Queueing Syst (2015) 79:391–426
DOI 10.1007/s11134-014-9429-3

Stochastic grey-box modeling of queueing systems: fitting birth-and-death processes to data

James Dong · Ward Whitt

Received: 8 January 2014 / Revised: 17 November 2014 / Published online: 2 December 2014

Abstract This paper explores grey-box modeling of queueing systems. A stationary birth-and-death (BD) process model is fitted to a segment of the sample path of the number in the system in the usual way. The birth (death) rates in each state are estimated by the observed number of arrivals (departures) in that state divided by the total time spent in that state. Under minor regularity conditions, if the queue length (number in the system) has a proper limiting steady-state distribution, then the fitted BD process has that same steady-state distribution asymptotically as the sample size increases, even if the actual queue-length process is not nearly a BD process. However, the transient behavior may be very different. We investigate what we can learn about the actual queueing system from the fitted BD process. Here we consider the standard $GI/GI/s$ queueing model with s servers, unlimited waiting room and general independent, non-exponential, interarrival-time and service-time distributions. For heavily loaded s-server models, we find that the long-term transient behavior of the original process, as partially characterized by mean first passage times, can be approximated by a deterministic time transformation of the fitted BD process, exploiting the heavy-traffic characterization of the variability.

J. Dong
School of Operations Research and Information Engineering, Cornell University,
287 Rhodes Hall, Ithaca, NY 14850, USA
e-mail: jd748@cornell.edu

W. Whitt (✉)
Department of Industrial Engineering and Operations Research,
Columbia University, New York, NY 10027, USA
e-mail: ww2040@columbia.edu

Keywords Birth-and-death processes · Grey-box modeling · Fitting stochastic models to data · Transient behavior · First passage times · Heavy traffic

Mathematics Subject Classification 60F17 · 60J25 · 60K25 · 62M09 · 90B25

1 Introduction

Queueing theory primarily involves *white-box modeling*, in which queueing models are defined from first principles and analyzed. To apply these models to analyze the performance of queueing systems, we check that the assumptions of the models are satisfied, for example, by performing statistical tests on system data, as in [7,28], and estimating model parameters, as in [4,5,36] and references therein.

An alternative to white-box modeling of queueing systems is *black-box modeling*, in which models (for example, statistical time-series methods) are employed, without using any structure of queueing models; see [42,59]. Black-box modeling is growing in popularity as the amount of available data grows, as well as our ability to extract useful information from it; see [22]. A modeling approach in between these two extremes, which might conceivably share the advantages of both, is *grey-box modeling*, which exploits some model structure together with learning from data; see [6,29].

We examine grey-box modeling applied to queueing processes. In this paper, the grey-box model is a general stationary birth-and-death (BD) process, which exploits the common property of many queue-length (number in system) processes that they almost surely make only unit transitions, corresponding to separate arrivals and departures. Indeed, BD queueing models such as the Markovian $M/M/s$ model are well studied and have been extensively used. Moreover, BD models have been found to be useful approximations for other more general but less tractable models, as illustrated by the BD approximation in [52] for the $M/GI/s+GI$ model having customer abandonment from the queue with non-exponential patience times (the $+GI$) as well as non-exponential service times.

The main idea here is to approach an application where we do not know what model is appropriate by fitting a stationary BD process to an observed segment of the sample path of a queue-length stochastic process, assuming only that it increases and decreases in unit steps. Just as is commonly done for estimating rates in a BD process [5,26,57], we estimate the birth rate in state k from arrival 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 we estimate the death rate in state k 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. As usual, the steady-state distribution of that fitted BD model, denoted by $\bar{\alpha}^{\mathrm{e}}_k \equiv \bar{\alpha}^{\mathrm{e}}_k(t)$ (with superscript e indicating the estimated rates), is well defined (under regularity conditions, see [54]) and is characterized as the unique probability vector satisfying the local balance equations,

$$\bar{\alpha}^{\mathrm{e}}_k \bar{\lambda}_k = \bar{\alpha}^{\mathrm{e}}_{k+1} \bar{\mu}_{k+1}, \quad k \geq 0. \tag{1.1}$$

Throughout this paper, we assume that limiting values of the rates as $t \to \infty$ exist and that our estimators are consistent, so we omit the t.

The BD process is appealing as a grey-box model for queueing systems, because the fitted BD steady-state distribution $\{\bar{\alpha}_k^{\mathrm{e}} : k \geq 0\}$ in (1.1) 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. Indeed, as has been known for some time (for example, see Chapter 4 of [19]), 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 periodic or non-Markovian; see [46].

If we directly fit a BD process to data as just described, we should not conclude without further testing that the underlying queue-length process actually is a BD process. In fact, the main point of [54] was to caution against drawing unwarranted positive conclusions from a close similarity in the steady-state distributions, because these two distributions are automatically closely related. Nevertheless, the fitted BD process might provide useful insight about the underlying system. The purpose of the present paper is to start investigating that idea.

In fact, our earlier paper [54] was largely motivated by exploratory data analysis carried out in Sect. 3 of [3]. They had fitted several candidate models to data on the number of patients in a hospital emergency department, and found that a BD process fit better than others. After writing our previous paper [54], cautioning against drawing unwarranted positive conclusions, we realized that fitting BD processes to data might indeed be useful, and started the present investigation. See Sect. 3 of [3] for more discussion about exploratory data analysis of an emergency department.

There also is a much earlier precedent for exploiting fitted BD processes. This idea was part of the program of operational analysis suggested by Buzen and Denning [8,9,15] in early performance analysis of computer systems. However, they expressed the view that there was no need for an associated "underlying" stochastic model. In contrast, as in Sects. 4.6–4.7 of [19], we think that an underlying stochastic model should play an important role. With that in mind, we want to see if the fitted BD model can provide useful insight into an unknown underlying stochastic model.

Our goal in the present paper is to better understand this fitted BD model. We primarily want to answer two questions:

(i) How are the fitted birth and death rates related to the key structural features of an actual queueing model?
(ii) How does the transient behavior of an actual queueing system differ from the transient behavior of the fitted BD process?

We start to address the first question by estimating the birth and death rates from simulation experiments for $GI/GI/s$ models. When the interarrival-time (service-time) distribution is exponential, then the fitted birth (death) rate captures the true behavior of the queue-length process, but otherwise it does not. For non-exponential distributions, we see that the fitted rates evidently have substantial structure, which we do our best to explain.

We start to address the second question by looking at first passage times in the underlying model and the fitted model. Perhaps our most interesting finding is a simple connection between the fitted BD process $Q_{\mathrm{f}} \equiv \{Q_{\mathrm{f}}(t) : t \geq 0\}$ and the queue-length (number in system) process $Q \equiv \{Q(t) : t \geq 0\}$ for a stationary $GI/GI/s$

model, which holds approximately under two conditions: (i) if we focus on the long-term transient behavior, as measured by the expected first passage time from the p^{th} percentile of the stationary distribution to the $(1-p)^{\text{th}}$ percentile and back again for p suitably small, for example, $p = 0.1$, and (ii) if we consider models with relatively high traffic intensity, so that heavy-traffic approximations are appropriate. Under these two conditions, we find that the two processes differ primarily in a one-dimensional way, by the speed that they move through the state space. In other words, we suggest the process approximation

$$\{Q(t) : t \geq 0\} \approx \{Q_{\mathrm{f}}(t/\omega) : t \geq 0\}, \tag{1.2}$$

where ω is the *speed ratio*, which admits the conventional heavy-traffic approximation (traffic intensity ρ near 1 with finitely many servers)

$$\omega \approx \frac{c_a^2 + c_s^2}{2}, \tag{1.3}$$

with c_a^2 and c_s^2 being the squared coefficients of variation (scv's, variance divided by the square of the mean) of an interarrival time and a service time, respectively; see Tables 2, 3, 4, and 5. The approximating speed ratio in (1.3) can be recognized as the heavy-traffic characterization of the variability in the $GI/GI/s$ queueing model. We also discuss how to estimate the speed ratio in s-server models without the strong independent assumptions; see Sect. 7.1.

We provide important mathematical support for (1.2) with (1.3) for both the $GI/GI/s$ model in the conventional heavy-traffic regime by Corollaries 5.1 and 5.2 (and the results leading up to them) and for the $GI/M/\infty$ model in the many-server heavy-traffic regime by Corollary 6.2 (and the results leading up to it). These corollaries establish that (1.2) with (1.3) is asymptotically correct in a heavy-traffic limit in considerable generality. In order to establish these results, we have to make quite strong assumptions on the fitted birth rates (consistent with our simulation results), but we conjecture that these properties of the fitted rates hold; see Conjectures 2.1, 2.2 and 5.1. The current conventional heavy-traffic results fully cover the $GI/M/s$ model by virtue of Theorem 3.1.

The proofs of these heavy-traffic limits are interesting because they depend on differences in the stochastic-process limits for the original queue-length process and the fitted BD process. The two processes necessarily have the same steady-state distribution, as we observed above. Since both processes are asymptotically characterized by diffusion limits, we see that *the common steady-state distribution is achieved by the non-Markov variability captured by the diffusion coefficient in the diffusion approximation for the original process being transferred into the drift of the diffusion limit of the fitted BD process.* The fitted BD process evidently captures less of the local variability in the original process. As a consequence, we get the speed ratio as in (1.2) above.

Consistent with extensive experience about the many-server heavy-traffic regime, we also discover that a very different story holds for large-scale many-server queueing systems with non-exponential service-time distributions, where the very different

many-server heavy-traffic approximations are appropriate. In particular, we find that the approximation in (1.2) is still appropriate for other $GI/GI/\infty$ models with non-exponential service-time distributions, but that (1.3) is not appropriate. Indeed, the speed ratio in the $M/GI/\infty$ model tends to depend on the scv c_s^2 of the non-exponential service-time distribution in way that is *inversely* related to (1.3); see Table 5.

Organization: We start in Sect. 2 by showing simulation estimates of the birth and death rates in $GI/GI/1$ single-server and $GI/GI/s$ multi-server models for $s = 40$ and $s = \infty$, for various interarrival-time and service-time distributions. Next, in Sect. 3 we establish mathematical results showing how the fitted BD rates depend on the model structure of stable $GI/GI/s$ models, trying to explain as much as we can of the structure we see in the fitted rates. In Sect. 4 we define the speed ratios in terms of first passage times and show simulation results for the speed ratio. Then in Sect. 5 we obtain insights about the fitted BD process and the speed ratio by making connections to conventional heavy-traffic limit theory. In Sect. 6 we make corresponding connections to the many-server heavy-traffic theory. In Sect. 7 we discuss estimation procedures. In Sect. 7.1 we discuss how to estimate the speed ratios, pointing out the possibility of exploiting indices of dispersion. In Sect. 7.2 we discuss how to estimate the birth and death rates, pointing out the advantages of smoothing. Finally, in Sect. 8 we draw conclusions. Additional details are provided in a short Appendix.

2 Fitted rates in simulation experiments

We start by reporting results of simulation experiments for $GI/GI/s$ models. For the non-exponential distributions, we use Erlang (E_k, sums of k i.i.d. exponentials with $c^2 = 1/k$, focusing on $k = 2$ and $k = 4$) and hyperexponential ($H_2(c^2)$, mixtures of two i.i.d. exponentials with scv $c^2 > 1$) to illustrate non-exponential distributions less and more variable than exponential. An H_2 cdf can be expressed as $F(x) \equiv p_1(1-\mathrm{e}^{-\lambda_1 x})+p_2(1-\mathrm{e}^{-\lambda_2 x})$. We fix one of the three parameters by assuming balanced means, $p_1/\lambda_1 = p_2/\lambda_2$, as on p. 137 of [44]. We primarily consider $H_2(2)$ with scv $c^2 = 2$.

For these models, it can be shown that the fitted rates are consistent estimators of finite limiting values; i.e., $\bar{\lambda}_k(t) \to \bar{\lambda}_k$, $k \geq 0$, and $\bar{\mu}_k(t) \to \bar{\mu}_k$, $k \geq 1$, as $t \to \infty$. (Justification can be by an application of the law of large numbers for renewal reward processes, as in Ch. 3 of [37]; arrivals to an empty system can be the renewal epochs.) We assume that the t in question is sufficiently large that we can regard our estimate as the limiting value. We estimated the rates from 30 independent replications of 1 million customers. This large sample size is sufficient for 95 % confidence intervals of the state-dependent rates to be within 1 % of the rates for states k with steady-state probability $\alpha_k \geq 0.01$. The width of the confidence intervals increases as a function of the variability parameters c_a^2 and c_s^2. Of course, even though our simulation runs are long, good estimates of the rates are only possible for states k that are frequently visited.

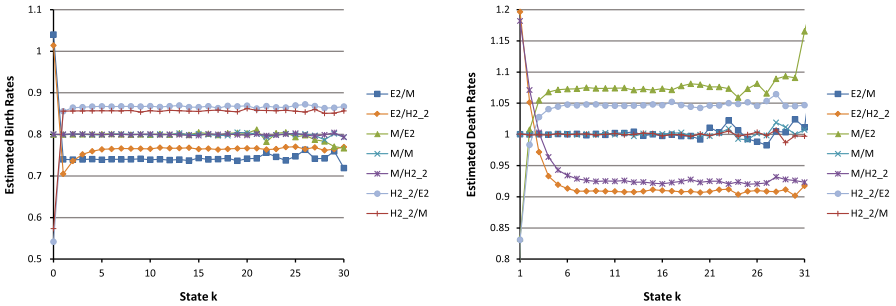

Fig. 1 Fitted birth rates $\bar{\lambda}_k$ (*left*) and death rates $\bar{\mu}_k$ (*right*) for seven $GI/GI/1$ models with $\rho = \lambda = 0.8$ and $\mu = 1$

2.1 Single-server queues

We first fit birth rates and death rates to data from simulations of various $GI/GI/1$ single-server models, all with traffic intensity $\rho = 0.8$. We let the mean service time be $1/\mu = 1$ in all cases, so that the overall arrival rate is $\lambda = \rho = 0.8$. Fig. 1 shows the fitted birth rates (left) and death rates (right) for seven different models. Some loss of statistical precision is seen in some of the plots for large values around $k = 30$ at the right end of these plots.

