Using a Birth-and-Death Process to Estimate the Steady-State Distribution of a Periodic Queue

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

December 7, 2015

Abstract

If the number of customers in a queueing system as a function of time has a proper limiting steady-state distribution, then that steady-state distribution can be estimated from system data by fitting a general stationary birth-and-death process model to the data and solving for its steady-state distribution by using the familiar local-balance steady-state equation for birth-and-death processes, even if the actual process is not a birth-and-death process. We show that this indirect way to estimate the steady-state distribution can be effective for periodic queues, because the fitted birth and death rates often have special structure allowing them to be estimated efficiently by fitting parametric functions with only a few parameters, e.g. 2. We focus on the multi-server $M_t/GI/s$ queue with a nonhomogeneous Poisson arrival process having a periodic time-varying rate function. We establish properties of its steady-state distribution and fitted birth and death rates. We also show that the fitted birth and death rates can be a useful diagnostic tool to see if an $M_t/GI/s$ model is appropriate for a complex queueing system.

Keywords: estimating steady-state distributions; periodic queues; birth-and-death processes; fitting models to data; grey-box stochastic model

1 Introduction

1.1 Steady-State Distributions and Birth-and-Death Processes

Let $\{Q(t) : t \geq 0\}$ be a stochastic processes taking values in the nonnegative integers, such as the number of customers in a queueing system at each time t. The stochastic process $\{Q(t) : t \geq 0\}$ has a proper (limiting) steady-state distribution if

$$\lim_{t\to\infty} P(Q(t) = k) = \alpha_k \quad \text{for all} \quad k \geq 0, \quad \text{where} \quad \sum_{k=0}^{\infty} \alpha_k = 1. \tag{1}$$

Given system data, i.e., a segment $\{Q(s) : 0 \leq s \leq t\}$ of the sample path, a standard way to estimate the steady-state probability vector α is to calculate the proportion of time spent in each state; i.e., if $T_k(t)$ is the total time spent in state k during $[0, t]$, then we estimate α_k by

$$\bar{\alpha}_k(t) = \frac{T_k(t)}{t}, \quad k \geq 0. \tag{2}$$

It may be surprising, but there usually is also another quite different way to estimate the steady-state probability vector α. The alternative way is to fit a general stationary birth-and-death (BD) process to the same sample path in the obvious way, as if the stochastic process $\{Q(t) : t \geq 0\}$ were a BD process, which we are *not assuming*.

In particular, let $A_k(t)$ and $D_k(t)$ be the number of arrivals and departures, respectively, observed in state k over $[0, t]$. State-dependent birth and death rates can be estimated by

$$\bar{\lambda}_k(t) = \frac{A_k(t)}{T_k(t)} \quad \text{and} \quad \bar{\mu}_k(t) = \frac{D_k(t)}{T_k(t)}. \tag{3}$$

This is the natural way to estimate the rates when $\{Q(t) : t \geq 0\}$ is a BD process [3, 20, 37]. In fact, these are the maximum likelihood estimators for these rates.

We then estimate the steady-state distribution by solving the local-balance equations for a BD process; i.e., we let $\bar{\alpha}^e(t) \equiv \{\bar{\alpha}^e_k(t) : k \geq 0\}$ (where $\equiv$ means "equality by definition") be the solution to the equation

$$\bar{\alpha}^e_k(t)\bar{\lambda}_k(t) = \bar{\alpha}^e_{k+1}(t)\bar{\mu}_k(t), \quad k \geq 0, \tag{4}$$

with the additional property that

$$\sum_{k=0}^{\infty} \bar{\alpha}^e_k(t) = 1. \tag{5}$$

We use the superscript e to denote that the vector $\bar{\alpha}^e_k(t)$ is obtained from the estimated BD rates in (3) via (4).

It turns out that the two empirical steady-state probability vectors $\bar{\alpha}(t)$ and $\bar{\alpha}^e(t)$ are intimately related. Indeed, if $Q(0) = Q(t)$, i.e., if the initial state $Q(0)$ coincides with the final state $Q(t)$ of the sample path, then the two probability vectors $\bar{\alpha}(t)$ and $\bar{\alpha}^e(t)$ are *identical*. More generally, they are stochastically ordered; see Theorem 1 of [33]. Moreover, under minor regularity conditions, $\bar{\alpha}^e_k(t)$ is a consistent estimator of α_k, i.e.,

$$\alpha_k \equiv \lim_{t\to\infty} \bar{\alpha}_k(t) = \lim_{t\to\infty} P(Q(t) = k) = \lim_{t\to\infty} \bar{\alpha}^e_k(t); \tag{6}$$

see Chapter 4 of [11] and Corollary 4.1 of [33].

Thus fitting a stationary BD process by (3) and solving the local balance equation (4) is a legitimate way to estimate the steady-state probability vector α, even though we are *not* assuming that the stochastic process $\{Q(t) : t \geq 0\}$ is a BD process. The fitted-BD approach to estimate the steady-state distribution has the advantage that the birth-rate and death-rate functions tend to have more structure, which makes the fitting easier. Thus, the structure in the fitted birth and death rates helps us make reasonable estimates in the tails of the steady-state distribution, where there tend to be relatively few data points.

We previously studied this indirect fitting approach for stationary non-Markovian $GI/GI/s$ models in [7]. We found that the fitted BD not only provides a way to calculate the steady-state distribution but the fitted birth-rate and death-rate functions reveal important structural properties of the original $GI/GI/s$ model; see §6.2 for further discussion.

1.2 Steady-State Distributions for Periodic Queues

In this paper we examine this indirect BD approach for estimating the steady-state distribution of a periodic queue, i.e., for a queue having a periodic arrival-rate function. Since the probabilities $P(Q(t) = k)$ should themselves be periodic functions of t, we first need to formulate the problem carefully. To do so, suppose that the period is c. The stochastic process $\{Q(t) : t \geq 0\}$ has a dynamic steady-state pmf $\alpha(t)$, $0 \leq t < c$ (a family of pmf's indexed by t), and an overall steady-state pmf α^c if the following limits are well defined probability vectors:

$$\begin{aligned}\alpha_k(t) &\equiv \lim_{n\to\infty} P(Q(nc+t) = k) = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} 1_{\{Q(jc+t)=k\}}, \quad 0 \leq t < c, \quad \text{and} \\ \alpha_k^c &\equiv \frac{1}{c}\int_0^c \alpha_k(t)\,dt = \lim_{t\to\infty}\frac{1}{t}\int_0^t 1_{\{Q(s)=k\}}\,ds, \quad k \geq 0. \end{aligned} \quad (7)$$

Moreover, α^c can be regarded as a special case of α in (1) if we randomize the initial state uniformly over the interval $[0, c]$.

Before proceeding, we note that *both* the dynamic steady-state pmf $\alpha(t)$ and the overall steady-state pmf α^c are of practical interest. For example, for a hospital emergency room, we may want to know the likelihood of states such as large queues, both averaged over all time and at fixed times. That is, we may want to know the average congestion or the average use of resources as well as the peak congestion or the peak use of resources.

Thus, in this paper we focus on the BD approach to estimating the steady-state pmf α^c. We do so for the special case of the many-server $M_t/GI/s$ queues with periodic arrival-rate functions, focusing especially on the case of sinusoidal arrival-rate functions, which are often used in studies of staffing algorithms for queues with time-varying arrival rates, e.g. see [16, 25].

The periodic $M_t/GI/s$ queueing models here have a nonhomogeneous Poisson process (NHPP, the M_t) as an arrival process, which is independent of independent and identically distributed (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, in our simulation examples we consider the stylized sinusoidal arrival rate function

$$\lambda(t) \equiv \bar{\lambda}\left(1 + \beta \sin\left(\gamma t\right)\right), \quad (8)$$

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 $c = 2\pi/\gamma$.

1.3 The Indirect Fitting Approach

For these models, we find that the indirect BD fitting approach tends to be more efficient because the fitted birth and death rates tend to have more structure than the steady-state distribution;

e.g., the fitted birth and death rates typically are nondecreasing functions; e.g., see §4. The fitted death rates are often very nearly piecewise-linear with

$$\bar{\mu}_k \equiv \bar{\mu}_k(\infty) \equiv \lim_{t\to\infty} \bar{\mu}_k(t) \approx \mu \min\{k, s\}, \tag{9}$$

where $1/\mu$ is the mean service time, just as for the exact formulas in the corresponding $M/M/s$ model. The fitted birth rates

$$\bar{\lambda}_k \equiv \bar{\lambda}_k(\infty) \equiv \lim_{t\to\infty} \bar{\lambda}_k(t) \tag{10}$$

tend to be approximately linear in the most frequently visited states, but are more complicated outside this region. Nevertheless, we find that they can be fit quite accurately by using a parametric function with only two parameters; see §5. In contrast, the direct steady-state distribution tends to be quite complicated. (Henceforth, we use $\bar{\mu}_k$ and $\bar{\mu}_k(\infty)$ interchangeably when we are concerned with the limiting values or values for very large t, and similarly for other variables.)

For the stationary $M/M/s$ model, the shape of the steady-state probability mass function (pmf) can be understood from heavy-traffic limits for the $M/M/s$ model in [17] and properties of BD and diffusion steady-state distributions discussed in [5]. The steady-state pmf of the number in system is approximately a normal pdf (where all servers are not busy) connected to an exponential pdf (where all servers are busy) joined at s. For associated non-stationary models, the steady-state pmf is even more complicated, e.g., see the lefthand plot in Figure 9.

1.4 Organization

We start in §2 by establishing structural results for the fitted birth and death rates and the steady-state distribution in the $M_t/GI/s$ queue. We have separate subsections for the $M_t/M/s$ and $M_t/GI/\infty$ special cases. Then in §3 we establish additional structural results for the special case of the sinusoidal arrival rate functions in (8).

We follow in §4 by showing the results of simulation experiments investigating what these fitted rates and steady-state distributions look like. In doing so, we confirm and illustrate the theoretical results in §2 and §3. From the special structure of the fitted rates, we see that it should not be difficult to fit parametric functions to the data. In §5 we show that the steady-state distribution of $M_t/M/s$ queues can be efficiently estimated by fitting parametric functions to the fitted birth and death rates; often only two parameters are needed.

Afterwards, in §6 we discuss how the fitted birth and death rates are promising to help diagnose what model is appropriate for a complex queueing system, as suggested in [7]. To illustrate how the present analysis can be used, we show that data from an emergency department are *consistent* with a periodic arrival-rate function (not a surprise), but are *inconsistent* with a sequence of i.i.d. length-of-stay random variables. It is natural to postulate the $M_t/GI/\infty$ model with i.i.d. length-of-stay random variables under the assumption that the length of stay should only depend on the patient's medical condition. The fitted death rates show that the length-of-stay distribution should be regarded as time-varying, which is consistent with the conclusions reached in [2, 29, 36]. In §6.4 we advocate the fitted BD process as a statistical test of the much-used $M_t/M/s + M$ Erlang-A model. We draw conclusions in §7.

