow in hospitals: a data-based queueing-science perspective. Working paper, New York University, http://www.stern.nyu.edu/om/faculty/armony/.

[3] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2005). Statistical analysis of a telephone call center: a queueing-science perspective. *J Amer Stat Assoc* 100:36–50.

[4] Browne, S. and Whitt, W. (1995). Piecewise-linear diffusion processes. In Dshalalow, J. (ed.), *Advances in Queueing*. Boca Raton, FL: CRC Press, pp. 463–480.

[5] Buzen, J. (1976). Fundamental operational laws of computer system performance. *Acta Informatika* 14:167–182.

[6] Buzen, J. (1978). Operational analysis: an alternative to stochastic modeling. In Ferarri, D. (ed.), *Performance of Computer Installations*. Amsterdam: North Holland, pp. 175–194.

[7] Chang, C. S., Chao, X. L. and Pinedo, M. (1991). Monotonicity results for queues with doubly stochastic Poisson arrivals: Ross's conjecture. *Advances in Applied Probability* 12(41):210–228.

[8] Choudhury, G. L., Mandelbaum, A., Reiman, M. I. and Whitt, W. (1997). Fluid and diffusion limits for queues in slowly changing random environments. *Stochastic Models* 13(1):121–146.

[9] Crescenzo, A. D. and Nobile, A. G. (1995). Diffusion approximation to a queueing system with time-dependent arrival and service rates. *Queueing Systems* 19:41–62.

[10] Davis, J. L., Massey, W. A. and Whitt, W. (1995). Sensitivity to the service-time distribution in the nonstationary Erlang loss model. *Management Sci* 41(6):1107–1116.

[11] Denning, P. J. and Buzen, P. J. (1978). The operational analysis of queueing network models. *Computing Surveys* 10:225–261.

[12] Dong, J. and Whitt, W. (2015). Stochastic grey-box modeling of queueing systems: fitting birth-and-death processes to data. *Queueing Systems* 79:391–426.

[13] Eick, S. G., Massey, W. A. and Whitt, W. (1993). $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci* 39:241–252.

[14] Eick, S. G., Massey, W. A. and Whitt, W. (1993). The physics of the $M_t/G/\infty$ queue. *Oper Res* 41:731–742.

[15] El-Taha, M. and Stidham, S. (1999). *Sample-Path Analysis of Queueing Systems*. Boston: Kluwer.

[16] Falin, G. I. (1989). Periodic queues in heavy traffic. *Advances in Applied Probability* 21:485–487.

[17] Gans, N., Liu, N., Mandelbaum, A., Shen, H. and Ye, H. (2010). Service times in call centers: Agent heterogeneity and learning with some operational consequences. *IMS Collections, Borrowing Strength: Theory Powering Applications A Festschrift for Lawrence D Brown* 6:99–123.

[18] Garnett, O., Mandelbaum, A. and Reiman, M. I. (2002). Designing a call center with impatient customers. *Manufacturing and Service Oper Management* 4(3):208–227.

[19] Goldberg, D. and Whitt, W. (2008). The last departure time from an $M_t/G/\infty$ queue with a terminating arrival process. *Queueing Systems* 58:77–104.

[20] Green, L. V. and Kolesar, P. J. (1991). The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci* 37:84–97.

[21] Green, L. V., Kolesar, P. J. and Whitt, W. (2007). Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper Management* 16:13–29.

[22] Gurvich, I., Huang, J. and Mandelbaum, A. (2014). Excursion-based universal approximatins for the erlang-a queue in steady-state. *Mathematics of Operations Research* 39:325–373.

[23] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29(3):567–588.

[24] Heyman, D. P. and Whitt, W. (1984). The asymptoic behavior of queues with time-varying arrival. *Journal of Applied Probability* 21(1):143–156.

[25] Ibrahim, R., L'Ecuyer, P., Regnard, N. and Shen, H. (2012). On the modeling and forecasting of call center arrivals. *Proceedings of the 2012 Winter Simulation Conference* 2012:256–267.

[26] Kim, S. and Whitt, W. (2014). Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing and Service Oper Management* 16(3):464–480.

[27] Kim, S. and Whitt, W. (2014). Choosing arrival process models for service systems: Tests of a nonhomogeneous Poisson process. *Naval Research Logistics* 17:307–318.

[28] Kim, S., Whitt, W. and Cha, W. C. (2015). A data-driven model of an appointment-generated arrival process at an outpatient clinic. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[29] Kim, S.-H., Vel, P., Whitt, W. and Cha, W. C. (2015). Poisson and non-Poisson properties in appointment-generated arrival processes: the case of an endrocrinology clinic. *Operations Research Letters* 43:247–253.

[30] Liu, Y. and Whitt, W. (2012). A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. *Oper Res Letters* 40:307–312.

[31] Liu, Y. and Whitt, W. (2012). Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper Res* 60(6):1551–1564.

[32] Liu, Y. and Whitt, W. (2014). Stabilizing performance in networks of queues with time-varying arrival rates. *Probability in the Engineering and Informational Sciences* 28:419–449.

[33] Mandelbaum, A., Massey, W. A. and Reiman, M. I. (1998). Strong approximations for Markovian service networks. *Queueing Systems* 30:149–201.

[34] Mandelbaum, A. and Zeltyn, S. (2007). Service engineering in action: The palm/erlang-a queue, with applications to call centers. *Advances in Services Innoovations* 20(1):33–64.

[35] Massey, W. A. and Whitt, W. (1993). Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* 13(1):183–250.

[36] Massey, W. A. and Whitt, W. (1996). Stationary-process approximations for the nonstationary Erlang loss model. *Oper Res* 44(6):976–983.

[37] Puhalskii, A. A. (2013). On the $M_t/M_t/K_t + M_t$ queue in heavy traffic. *Math Methods Oper Res* 78:119–148.

[38] Rolski, T. (1989). Queues with nonstationary inputs. *Queueing Systems* 5:113–130.

[39] Rothkopf, M. H. and Oren, S. S. (1979). A closure approximation for the nonstationary $M/M/s$ queue. *Management Science* 25(6):522–534.

[40] Shi, P., Chou, M. C., Dai, J. G. and Sim, J. (2015). Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Science* articles in advance:doi10.1287/mnsc.2014.2112.

[41] Whitt, W. (1984). Departures from a queue with many busy servers. *Mathematics of Operations Research* 9(4):534–544.

[42] Whitt, W. (1991). The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Science* 37(3):307–314.

[43] Whitt, W. (2012). Fitting birth-and-death queueing models to data. *Statistics and Probability Letters* 82:998–1004.

[44] Whitt, W. (2014). Heavy-traffic limits for queues with periodic arrival processes. *Operations Research Letters* 42:458–461.

[45] Whitt, W. (2014). The steady-state distribution of the $M_t/M/\infty$ queue with a sinusoidal arrival rate function. *Operations Research Letters* 42:311–318.

[46] Whitt, W. and Zhang, X. (2015). A data-generated queueing model of an emergency department. In preparation, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[47] Wolff, R. W. (1965). Problems for statistical inference for birth and death queueing models. *Operations Research* 13:343–357.