From Fig. 1, the fitted rates seem to satisfy, at least approximately, the property

$$\bar{\lambda}_k = \bar{\lambda} \quad \text{and} \quad \bar{\mu}_k = \bar{\mu} \quad \text{for all} \quad k \geq k_0, \tag{2.1}$$

for some k_0, for example, $k_0 = 5$. We show that the relations in (2.1) hold exactly in the $GI/M/s$ model (Theorem 3.1). We show that the fitted death rates differ in the $M/GI/1$ model for any two different service-time distributions (Theorem 3.3). Thus, the relation (2.1) evidently does not hold exactly for any non-exponential distribution in the $M/GI/1$ model. Evidently limits exist as $k \to \infty$ for a large class of models.

We can make inferences about the interarrival-time and service-time distributions from Fig. 1. First, from the queue-length sample path we can also estimate the arrival rate λ and the service rate μ, which of course are known in advance in the simulation experiments. Our estimated arrival (service) rate is $A(t)/t$ ($D(t)/B(t)$), where $A(t)$ ($D(t)$) is the total number of arrivals (departures) over the interval $[0, t]$ (with $D(t) \approx A(t)$ if t is large) and $B(t)$ is the total time that the server is busy in $[0, t]$.

Of particular importance for understanding performance is the *fitted traffic intensity* $\bar{\rho} \equiv \bar{\lambda}/\bar{\mu}$, which we would estimate or calculate directly as the limit of $\bar{\rho}_k \equiv \bar{\lambda}_k/\bar{\mu}_{k+1} = \alpha_{k+1}/\alpha_k$ as $k \to \infty$; see Sect. 3.3. For its impact on congestion, we would focus on $(1 - \bar{\rho})^{-1}$ and its relation to $(1 - \rho)^{-1}$ via $\bar{\omega} \equiv (1 - \rho)/(1 - \bar{\rho})$. An important observation is that all these quantities—$\bar{\lambda}$, $\bar{\mu}$, $\bar{\rho}$, and $\bar{\omega}$—are actually functions of ρ. Theorems 3.5 and 5.1 show that for a large class of $GI/GI/s$ models

$$\bar{\omega}(\rho) \equiv \frac{1 - \rho}{1 - \bar{\rho}(\rho)} \approx \omega \equiv \frac{c_a^2 + c_s^2}{2} \tag{2.2}$$

Table 1 Estimates of the asymptotic fitted birth rate $\bar{\lambda}$, death rate $\bar{\mu}$, traffic intensity $\bar{\rho}$, and speed ratio $\bar{\omega}$ via (2.2) for the nine $GI/GI/1$ models with $\rho = 0.8$ in Fig. 1

Model	$\bar{\lambda}$	$\bar{\mu}$	$\bar{\rho}$	$\bar{\omega}$	$\omega \equiv (c_a^2 + c_s^2)/2$
$E_2/E_2/1$	0.700 ± 0.002	1.138 ± 0.026	0.603 ± 0.044	0.520 ± 0.035	0.500
$E_2/M/1$	0.741 ± 0.001	1.002 ± 0.003	0.740 ± 0.001	0.770 ± 0.004	0.750
$E_2/H_2(2)/1$	0.766 ± 0.001	0.910 ± 0.001	0.843 ± 0.001	1.271 ± 0.010	1.250
$M/E_2/1$	0.801 ± 0.001	1.074 ± 0.002	0.746 ± 0.003	0.788 ± 0.008	0.750
$M/M/1$	0.800 ± 0.001	1.000 ± 0.001	0.800 ± 0.002	1.000 ± 0.011	1.000
$M/H_2(2)/1$	0.799 ± 0.001	0.926 ± 0.002	0.865 ± 0.001	1.478 ± 0.014	1.500
$H_2(2)/E_2/1$	0.867 ± 0.001	1.047 ± 0.001	0.828 ± 0.001	1.161 ± 0.009	1.250
$H_2(2)/M/1$	0.857 ± 0.001	0.945 ± 0.003	0.857 ± 0.001	1.399 ± 0.011	1.500
$H_2(2)/H_2(2)/1$	0.843 ± 0.001	0.945 ± 0.003	0.893 ± 0.002	1.877 ± 0.027	2.000

when ρ is not too small, where ω is the theoretical speed ratio in (1.3).

Table 1 shows the estimates of the asymptotic BD rates $\bar{\lambda}$ and $\bar{\mu}$ with 95 % confidence intervals and the associated estimates of the fitted traffic intensity $\bar{\rho}$ and the speed ratio $\bar{\omega}$ for the models with $\rho = 0.8$ in Fig. 1. This last estimate is compared to the approximation in (2.2). These statistical estimates are obtained directly from the final estimated rates $\bar{\lambda}_k$ and $\bar{\mu}_k$ by treating the values with $5 \le k \le 24$ as a sample from 20 i.i.d. normally distributed random variables with unknown variance, so that the confidence intervals are constructed with the Student t distribution having 19 degrees of freedom. Of course, in this case the statistical precision is compounded with the actual deterministic variation in the rates, but Table 1 shows that both sources of variability are not great.

Figure 1 and Table 1 show that $\bar{\lambda}$ ($\bar{\mu}$) is increasing in c_a^2 (c_s^2) with $\bar{\lambda} = \lambda$ ($\bar{\mu} = \mu$) in the M case when $c_a^2 = 1$ ($c_s^2 = 1$). We provide theoretical support for this conclusion in the $GI/M/s$ model in Theorem 3.2. The last two columns of Table 1 provide support for the approximation in (2.2). The analogs of Fig. 1 and Table 1 for $\rho = 0.9$ are given in the Appendix.

2.2 Many-server models

We next look at many-server models. Specifically, we consider seven $GI/GI/s$ models with $s = 40$ and seven $GI/GI/\infty$ models. These all have the parameters $\lambda = 39$ and $\mu = 1$. The fitted birth and death rates are shown in Figs. 2 and 3. By Little's law, the mean number of busy servers is always $\rho \equiv \lambda/\mu = 39$. Since the number in the system tends to concentrate in the interval [20, 60], we show only that part. We would see the impact of unreliable estimation outside of that interval.

The most striking behavior in Figs. 2 and 3 is the systematic structure of the death rates within the interval [20, 60] of frequently visited states. We evidently have, at least approximately,

$$\bar{\mu}_k = (k \wedge s)\mu = (k \wedge s), \quad k \ge 1, \tag{2.3}$$

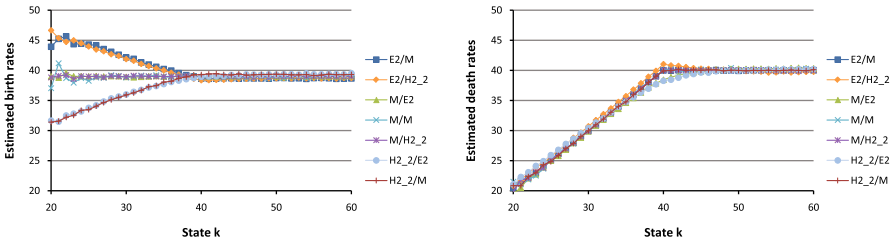

Fig. 2 Fitted birth and death rates for seven $G/G/40$ models with $\lambda = 39$ and $\mu = 1$

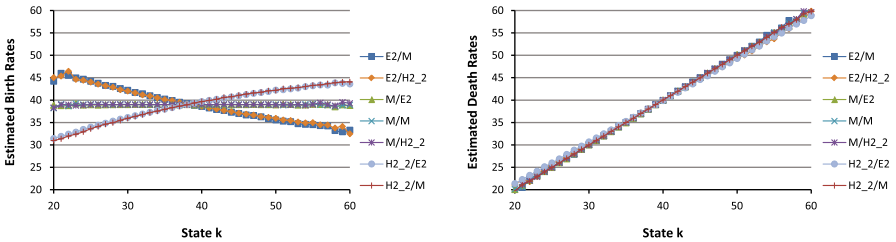

Fig. 3 Fitted birth and death rates for seven $G/G/\infty$ models with $\lambda = 39$ and $\mu = 1$

for all interarrival-time and service-time distributions, where $a \wedge b \equiv \min\{a, b\}$. We show that relation (2.3) holds exactly in the $M/GI/\infty$ model (Theorem 3.4) by applying the classical insensitivity property for the $M/GI/\infty$ model, stating that the steady-state queue length depends on the service-time distribution only through its mean.

The observed relation in (2.3) leads us to conjecture that it is asymptotically correct in a many-server heavy-traffic limit. However, it is not immediate how to properly formulate such a limit, because the frequently visited states evidently fall in an interval $[\rho - c\sqrt{\rho}, \rho + c\sqrt{\rho}]$ for some positive constant c, where $\rho \equiv \lambda/\mu \to \infty$. We give the following specific conjecture. Let $\bar{\mu}_k(\rho)$ be the death rate in state k of model ρ and let $\lfloor x \rfloor$ be the greatest integer less than or equal to x. Recall that $f(\epsilon) = o(\epsilon)$ $(O(\epsilon))$ if $f(\epsilon)/\epsilon \to 0$ (remains bounded) as $\epsilon \downarrow 0$.

Conjecture 2.1 (many-server HT limit for fitted death rates) *Consider a family of $GI/GI/s$ models indexed by $\rho > 0$ with fixed service-time distribution having finite mean $1/\mu$ and variance, and arrival process $A_\lambda(t) \equiv A(\lambda t)$, where A is a rate-1 renewal process whose interarrival time has a finite variance, and $\lambda \equiv \rho\mu$. If there exists an $\epsilon > 0$ such that $s_\rho > (1 + \epsilon)\rho$ for each ρ, then $\bar{\mu}_{\lfloor \rho + c\sqrt{\rho} \rfloor}(\rho) = (\rho + c\sqrt{\rho})\mu + o(\sqrt{\rho})$ as $\rho \to \infty$ for all real numbers c; i.e., for any real number c,*

$$\frac{\bar{\mu}_{\lfloor \rho + c\sqrt{\rho} \rfloor}(\rho) - (\rho + c\sqrt{\rho})\mu}{\sqrt{\rho}} \to 0 \quad as \quad \rho \to \infty. \tag{2.4}$$

Support. Conjecture 2.1 is consistent with (but not directly implied by) the Poisson limit for the departure process as λ and s grow for fixed μ, as established in [47].

Figure 3 also shows regularity in the arrival rates in infinite-server models. For infinite-server models with individual service rate $\mu = 1$ and arrival rate ρ, we see that, again at least approximately, the fitted birth rate in state k is

$$\lambda_{\rho,k} \equiv 0 \vee [\rho + b(k - \rho)], \quad k \geq 0, \tag{2.5}$$

where the constant b depends on the interarrival-time distribution, satisfying $-\infty < b < 1$, and $a \vee b \equiv \max\{a, b\}$. (The maximum is needed to prevent the arrival rate from becoming negative as k increases if $b < 0$.)

Paralleling Conjecture 2.1, we conjecture that the conclusions of (2.5) hold more generally.

Conjecture 2.2 (HT limit for fitted arrival rates in infinite-server models) *In addition to Conjecture 2.1, for $GI/GI/\infty$ infinite-server models,*

$$\frac{\lambda_{\rho,k} - 0 \vee [\rho + b(k - \rho)]}{\sqrt{\rho}} \to 0 \quad as \quad \rho \to \infty \tag{2.6}$$

for all integers k.

Formula 2.5 holds exactly with $b = 0$ for exponential interarrival times; otherwise Fig. 3 shows that $b > (<)0$ when $c_a^2 > (<)1$. Formula (2.5) also is appropriate for $k \leq s$ in s-server systems provided that s is suitably large, at least $s > n$, as in Fig. 2.

2.3 Detecting deviations from the $GI/GI/s$ model

In applications, estimating birth and death rates may be especially useful to detect deviations from the $GI/GI/s$ model. Figs. 1 and 2 suggest that such deviations may be easy to detect, because these figures show remarkable consistency in the estimated birth and death rates. For states $k > s$, the rates are nearly constant in all cases. We should thus be able to detect model deviations in which customers are less likely to join if there is a long queue (causing birth rates to decrease in k) or impatient customers abandon the queue (causing death rates to increase in k) as k increases above s. A specific model with these features is the modification of the $M/M/s$ queue with state-dependent balking and abandonment, having parameters

$$\lambda_k \equiv \lambda e^{-(k-s)^+} \quad \text{and} \quad \mu_k \equiv \mu(k \wedge s) + \theta(k - s)^+, \quad k \geq 0. \tag{2.7}$$

where $(x)^+ \equiv \max\{x, 0\}$. Figures 1 and 2 show that those effects are highly unlikely to be caused by just having a $GI/GI/s$ model with non-exponential distributions.

The consistent piecewise-linear structure for larger s in Fig. 2 shows that we should be able to detect if the actual number of servers varies over time, as in [53]. For example, if there were s_1 servers over the interval $[0, t/2)$ and s_2 servers over the interval $[t/2, t]$, where $s_1 < s_2$, then we would have the average of two of the plots in Fig. 2. Unlike Fig. 2, the estimated birth and death rates would have three separate linear pieces, over the intervals $[0, s_1]$, $[s_1, s_2]$, and $[s_2, \infty)$.

For the $M/GI/s + GI$ model, having customer abandonment from the queue, where customers have i.i.d. patience times with a general distribution, fitting the death rates provides a direct statistical approach to constructing an approximating BD model, paralleling the analytical BD approximation developed for this model in [52]. (In a separate study, we found that the estimated death rates agree quite closely with the analytical approximations of the death rates developed in [52].)

3 Rate properties

We now establish properties of the fitted rates $\bar{\lambda}_k$ and $\bar{\mu}_k$ in the stationary stable $G/G/s$ model with arrival rate λ and individual service rate μ. The G instead of GI means that we allow general stationary arrival and service processes, as in [38], but we assume that the service times are independent of the arrival process. We assume that $P(Q(t) = k) \to \alpha_k$ as $t \to \infty$, where $\sum_{k=0}^{\infty} \alpha_k = 1$.