2 The Periodic $M_t/GI/s$ Queueing Model

We start by develop supporting theory. 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

$$\bar{\lambda} \equiv \frac{1}{c}\int_0^c \lambda(s)\, ds = \lim_{t\to\infty} \frac{A(t)}{t}. \tag{11}$$

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 for the periodic $M_t/M/s$ model in §3 of [18], because of the NHPP arrival process, the stochastic process $\{Q(nc+t) : n \geq 0\}$ is a regenerative stochastic process for any fixed t, $0 \leq t < c$, with the events $\{Q(nc+t) = 0\}$, $n \geq 1$, being regenerative events. For applications of this regenerative structure, we also require that the interval between such emptiness epochs has finite mean. That was proved for M service in §3 of [18] and for various other service distributions in §6 of [18]. In this paper, *we make the assumption for the general GI service that the interval between such emptiness epochs has finite mean.* Under this assumption, we have a well defined periodic steady-state distribution when $\bar{\rho} < 1$.

Theorem 2.1 (*periodic steady-state distribution*) *If $\bar{\rho} < 1$ in the regenerative periodic $M_t/GI/s$ queueing model, then a dynamic steady-state pmf $\alpha(t)$, $0 \leq t < c$, and an overall steady-state pmf α^c are well defined probability vectors as in (7).*

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

$$\bar{\lambda}_k(\infty) = \frac{\int_0^c \alpha_k(t)\lambda(t)\, dt}{\int_0^c \alpha_k(t)\, dt} = \frac{\int_0^c \alpha_k(t)\lambda(t)\, dt}{c\alpha_k^c} \tag{12}$$

and

$$\bar{\mu}_{k+1}(\infty) = \frac{\alpha_k^c \bar{\lambda}_k(\infty)}{\alpha_{k+1}^c} = \frac{\int_0^c \alpha_k(t)\lambda(t)\, dt}{c\alpha_{k+1}^c}. \tag{13}$$

for $\alpha_k(t)$ and α_k^c in (7).

Proof. We use the regenerative structure to focus on (i) the expected number of arrivals in state k per regenerative cycle divided by the expected length of a regenerative cycle and (ii) the expected time spent in state k per regenerative cycle divided by the expected length of a regenerative cycle. We get (12) involving a single periodic cycle by looking at the ratio. Since the arrival rate depends only on time, we have (12). We then can apply the detailed balance equation in (4) to get (13). ■

Theorems 2.1 and 2.2 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 (4) 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 2.1. Let $\bar{\alpha}^e \equiv \bar{\alpha}^e(\infty)$ be the steady-state probability vector of the fitted BD process obtained from (4). Since $\bar{\alpha}^e$ coincides with α^c in (7), we can use the fitted BD model to calculate the steady-state distribution α^c in (7). To do so, we estimate the birth and death rates and then apply the detailed balance equation in (4). Moreover, by developing analytical approximations for the fitted birth and death rates, we succeed in developing an approximation for α^c.

We can immediately apply Theorem 2.2 to obtain bounds on the fitted birth rates. Since formula (12) expresses $\bar{\lambda}_k(\infty)$ as an average of the arrival rate function over one cycle, we can immediately deduce

Corollary 2.1 (*bounds on the fitted birth rates*) *In the periodic $M_t/GI/s$ queueing model starting empty in the distant past,*

$$\lambda_L \equiv \inf_{0 \le t < c} \lambda(t) \le \bar{\lambda}_k(\infty) \le \sup_{0 \le t < c} \lambda(t) \equiv \lambda_U. \tag{14}$$

2.1 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 [7] also holds for the fitted death rates in the present time-varying case.

Theorem 2.3 (*explicit death rates*) *For the periodic $M_t/M/s$ model with $\bar{\rho} < 1$,*

$$\bar{\mu}_k(\infty) = \min\{k, s\}\mu, \quad k \ge 0, \tag{15}$$

so that

$$\bar{\lambda}_k(\infty) = \frac{\alpha^c_{k+1} \min\{k+1, s\}\mu}{\alpha^c_k}, \quad k \ge 0, \tag{16}$$

for α^c_k in (7).

Proof. As for Theorem 3.1 of [7], (15) follows from the lack of memory property of the exponential distribution. We then apply (4) to get (16). However, we now show that it is also possible to directly apply Theorem 3.1 of [7] 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. [18]. 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 [7] (a). But then we observe that the randomization does not alter the limit (15). ■

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

$$\alpha_k \sim \zeta \sigma^k \quad \text{as} \quad k \to \infty, \tag{17}$$

for positive constants σ and ζ, i.e., if the ratio of the two sides in (17) converges to 1 as $k \to \infty$; see §3.3 of [7].

Theorem 2.4 (*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_k^c in (7) possess a geometric tail as in (17) with the same decay rate as in the associated stationary $M/M/s$ model with arrival rate $\bar{\lambda}$; i.e.,*

$$\alpha_k(t) \sim \zeta_t \sigma_t^k \quad as \quad k \to \infty \quad for\ each \quad t, \quad 0 \le t < c, \tag{18}$$

and

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

where

$$\sigma_c = \sigma_t = \sigma = \rho \equiv \frac{\bar{\lambda}}{s\mu}, \quad \zeta_t \ge \zeta \ge (1-\rho) \quad and \quad \zeta^c \ge \zeta \ge (1-\rho) \tag{20}$$

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,

$$\bar{\lambda}_k(\infty) \to \bar{\lambda} \quad as \quad k \to \infty. \tag{21}$$

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 (17) 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 [6] then implies that $\beta_t \ge \beta$. For the final inequality in (20), 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 (21) follows from Theorem 2.3 and (19), 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.

2.2 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 [9, 10, 27]. If we let the model start empty in the indefinite past with a fixed periodic arrival-rate function, then it can be regarded as in periodic steady-state at time 0. However, to directly show the convergence to a periodic steady state (prove Theorem 2.1) in this case, we assume it starts empty at time 0.

By Theorem 1 of [10], the number in system has a Poisson distribution for each t with mean function $m(t)$, where

$$m(t) = E[S] \int_0^t \lambda(t-s) dG_e(s), \quad t \ge 0, \tag{22}$$

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

$$G_e(t) \equiv P(S_e \le t) \equiv \frac{1}{E[S]} \int_0^t (1 - G(s))\, ds, \quad t \ge 0. \tag{23}$$

Moreover, the departure process in the $M_t/GI/\infty$ model is a Poisson process with rate function where

$$\delta(t) = \int_0^t \lambda(t-s) dG(s), \quad t \geq 0. \tag{24}$$

For a periodic arrival-rate function with period c, we have

$$m(nc+t) = E[S] \int_0^{nc+t} \lambda(nc+t-s) dG_e(s) = E[S] \int_0^{nc+t} \lambda(c+t-s) dG_e(s) \tag{25}$$

and

$$\delta(nc+t) = \int_0^{nc+t} \lambda(nc+t-s) dG(s) = \int_0^{nc+t} \lambda(c+t-s) dG(s), \quad t \geq 0, \tag{26}$$

Because these integrals are nondecreasing functions of n, they converge to limits as $n \to \infty$, thus directly proving Theorem 2.1. The periodic steady state is obtained directly by starting empty in the indefinite past. As stated in Theorem 1 of [10], the number in system has a Poisson distribution for each t with periodic mean function $m(t)$, with the same period c, where

$$m(t) = E[\lambda(t-S_e)]E[S] = E[S] \int_0^{\infty} \lambda(t-s) dG_e(s), \quad t \geq 0, \tag{27}$$

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

$$\delta(t) = E[\lambda(t-S)] = \int_0^{\infty} \lambda(t-s) dG(s), \quad t \geq 0. \tag{28}$$

For the special case of a sinusoidal arrival rate function, an explicit expression for $m(t)$ is given in Theorem 4.1 of [9].

As a consequence, we have the following corollary.

Corollary 2.2 (*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*

$$\alpha_k(t) = \pi_k(m(t)), \quad 0 \leq t < c, \quad \textit{and} \quad \alpha_k^c = \frac{1}{c} \int_0^c \pi_k(m(t))\, dt, \tag{29}$$

for $m(t)$ in (27), 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 [14], 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 2.5 (*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 2.5 to obtain the following general result about the fitted death rates.

Theorem 2.6 (*fitted death rates*) *Consider the $M_t/GI/\infty$ queue with a periodic arrival rate function in the setting of Theorem 2.5. Conditional on $Q(t) = k$, the departure rate at time t is*

$$\delta_k(t) = k\delta_1(t) = kg_{k,t}(0) = \frac{k\mu E[\lambda(t-S)]}{E[\lambda(t-S_e)]} = \frac{k\delta(t)}{m(t)}. \tag{30}$$

Hence, paralleling the fitted birth rate in (12),

$$\bar{\mu}_k(\infty) = \frac{\int_0^c \alpha_k(t)\delta_k(t)\,dt}{c\alpha_k^c} = \frac{k\int_0^c \alpha_k(t)(\delta(t)/m(t))\,dt}{c\alpha_k^c}, \quad k \geq 1, \tag{31}$$

where $\alpha_k(t)$, α_k^c, $m(t)$ and $\delta(t)$ are given in (29), (27) and (28).

Proof. First, we get (30) directly from Theorem 2.5 and formulas (27) and (28). The first term in (31) can be taken as a definition. Then we apply (30). ∎

Paralleling Corollary 2.1, Theorem 2.6 implies bounds for the fitted death rates.

Corollary 2.3 (*bounds on the fitted death rates*) *In the periodic $M_t/GI/\infty$ queueing model starting empty in the distant past,*

$$\mu_L \equiv \inf_{0\leq t<c}\{\delta(t)/m(t)\} \leq \frac{\bar{\mu}_k(\infty)}{k} \leq \sup_{0\leq t<c}\{\delta(t)/m(t)\} \equiv \mu_U. \tag{32}$$

for $m(t)$ in (27) and $\delta(t)$ in (28).

Proof. Theorem 2.6 expresses $\bar{\mu}_k(\infty)/k$ as an average of $\delta(t)/m(t)$ over one cycle. ∎

We get equality of the upper and lower bounds in (32), recovering (15) for $s = \infty$, if S is exponential, because then $\delta(t) = m(t)\mu$ since S_e is distributed the same as S. Indeed, we also get an associated negative result, because it is known that S_e is distributed the same as S if and only if S is exponential. Thus, we have the following consequence.

Corollary 2.4 (*direct proportionality of the fitted death rates*) *In the periodic $M_t/GI/\infty$ queueing model with service-time S having mean $E[S] = 1/\mu$,*

$$\bar{\mu}_k = k\mu \quad \textit{for all} \quad k \geq 0 \tag{33}$$

and for all periodic arrival-rate functions λ if and only if S is exponential.

We now apply Theorem 2.5 to deduce a rate conservation property for this $M_t/GI/\infty$ model in each state over a periodic cycle.