We first make the elementary observation (by applying the lack of memory property of the exponential distribution) that the fitted rates are asymptotically correct in the Markovian (M) cases (which includes i.i.d.). For the $G/M/s$ model, $\bar{\mu}_k = \min\{k, s\}\mu$, $k \geq 1$, and for the $M/G/s$ model, $\bar{\lambda}_k = \lambda$, $k \geq 0$. We also observe that there is a basic rate conservation, which is an elementary consequence of the limits of the averages (assumed in Sect. 1):

$$\lambda = \sum_{k=0}^{\infty} \alpha_k \bar{\lambda}_k = \sum_{k=0}^{\infty} \alpha_k \bar{\mu}_k. \tag{3.1}$$

3.1 $GI/M/s$ models

Additional structure allows us to say more. We can apply standard results for $GI/M/s$ models to deduce the following results.

Theorem 3.1 (explicit expression for the fitted rates in $GI/M/s$) *In the $GI/M/s$ model,*

$$\bar{\mu}_k = (k \wedge s)\mu, \quad k \geq 1, \quad \textit{and} \quad \bar{\lambda}_k = s\mu\sigma, \quad k \geq s, \tag{3.2}$$

where σ is the unique root of the equation

$$\phi_a((1-\sigma)\mu s) = \sigma, \tag{3.3}$$

where $\phi_a(s)$ is the Laplace–Stieltjes transform of an interarrival time, i.e., $\phi_a(s) = E[\mathrm{e}^{-sU}] = \int_0^{\infty} \mathrm{e}^{-st} \mathrm{d}F_a(t)$, where $F_a(t) \equiv P(U \leq t)$ is the cdf of an interarrival time.

Proof Use (1.1) together with Sect. 5.14 of [13]. □

Corollary 3.1 (the case $s = 1$) *In the $GI/M/1$ model,*

$$\bar{\lambda}_0 = \frac{\rho\mu(1-\sigma)}{1-\rho} \quad \textit{and} \quad \bar{\lambda}_k = \sigma\mu, \quad k \geq 1, \tag{3.4}$$

where $\mu = \bar{\mu}_k$ for all k and σ is the unique root to the equation (3.3) in the case $s = 1$.

Proof We can apply Little's law with the server to get $1 - \alpha_0 = \rho$. We then can apply the rate-conservation formula in (3.1) and Theorem 3.1. □

Example 3.1 (the case of $H_2/M/40$) We illustrate by considering the $H_2/M/40$ model with $\lambda = 39$, $\mu = 1$, and an H_2 interarrival-time cdf with scv $c_a^2 = 2$. Here $\rho = 39/40 = 0.975$. Our simulation experiment confirms that the conclusions of Theorem 3.1 hold with $\bar{\lambda}_k = 39.35, k \geq 40$, and $\sigma = \bar{\lambda}_k/40 = 39.35/40 = 0.98375$. (Again we can obtain high accuracy by fitting a linear or constant rate function for $k \geq 1$.) For this model, the one-term (order $O(1-\rho)$) and two-term (order $O((1-\rho)^2)$) heavy-traffic approximations for the root σ in (5.3) are, respectively, $\sigma \approx 1 - (0.05)/3 = 0.9833$ and $\sigma \approx 0.98370$.

To compare the fitted birth rates in $GI/M/s$ models with different interarrival-time distributions, we can use convex stochastic order. We say that one random variable X_1 is less than or equal to another, X_2, in convex stochastic order and write $X_1 \leq_c X_2$ if $E[g(X_1)] \leq E[g(X_2)]$ for all convex real-valued functions g; see Chapter 9 of [37]. Convex stochastic order is a variability ordering; in fact it implies that the two random variables necessarily have the same mean (because $g(-x)$ is convex if $g(x)$ is convex). Useful examples are

$$D(m) \leq_c E_k(m) \leq_c M(m) \leq_c H_k(m), \tag{3.5}$$

where $D(m)$, $E_k(m)$, $M(m)$, and $H_k(m)$ denote, respectively, a deterministic, Erlang of order k (sum of k i.i.d. exponentials), exponential and hyperexponential of order k (mixture of k independent exponentials) random variable, all with mean m.

Theorem 3.2 (the implication of convex order for interarrival-time distributions) *Consider two $GI/M/s$ queueing models with the same s, λ, and μ, and thus also $\rho \equiv \lambda/s\mu < 1$, but different interarrival-time distributions U_1 and U_2. If $U_1 \leq_c U_2$, then $\sigma_1 \leq \sigma_2$, where σ_i is the root for model i in (3.3) and the limiting values of the BD fitted arrival rates satisfy*

$$\bar{\lambda}_k^{(1)} = \bar{\lambda}^{(1)} = s\mu\sigma_1 \leq s\mu\sigma_2 = \bar{\lambda}^{(2)} = \bar{\lambda}_k^{(2)}, \quad k \geq s, \tag{3.6}$$

where $\bar{\lambda}^{(i)}$ denotes the limiting value of the fitted birth rate in model i as $k \to \infty$. Since $\bar{\lambda}_k = \lambda$ for an exponential interarrival-time distribution, i.e., for the $M/M/s$ model, we have $\bar{\lambda} \leq (\geq)\lambda$ whenever $U \leq_c (\geq_c) M$.

Proof Observe that the Laplace–Stieltjes transform that determines the root σ of the equation in (3.3) is the expectation of a convex function. □

More generally, Theorem 3.2 tells us that we expect to have $\bar{\lambda}_k \leq (\geq)\lambda$ for $k \geq s$ whenever the arrival process is less (more) variable or bursty than a Poisson process. This ordering helps us interpret fitted birth rates.

3.2 $M/GI/s$ models

For $M/GI/s$ models, the cases $s = 1$ and $s = \infty$ are very different in what the fitted rates tell us about the original model, i.e., about the service-time distribution. For $s = 1$, at least in principle, the fitted death rates tell us everything; for $s = \infty$, the fitted death rates tell us nothing.

Theorem 3.3 (the $M/GI/1$ model) *For the $M/GI/1$ model, $\bar{\lambda}_k = \lambda$ for all $k \geq 0$ and there is a one-to-one correspondence between the fitted BD death rates $\bar{\mu}_k$ and the service distributions with mean $1/\mu$. Hence, the service-time cdf is exponential if and only if $\bar{\mu}_k = \mu$, $k \geq 1$.*

Proof It is well known that there is a one-to-one correspondence between the steady-state distribution $\{\alpha_k\}$ and the service-time distribution, via the classical Pollaczek-Khintchine generating function, see Sect. 5.8 of [13]. Since there is a one-to-one correspondence between the fitted BD rates and the steady-state distribution, the final conclusion follows. □

Theorem 3.4 (the $M/GI/\infty$ model) *For the $M/GI/\infty$ model, the fitted BD rates satisfy $\bar{\lambda}_k = \lambda$ and $\bar{\mu}_k = k\mu$ for all $k \geq 0$, just as in the $M/M/\infty$ model.*

Proof By the well-known insensitivity property, the steady-state distribution of the $M/GI/\infty$ model is Poisson with mean λ/μ for all service-time distributions with mean $1/\mu$. Hence, the fitted BD rates are the same as for the $M/M/\infty$ model. □

3.3 Tail-probability asymptotics: limits as $k \to \infty$

In many $G/G/s$ queueing models, the steady-state distribution of $Q(t)$ has an asymptotically geometric tail, i.e.,

$$\alpha_k \sim \beta\sigma^k \quad \text{as} \quad k \to \infty, \tag{3.7}$$

i.e., $\alpha_k/\beta\sigma^k \to 1$ as $k \to \infty$. Hence, for k not too small, the approximation $\alpha_k \approx \beta\sigma^k$ is effective and useful; see [2,11] and references therein. However, it is important to note that (3.7) does not cover all possibilities. For example, with heavy-tail service-time cdf's we can have

$$\alpha_k \sim Ak^{-p} \quad \text{as} \quad k \to \infty \tag{3.8}$$

for positive constants A and p.

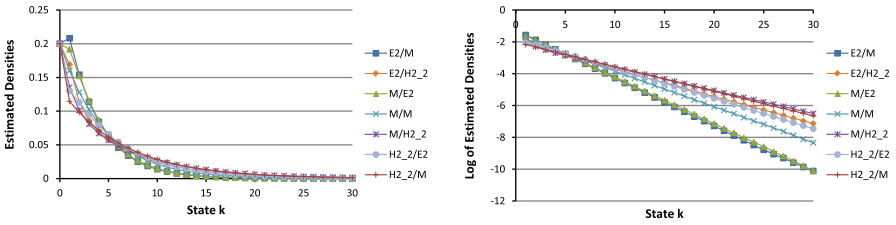

Fig. 4 Estimated steady-state probabilities $\bar{\alpha}_k$ (*left*) and their logarithms $\log_e \bar{\alpha}_k$ (*right*) for seven $GI/GI/1$ models with $\rho = \lambda = 0.8$ and $\mu = 1$

Theorem 3.5 (limits for the fitted rate ratios) *If the steady-state distribution satisfies* (3.7)*, then the limiting fitted BD rate ratios satisfy*

$$\bar{\rho}_k \equiv \frac{\bar{\lambda}_k}{\bar{\mu}_{k+1}} \to \sigma \quad as \quad k \to \infty. \tag{3.9}$$

If the arrival process is Poisson, then $\bar{\mu}_k \to \lambda/\sigma$ *as* $k \to \infty$*. If instead the steady-state distribution satisfies* (3.8)*, then the limiting fitted BD rate ratios satisfy*

$$(k+1)\log \bar{\rho}_k \to -p \quad as \quad k \to \infty. \tag{3.10}$$

Proof For (3.9), use the local balance equation (1.1) with (3.7). For (3.10), use (1.1) with (3.8) to get $\bar{\rho}_k = p \log(1 - [1/(1+k)])$ and then $\log(1+\epsilon)/\epsilon \to 1$ as $\epsilon \downarrow 0$. □

We first illustrate the main case in (3.7) in Fig. 4 by showing the estimated steady-state probability mass function (left) and its logarithm (right) for the seven $GI/GI/1$ models with $\lambda = \rho = 0.8$ and $\mu = 1$ in Fig. 1. The near-linear plots on the right illustrate the relation (3.7) which is known to hold for all these models. We will exploit this geometric-tail structure extensively in our conventional heavy-traffic analysis in Sect. 5.

We also illustrate the power-tail case by considering the $M/GI/1$ queue with a service-time pdf $g(x) = Bx^{-q}$, $x \geq c$, for $B = c^{q-1}/(q-1)$ and c chosen so that it has mean $1/\mu = 1$. In the case $q = 3$, this power-tail pdf has an infinite variance. We show the fitted death rates and the estimated values of $(k+1)\log \bar{\rho}_k$ in Fig. 5 for three cases: $q = 3$, 5, and 7. Figure 5 shows that the estimated death rate is decaying and that (3.8) indeed evidently holds with $p = q - 1 = 2$ for the case $q = 3$, but that (3.7) holds for $q = 5$ and $q = 7$. In particular, we see that a power-tail service-time distribution affects the steady-state queue-length distribution differently from the steady-state waiting time, which always has a power tail in these cases; see [1,12,30] and citations to these papers. We intend to discuss this phenomenon in a future paper.

4 Transient behavior: the speed ratio

We wish to understand how the transient behavior of the fitted BD process differs from the transient behavior of the original model. To approximately characterize the

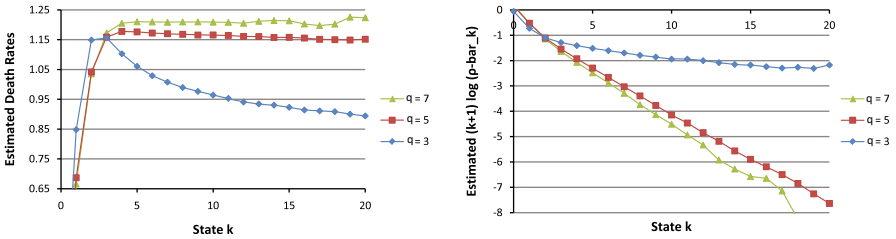

Fig. 5 Estimated death rates $\bar{\mu}_k$ (*left*) and the scaled logarithms $(k+1)\log\bar{\rho}_k$ (*right*) for the $M/GI/1$ model with $\rho = \lambda = 0.8$ and the power-tail service-time pdf $g(x) = Ax^{-q}$, $x \geq c$, with $\mu = 1$ and three values of q: 3, 5, and 7

transient behavior, we focus on first passage times. Let $T(p)$ be the first passage time from the p^{th} percentile of the steady-state distribution to the $(1-p)^{\text{th}}$ percentile of the steady-state distribution and then back again (round trip) in the original process, and let $T_{\text{f}}(p)$ be the corresponding round-trip first passage time in the fitted BD process. (Since the steady-state distribution is discrete, we use the state closest to the percentile, which will be the same for both processes because they have the same steady-state distribution.)

These first passage times are fully specified for the fitted BD process because it is a Markov process, but they are not fully 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 will in fact estimate the expected first passage times for the original process from simulations, by considering successive alternating visits to the p^{th} and $(1-p)^{\text{th}}$ percentiles of the steady-state distribution. As an approximation, which we regard as reasonable as long as p is not too close to $1/2$, 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

$$r(p) \equiv \frac{1}{E[T(p)]} \quad \text{and} \quad r_{\text{f}}(p) \equiv \frac{1}{E[T_{\text{f}}(p)]} \tag{4.1}$$

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

$$\omega(p) \equiv \frac{r(p)}{r_{\text{f}}(p)} = \frac{E[T_{\text{f}}(p)]}{E[T(p)]}. \tag{4.2}$$

If the fitted BD process had the same transient behavior as the original process, then we would find that $\omega(p) \approx 1$ for all p. However, in simulation experiments we find that these speed ratios differ significantly from 1 when p is not close to $1/2$. We see regularity in $\omega(p)$ as $p \downarrow 0$ and as the load increases (always assuming stability, i.e., $\rho < 1$).