Theorem 2.7 (*arrival and departure rates over a cycle*) *For the periodic $M_t/GI/\infty$ queueing model starting empty in the distant past,*

$$\int_0^c \alpha_k(t)\lambda(t)\,dt = \int_0^c \alpha_k(t)\delta(t)\,dt \quad \textit{for each} \quad k \geq 0 \tag{34}$$

for $\alpha_k(t)$ in (29), so that

$$\int_0^c \lambda(t)\,dt = \int_0^c \delta(t)\,dt. \tag{35}$$

Proof. Since the arrival rate at time t is $\lambda(t)$, independent of the state k, we can apply first (4) and then (30) to obtain

$$\begin{aligned}\int_0^c \alpha_k(t)\lambda(t)\,dt &= c\alpha_k^c\bar{\lambda}_k(\infty) = c\alpha_{k+1}^c\bar{\mu}_{k+1}(\infty) = \int_0^c \alpha_{k+1}(t)\delta_{k+1}(t)\,dt \\ &= \int_0^c \alpha_{k+1}(t)(k+1)[\delta(t)/m(t)]\,dt = \int_0^c \alpha_k(t)\delta(t)\,dt, \end{aligned} \tag{36}$$

as in (34). We add over k to get (35). ∎

3 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 (8). For this model we draw on previous results established in [9].

3.1 A General Service-Time Distribution

We can apply Corollary 2.1 to obtain explicit bounds on the fitted birth rates.

Corollary 3.1 (*bounds on the fitted birth rates*) *For the $M_t/GI/\infty$ model with sinusoidal arrival rate function in* (8) *having $0 < \beta < 1$, starting empty in the distant past,*

$$0 < (1-\beta) \le \frac{\bar{\lambda}_k(\infty)}{\bar{\lambda}} \le (1+\beta) < 2 \quad \textit{for all} \quad k \ge 0. \tag{37}$$

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 [30].

Theorem 3.1 (*short cycles*) *For the $M_t/GI/\infty$ model with sinusoidal arrival rate function in* (8),

$$\bar{\lambda}_k(\infty;\gamma) \to \bar{\lambda} \quad \textit{and} \quad \bar{\mu}_{k+1}(\infty;\gamma) \to (k+1)\mu \quad \textit{as} \quad \gamma \uparrow \infty \quad \textit{for all} \quad k \ge 0. \tag{38}$$

Proof. First, it is helpful to rewrite (12) so that the integrals are over a fixed interval, independent of γ. By making a change of variables $s = \gamma t$, we obtain

$$\bar{\lambda}_k(\infty;\gamma) = \frac{\int_0^{2\pi/\gamma} \alpha_k(t)\lambda(t)\,dt}{\int_0^{2\pi/\gamma} \alpha_k(t)\,dt} = \frac{\int_0^{2\pi} \alpha_k(s/\gamma)\lambda(s/\gamma)\,ds}{\int_0^{2\pi} \alpha_k(s/\gamma)\,ds} \tag{39}$$

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 [9], $m(t;\gamma) \to \bar{\lambda}/\mu$ 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 (4) to write

$$\bar{\mu}_{k+1}(\infty;\gamma) = \frac{\bar{\lambda}_k(\infty;\gamma)\alpha_{k;\gamma}^c}{\alpha_{k+1;\gamma}^c} \to \frac{\bar{\lambda}\alpha_{k;\infty}^c}{\alpha_{k+1;\infty}^c} = (k+1)\mu \quad \text{as} \quad \gamma \uparrow \infty, \tag{40}$$

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 [31]

Theorem 3.2 *(long cycles) For the $M_t/GI/\infty$ model with sinusoidal arrival rate function in (8),*

$$\bar{\lambda}_k(\infty;\gamma) \to \frac{(k+1)\mu\alpha^c_{k+1;0}}{\alpha^c_{k;0}} \quad \textit{and} \quad \bar{\mu}_{k+1}(\infty;\gamma) \to (k+1)\mu \quad \textit{as} \quad \gamma \downarrow 0 \tag{41}$$

for all $k \geq 0$, 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 $\lambda_1(t) = 1 + \beta\sin(t)$, $0 \leq t \leq 2\pi$.

Proof. By Theorem 4.4 of [9], $m(t/\gamma) \to \lambda(t)/\mu$ 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 (39), 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. ■

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

In this section we determine tight bounds for the fitted birth rates for the $M_t/M/\infty$ model with sinusoidal arrival rate. Tightness is verified by showing that these bounds are approached in the heavy-traffic limit.

As shown in [9], the $M_t/M/\infty$ model with sinusoidal arrival rate function in (8) is especially tractable. From (15) of [9], the number in system, $Q(t)$, has a Poisson distribution for each t with mean

$$m(t) \equiv E[Q(t)] = \bar{\lambda}(1 + s(t)), \quad s(t) = \frac{\beta}{1+\gamma^2}\left(\sin(\gamma t) - \gamma\cos(\gamma t)\right). \tag{42}$$

Moreover,

$$s^U \equiv \sup_{t \geq 0} s(t) = \frac{\beta}{\sqrt{1+\gamma^2}} \tag{43}$$

and

$$s(t^m_0) = 0 \quad \text{and} \quad \dot{s}(t^m_0) > 0 \quad \text{for} \quad t^m_0 = \frac{\cot^{-1}(1/\gamma)}{\gamma}. \tag{44}$$

The function $s(t)$ increases from 0 at time t^m_0 to its maximum value $s^U = \beta/\sqrt{1+\gamma^2}$ at time $t^m_0 + \pi/(2\gamma)$. The interval $[t^m_0, t^m_0 + \pi/(2\gamma)]$ 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,

$$P(Z = k) = \frac{\gamma}{2\pi}\int_0^{2\pi/\gamma} P(Q(t) = k)\,dt, \quad k \geq 0, \tag{45}$$

The moments of Z are given by the corresponding mixture

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

so that $E[Z] = \bar{\lambda}$. For more details, see [34].

We use (42) to improve the bounds in Corollary 3.1 and obtain a simple proof of Theorem 3.1 in this case.

Theorem 3.3 (*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* (8), *starting empty in the distant past,*

$$\bar{\lambda}\left(1 - \frac{\beta}{\sqrt{1+\gamma^2}}\right) \le \bar{\lambda}_k(\infty) \le \bar{\lambda}\left(1 + \frac{\beta}{\sqrt{1+\gamma^2}}\right) \quad \textit{for all} \quad k \ge 0. \tag{46}$$

and

$$\bar{\lambda}_k(\infty) \to \bar{\lambda} \quad \textit{as} \quad \gamma \to \infty \quad \textit{for all} \quad k \ge 0. \tag{47}$$

Proof. We apply (4) to obtain the expression

$$\frac{(k+1)\bar{\lambda}_k(\infty)}{\bar{\mu}_{k+1}(\infty)} = \frac{\alpha^c_{k+1}}{\alpha^c_k}, \quad k \ge 0. \tag{48}$$

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

$$\bar{\lambda}_k(\infty) = \frac{(k+1)\mu\alpha^c_{k+1}}{\alpha^c_k}, \quad k \ge 0. \tag{49}$$

Since the integrand in the integral representation of α^c_{k+1} in (29) 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 [9] to obtain (46). Clearly, (47) follows from (46). ∎

We conclude by deriving a heavy-traffic limit showing that the lower and upper bounds established in Theorem 3.3 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 in the first limit in (50) 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 3.4 (*heavy-traffic limits*) *In the $M_t/M/\infty$ IS queueing model with periodic arrival rate function, starting empty in the distant past,*

$$\begin{aligned}\frac{\bar{\lambda}_k(\infty)}{\bar{\lambda}} &\to 1 - \frac{\beta}{\sqrt{1+\gamma^2}} \quad \textit{as} \quad \bar{\lambda} \to \infty \quad \textit{and} \\ \frac{\bar{\lambda}_{\lfloor m\bar{\lambda}\rfloor + k}(\infty)}{\bar{\lambda}} &\to 1 + \frac{\beta}{\sqrt{1+\gamma^2}} \quad \textit{as} \quad \bar{\lambda} \to \infty \quad \textit{for} \quad m > 1/\log_e 2 \approx 1.44.\end{aligned} \tag{50}$$

Proof. We expand (49), writing

$$\bar{\lambda}_k(\infty) = \frac{\alpha^c_{k+1}(k+1)\mu}{\alpha^c_k} = \frac{\mu \int_0^c e^{-m(t)} m(t)^{k+1}\, dt}{\int_0^c e^{-m(t)} m(t)^k\, dt} \tag{51}$$

In each case of (50), we apply Laplace's method to the numerator and denominator of (51), 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 (42), 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 [9]. (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 (51) 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 [9]. (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 (51) again approaches $1 + s(x_0)$. ■

4 Simulation Experiments

We now investigate whtat the fitted birth rates, death rates and steady-state distributions look like in these $M_t/GI/s$ models by conducting simulation experiments. In doing so, we illustrate and confirm the theoretical results in §2 and §3.

4.1 Designing the Simulation Experiments

Our base model is the $M_t/M/\infty$ model with the sinusoidal arrival-rate function in (8), which is the special case of the $M_t/GI/s$ model in which $s = \infty$, S has an exponential distribution and $\beta = 10/35$. 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.

4.2 The $M_t/M/s$ Models

Figures 1 and 2 show the estimated birth and death rates for the $M_t/M/s$ models with $s = \infty$ and $s = 40$, respectively.

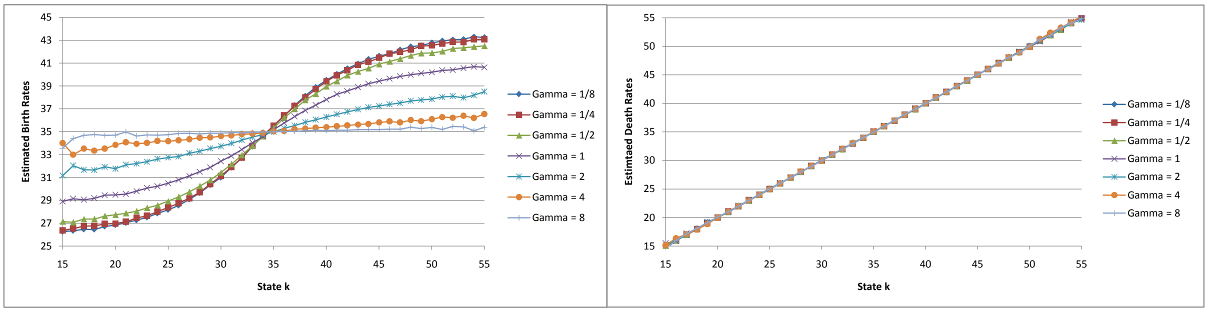

Figure 1: Fitted birth rates (left) and fitted death rates (right) for the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 35$ and $\beta\bar{\lambda} = 10$ and 7 values of γ ranging from 1/8 to 8.

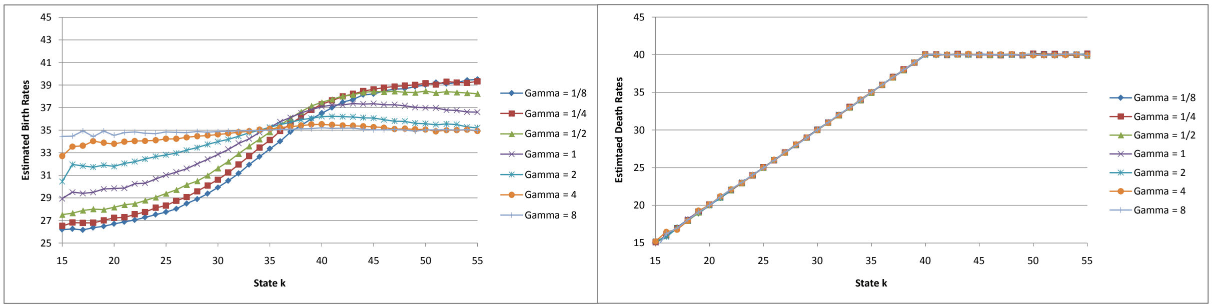

Figure 2: Fitted birth rates (left) and fitted death rates (right) for the $M_t/M/40$ queue with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 35$ and $\beta\bar{\lambda} = 10$ and 7 values of γ ranging from 1/8 to 8.

The estimated BD rates yield corresponding estimates of the steady-state distribution by solving the local balance equation (4). The estimated steady-state distributions for the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 35$ and $\beta = 10/35$ for different ranges of γ are shown in Figure 3. 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 $1/8 \le \gamma \le 8$ for these parameters $\bar{\lambda} = 35$ and $\beta = 10/35$, with the limits serving as effective approximations outside this interval.

4.3 Different Service Distributions: Near Insensitivity

We have also conducted corresponding simulation experiments for $M_t/GI/s$ models with non-exponential service-time distributions. Figures 4 and 5 show the fitted rates for the $M_t/GI/\infty$ model with H_2 and E_2 service distributions with scv $c^2 = 2$ and $c^2 = 1/2$, respectively, just as in §2 of [7]. Figure 6 shows the associated steady-state mass functions for H_2 and E_2 service times, which also look similar.

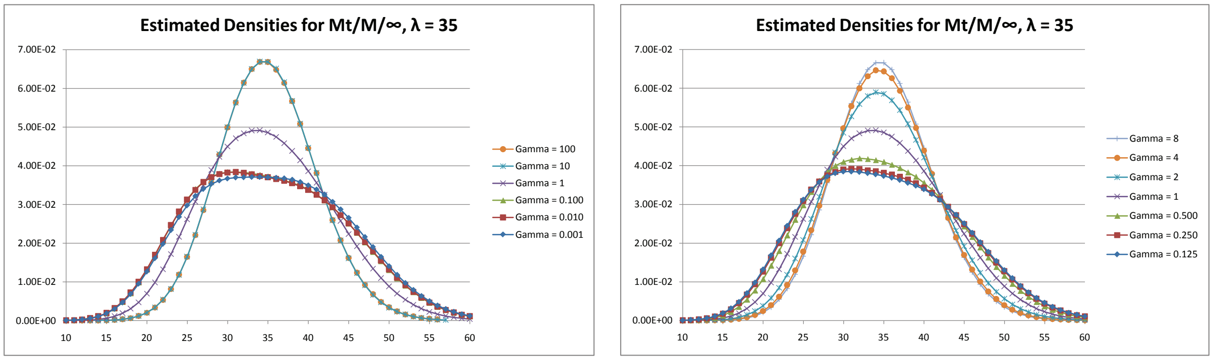

Figure 3: the estimated steady-state number in the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 35$ and $\beta = 10/35$ for different ranges of γ.

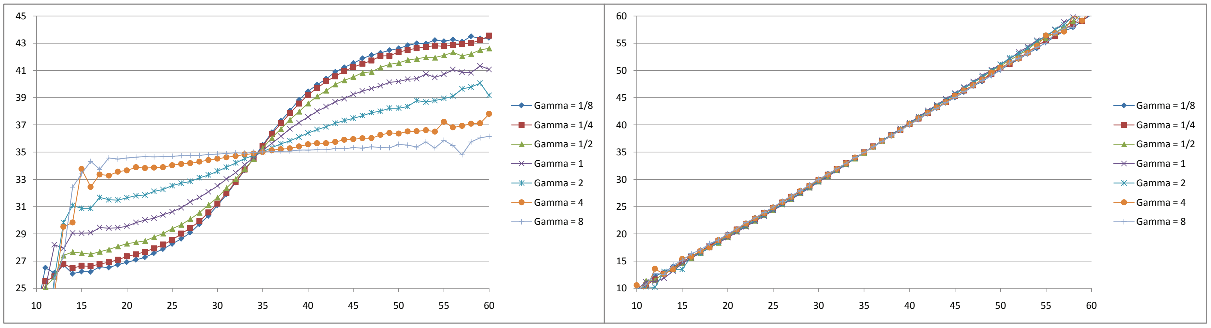

Figure 4: Fitted birth rates (left) and fitted death rates (right) for the $M_t/H_2/\infty$ model with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 35$ and $\beta\bar{\lambda} = 10$ and 7 values of γ ranging from 1/8 to 8. (The service scv is $c^2 = 2$.)

Indeed, the agreement is so good for these infinite-server models that it is natural to wonder whether the fitted birth rate, fitted death rate and steady-state pmf with $s = \infty$ 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. We do see that the insensitivity property does hold asymptotically as $\gamma \downarrow 0$ for $s = \infty$, which is to be expected. In that limit the PSA approximation is valid [32], 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 $\lambda(t)$.

Figures 7-9 illustrate the full range of possibilities by showing the estimated birth rates, death rates and steady-state pmf's for the $M_t/GI/\infty$ model with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 100$ and $\beta = 10/35$ and 4 different service-time distributions for $\gamma = 1/2$ (left) and $\gamma = 2$ (right). For $\gamma = 1/2$, we again see near-insensitivity, but for $\gamma = 2$, we see significant dependence upon the service-time distribution. Further study shows that the difficulty primarily arises for large values of γ, which are not so common in applications. We also see that changing only the service-time pmf can affect the estimated birth rates as well as the estimated death rates, and then the final steady-state pmf. The interesting bimodal form of the steady-state pmf in some cases is investigated in [34]. The two modes roughly correspond to the two extremes of the pmf; the process $Q(t)$ tends to spend more time near these extremes.

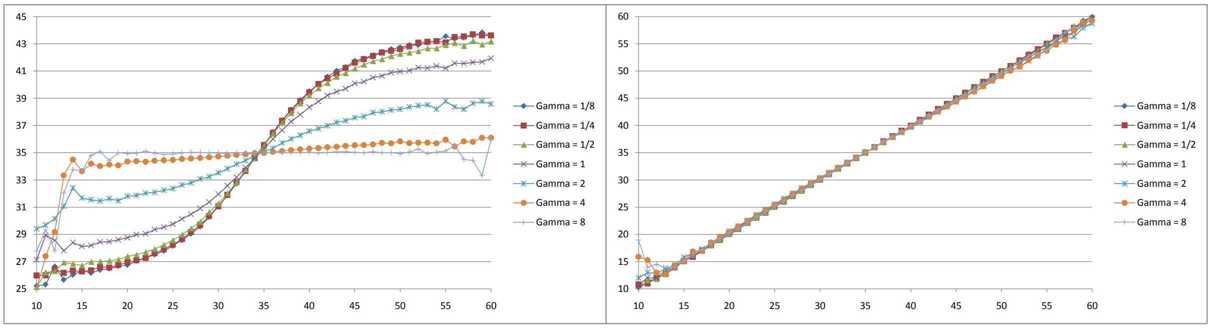

Figure 5: Fitted birth rates (left) and fitted death rates (right) for the $M_t/E_2/\infty$ model with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 35$ and $\beta\bar{\lambda} = 10$ and 7 values of γ ranging from 1/8 to 8.

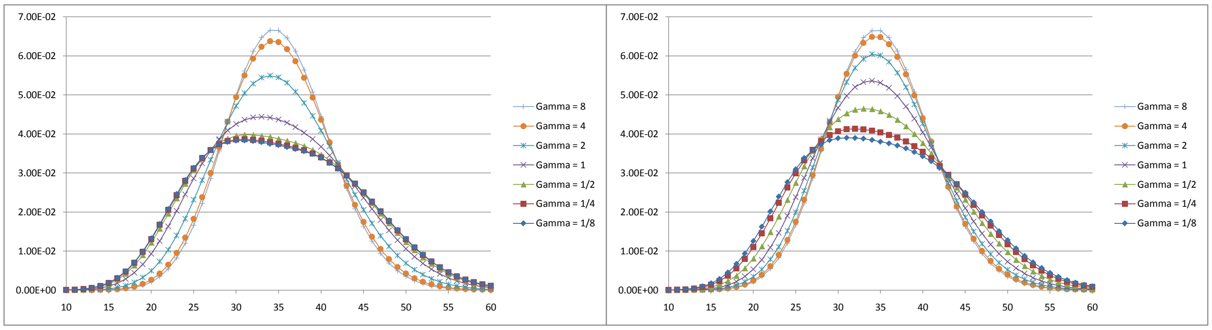

Figure 6: Fitted steady-state mass functions for the $M_t/H_2/\infty$ model (left) and the $M_t/E_2/\infty$ model (right) for with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 35$ and $\beta\bar{\lambda} = 10$ and 7 values of γ ranging from 1/8 to 8.

4.4 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 10, 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 (8) 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.

However, the sample paths are not always so strikingly different. Indeed, the sample paths get less different as γ increases. Figures 11 and 12 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 11 and 12, 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 [23, 24] and references therein.

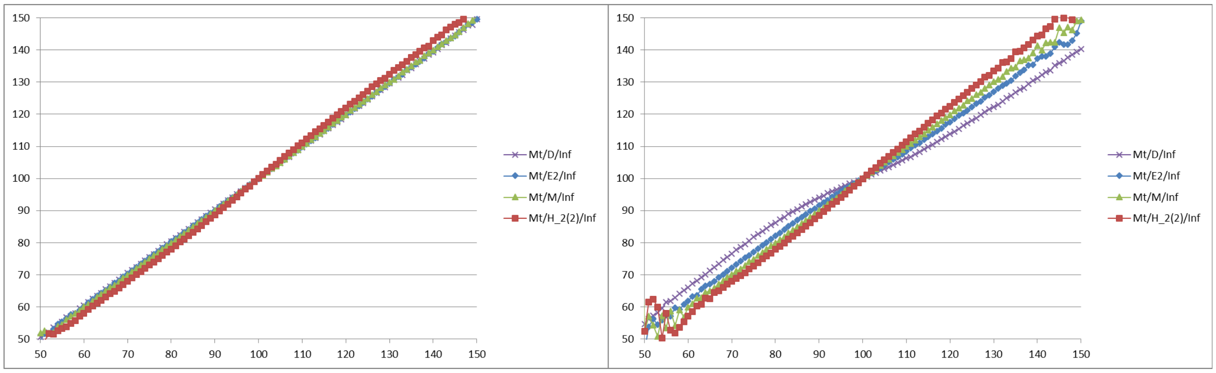

Figure 7: Estimated death rates for the $M_t/GI/\infty$ model with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 100$ and $\beta = 10/35$ and 4 different service-time distributions for $\gamma = 1/2$ (left) and $\gamma = 2$ (right).