4.1 Examples in the conventional heavy-traffic regime

We start with very heavily loaded $GI/GI/1$ single-server models. We let the mean service time be 1 and the arrival rate be 0.975, so that the traffic intensity is $\rho = 0.975$. This is an extremely high traffic intensity, so that heavy-traffic approximations are clearly appropriate. Table 2 shows, first, simulation estimates of the average round-trip first passage times $E[T(p)]$ for $p = 0.1$ and associated half widths of 95 % confidence intervals for $\rho = 0.975$, assuming (as a reasonable approximation) that successive first passage times are i.i.d. and, second, the estimated speed ratios $\bar{\omega}(p)$ for $p = 0.1$. The speed ratios are estimated by the reciprocal of the ratio of the first passage times, as in (4.2). The final two columns give the associated estimated speed ratio for $\rho = 0.8$ with $p = 0.1$ and the associated approximation for the speed ratio in (1.3). These first passage times were simulated using (i) a birth–death process using the birth and death rates obtained from 30 i.i.d. replications of 1 million customers and (ii) the true process, each based on 8 million simulated customers (independent of the data used to estimate the parameters). Process (i) utilizes the average rates from the 30 i.i.d. replications. In order to lessen the effects of noisy data, we prevented the birth–death

Table 2 Estimated speed ratios of transitions between the 10th and 90th percentile of the steady-state distributions, i.e., $\omega(p)$ in (4.2) for $p = 0.1$, in nine $GI/GI/1$ single-server queues for $\rho = 0.975$ and $\rho = 0.800$, with estimation details only for $\rho = 0.975$

Speed Ratios of 10th-to-90th Percentile Transitions for $GI/GI/1$ Single-Server Models					
Traffic intensity $\rho = 0.975$				$\rho = 0.800$	Approx.
Model	$E[T(p)]$, $p = 0.1$	CI halfwidth	$\bar{\omega}(0.1)$	$\bar{\omega}(0.1)$	$\frac{c_a^2+c_s^2}{2}$
$E_2/E_2/1$ True	7,010.2	425.4	0.51	0.63	0.5
$E_2/E_2/1$ BD	3,546.3	140.9			
$E_2/M/1$ True	10,503.3	969.4	0.77	0.80	0.75
$E_2/M/1$ BD	8,113.6	399.3			
$M/E_2/1$ True	10,744.8	1,099.2	0.76	0.81	0.75
$M/E_2/1$ BD	8,181.2	645.5			
$M/M/1$ True	14,048.0	544.9	1.02	1.00	1.00
$M/M/1$ BD	14,276.8	658.9			
$E_2/H_2(2)/1$ True	17,223.1	544.7	1.30	1.21	1.25
$E_2/H_2(2)/1$ BD	22,373.2	1,883.2			
$H_2(2)/E_2/1$ True	18,449.3	1,320.7	1.17	1.07	1.25
$H_2(2)/E_2/1$ BD	21,481.1	1,573.7			
$H_2(2)/M/1$ True	21,846.4	550.3	1.43	1.29	1.50
$H_2(2)/M/1$ BD	31,203.4	2,267.4			
$M/H_2(2)/1$ True	22,086.6	1,200.0	1.43	1.40	1.50
$M/H_2(2)/1$ BD	31,613.4	749.6			
$H_2(2)/H_2(2)/1$ True	30,197.6	2,279.0	1.85	1.74	2.00
$H_2(2)/H_2(2)/1$ BD	55,786.8	6,677.6			

process from going above the 99.5 percentile. The process is initialized in steady state by choosing the initial state from the estimated steady-state distribution. The process then ran for 16 million transitions. In process (i) and process (ii), the system observes 1 million and 8 million customers, respectively. The simulation concluded when all customers had been served.

Table 2 shows that approximation (1.2) with (1.3) is well justified for single-server queues. The estimated speed ratios should be compared to the reference case of 1.00 for the $M/M/1$ model, where the queue-length process is actually a BD process, and the proposed approximation in (1.3) is exactly 1. In all cases, we see that the estimated speed ratio $\bar{\omega}(0.1)$ differs from 1 in the right direction. In all cases except one (the $E_2/H_2(2)/1$ model for $\rho = 0.975$), the approximation in (1.3) slightly overestimates the difference from 1; i.e., $(c_a^2 + c_s^2)/2$ overestimates (underestimates) ω when $(c_a^2 + c_s^2)/2 > 1$ (< 1).

Of course, it is well known that extremely long simulation runs are required to obtain accurate estimates of the steady-state performance in single-server queues when the traffic intensity is close to 1; for example, see [48]. Thus, the example with $\rho = 0.975$ in Table 2 is problematic. Hence we also include another example with $\rho = 0.9$ in Table 3, which still is challenging, but less so. We also more carefully examine the statistical precision. To do so, we perform 5 independent replications of the previous simulation experiment with $\rho = 0.9$ and report the speed ratios for each replication, and then construct confidence intervals from the 5 independent replications using the Student t distribution with 4 degrees of freedom. The halfwidth of the confidence intervals is typically around 1 % with the largest one being less than 3 %. The halfwidth would be approximately $\sqrt{5} \approx 2.24$ times greater for a single replication, as in Table 2.

Next, Table 4 shows corresponding results for $GI/GI/s$ multi-server queues with $s = 40$. Many-server queues are more complicated, admitting a greater variety of behavior, including conventional heavy traffic for any s if the traffic intensity is high enough (but still less than 1). We again let the mean service time be 1. For $s = 40$,

Table 3 Estimated speed ratios of transitions between the 10th and 90th percentile of the steady-state distributions, i.e., $\omega(p)$ in (4.2) for $p = 0.1$, in nine $GI/GI/1$ single-server queues for $\rho = 0.9$

Single-server queue, $\rho = 0.9$								
	$\omega_1(0.1)$	$\omega_2(0.1)$	$\omega_3(0.1)$	$\omega_4(0.1)$	$\omega_5(0.1)$	$\bar{\omega}(0.1)$	CI halfwidth	$\frac{c_a^2+c_s^2}{2}$
$E_2/E_2/1$	0.548	0.555	0.550	0.551	0.551	0.551	0.003	0.500
$E_2/H_2(2)/1$	1.270	1.244	1.215	1.266	1.217	1.242	0.032	1.250
$E_2/M/1$	0.770	0.776	0.770	0.766	0.771	0.771	0.005	0.750
$H_2(2)/E_2/1$	1.156	1.136	1.168	1.155	1.159	1.155	0.015	1.250
$H_2(2)/H_2(2)/1$	1.841	1.894	1.878	1.846	1.902	1.872	0.034	2.000
$H_2(2)/M/1$	1.397	1.384	1.434	1.405	1.390	1.402	0.025	1.500
$M/E_2/1$	0.781	0.773	0.777	0.761	0.773	0.773	0.009	0.750
$M/H_2(2)/1$	1.437	1.495	1.486	1.470	1.484	1.474	0.029	1.500
$M/M/1$	0.998	1.003	1.013	0.990	0.988	0.998	0.013	1.000

Table 4 Estimated speed ratios of transitions between the 10th and 90th percentile of the steady-state distributions, i.e., $\omega(p)$ in (4.2) for $p = 0.1$, in nine $GI/GI/40$ queues with $\rho = 0.975$

Speed ratios of 10th-to-90th percentile transitions for $GI/GI/40$ models				
Traffic intensity $\rho = 0.975$				Approx.
Model	$E[T(p)]$, $p = 0.1$	CI halfwidth	$\bar{\omega}(0.1)$	$\frac{c_a^2+c_s^2}{2}$
$E_2/E_2/40$ True	164.7	6.2	0.53	0.50
$E_2/E_2/40$ BD	88.1	3.4		
$E_2/M/40$ True	266.9	11.5	0.76	0.75
$E_2/M/40$ BD	201.7	6.9		
$M/E_2/40$ True	252.7	11.8	0.78	0.75
$M/E_2/40$ BD	196.6	6.4		
$M/M/40$ True	350.8	36.7	1.00	1.00
$M/M/40$ BD	351.7	24.9		
$E_2/H_2(2)/40$ True	461.8	39.8	1.13	1.25
$E_2/H_2(2)/40$ BD	519.4	44.4		
$H_2(2)/E_2/40$ True	427.4	31.0	1.21	1.25
$H_2(2)/E_2/40$ BD	515.7	33.6		
$H_2(2)/M/40$ True	522.2	34.2	1.47	1.50
$H_2(2)/M/40$ BD	768.3	65.2		
$M/H_2(2)/40$ True	536.0	27.4	1.45	1.50
$M/H_2(2)/40$ BD	775.4	82.7		
$H_2(2)/H_2(2)/40$ True	753.8	56.8	1.81	2.00
$H_2(2)/H_2(2)/40$ BD	1, 363.5	104.2		

we let the arrival rate be 39, so that the traffic intensity is again $\rho = 0.975$. As discussed in [49], the typical traffic intensities in s-server queues tend to increase with s. Formula (1) of [49] shows that $\rho = 0.975$ for $s = 40$ should be roughly equivalent to $\rho = 0.8$ for $s = 1$. That high traffic intensity still supports a conventional heavy-traffic approximation.

The results in Table 4 are very similar to the previous results in Table 2, once again showing that approximation (1.2) with (1.3) is well justified. We see that the mean first passage times are less with 40 servers in Table 4 than for $s = 1$ in Table 2. That is to be expected because the arrival rate is approximately 40 times greater when $s = 40$; this explains the difference we see.

From many experiments for $GI/GI/s$ queues in the conventional heavy-traffic regime, like above, we find that $\bar{\omega}(p)$ tends to move away from 1 toward (1.3) as p decreases below 0.5 toward 0. When ρ is suitably large (in the conventional heavy-traffic regime) and p is suitably small, we find that the speed ratio can be approximated by (1.3).

For given ρ, we find that the speed ratio approaches 1 as p approaches 0.5. For given p, we find that the speed ratio approaches (1.3) as ρ increases. For example, for the $H_2(2)/M/1$ model with $\rho = 0.8$, the estimated speed ratios for $p = 0.1, 0.2,$

0.35, and 0.45 were, respectively, 1.29, 1.25, 1.16, and 1.01. On the other hand, for $\rho = 0.975$ all these estimated speed ratios fell in the interval [1.40, 1.50]. Similarly, for the $M/E_2/1$ model with $\rho = 0.8$, the estimated speed ratios for $p = 0.1$, 0.2, 0.35, and 0.45 were, respectively, 0.81, 0.84, 0.92, and 1.00. On the other hand, for $\rho = 0.975$ all these estimated speed ratios fell in the interval [0.75, 0.78].

As should be anticipated, given this systematic behavior as ρ increases, we find that insight can be gained from conventional heavy-traffic theory, which indicates that, if ρ is sufficiently high, then the stochastic process $\{Q(t) : t \geq 0\}$ in the $G/G/s$ model can be approximated by reflected Brownian motion (a Markov process); we discuss supporting heavy-traffic theory in Sect. 5.

4.2 Examples in the many-server heavy-traffic regime

To show the importance of the heavy-traffic regime, we illustrate with infinite-server queues having the same arrival rate $\lambda = 39$ as in Table 4. Given that the mean service time is fixed at $1/\mu = 1$, the arrival rate of $\lambda = 39$ is already quite large, so we shift from the conventional heavy-traffic regime to the many-server heavy-traffic regime as we increase the number of servers. The many-server heavy-traffic regime is characterized by $\lambda \to \infty$ and $s \to \infty$ with fixed μ such that $\sqrt{s}(1 - \rho) \to \beta < \infty$; see [21,31,35,51]. The essential character is seen by considering $s = \infty$. (It can be important to consider subcases, but we do not do so here.)

Table 5 shows the results of the simulation experiment in the infinite-server cases. First, we see that the mean first passage times between the percentiles are less in Table 5 than in Table 4, even though the arrival rate is $\lambda = 39$ in both cases. That can be explained because the steady-state distribution is more concentrated (less variable) with infinitely many servers. The extra servers prevent excursions to very large values.

As before, the estimated speed ratios $\bar{\omega}(p)$ differ from 1 as p moves away from 0.5 toward 0. We find that the speed ratios in the infinite-server case do match the formulas above for the $GI/M/\infty$ models, but not for the other queues with non-exponential service distributions. For example, for the $H_2(2)/M/s$ model with $\lambda = 39$ and $\mu = 1$, we have $\bar{\omega}(0.1) = 1.49$ when $s = 40$ and 1.39 when $s = \infty$. In striking contrast, for the $M/H_2(2)/s$ model with $\lambda = 39$ and $\mu = 1$, we have $\bar{\omega}(0.1) = 1.42$ when $s = 40$ and 0.87 when $s = \infty$. In both cases, approximation (1.3) yields $\omega = (c_a^2 + c_s^2)/2 = 1.5$. For the $M/H_2(2)/\infty$ case, the estimated speed ratio is not close to 1.5, but less than 1, and thus in the wrong direction away from 1.

We do see impressive regularity in Table 5 beyond the $GI/M/\infty$ model for which approximation (1.3) remains good. More generally, we observe that the estimated speed ratios $\bar{\omega}(0.1)$ are consistently increasing in c_a^2 but *decreasing* in c_s^2.

For infinite-server queues and associated many-server queues in the many-server heavy-traffic regime, we again suggest the approximation in (1.2), but with (1.3) only in the case of $G/M/\infty$ and $G/M/s$ queues. Insight into the good performance of (1.2) with (1.3) for $G/M/\infty$ and $G/M/s$ queues can be gained from many-server heavy-traffic limits, indicating that the stochastic process $\{Q(t) : t \geq 0\}$ can be approximated by $\rho + \sqrt{\rho}X(t)$, where $\rho \equiv \lambda/\mu$ and $X(t)$ is an Ornstein-Uhlenbeck (OU) diffusion process in the infinite-server case; see [45]; we discuss in Sect. 6. We also explain the

Table 5 Estimated speed ratios of transitions between the 10th and 90th percentile of the steady-state distributions, i.e., $\omega(p)$ in (4.2) for $p = 0.1$, in nine infinite-server queues with $\lambda = 39$ and $\mu = 1$

Speed ratios of 10th-to-90th percentile transitions for $GI/GI/\infty$ models				
$\lambda = 39$				Approx.
Model	$E[T(p)]$, $p = 0.1$	CI halfwidth	$\bar{\omega}(0.1)$	$\frac{c_a^2+c_s^2}{2}$
$E_2/E_2/\infty$ True	7.3	0.02	0.87	0.50
$E_2/E_2/\infty$ BD	6.4	0.03		
$E_2/M/\infty$ True	8.5	0.06	0.78	0.75
$E_2/M/\infty$ BD	6.6	0.03		
$M/E_2/\infty$ True	7.4	0.03	1.17	0.75
$M/E_2/\infty$ BD	8.7	0.02		
$M/M/\infty$ True	8.7	0.01	1.00	1.00
$M/M/\infty$ BD	8.7	0.04		
$E_2/H_2(2)/\infty$ True	9.7	0.07	0.67	1.25
$E_2/H_2(2)/\infty$ BD	6.5	0.04		
$H_2(2)/E_2/\infty$ True	7.9	0.02	1.74	1.25
$H_2(2)/E_2/\infty$ BD	13.7	0.17		
$H_2(2)/M/\infty$ True	9.8	0.02	1.39	1.50
$H_2(2)/M/\infty$ BD	13.6	0.06		
$M/H_2(2)/\infty$ True	10.0	0.07	0.87	1.50
$M/H_2(2)/\infty$ BD	8.7	0.07		
$H_2(2)/H_2(2)/\infty$ True	9.6	0.08	1.25	2.00
$H_2(2)/H_2(2)/\infty$ BD	12.0	0.10		

anomalous behavior with non-exponential service times in Sect. 6, which is related to previous observations, for example, [33,58], and can be explained by the more complex heavy-traffic limit, which is only Markov when viewed in a higher dimension [32,35].

5 Conventional heavy-traffic approximations

In this section we show that the regularity observed in Sect. 4.1 can be explained by conventional heavy-traffic (HT) limits for the $GI/GI/s$ model, as in [2,11,24,25,50]. We create a family of rate-λ arrival processes by scaling time in a fixed rate-1 arrival process A; i.e., we let

$$A_\lambda(t) \equiv A(\lambda t), \quad t \geq 0. \tag{5.1}$$

We leave the service process unchanged. We assume that the fixed interarrival-time and service-time distributions have finite variance. We now assume that $\lambda \uparrow s\mu$, where μ is the fixed individual service rate, so that $\rho \equiv \rho(\lambda) \equiv \lambda/s\mu \uparrow 1$. Hence, the systems can be indexed by ρ, where $\rho \uparrow 1$.

As indicated in Sect. 2.1, it is important to note that the key parameters $\bar{\lambda}$, $\bar{\mu}$, $\bar{\rho}$, and $\bar{\omega}$ estimated in Table 1, as well as the fitted birth rates $\bar{\lambda}_k$ and the fitted death rates $\bar{\mu}_k$, depend on ρ. (This can be seen by comparing the cases $\rho = 0.8$ and $\rho = 0.9$ in Tables 1 and 6.)

We now need to understand how $\bar{\rho}(\rho) \equiv \bar{\lambda}(\rho)/\bar{\mu}(\rho)$ changes with ρ. For that purpose, we assume that the geometric tail in (3.7) is valid for each ρ. We also assume that the estimated rates have limits as $k \to \infty$, i.e.,

$$\bar{\lambda}_k(\rho) \to \bar{\lambda}(\rho) \quad \text{and} \quad \bar{\mu}_k(\rho) \to \bar{\mu}(\rho) \quad \text{as} \quad k \to \infty \quad \text{for each} \quad \rho. \tag{5.2}$$

The value of the asymptotic relation (3.7) is enhanced by heavy-traffic expansions given in [2,11], which we assume are valid here as well. We state the one for the $GI/GI/s$ model from [2]. Let u_k and v_k be the k^{th} moments of the mean-1 random variables $U/E[U] = \lambda U$ and $V/E[V] = \mu V$, where U is a generic interarrival time and V is a generic service time. Under quite general regularity conditions, as $1-\rho \downarrow 0$,

$$\sigma(\rho) = 1 - \frac{2(1-\rho)}{c_a^2 + c_s^2} + \left(\frac{8(d_s - d_a)}{(c_a^2 + c_s^2)^3} - \frac{2(c_a^2 - 1)}{(c_a^2 + c_s^2)^2} \right) (1-\rho)^2 + O((1-\rho)^3), \tag{5.3}$$

where $c_a^2 \equiv u_2 - u_1^2 = u_2 - 1$ and $c_s^2 \equiv v_2 - v_1^2 = v_2 - 1$ are the squared coefficients of variation (scv's, variances divided by the square of the mean) of the interarrival time λU and service time μV, respectively, while d_a and d_s are parameters based on the first three moments of λU and μV, namely,

$$d_a \equiv \frac{u_3 - 3c_a^2(c_a^2 + 1) - 1}{6} \quad \text{and} \quad d_s \equiv \frac{v_3 - 3c_s^2(c_s^2 + 1) - 1}{6}. \tag{5.4}$$

We thus can combine previous results to obtain the asymptotic behavior of $\bar{\rho}(\rho) \equiv \bar{\lambda}(\rho)/\bar{\mu}(\rho)$, which we regard as the traffic intensity of the fitted BD model.

Theorem 5.1 (HT limit for the fitted traffic intensity of the fitted BD model) *If a family of $G/G/s$ models is created using (5.1), where (5.2), (3.7), and (5.3) are valid, then*

$$\bar{\rho}(\rho) \equiv \frac{\bar{\lambda}(\rho)}{\bar{\mu}(\rho)} = \sigma(\rho) \quad \textit{for each} \quad \rho, \tag{5.5}$$

where $\sigma(\rho)$ is determined by (3.7) and

$$\bar{\omega}(\rho) \equiv \frac{1-\rho}{1-\bar{\rho}(\rho)} = \omega + O(1-\rho) \quad \textit{as} \quad \rho \uparrow 1. \tag{5.6}$$

for $\omega \equiv (c_a^2 + c_s^2)/2$ in (1.3).

Proof Combine (1.1) with (3.7) to get (5.5). Then apply (5.3) to get (5.6). □

We anticipate that, under regularity conditions, the entire fitted BD process should have the same HT limit as the queue-length process in an $M/M/1$ queue with constant arrival rate $\bar{\lambda}(\rho)$, constant service rate $\bar{\mu}(\rho)$, and traffic intensity $\bar{\rho} \equiv \bar{\rho}(\rho)$, when we let $\bar{\rho}(\rho) \uparrow 1$ as $\rho \uparrow 1$ as in (5.6). In addition to (5.2), we now also need an additional condition, which we formulate in the following conjecture. It is a generalization of (2.1), which holds for all $GI/M/s$ models by Theorem 3.1.

Conjecture 5.1 (conventional HT limit for fitted birth and death rates) *If the conditions of Theorem 5.1 hold, then there exist finite constants $\dot{\lambda}(1-)$, $\dot{\mu}(1-)$ and k_0, depending on the arrival and service processes, such that $\dot{\lambda}(1-) > \dot{\mu}(1-)$,*

$$\sup_{k:k\ge k_0} \{|\bar{\lambda}_{\rho,k} - [\mu s - \dot{\lambda}(1-)(1-\rho)]|\} = o(1-\rho) \quad as \quad \rho \uparrow 1, \quad and$$
$$\sup_{k:k\ge k_0} \{|\bar{\mu}_{\rho,k} - [\mu s - \dot{\mu}(1-)(1-\rho)]|\} = o(1-\rho) \quad as \quad \rho \uparrow 1. \tag{5.7}$$

The leading term $s\mu$ of both functions $\bar{\lambda}_{\rho,k}$ and $\bar{\mu}_{\rho,k}$ in (5.7) is the overall maximum possible service rate. The notation in (5.7) suggests that the functions $\bar{\lambda}(\rho)$ and $\bar{\mu}(\rho)$ in (5.2) are differentiable functions of ρ for $0 < \rho < 1$, with the derivatives having finite left limits $\dot{\lambda}(1-)$ and $\dot{\mu}(1-)$ as $\rho \uparrow 1$, which we also conjecture is the case, but that structure is not required.

We now establish a heavy-traffic limit under the conditions of Conjecture 5.1. Let D be the standard function space of right-continuous functions on $[0, \infty)$ with left limits everywhere and with the Skorohod topology and let $\Rightarrow$ denote convergence in distribution in the function space D, as in [50]. Let $R \equiv \{R(t; a, b) : t \ge 0\}$ be a reflected Brownian motion (RBM) with drift coefficient a and diffusion coefficient b.

Theorem 5.2 (HT limit for fitted BD process) *If the conditions and the conclusion of Conjecture 5.1 hold, then*

$$\{(1-\rho)Q_{f,\rho}((1-\rho)^{-2}t) : t \ge 0\}$$
$$\Rightarrow \{R(t; -\mu s/\omega, 2\mu s) : t \ge 0\} \quad in \quad D \quad as \quad \rho \uparrow 1. \tag{5.8}$$

Proof We do the proof in three steps. First, we do the proof for the case in which

$$\bar{\lambda}_{\rho,k} = \mu s - \dot{\lambda}(1-)(1-\rho) \quad \text{and} \quad \bar{\mu}_{\rho,k} = \mu s - \dot{\mu}(1-)(1-\rho) \quad \text{for all} \quad k \tag{5.9}$$

and then in the remaining two steps we show that the limit has to be the same as under condition (5.9). The condition (5.9) corresponds to a family of $M/M/1$ models with the specified arrival and service rates, for which established heavy-traffic limits in [24,25,50] imply convergence to RBM with drift coefficient $-(\dot{\lambda}(1-) - \dot{\mu}(1-))$ and variance coefficient $2\mu s$. However, Theorem 5.1 and simple algebra imply that $\dot{\lambda}(1-) - \dot{\mu}(1-) = \mu s/\omega$. Hence, we get the limit in (5.8).

For the second step, we assume that the relation in (5.7) holds for all k, not just for $k \geq k_0$. Under this strengthened version of condition (5.7), we can bound the fitted birth rates and death rates above and below arbitrarily closely, in the sense that, for any $\epsilon > 0$, we can find $\rho(\epsilon)$, $\dot{\lambda}_L(1-)$, $\dot{\lambda}_U(1-)$, $\dot{\mu}_L(1-)$, and $\dot{\mu}_U(1-)$ such that $0 \leq \rho(\epsilon) < 1$, $\dot{\lambda}_U(1-) - \dot{\lambda}_L(1-) < \epsilon$, $\dot{\mu}_U(1-) - \dot{\mu}_L(1-) < \epsilon$,

$$\begin{aligned} s\mu - \dot{\lambda}_U(1-)(1-\rho) \leq \bar{\lambda}_{\rho,k} \leq s\mu - \dot{\lambda}_L(1-)(1-\rho) \quad \text{and} \\ s\mu - \dot{\mu}_U(1-)(1-\rho) \leq \bar{\mu}_{\rho,k} \leq s\mu - \dot{\mu}_L(1-)(1-\rho) \\ \text{for all} \quad k \geq 1 \quad \text{and} \quad \rho > \rho(\epsilon). \end{aligned} \tag{5.10}$$

We can then construct upper and lower bounds for the family of BD processes $Q_{f,\rho}(t)$ in sample path stochastic order; i.e., we can construct three processes $\tilde{Q}_{f,\rho}(t)$, $\tilde{Q}_{f,\rho,L}(t)$, and $\tilde{Q}_{f,\rho,U}(t)$ such that

$$\tilde{Q}_{f,\rho,L}(t) \leq \tilde{Q}_{f,\rho}(t) \leq \tilde{Q}_{f,\rho,U}(t) \quad \text{for all} \quad t \quad \text{w. p. 1},$$

where all three processes have the proper distributions as BD processes, with $\{\tilde{Q}_{f,\rho}(t) : t \geq 0\}$ distributed the same as $\{Q_{f,\rho}(t) : t \geq 0\}$. This is done by generating all transitions from common Poisson processes. We let higher processes have births whenever lower ones do; we let lower processes have deaths whenever higher processes do; in this way the sample paths stay ordered. By choosing ϵ suitably small, the three processes can be made to have identical sample paths over any finite interval $[0, t]$ with probability arbitrarily close to 1. As in [43], the upper-bound BD process has the upper-bound birth rates $s\mu - \dot{\lambda}_L(1-)(1-\rho)$ and the lower-bound death rates $s\mu - \dot{\mu}_U(1-)(1-\rho)$, while the lower-bound BD process has the lower-bound birth rates $s\mu - \dot{\lambda}_U(1-)(1-\rho)$ and the upper-bound death rates $s\mu - \dot{\mu}_L(1-)(1-\rho)$. These bounding BD processes satisfy the assumptions of the first part, but with different parameters, and so have heavy-traffic limits. By letting $\epsilon \downarrow 0$, we can sandwich the process $\tilde{Q}_{f,\rho}(t)$ between these bounding processes and obtain the limit in the first step.

For the third step, we observe that the limit under condition (5.7) must be the same as in the first two steps because the processes $Q(t)$ and $Q_{\mathrm{f}}(t)$ spend asymptotically negligible time in the states with $k \leq k_0$. The space scaling thus makes the difference asymptotically negligible; i.e., we can apply the "convergence-together theorem," Theorem 11.4.7 of [50]. □

Recall from [25] that the heavy-traffic limit of the scaled queue-length process in the $G/G/s$ model is

$$\{(1-\rho)Q((1-\rho)^{-2}t) : t \geq 0\} \Rightarrow \{\omega R(t; -s\mu, 2s\mu\omega) : t \geq 0\} \quad \text{in} \quad D, \tag{5.11}$$

which differs from (5.8) by having the variability parameter ω as part of the diffusion coefficient instead of part of the drift.

We now show that these approximating RBM's can be related as in (1.2). To see this directly, it is convenient to rescale these limiting RBM's to canonical RBM's (with drift coefficient -1 and diffusion coefficient 1), for example, using (25) of [48].