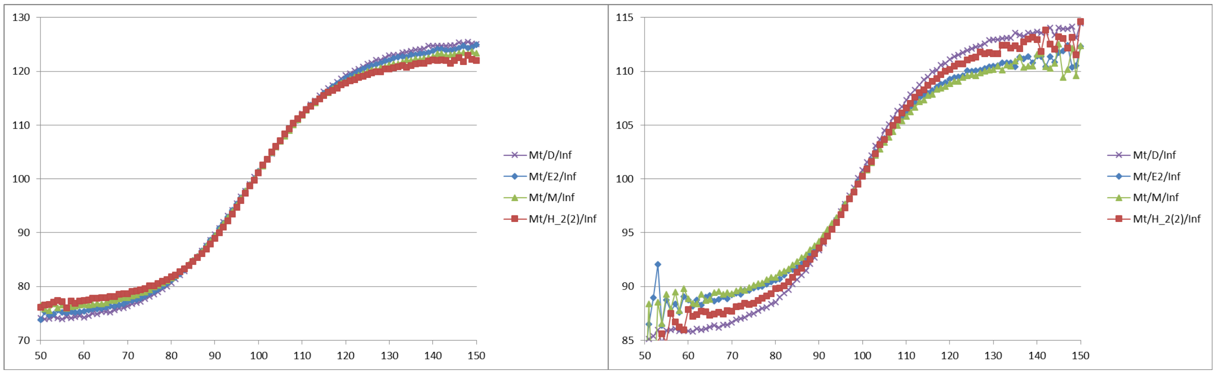

Figure 8: Estimated birth rates for the $M_t/GI/\infty$ model with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 100$ and $\beta = 10/35$ and 4 different service-time distributions for $\gamma = 1/2$ (left) and $\gamma = 2$ (right).

4.5 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 [30] and references therein; [35]. As a consequence, 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 1.

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 [15], which is asymptotically correct for the $M_t/M/\infty$ model as $\gamma \downarrow 0$. That follows from Theorem 1 of [31]. 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

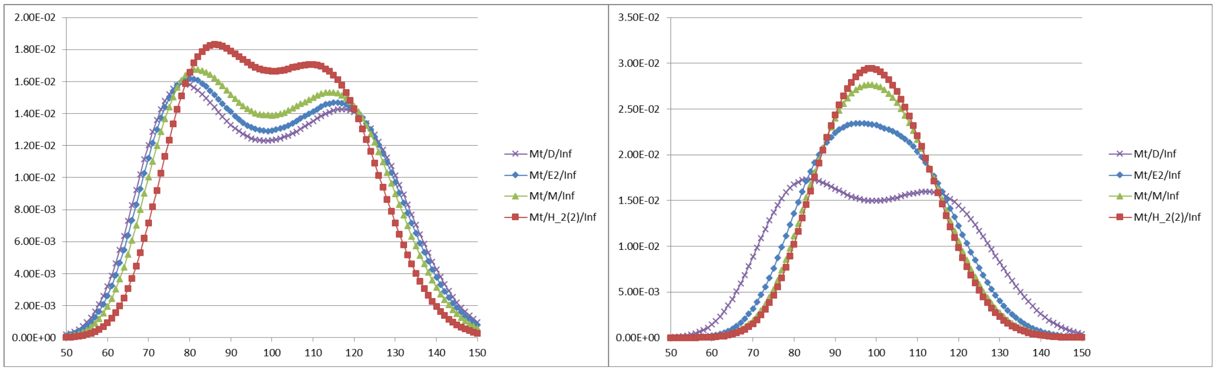

Figure 9: Estimated steady-state pmf's for the $M_t/GI/\infty$ model with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 100$ and $\beta = 10/35$ and 4 different service-time distributions for $\gamma = 1/2$ (left) and $\gamma = 2$ (right).

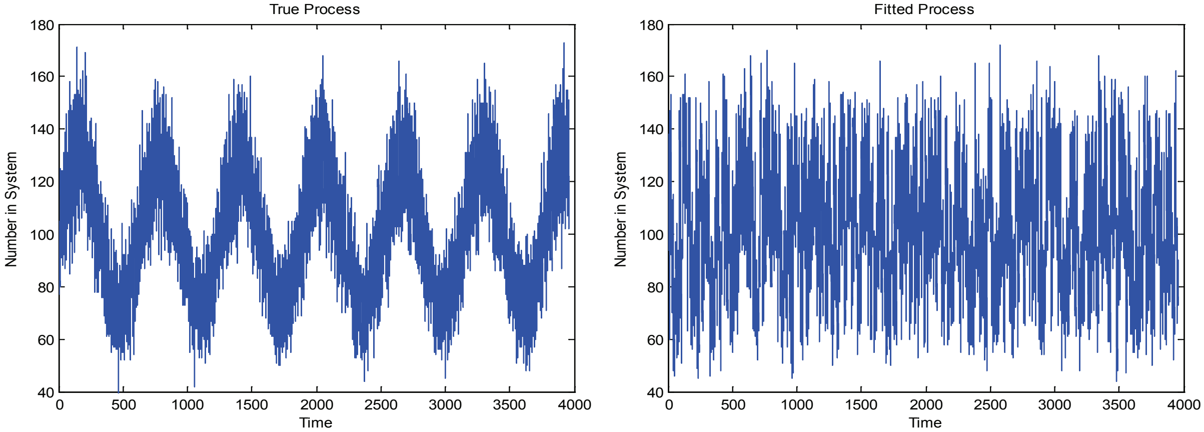

Figure 10: sample paths of the number in system for the original process (left) and the fitted BD process (right) for the $M_t/M/\infty$ queue with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 100$ and $\beta = 10/35$ and $\gamma = 0.01$.

at each time t for t in a sinusoidal cycle, and that is precisely what we see as γ increases in Figure 1. 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 (45), we see that

$$P(Z_0 = k) = \frac{1}{2\pi} \int_0^{2\pi} P(Q_0(t) = k)\, dt, \quad k \geq 0, \tag{52}$$

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 13 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 (8) 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.

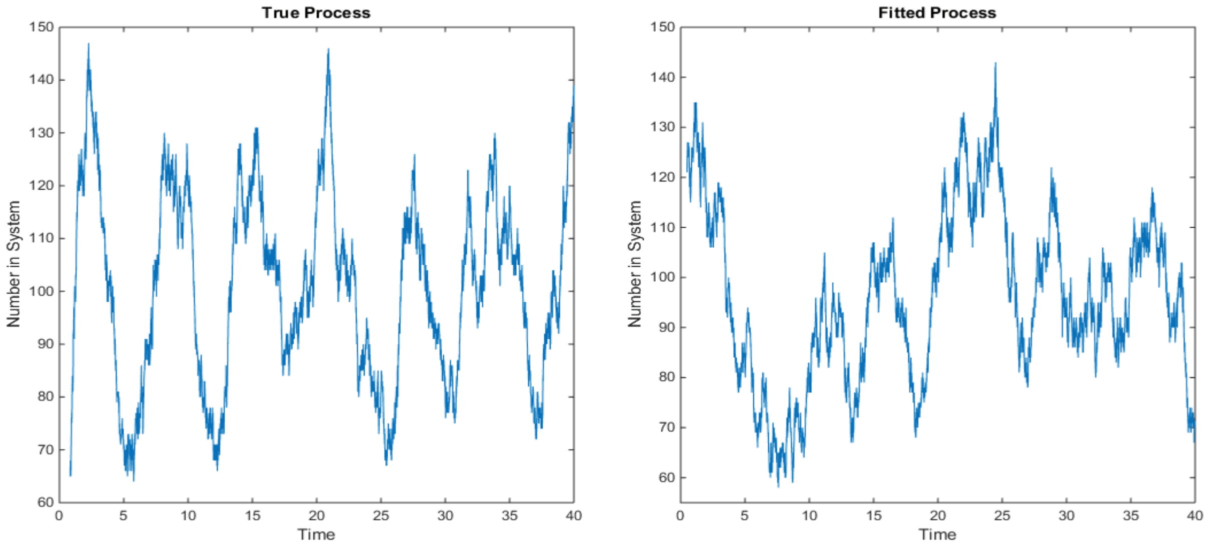

Figure 11: sample paths of the number in system for the original process (left) and the fitted BD process (right) for the $M_t/M/\infty$ queue with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 100$ and $\beta = 10/35$ and $\gamma = 1.0$.

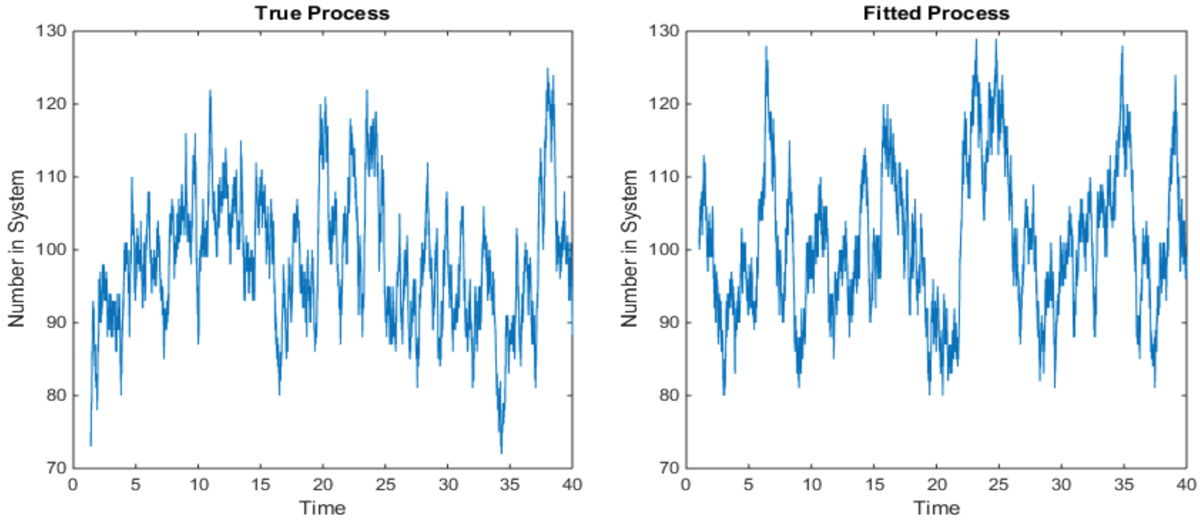

Figure 12: sample paths of the number in system for the original process (left) and the fitted BD process (right) for the $M_t/M/\infty$ queue with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 100$ and $\beta = 10/35$ and $\gamma = 10$.

5 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 equation (4). We illustrate what should be possible in applications by considering the base $M_t/M/\infty$ IS model with a sinusoidal arrival-rate function. Figures 1-6 in §4.2 show what to expect. In applications it is likely to be appropriate to use different fitting functions, but the general approach should be the same.

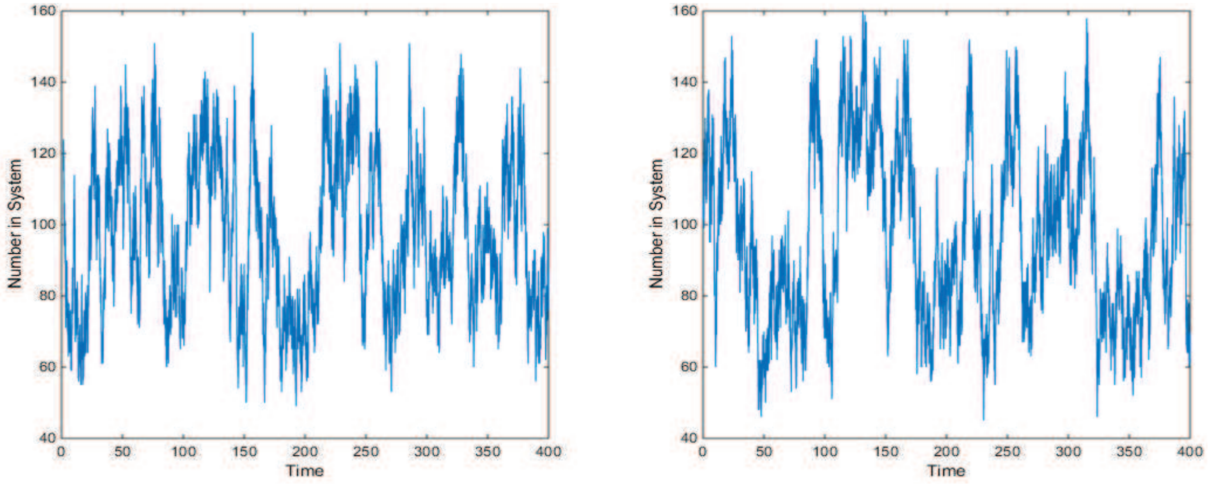

Figure 13: sample paths of the number in system for the fitted BD process in the $M_t/M/\infty$ queue with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 100$ and $\beta = 10/35$ and $\gamma = 0.1$ (left) and $\gamma = 0.01$ (right).

5.1 The Base Infinite-Server Model with Sinusoidal Arrival Rate Function

First, for $M_t/GI/\infty$ IS model with $E[S] = 1$, we do not need to consider the death rates, because we have $\bar{\mu}_k \approx k$ throughout; for M service, we have equality by Theorem 2.3. Hence, we concentrate on the birth rates, initially using the same large sample as before. For larger values of γ, a linear function works well for the fitted birth rate, but not for smaller values of γ. As our parametric function, we choose

$$\lambda_k^p = a \arctan b(k - c) + d, \tag{53}$$

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 14, 15 and 16 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,

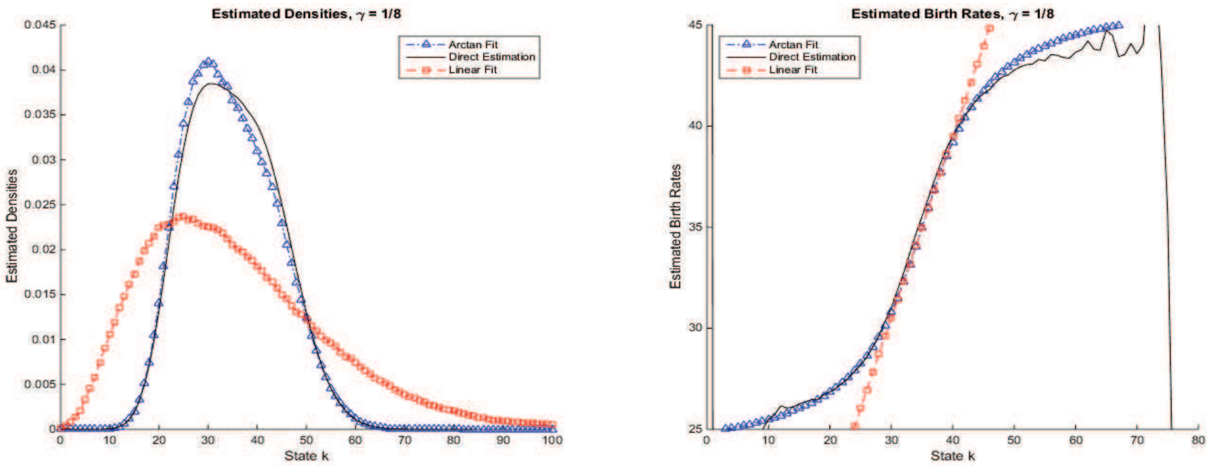

Figure 14: Fitted mass function (left) and birth rates (right) for the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 35$, $\beta\bar{\lambda} = 10$ and $\gamma = 0.125$

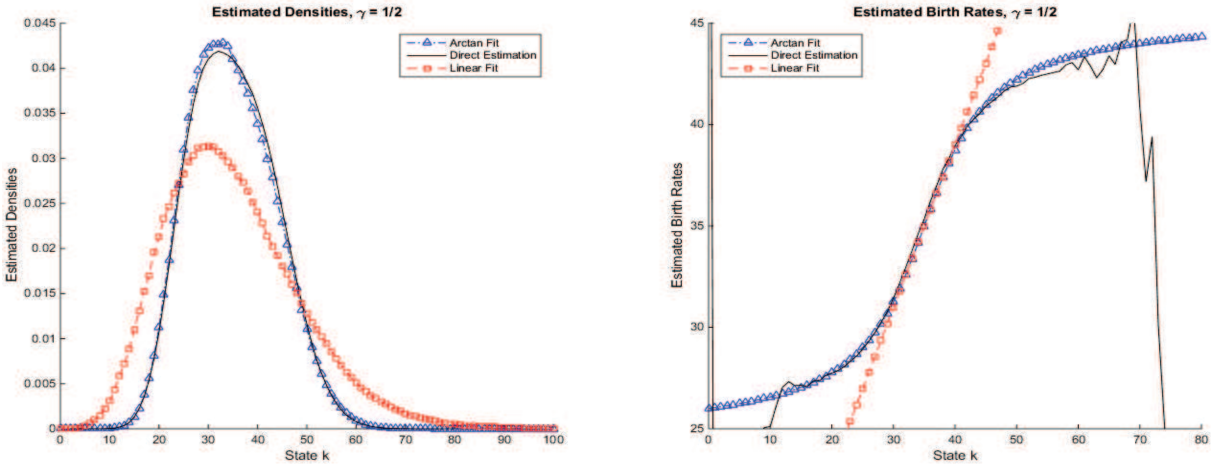

Figure 15: Fitted mass function (left) and birth rates (right) for the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 35$, $\beta\bar{\lambda} = 10$ and $\gamma = 0.5$

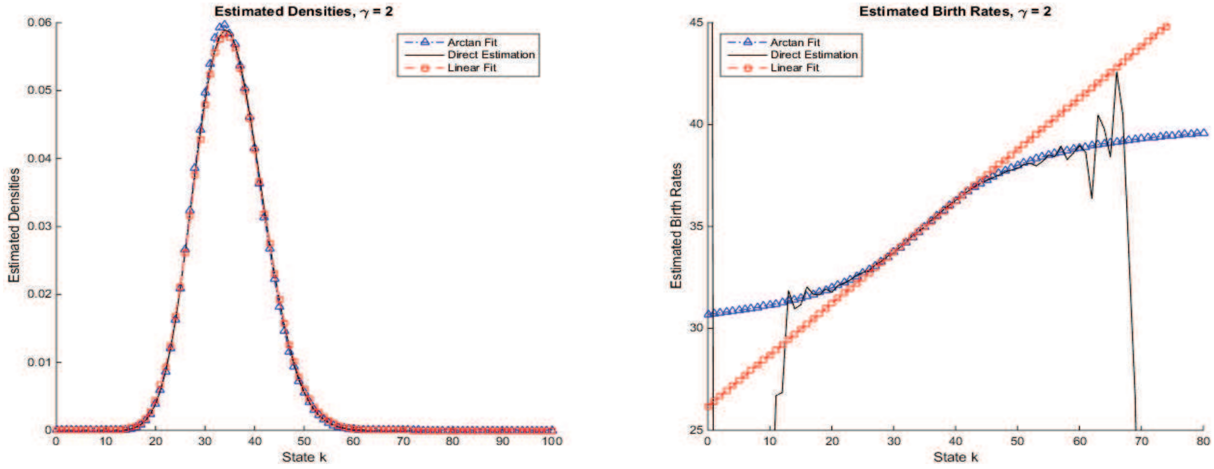

Figure 16: Fitted mass function (left) and fitted birth rates (right) for the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 35$, $\beta\bar{\lambda} = 10$ and $\gamma = 2.0$

which fits by least squares. The figures show that the special arctangent function in (53) 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. The regular structure of the birth and death rates make it possible to obtain relatively good (at least reasonable) estimates of the steady-state in distribution in the tails where there tend to be few data points.

We make some further comments about Figures 14-16. As $\gamma \to \infty$, the cycles become very short, so that the arrival-rate function oscillates very rapidly, but the cumulative arrival rate function approaches the function $\bar{\lambda}t$, and so the arrival process approaches a stationary Poisson process as observed in §4.5. Accordingly, the steady-state distribution is asymptotically Poisson, which is approximately normal. We see that already in Figure 16 for $\gamma = 2$. The distributions with smaller γ (longer cycles) are the average of Poisson distributions, but Figure 14 show that they are more complicated; see [34].

5.2 Small Samples

To illustrate the advantages of the arctan BD fit, we also conducted experiments evaluating the estimators with smaller sample sizes. Figures 17 and 18 illustrate for the model with $\gamma = 0.125$ in Figure 14, but with the sample size greatly reduced.

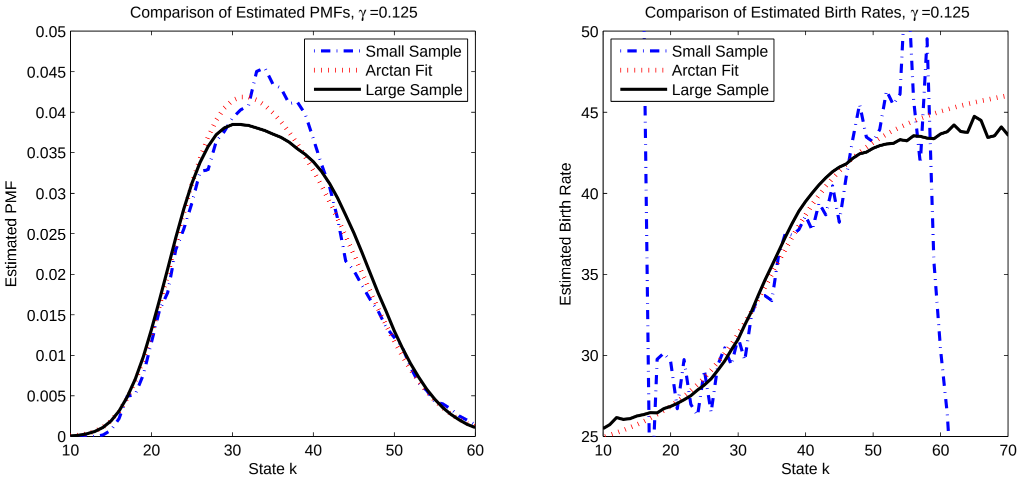

Figure 17: A comparison of the arctan BD fit to a small-sample direct fit of the steady-state probability mass function (left) and birth rates (right) for the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (8) having parameters $\bar{\lambda} = 35$, $\beta\bar{\lambda} = 10$ and $\gamma = 0.125$. Data are collected from 10 cycles as indicated. The previous large-sample fit is used to represent the "true value."