Corollary 5.1 (simple time transformation) *The two approximating RBM's in* (5.8) *and* (5.11) *correspond to the following scaled canonical RBM's:*

$$\{R(t; -\mu s/\omega, 2\mu s) : t \geq 0\} \stackrel{\mathrm{d}}{=} \{2\omega R(\mu st/(2\omega^2); -1, 1) : t \geq 0\} \quad \textit{and}$$
$$\{\omega R(t; -s\mu, 2s\mu\omega) : t \geq 0\} \stackrel{\mathrm{d}}{=} \{2\omega R(\mu st/(2\omega); -1, 1)) : t \geq 0\}. \tag{5.12}$$

The representation (5.12) *implies that the two RBM's have a common exponential steady-state distribution with mean* ω *and that, under the conditions of Theorem 5.2, the simple time transformation in* (1.2) *is asymptotically correct* (*in distribution*) *in the heavy-traffic limit.*

Proof Since $RBM(t; -a, b)$ with $a, b > 0$ has an exponential steady-state distribution with mean $b/2a$, the RBM's in (5.8), (5.11), and (5.12) all have an exponential steady-state distribution with mean ω, consistent with the fact the the fitted BD process must have the same steady-state distribution as the original queue-length process. For the transient behavior, we see that the two RBM's in (5.12) are identical except that the RBM limit for the fitted BD process has an extra ω in the denominator of the time scaling. □

We have numerically studied this effect via the speed ratios. Let $T_\rho(p)$ be the first passage time from percentile p of the stationary distribution to percentile $1 - p$ of the stationary distribution and then back in the original model, and let $T_{f,\rho}(p)$ be the corresponding round-trip first passage time in the fitted BD process. We can apply the limits in (5.8) and (5.11) with (5.12) to obtain corresponding limits for the first passage times and then the speed ratios. The results above immediately imply the following corollary.

Corollary 5.2 (HT limit for the speed ratios) *If the limits in* (5.8) *and* (5.11) *hold, then*

$$\frac{T_{f,\rho}(p)}{T_\rho(p)} \Rightarrow \frac{c_a^2 + c_s^2}{2} \equiv \omega \quad as \quad \rho \uparrow 1 \tag{5.13}$$

and, assuming associated uniform integrability,

$$\frac{E[T_{f,\rho}(p)]}{E[T_\rho(p)]} \to \frac{c_a^2 + c_s^2}{2} \equiv \omega \quad as \quad \rho \uparrow 1 \tag{5.14}$$

for all p *with* $0 < p < 1/2$ *as* $\rho \uparrow 1$.

Proof We can establish limits for the first passage times of each process separately using the standard continuous mapping theorem argument, provided that the state space is scaled consistently with the HT scaling; see Sect. 5.7.5 of [50]. We achieve that scaling of the state space by looking at the first passage times between percentiles of the stationary distribution, which is the same for both processes. □

We conclude this section by giving two numerical examples.

Example 5.1 (the case of $M/H_2/1$) We now illustrate by considering the $M/H_2/1$ model with $\lambda = 0.8$, $\mu = 1$, and an H_2 service-time cdf $F_s(x)$ with scv $c_s^2 = 2$ and balanced means. For this model, the steady-state distribution is given by the Pollaczek-Khintchine generating function; see Sect. 5.8 of [13]. The mean has the well-known explicit form $E[Q(\infty)] = \rho + (\rho^2(c_s^2 + 1)/2(1 - \rho))$. For $M/M/1$ the mean is 4.8; for $M/H2/1$ with $c_s^2 = 2$, the mean is 5.6. For $M/H2/1$, α_k decays somewhat more slowly.

We simulated this $M/H_2/1$ model. The estimated birth rates are essentially exact, fitting a constant function well, while the estimated death rates were $\bar{\mu}_1 = 1.181$, $\bar{\mu}_2 = 1.072$, $\bar{\mu}_3 = 0.999$, $\bar{\mu}_4 = 0.965$ and $\bar{\mu}_k$ approaching the constant value of about 0.925. Thus, the estimated value of σ in Theorem 3.5 is $\bar{\lambda}_k/\bar{\mu}_{k+1} \approx 0.8/0.925 \approx 0.865$, which is consistent with the first term in the heavy-traffic approximation, 0.867.

Example 5.2 (the case of $M/H_2/40$) We also illustrate an example with $s > 1$ by considering the $M/H_2/40$ model with $\lambda = 39$, $\mu = 1$, and the same H_2 service-time cdf as in Example 5.1. Here $\rho = 39/40 = 0.975$. we simulated this $M/H_2/40$ model, as described in the next section. Our simulation experiment confirms that the fitted birth rates are exact. Over the region most frequently visited, i.e., [20, 100], the death rates $\bar{\mu}_k(\infty)$ were first slightly greater than the corresponding $M/M/s$ death rates but then eventually lower, with the crossover point being $k = 52$. For example, $\bar{\mu}_{30}(\infty) = 30.31 > 30$, $\bar{\mu}_{35}(\infty) = 35.59 > 35$, $\bar{\mu}_{40}(\infty) = 40.318 > 40$ with a peak at $\bar{\mu}_{42}(\infty) = 40.758$ and then steadily but gradually declining with $\bar{\mu}_{52}(\infty) \approx 40.000$ and approaching $\bar{\mu}_k(\infty) = 39.67$ for large k. Thus, the estimated value of σ in Theorem 3.5 is $\bar{\lambda}_k(\infty)/\bar{\mu}_{k+1}(\infty) \approx 39.0/39.67 \approx 0.9831$, which is consistent with the first term in the heavy-traffic approximation, 0.9833.

6 Heavy-traffic approximations for infinite-server models

We now discuss the results for infinite-server queues in Table 5.

6.1 Heavy-traffic limit for the fitted BD process

We have seen that the fitted birth and death rates in a $GI/GI/\infty$ infinite-server model tend to have, at least approximately, the structure in equations (2.3) and (2.5). To gain insight, we establish many-server heavy-traffic limits for such BD processes. Let $\{Q_n(t) : t \geq 0\}$ be the BD stochastic process as a function of n, $n \geq 1$, and consider the scaled stochastic processes defined by

$$\hat{Q}_n(t) \equiv n^{-1/2}(Q_n(t) - n), \quad t \geq 0, \quad n \geq 1. \tag{6.1}$$

Theorem 6.1 (HT limit for BD processes with asymptotically linear birth rate) *Consider a sequence of fitted BD processes associated with a $GI/GI/\infty$ model indexed by the arrival rate n, satisfying Conjectures 2.1 and 2.2, which includes the special*

case in (2.3) *and* (2.5). *If*

$$\hat{Q}_n(0) \Rightarrow L \quad in \quad \mathbb{R} \quad as \quad n \to \infty, \tag{6.2}$$

where $P(L < \infty) = 1$, *then*

$$\{\hat{Q}_n(t) : t \geq 0\} \Rightarrow \{OU(t; 1-b, 2) : t \geq 0\} \quad in \quad D, \tag{6.3}$$

where $OU(0) \stackrel{\mathrm{d}}{=} L$ *and* $\{OU(t; 1-b, 2) : t \geq 0\}$ *is an Ornstein-Uhlenbeck (OU) diffusion process evolving independently conditional on* $OU(0)$, *with drift function* $\mu(x) = -(1-b)x$ *and diffusion function* $\sigma^2(x) = \sigma^2 = 2$, *which has a Gaussian steady-state distribution with mean* 0 *and variance* $1/(1-b)$.

Proof We follow the approach used by Iglehart [23] for the $M/M/\infty$ model and apply the limit theorem for BD processes developed by Stone [40]. To apply that theorem here, it suffices to show that the infinitesimal means and variances of the BD processes converge to the drift and diffusion functions of the OU diffusion process. For the infinitesimal mean, let x_n be a possible value of $\hat{Q}_n(t)$ such that $x_n \to x$ as $n \to \infty$. Then

$$\begin{aligned} m_n(x_n) &\equiv \lim_{h \to 0} E[(\hat{Q}_n(t+h) - \hat{Q}_n(t))/h | \hat{Q}_n(t) = x_n] \\ &= \lim_{h \to 0} E[(Q_n(t+h) - Q_n(t))/h\sqrt{n} | Q_n(t) = n + \sqrt{n}x_n] \\ &= \frac{\lambda_{n,n+\sqrt{n}x_n} - \mu_{n+\sqrt{n}x_n}}{\sqrt{n}} \\ &= \frac{n + b(n + \sqrt{n}x_n - n) - (n + \sqrt{n}x_n) + o(\sqrt{n})}{\sqrt{n}} \\ &= (b-1)x_n + o(1) \to (b-1)x \quad \text{as} \quad n \to \infty. \end{aligned} \tag{6.4}$$

For the infinitesimal variance,

$$\begin{aligned} \sigma_n^2(x_n) &\equiv \lim_{h \to 0} E[(\hat{Q}_n(t+h) - \hat{Q}_n(t))^2/h | \hat{Q}_n(t) = x_n] \\ &= \lim_{h \to 0} E[(Q_n(t+h) - Q_n(t))^2/hn | Q_n(t) = n + \sqrt{n}x_n] \\ &= \frac{\lambda_{n,n+\sqrt{n}x_n} + \mu_{n+\sqrt{n}x_n}}{n} \\ &= \frac{n + b(n + \sqrt{n}x_n - n) + (n + \sqrt{n}x_n) + o(\sqrt{n})}{n} \to 2 \quad \text{as} \quad n \to \infty. \end{aligned} \tag{6.5}$$

It is well known that the steady-state distribution of the OU process is Gaussian with mean 0 and variance $\sigma^2/2|\mu|$, which here takes the value $1/(1-b)$. □

Corollary 6.1 (identifying the constant b) *To have the same steady-state distribution as the $GI/M/\infty$ model, the drift of the limiting OU process in Theorem 6.1 must be $\mu(x) = -1/\omega$, where $\omega \equiv (c_a^2 + 1)/2$ is the speed ratio.*

Proof Theorem 1 of [45] implies that the HT limit for the $GI/M/\infty$ model is also an OU process with mean 0 but variance equal to $\omega \equiv (c_a^2 + 1)/2$. Since the steady-state distribution of the fitted BD process coincides with the steady-state distribution of the $GI/M/\infty$ model as the sample size increases, if the birth rates are as in (2.5), then we must have

$$\frac{1}{1-b} = \omega \quad \text{or} \quad b - 1 = \frac{1}{\omega} = \frac{2}{c_a^2 + 1}. \tag{6.6}$$

□

Since the speed ratio is defined in terms of the first passage times, we need to exploit associated HT limits for the first passage times. We first need to establish a corresponding limit for the steady-state distributions.

Theorem 6.2 (HT limit for steady-state distributions) *Under the conditions of Theorem 6.1,*

$$\hat{Q}_n(t) \Rightarrow \hat{Q}_n(\infty) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad t \to \infty, \tag{6.7}$$

where $\hat{Q}_n(\infty)$ is a proper random variable for each n and

$$\hat{Q}_n(\infty) \Rightarrow OU(\infty; 1-b, 2) \stackrel{\mathrm{d}}{=} N(0, 1/(1-b)) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty. \tag{6.8}$$

Proof We follow the argument used to prove Theorem 2 of [45]. We first prove that (6.7) holds and show that the sequence $\{\hat{Q}_n(\infty) : n \geq 1\}$ is tight. We give a detailed derivation under the simplified structure in (2.3) and (2.5). The same conclusions can be shown to hold under Conjectures 2.1 and 2.2. Toward that end, note that, from (2.5) and (2.3), the stochastic process $|Q_n(t) - n|$ is directly a BD process on the nonnegative integers for each n with birth rate $b_{n,k} \equiv n + bk$ and death rate $d_{n,k} \equiv n + k$, where $-\infty < b < 1$, so that we can directly calculate its steady-state distribution using the familiar BD formulas

$$\alpha_{n,k} \equiv \lim_{t\to\infty} P(Q_n(t) = k) = \frac{r_{n,k}}{\sum_{j=0}^{\infty} r_{n,j}}, \tag{6.9}$$

where $r_{n,0} \equiv 1$ and

$$r_{n,k} = \frac{b_{n,0} \times \cdots \times b_{n,k-1}}{d_{n,1} \times \cdots \times d_{n,k}} = \frac{n(n+b) \times \cdots \times (n+k-1)}{(n+1) \times \cdots \times (n+k)} \tag{6.10}$$

for $k \geq 1$. Hence,

$$\log_{\mathrm{e}} r_{n,k} = \sum_{j=0}^{k-1} \log_{\mathrm{e}} (1 + (jb/n)) - \sum_{j=1}^{k} \log_{\mathrm{e}} (1 + (j/n)), \tag{6.11}$$

so that

$$\log_{\mathrm{e}} r_n, k \sim \sum_{j=0}^{k-1} (jb/n) - \sum_{j=1}^{k} (j/n) = \frac{b(k-1)k}{2n} - \frac{k(k+1)}{2n} \quad \text{as} \quad n \to \infty \tag{6.12}$$

and

$$r_{n,c\sqrt{n}} \sim \mathrm{e}^{-(1-b)c^2/2} \quad \text{for} \quad b < 1. \tag{6.13}$$

From (6.13), we can deduce that $|Q_n(t) - n|$ has a proper steady-state limit $|Q_n(\infty) - n|$ for all n sufficiently large, and that the sequence $\{n^{-1/2}|Q_n(\infty) - n| : n \geq 1\}$ is tight in $\mathbb{R}$. That implies that the sequence $\{\hat{Q}_n(\infty) : n \geq 1\}$ is tight. By symmetry, we see that $Q_n(t) - n$ has a steady-state distribution for each n, which implies that $Q_n(t)$ does as well. From the tightness of $\{\hat{Q}_n(\infty) : n \geq 1\}$, we know that there is a convergent subsequence. We consider such a convergent subsequence and make those terms the initial values $\hat{Q}_n(0)$ in (6.2). Those initial values make the scaled BD processes strictly stationary processes for each n. However, we can also invoke Theorem 6.1. That allows us to identify the limit of the convergent subsequence of $\{\hat{Q}_n(\infty) : n \geq 1\}$. Since every convergent subsequence must have that same limit, we actually have convergence, i.e., we have proved (6.8). □

Again, let T_p be the round-trip first passage time from the p^{th} percentile of the steady-state distribution to the $(1-p)^{\text{th}}$ percentile and back. Just as for Corollary 5.2 in the conventional heavy-traffic regime, the asymptotics for the speed ratio here in the many-server heavy-traffic regime depend on key differences in the heavy-traffic diffusion stochastic-process limits for the fitted BD process and the original queueing process. We have already observed that both limit processes are OU diffusion processes, but the parameters of those OU processes are different. For the fitted BD process, the drift and diffusion parameters were $-(1-b) = -2/(1+c_a^2)$ and 2, respectively. From Theorem 1 of [45], we see that for the original $GI/M/\infty$ queue-length process, both of these parameters are multiplied by $\omega \equiv (1+c_a^2)/2$; the drift and diffusion parameters $GI/M/\infty$ queue-length process are -1 and $(1+c_a^2)$. This change makes the asymptotic value of the speed ratio ω, as we now show.

Corollary 6.2 *Under the conditions of Theorem (6.1),*

$$T_p(Q_n) \Rightarrow T_p(OU) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty, \tag{6.14}$$

where Q_n is the BD process with the fitted rates in (2.5) and (2.3), $OU \equiv \{OU(t; 1-b, 2) : t \geq 0\}$ and

$$E[T_p(OU(\cdot, \mu, \sigma^2)] = \frac{E[T_p(OU(\cdot, 1, 2))]}{1-b}. \tag{6.15}$$

As a consequence, the speed ratios in the $G/M/\infty$ model satisfy

$$\omega_n \to \omega \equiv \frac{c_a^2 + 1}{2} \quad as \quad n \to \infty.$$

Proof By Theorem 6.2, the steady-state distributions converge. Hence, the percentiles of the scaled process converge to the percentiles of the limit process. Thus, the limit in (6.14) follows from the continuous mapping theorem, as on p. 447 of [50]. Equation (6.15) follows from [10]; see also (2.3) of [41]. By focusing on the first passage times from one percentile to another, we automatically achieve the space and time transformation in (3.4)–(3.7) of [10]. □

6.2 Supporting theory for $M/GI/\infty$ models

We now discuss the anomalous behavior observed in Table 5 for the $M/GI/\infty$ models. First, Theorem 3.4 implies for $M/GI/\infty$ models that the birth and death rates coincide with those in the associated $M/M/\infty$ BD model. In contrast, for the $GI/GI/s$ model in the conventional heavy-traffic regime, the mean steady-state number in the system tends to be directly proportional to the speed ratio $\omega = (c_a^2 + c_s^2)/2$. Thus, the distance between the percentiles of the steady-state distribution is radically different for the $M/GI/\infty$ model.

The transient behavior of the $M/GI/\infty$ model with large arrival rate is also more complicated than for the other models. For the $GI/M/\infty$ model and for the $GI/GI/s$ model in the conventional heavy-traffic regime, the heavy-traffic approximations are Markov processes. In contrast, the transient behavior of the $M/GI/\infty$ model depends strongly on the elapsed service times (ages) of the customers in service. Simple approximations are less likely to be effective because the heavy-traffic approximating processes are not Markov unless the dimension of the state is increased, as shown in [32,35] and references therein. Our approach based on first passage times does not keep track of that full expanded state.

Nevertheless, the transient behavior of the $M/GI/\infty$ model can be understood and approximated, as shown in [17,18], but it leads to a different story. It does *not* support approximation (1.2) with (1.3). Suppose that S denotes a service time with cdf G. First, we have just noted that the percentiles of the stationary distribution are independent of the service-time distribution beyond its mean. However, as indicated in Theorem 1 of [18] and Sect. 2 of [17], the transient behavior depends largely on the service-time cdf G through the stationary-excess cdf associated with G, denoted by G_e. Let S_e denote a random variable with cdf G_e. Then

$$G_{\rm e}(t) \equiv P(S_{\rm e} \le t) = \frac{1}{E[S]} \int_0^t P(S > u)\, {\rm d}u \quad \text{with} \quad E[S_{\rm e}] = \frac{E[S](1+c_s^2)}{2}. \tag{6.16}$$

Thus, if $c_s^2 > 1$ as for the H_2 cdf, $E[S_{\rm e}] > E[S]$, while if $c_s^2 < 1$ as for the E_2 cdf, $E[S_{\rm e}] < E[S]$.

Understanding of the transient behavior of these $G/GI/\infty$ infinite-server models is enhanced by recognizing that new arrivals can be treated separately from old content. First, suppose that we consider the evolution of the customers in the system, given that we consider the system in steady state under the condition that we start at a high percentile of the stationary distribution (for example, the 90$^{\rm th}$ percentile). Theorem 1 [17] implies that the remaining service times of the customers in service are distributed as i.i.d. random variables distributed according to $S_{\rm e}$. This implies that if the service-time distribution is H_2 with $c_s^2 > 1$, then the time for the excess number of customers to depart will tend to be longer than if the service time were exponential. If we start with n busy servers, then the mean number of these servers still busy with the same customers at time t is $nG_{\rm e}(t)$. These results indicate that, from this perspective, the transient behavior of the $M/GI/\infty$ model with H_2 service will be *slower* than if the service time were exponential, consistent with our simulation results.

Now consider the new arrivals, starting empty. That is unambiguously defined with a Poisson arrival process. The mean number of busy servers (which coincides with the number of customers in the system) increases toward its steady-state limit $\lambda E[S]$. As shown in (20) of [18], the expected proportion of this mean achieved by time t turns out be equal to $G_{\rm e}(t)$. Since this cdf has mean $E[S_{\rm e}] = E[S](c_s^2+1)/2$, we see that the mean increases more slowly than would be the case if $c_s^2 = 1$.

This analysis of the $M/GI/\infty$ model does not apply directly to the first passage times, which tend to depend on the ages of the customers in service. Nevertheless, this analysis leads to the conclusion that, roughly, the process $Q(t)$ with H_2 service moves more slowly than the corresponding process with M service by a factor of $(c_s^2+1)/2$, which is exactly the speed ratio in (1.3). That supports approximation (1.2) but with a speed ratio equal to the *reciprocal* of the formula in (1.3), which is roughly consistent with the simulation results in Table 5.

It remains to consider more general $G/GI/\infty$ models. Given research on those models in [32–34], we anticipate the story will be more complicated.

7 Estimation methods

We now discuss two estimation issues: (i) estimating the speed ratios and (ii) estimating the birth and death rates of the fitted model.

7.1 Estimating the speed ratios

We clearly can estimate the mean round-trip first passage times $E[T(p)]$ and $E[T_{\rm f}(p)]$ in Sect. 4 by the sample averages of successive observed passage times, just as we

did in the simulation experiments. That leads directly to estimates of the speed ratios $\omega(p)$ in (4.2). However, there is an alternative approach via the arrival and service counting processes $A(t)$ and $S(t) \equiv D(t)/B(t)$ defined in Sect. 2.1 that allows for dependence among interarrival times and service times. The key is the observation that the heavy-traffic limit theorems characterizing the asymptotic speed ratio ω actually depend on the arrival process through the FCLT for $A(t)$, and similarly for the service process $S(t)$ when $s < \infty$; see [25,32].

To account for the dependence, it is natural to work with indices of dispersion, as in [14,20,39] and Sect. 9.6 of [50]. Given a stationary sequence of interarrival times, the *index of dispersion for intervals* (IDI) is defined as

$$I_i(n) \equiv \frac{Var(U_1 + \cdots + U_n)}{nE[U_n]^2}, \quad n \geq 1, \tag{7.1}$$

while the associated *index of dispersion for counts* (IDC) is defined as

$$I_c(t) \equiv \frac{\mathrm{Var}(A(t))}{E[A(t)]}, \quad t \geq 0, \tag{7.2}$$

The heavy-traffic limits correspond to the limiting values, i.e.,

$$c_a^2 = \lim_{n\to\infty} I_i(n) = \lim_{t\to\infty} I_c(t). \tag{7.3}$$

Similar results hold for c_s^2. Hence, we should want to estimate the IDI for relatively large n and the IDC for relatively large t.

7.2 Rate estimation methods

In this section we discuss ways to enhance the rate estimation.

7.2.1 Piecewise-linear fits

If we are confident that we have a $GI/GI/s$ model, or if we perform a rate estimation and obtain results as in Sect. 2, we can obtain more reliable estimates for individual rates if we fit piecewise-linear functions. In particular, for the single-server models with $\rho = 0.8$ in Sect. 2.1, we can exploit (2.1) and fit linear or constant functions for $k \geq s$. For s-server models with relatively large s, like 40, we can exploit (2.3) and (2.5) and fit functions of that piecewise-linear form. For system data, we presumably do *not* want to do this initially, because a major goal may be to detect deviations from the $G/G/s$ model, as discussed in Sect. 2.3.

7.2.2 Smoothing: binomial averaging

More generally, more reliable estimates for individual rates can be obtained using smoothing techniques. The state-dependent birth and death rates in each state are

likely to be similar to those in neighboring states. Thus, smoothing can be utilized to refine the rate estimation and to reduce the amount of data needed to accurately estimate the birth and death rates. In our simulation experiments, we used very large sample sizes, but in applications that may not be possible, because the model may change over different time intervals of each day and over different days; for example, see [27].

In our simulation experiments we investigated various smoothing methods. We remark that we found binomial averaging to be an effective smoothing method. Let $\bar{r}_k$ be the directly estimated rate at state k (birth or death). The binomial average $\bar{r}_k^*$ is obtained as a binomial average of the rates $\bar{r}_j$ in an interval of states centered at k. In particular, for $k \geq m$, we used binomial probability weights $b(k; 2m, p)$ for $p = 1/2$; i.e.,

$$\bar{r}_k^* = \sum_{j=0}^{2m} (2m!/j!(2m-j)!)(1/2)^{2m} \bar{r}_{k-m+j}. \tag{7.4}$$

We found that binomial averaging with $m = 5$ reduced the amount of data needed for good rate estimates by as much as a factor of three.

8 Conclusions

In this paper we studied the birth-and-death (BD) process obtained by directly fitting it to data from a queue-length process over a finite interval $[0, t]$, in the usual manner as if the process were a BD process, but without performing any statistical tests. The intended application is to data obtained from a complex queueing system, such as a hospital ward, as in Sect. 3 of [3], or a computer system, as was done with operational analysis in [8,9,15].

We started in Sect. 2 by reporting results from extensive simulation experiments fitting BD processes to data from $GI/GI/s$ queueing systems. Figures 1, 2 and 3 show that the rates have consistent structure, as summarized in (2.1), (2.3), and (2.5). In Sect. 2.3 we observed that this consistency should help detect departures from the standard s-server model, for example, due to customer balking, customer abandonment from the queue and a variable number of servers. In Sect. 3 we also established theoretical results about the fitted birth and death rates for $GI/GI/s$ models.

The main focus of the paper was on the transient behavior of the BD process. In particular, we focused on the round-trip first passage times from percentile p of the stationary distribution to percentile $1 - p$ and back. In Sect. 4 we reported simulation results showing that, even though the steady-state behavior of the fitted BD process is the same as for the original process, the transient behavior of the fitted BD process can be quite different, amplifying the point made in [46]. Extensive simulation experiments showed that the ratio of the long-run average first passage times exhibits remarkably regularity as the traffic intensity ρ increases and p decreases. That led to the definition of the *p-speed ratio* defined in (4.2) and the associated one-dimensional process approximation in (1.2) and (1.3).

In Sects. 5 and 6 we provided strong mathematical support for the speed ratio approximation in (1.2) and (1.3). The connection was established in both conventional and many-server (for M service) heavy-traffic limits, as in Corollaries 5.1, 5.2, and 6.2. It is significant that the stochastic-process limits for the fitted BD process and the underlying queueing process differ in a systematic way, as exposed in Sects. 5 and 6, and as summarized in (1.2) and (1.3). The theory shows that the non-Markov variability in the original queue-length process captured by the diffusion coefficient in its limiting diffusion process is transformed to the drift of the diffusion process limit of the fitted BD process. This transformation is constrained by the requirement that it must leave the steady-state distribution unchanged. It is also significant that the speed ratio in (1.3) coincides with the variability parameter in these conventional heavy-traffic limits, and so can be estimated in settings for which the common independence assumptions are in doubt using indices of dispersion, as suggested much earlier by [14,20,39]; see Sect. 9.6 of [50] and Sect. 7.1 here.

Consistent with heavy-traffic theory, we found that the speed ratio is very different for $GI/GI/s$ queues having large (s, λ) and non-exponential service-time distributions. The difficulties can be understood because the established many-server heavy-traffic limits for the queue-length process with non-exponential service-time distributions are non-Markov processes, which only can be related to Markov processes by increasing the dimension, as in [32,35]. The very different behavior can be understood by considering the $M/GI/\infty$ model, as discussed in Sect. 6.2.

There are many remaining problems: First, it remains to establish stronger results about the asymptotic behavior of the estimated rates $\bar{\lambda}_k$ and $\bar{\mu}_k$, both as $k \to \infty$ and as ρ increases. Specific conjectures have been stated in Conjectures 2.1, 2.2, and 5.1. Second, it remains to extend the heavy-traffic limits established in Sects. 5 and 6 to such more general estimated rates, as in the power-tail case in (3.8) and (3.10). All the results in Sect. 5 depend on the asymptotic geometric tail assumed in (3.7) and its associated heavy-traffic expansion in (5.3). It remains to explore what happens when these conditions are not satisfied.

It remains to investigate fitted BD processes for other models. We ourselves have started studying queueing models with periodic arrival rates that still have dynamic steady-state distributions [16]; supporting work appears in [55,56]. It remains to study the method applied to queueing system data.

Acknowledgments 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.

Appendix: Additional simulation results for the $GI/GI/1$ queue

We now supplement Fig. 1 and Table 1 in Sect. 2.1, which display estimated birth rates $\bar{\lambda}_k$ and estimated death rates $\bar{\mu}_k$ for several $GI/GI/1$ queues with traffic intensity $\rho = 0.8$. Here Fig. 6 and Table 6 show the corresponding estimates of birth and death rates for $\rho = 0.9$.