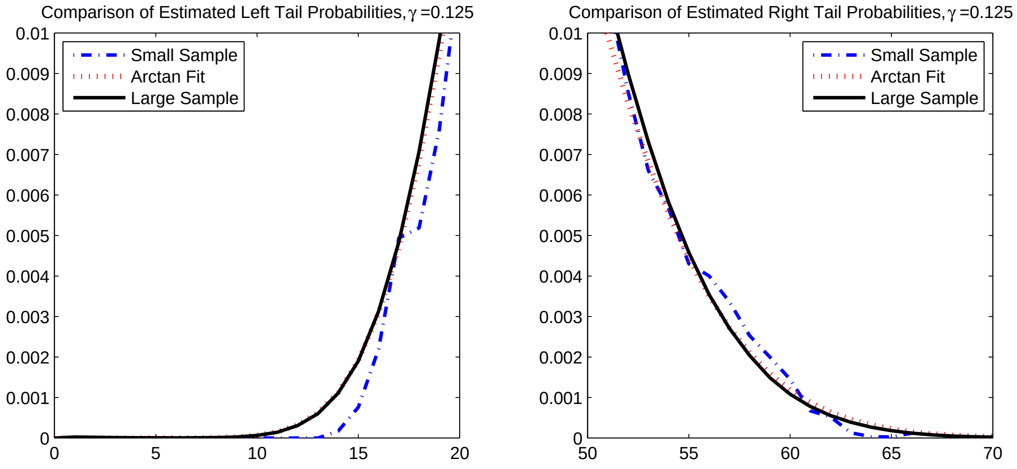

Figure 18: A comparison of the arctan BD fit to a small-sample direct fit of the steady-state tail probabilities (left) and (right) for the $M_t/M/\infty$ model in Figure 17. The large-sample fit is used to represent the "true value."

Now we consider a smaller experiment with about 25,000 potential arrivals instead of the 45 million potential arrivals (resulting in about 35 million actual arrivals) in our previous experiments. In particular, we simulate 10 replications of the model over a full cycle. Since $\gamma = 1/8$, a cycle length is $2\pi/\gamma \approx 50$. Since the average arrival rate is $\bar{\lambda} = 35$, there are about $35 \times 50 = 1750$ arrivals per cycle.

We first take measures to ensure that the system starts approximately in dynamic periodic steady state. To do that, we extract our data from a longer run that has already reached steady state, which for our $\gamma = 1/8$ and all γ not too large, is in the first cycle, after starting empty. we start the data collection at the first time t such that $Q(t) = 15$ after a time t_0 for which $Q(t_0) = 45$ and we terminate the first time that condition is repeated, which invariably is approximately a full cycle after the data collection began. (We remark that the sampling procedure is similar to the way that the speed ratios were estimated in [7].)

That procedure satisfies three requirements: (i) the system starts approximately in dynamic periodic steady state; (ii) the terminal state coincides with the initial state, so that the two empirical mass functions $\bar{\alpha}$ and $\bar{\alpha}^e$ constructed from the directly estimated BD rates coincide, as indicated in §1.1, and (iii) we average over roughly a full cycle (or more generally, multiple full cycles). In addition, we get a smoother fit by fitting the parametric function in (53) to the estimated birth rates.

Figure 17 shows that both estimated steady-state pmf's are quite good considering the small sample size. Our sampling procedure clearly helps for both the direct estimation and the arctan BD fit. The advantage of the arctan fit is perhaps most evident in Figure 18, which shows the estimated tail probabilities. Evidently, the arctan BD fit can "extrapolate" reasonably to regions where there are few data points or none at all. However, as in all such extrapolations, caution is required. Justification of such extrapolations may in part depend on extra information about what we know about the system.

The observations above are amplified by considering even smaller sample sizes. To illustrate, we now repeat the experiment using the data from only a single cycle. (The sample size is now further reduced by a factor of 10.) Figures 19 and 20 show the analogs of Figures 17 and 18.

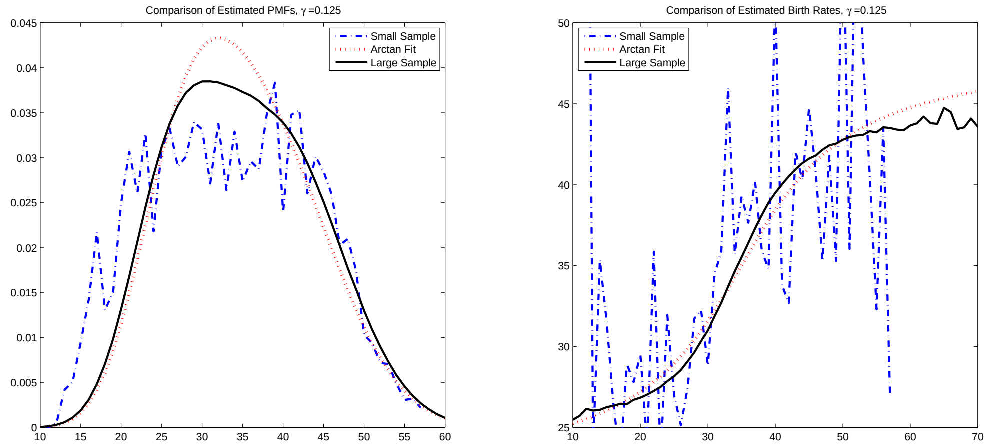

Figure 19: A repeat of the experiment in 17 with data from one cycle instead of ten cycles. Estimates are shown of the steady-state probability mass function (left) and birth rates (right)

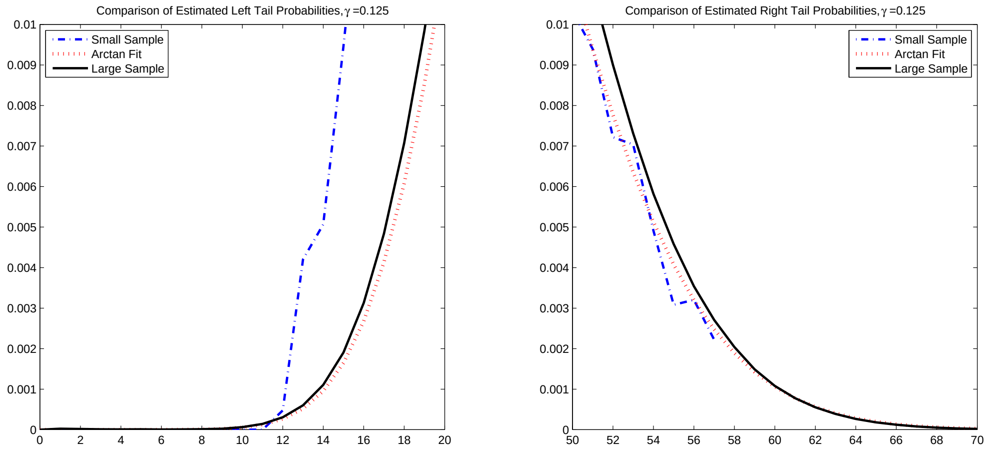

Figure 20: A comparison of the arctan BD fit to a small-sample direct fit of the steady-state tail probabilities (left) and (right) with data from one cycle instead of ten cycles.

6 Stochastic Grey-Box Modeling of Queueing Systems

In addition to investigating an alternative way to estimate a steady-state distribution from data, as illustrated by §5, this paper further explores how fitting a BD process to queueing system data can be used as a grey-box model, which was begun in [7].

6.1 A Diagnostic Tool to Determine What Model Is Appropriate

The goal is to have a diagnostic tool to help evaluate what stochastic queueing model is appropriate for complex queueing systems. 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 [7] we referred to this as *grey-box stochastic modeling*. In [7] 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 have considered many-server $M_t/GI/s$ queueing models with sinusoidal periodic arrival rate functions. We find that the fitted death rates usually 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.

6.2 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 [7] and Theorem 2.3 here.) However, the fitted birth rates are revealing.

In [7] 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

$$\bar{\lambda}_{n,k} \approx (n + b(k-n)) \vee 0, \tag{54}$$

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 21, 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 21 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.

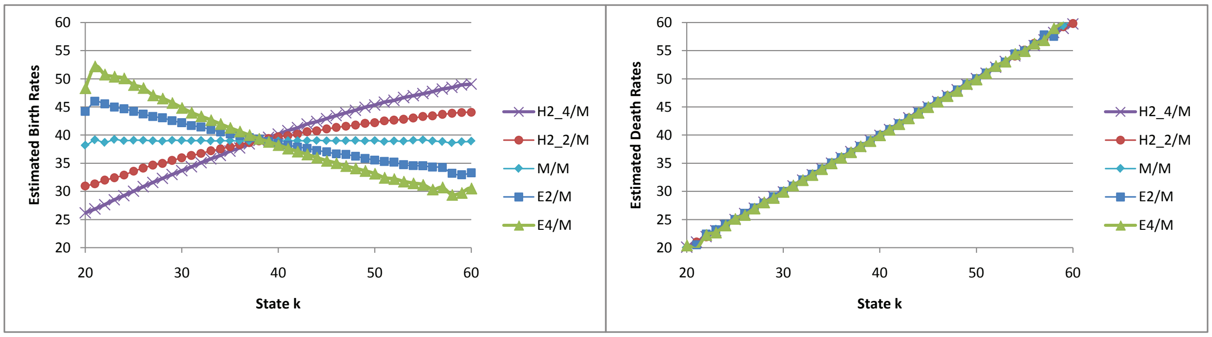

Figure 21: Fitted birth rates and death rates for five $G/M/\infty$ models with $\lambda = 39$ and $\mu = 1$.

Figure 21 should be compared to Figure 1 in §4.2, which shows the fitted rates for the $M_t/M/\infty$ IS model with the sinusoidal arrival rate function in (8). 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 [28]. That means we expect to see something like the fitted birth rates with increasing linear slopes in Figure 21. And indeed that is exactly what we do see in Figure 1, but restricted to a subinterval centered at the long-run average $\lambda_{n,n} = n$.