As in Sect. 2.1, we estimated the rates from 30 independent replications of 1 million customers. This large sample size is sufficient for 95 % confidence intervals of the

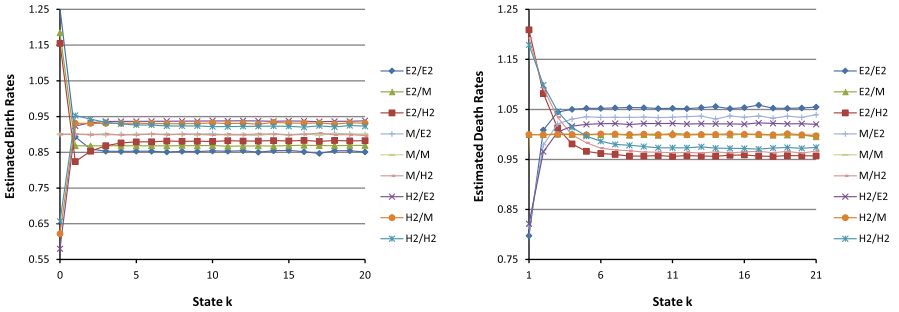

Fig. 6 Fitted birth rates $\bar{\lambda}_k$ (*left*) and death rates $\bar{\mu}_k$ (*right*) for nine $GI/GI/1$ models with $\rho = \lambda = 0.9$ and $\mu = 1$

Table 6 Estimates of the asymptotic fitted birth rate $\bar{\lambda}$, death rate $\bar{\mu}$, traffic intensity $\bar{\rho}$ and speed ratio $\bar{\omega}$ via (2.2) for the nine $GI/GI/1$ models with $\rho = 0.9$ in Fig. 6

Model	$\bar{\lambda}$	$\bar{\mu}$	$\bar{\rho}$	$\bar{\omega}$	$\omega \equiv (c_a^2 + c_s^2)/2$
$E_2/E_2/1$	0.8529 ± 0.0010	1.0530 ± 0.0015	0.8100 ± 0.0021	0.5266 ± 0.0058	0.500
$E_2/M/1$	0.8688 ± 0.0004	0.9997 ± 0.0008	0.8690 ± 0.0010	0.7638 ± 0.0060	0.750
$E_2/H_2/1$	0.8812 ± 0.0006	0.9578 ± 0.0012	0.9205 ± 0.0010	1.2592 ± 0.0149	1.250
$M/E_2/1$	0.8998 ± 0.0008	1.0348 ± 0.0009	0.8693 ± 0.0009	0.7654 ± 0.0053	0.750
$M/M/1$	0.8997 ± 0.0004	0.9998 ± 0.0006	0.8998 ± 0.0009	0.9983 ± 0.0085	1.000
$M/H_2/1$	0.8999 ± 0.0007	0.9662 ± 0.0022	0.9323 ± 0.0014	1.4803 ± 0.0296	1.500
$H_2/E_2/1$	0.9373 ± 0.0004	1.0213 ± 0.0005	0.9177 ± 0.0007	1.2151 ± 0.0103	1.250
$H_2/M/1$	0.9308 ± 0.0006	0.9997 ± 0.0005	0.9311 ± 0.0007	1.4517 ± 0.0146	1.500
$H_2/H_2/1$	0.9233 ± 0.0008	0.9757 ± 0.0028	0.9474 ± 0.0013	1.9068 ± 0.0444	2.000

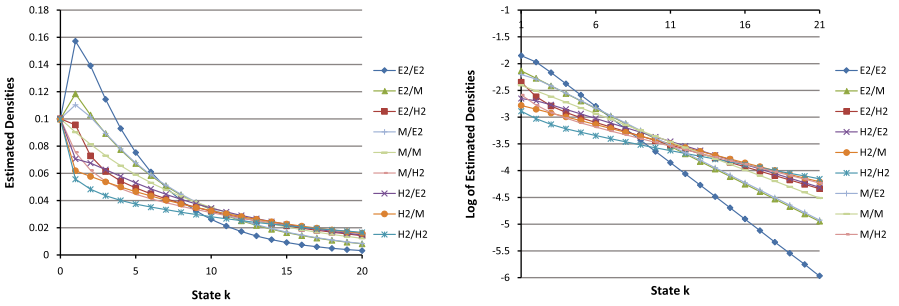

Fig. 7 Estimated steady-state probabilities $\bar{\alpha}_k$ (*left*) and their logarithms $\log_e \bar{\alpha}_k$ (*right*) for nine $GI/GI/1$ models with $\rho = \lambda = 0.9$ and $\mu = 1$

state-dependent rates to be within 1 % of the rates for states k with steady-state probability $\alpha_k \geq 0.01$.

Paralleling Fig. 4 in Sect. 3.3, we also estimated the steady-state queue-length probabilities α_k and their logarithms. The statistical precision is less with the higher

traffic intensity $\rho = 0.9$ instead of $\rho = 0.8$, as expected from [48]. To illustrate, the estimate of α_{10} in the $H_2/M/1$ model with $\rho = 0.9$ was $\bar{\alpha}_{10} = 0.03251$. The sample standard deviation from the 30 replications was 0.000351. Using the Student t distribution with 29 degrees of freedom, the t value for a two-sided 95 % confidence interval is 2.045. Thus, the halfwidth of the 95 % confidence interval is $(2.045 \times 0.000351)/\sqrt{30} = 0.37340 \times 0.000351 = 0.000131$, which is less than 0.5 % of the estimated value (Fig. 7).

This is contrasted with the case with traffic intensity $\rho = 0.8$. The estimate of α_{10} in the $H_2(2)/M/1$ model with $\rho = 0.8$ was $\bar{\alpha}_{10} = 0.02839$. The sample standard deviation from the 30 replications was 0.000282. Using the procedure as above, the halfwidth of the 95 % confidence interval was found to be 0.000105, which is less than 0.4 % of the estimated value, about 0.37 %.

References

1. Abate, J., Choudhury, G.L., Whitt, W.: Waiting-time tail probabilities in queues with long-tail service-time distributions. Queueing Syst. **16**, 311–338 (1994)
2. Abate, J., Whitt, W.: A heavy-traffic expansion for the asymptotic decay rates of tail probabilities in multi-channel queues. Op. Res. Lett. **15**, 223–230 (1994)
3. Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y., Tseytlin, Y., Yom-Tov, G.: Patient flow in hospitals: a data-based queueing-science perspective. New York University. Available at http://ie.technion.ac.il/serveng/References/ (2014). Accessed 19 Oct 2014
4. Bhat, U.N., Miller, G.K., Rao, S.S.: Statistical analysis of queueing systems. In: Dshalalow, J.H. (ed.) Frontiers in Queueing Theory, pp. 351–394. CRC Press, Boca Raton, FL (1997)
5. Billingsley, P.: Statistical Inference for Markov Processes. University of Chicago Press, Chicago (1961)
6. Bohlin, T.: Practical Grey-Box Process Identification. Springer, London (2006)
7. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., Zhao, L.: Statistical analysis of a telephone call center: a queueing-science perspective. J. Am. Stat. Assoc. **100**, 36–50 (2005)
8. Buzen, J.: Fundamental operational laws of computer system performance. Acta Inform. **14**, 167–182 (1976)
9. Buzen, J.: Operational analysis: an alternative to stochastic modeling. In: Ferarri, D. (ed.) Performance of Computer Installations, pp. 175–194. North Holland, Amsterdam (1978)
10. Cerbone, G., Ricciardi, L.M., Sacerdote, L.: Mean, variance and the skewness of the first passage time for the Ornstein–Uhlenbeck process. Cybern. Syst. **14**(2), 395–429 (1981)
11. Choudhury, G.L., Whitt, W.: Heavy-traffic asymptotic expansions for the asymptotic decay rates in the $BMAP/G/1$ queue. Stoch. Models **10**(2), 453–498 (1994)
12. Cohen, J.W.: Some results on regular variation for distributions in queueing and fluctuation theory. J. Appl. Probab. **10**(2), 343–353 (1973)
13. Cooper, R.B.: Introduction to Queueing Theory, 2nd edn. North Holland, Amsterdam (1982)
14. Cox, D.R., Lewis, P.A.W.: The Statistical Analysis of Series of Events. Methuen, London (1966)
15. Denning, P.J., Buzen, P.J.: The operational analysis of queueing network models. Comput. Surv. **10**, 225–261 (1978)
16. Dong, J., Whitt, W.: Stationary birth-and-death processes fit to queues with periodic arrival rate functions. In preparation (2014)
17. Duffield, N.G., Whitt, W.: Control and recovery from rare congestion events in a large multi-server system. Queueing Syst. **26**, 69–104 (1997)
18. Eick, S.G., Massey, W.A., Whitt, W.: The physics of the $M_t/G/\infty$ queue. Oper. Res. **41**, 731–742 (1993)
19. El-Taha, M., Stidham, S.: Sample-Path Analysis of Queueing Systems. Kluwer, Boston (1999)
20. Fendick, K.W., Whitt, W.: Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. Proc. IEEE **71**(1), 171–194 (1989)
21. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Op. Res. **29**(3), 567–588 (1981)

22. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer, New York (2009)
23. Iglehart, D.L.: Limit diffusion approximations for the many-server queue and the repairman problem. J. Appl. Probab. **2**, 429–441 (1965)
24. Iglehart, D.L., Whitt, W.: Multiple channel queues in heavy traffic. I. Adv. Appl. Probab. **2**(1), 150–177 (1970)
25. Iglehart, D.L., Whitt, W.: Multiple channel queues in heavy traffic, II: sequences, networks and batches. Adv. Appl. Probab. **2**(2), 355–369 (1970)
26. Keiding, N.: Maximum likelihood estimation in the birth-and-death process. Ann. Stat. **3**, 363–372 (1975)
27. Kim, S., Whitt, W.: Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? Manuf. Serv. Op. Manag. **16**(3), 464–480 (2014)
28. Kim, S., Whitt, W.: Choosing arrival process models for service systems: tests of a nonhomogeneous Poisson process. Nav. Res. Logist. **17**, 307–318 (2014)
29. Kristensen, N.R., Madsen, H., Jorgensen, S.B.: Parameter estimation in stochastic grey-box models. Automatica **40**(2), 225–237 (2004)
30. Pakes, A.G.: On the tails of waiting-time distributions. J. Appl. Probab. **12**, 555–564 (1975)
31. Pang, G., Talreja, R., Whitt, W.: Martingale proofs of many-server heavy-traffic limits for Markovian queues. Probab. Surv. **4**, 193–267 (2007)
32. Pang, G., Whitt, W.: Two-parameter heavy-traffic limits for infinite-server queues. Queueing Syst. **65**, 325–364 (2010)
33. Pang, G., Whitt, W.: The impact of dependent service times on large-scale service systems. Manuf. Serv. Op. Manag. **14**(2), 262–278 (2012)
34. Pang, G., Whitt, W.: Two-parameter heavy-traffic limits for infinite-server queues with dependent service times. Queueing Syst. **73**(2), 119–146 (2013)
35. Puhalskii, A.A., Reiman, M.I.: The multiclass $GI/PH/N$ queue in the Halfin–Whitt regime. Adv. Appl. Probab. **32**, 564–595 (2000)
36. Ross, J.V., Taimre, T., Pollett, P.K.: Estimation for queues from queue-length data. Queueing Syst. **55**, 131–138 (2007)
37. Ross, S.M.: Stochastic Processes, 2nd edn. Wiley, New York (1996)
38. Sigman, K.: Stationary Marked Point Processes: An Intuitive Approach. Chapman and Hall/CRC, New York (1995)
39. Sriram, K., Whitt, W.: Characterizing superposition arrival processes in packet multiplexers for voice and data. IEEE J. Sel. Areas Commun. **SAC–4**(6), 833–846 (1986)
40. Stone, C.J.: Limit theorems for random walks, birth and death processes and diffusion processes. Ill. J. Math. **4**, 638–660 (1963)
41. Thomas, M.U.: Some mean first-passage time approximations for the Ornstein–Uhlenbeck process. J. Appl. Probab. **12**(3), 600–604 (1975)
42. Vasilakis, C., Marshall, A.H.: Modelling nationwide hospital length of stay: opening the black box. J. Op. Res. Soc. **56**(3), 862–869 (2005)
43. Whitt, W.: Comparing counting processes and queues. Adv. Appl. Probab. **13**, 207–220 (1981)
44. Whitt, W.: Approximating a point process by a renewal process: two basic methods. Op. Res. **30**, 125–147 (1982)
45. Whitt, W.: On the heavy-traffic limit theorem for $GI/G/\infty$ queue. Adv. Appl. Probab. **14**(1), 171–190 (1982)
46. Whitt, W.: Untold horrors of the waiting room. What the equilibrium distribution will never tell about the queue-length process. Manag. Sci. **29**(4), 395–408 (1983)
47. Whitt, W.: Departures from a queue with many busy servers. Math. Op. Res. **9**(4), 534–544 (1984)
48. Whitt, W.: Planning queueing simulations. Manag. Sci. **35**(11), 1341–1366 (1989)
49. Whitt, W.: Understanding the efficiency of multi-server service systems. Manag. Sci. **38**(5), 708–723 (1992)
50. Whitt, W.: Stochastic-Process Limits. Springer, New York (2002)
51. Whitt, W.: A diffusion approximation for the $G/GI/n/m$ queue. Op. Res. **52**(6), 922–941 (2004)
52. Whitt, W.: Engineering solution of a basic call-center model. Manag. Sci. **51**, 221–235 (2005)
53. Whitt, W.: Staffing a call center with uncertain arrival rate and absenteeism. Prod. Op. Manag. **15**(1), 88–102 (2006)
54. Whitt, W.: Fitting birth-and-death queueing models to data. Stat. Probab. Lett. **82**, 998–1004 (2012)

55. Whitt, W.: Heavy-traffic limits for queues with periodic arrival rates. Op. Res. Lett. **42**, 458–461 (2014)
56. Whitt, W.: The steady-state distribution of the $M_t/M/\infty$ queue with a sinusoidal arrival rate function. Op. Res. Lett. **42**, 311–318 (2014)
57. Wolff, R.W.: Problems for statistical inference for birth and death queueing models. Op. Res. **13**, 343–357 (1965)
58. Wolff, R.W.: The effect of service-time regularity on system performance. In: Chandy, K.M., Reiser, M. (eds.) Comput. Perform., pp. 297–304. North-Holland, Amsterdam (1977)
59. Yin, L., Uttamchandani, S., Katz, R.: An empirical exploration of black-box performance models for storage systems. In: Proceedings of the 14th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, MASCOTS 2006. IEEE, pp. 433–440 (2006)