The evolution of a BD queue primarily depends on the birth and death rates λ_k and μ_k through their difference, the drift $\delta_k \equiv \lambda_k - \mu_{k+1}$, $k \geq 0$. To see the relevance of the drift, note that

$$\frac{\alpha_{k+1}}{\alpha_k} > (=<) 1 \quad \text{if and only if} \quad \delta_k > (-<) 0. \tag{55}$$

Hence we see the modes of α through the zeroes of δ.

We plot the drift functions associated with the $G/M/\infty$ and $M_t/M/\infty$ models in Figures 21 and 1 in Figure 22. These show that there is drift toward the overall mean in all cases, which is stronger when there is less variability.

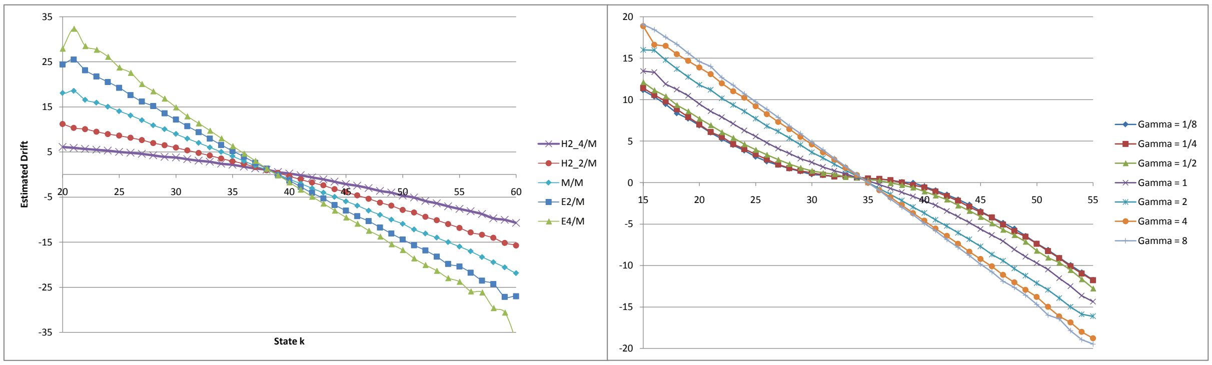

Figure 22: The estimated drift functions (birth rates minus death rates) in *(55)* for the $G/M/\infty$ model in Figure 21 (left) and the $M_t/M/\infty$ model in Figure 1 (right).

Similar results hold for models with finitely many servers. We show the results paralleling Figure 1 for the case of 40 servers in Figure 2. Figure 2 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 [7] displays similar plots for $GI/GI/s$ queues. However, the estimated birth rates in Figures 1 and 2 are unlike those of any $GI/GI/s$ queue. Theorems 3.3 and 3.4 establish finite bounds and heavy-traffic limits for the fitted birth rates, consistent with these figures.

6.3 An Emergency Room Example

To illustrate how the results here can be applied, we show the fitted BD rates obtained from 25 weeks of data from an Israeli emergency department studied in [36]. Figure 23 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 [36] included about 25,000 patient visits to the internal unit of the ED over a 25 week period from December 2004 to May 2005.

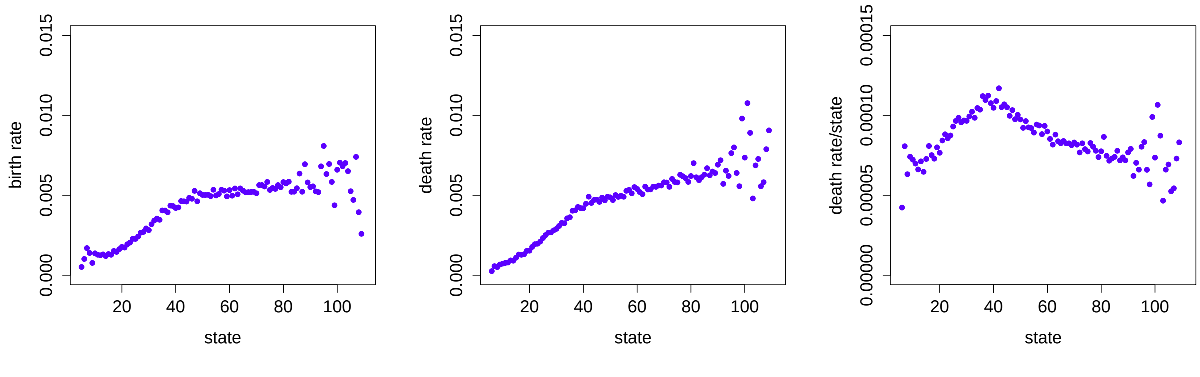

Figure 23: The fitted state-dependent birth rate $\bar{\lambda}_k$ (left), death rate $\bar{\mu}_k$ (center) and death rate divided by the state, $\bar{\mu}_k/k$ (right) obtained from arrival and departure data in an Israeli emergency department over 25 weeks, taken from [36]

It is well known that the arrivals to an ED vary strongly over time, just as in most service systems; see Figure 9 of [16]. Thus, a natural candidate rough aggregate model for an ED is the

$M_t/GI/\infty$ queue, which has a nonhomogeneous Poisson process (NHPP) as its arrival process, 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 assumption that the length-of-stay random variables are i.i.d. might be postulated under the assumption that the length of stay should only depend on the patient's medical condition.

The fitted BD process is important as a diagnostic tool because it shows that the data are inconsistent with the $M_t/GI/\infty$ model. The present paper is essential for interpreting Figure 23. First, Figures 1, 2 and 4 in §6.2 and §4.3 provide strong support for two conclusions: First, as anticipated, 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 23. 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 [36]. The data analysis in [36] 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, 29] and references there. The fitted BD is convenient because it quickly exposes the difficulty with a model using i.i.d. service times.

6.4 Fitting the Erlang-A Model to Data

Perhaps the most frequently applied stochastic queueing model is the $M/M/s+M$ Erlang-A model. 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 [13, 26] 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 $\theta = 0$.

The fitted BD process may 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 typically fit within an *assumed* $GI/GI/s + GI$ model framework, assuming that the interarrival times, service times and patience times come from mutually independent i.i.d. sequences of i.i.d. random variables. Assuming this framework, the Erlang-A model is typically fit by estimating the distributions of the interarrival times and service times to see if they are nearly exponential. It has been recognized that abandonment is more complicated because of censoring; thus other estimation methods have been proposed for estimating the customer patience distribution via the hazard rate of the customer patience distribution; see [4].

However, the assumed $GI/GI/s + GI$ framework need not hold. 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 [12]. Indeed, careful statistical analysis of service system data can be quite complicated, e.g., see [2, 4, 19, 21, 22].

Thus, the fitted birth and death rates provides another framework to test and 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 approximately the simple linear structure in the estimated birth and death rates. With enough data, we will see that

$$\bar{\lambda}_k \approx \lambda, \quad k \geq 0, \quad \text{and} \quad \bar{\mu}_k \approx (k \wedge s)\mu + (k-s)^+\theta, \quad k \geq 1, \tag{56}$$

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 (56) 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.

6.5 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 (56)? 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 (56).

We may also consider directly applying the fitted BD process even if we do not see the Erlang-A structure in (56), 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 [5], which is not limited to the classical Erlang models in [13, 17]. 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].

7 Conclusions

After observing that the steady-state distribution of any stochastic process on the integers with only unit-step transitions can be estimated from data by fitting birth and death rates and solving the BD local-balance equations in (4), we investigated the application of this approach to the $M_t/GI/s$ queueing model with a periodic arrival-rate function, primarily focusing on the case of the sinusoidal arrival-rate function in (8). In §2 and §3 we established structural results for the steady-state distribution and the fitted rates in this model.

In §4 we conducted simulation experiments showing the results of the fitting procedure. Figures 4-6 show near insensitivity to the service-time distribution beyond its mean in the $M_t/GI/\infty$ model for some cases with moderate scale (average arrival rate $\bar{\lambda} = 35$), but Figures 7-9 show significant deviations from insensitivity at larger scale $\bar{\lambda} = 100$). As expected, we see that the transient behavior of the fitted BD process can be very different from the original process, but we see that the difference decreases as the cycle length of the periodic function decreases (relative to the mean service time). Figure 10 shows great differences in the sample paths for the long cycles with $\gamma = 0.01$ in (8), but Figures 11 and 12 show striking similarities with shorter cycles based on $\gamma = 1.0$ and 10 in (8).

In §5 we showed that estimating the fitted rates and then solving the local-balance equations in (4) can be an efficient way to estimate the steady-state distribution because the birth-rate and death-rate functions tend to be more elementary functions. We found that good estimates of the steady-state distribution in the $M_t/M/\infty$ model can be obtained in this way by fitting only two parameters in the four-parameter arctangent function in (53).

Finally, in §6 we see how the fitted BD model can serve as a grey-box model to help diagnose what is an appropriate queueing model for complex queueing applications, as suggested in [7]. In §6.2 we compared the BD models fit to stationary $GI/GI/s$ models in [7] to the BD models fit to the $M_t/GI/s$ models in this paper. For large s, we see that the fitted death rates have the same piecewise-linear structure found in the $M/M/s$ model, but there are significant differences in the fitted birth rates. However, the birth rates have the same linear structure around the steady-state mean number of busy servers.

In §6.3 we illustrated how the grey-box modeling approach can be applied. Figure 23 dsplaying the fitted birth and death rates in an Israeli emergency department from [2, 36] 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 (length-of-stay) times. Instead, the length-of-stay distribution should be time-varying, consistent with observations in [2, 29, 36]. In §6.4 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 an additional statistical test.

There are many directions for future research. As indicated before Theorem 2.1 in §2, it remains to prove that the interval between successive emptiness epochs in the stochastic process $\{Q(nc+t) : n \geq 0\}$ have finite mean for any periodic $M_t/GI/s$ model with $E[S] < \infty$ and $\bar{\rho} < 1$. It also remains to derive explicit formulas and asymptotic approximations for the fitted rates in these and other models. In forthcoming [8], we establish many-server heavy-traffic limits for the fitted birth rates and death rates in the $M_t/GI/\infty$ model.

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 Columbia University doctoral student Xiaopei Zhang for preparing Figure 23 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. Stochastic Systems, published online, DOI-10.1214/14-SSY153.

[3] Billingsley, P. (1961). *Staistical Inference for Markov Processes*. Chicago: University of Chicago press.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Dong, J. and Whitt, W. (2016). Many-server heavy-traffic limits for fitted birth and death rates in time-varying infinite-server queues. In preparation.

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

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

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

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

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

[19] 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.

[20] Keiding, N. (1975). Maximum likelhood estimation in the birth-and-death process. *Ann Statist* 3:363–372.

[21] 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.

[22] 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.

[23] 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.

[24] Kim, S.-H., 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.

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

[26] 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.

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

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

[29] 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.

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

[31] 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.

[32] Whitt, W. (1991). A review of $L = \lambda W$. *Queueing Systems* 9:235–268.

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

[34] 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.

[35] Whitt, W. (2015). A Poisson limit for the departure process from a queue with many slow servers. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[36] 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.

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