Adv. Appl. Prob. **51,** 268–299 (2019)
*doi:*10.1017/apr.2019.11

NEW PERSPECTIVES ON THE ERLANG-A QUEUE

ANDREW DAW,* ** AND
JAMOL PENDER,* *** *Cornell University*

Abstract

The nonstationary Erlang-A queue is a fundamental queueing model that is used to describe the dynamic behavior of large-scale multiserver service systems that may experience customer abandonments, such as call centers, hospitals, and urban mobility systems. In this paper we develop novel approximations to all of its transient and steady state moments, the moment generating function, and the cumulant generating function. We also provide precise bounds for the difference of our approximations and the true model. More importantly, we show that our approximations have *explicit stochastic representations as shifted Poisson random variables*. Moreover, we are also able to show that our approximations and bounds also hold for nonstationary Erlang-B and Erlang-C queueing models under certain stability conditions.

Keywords: Multiserver queue; abandonment; dynamical system; asymptotics; time-varying rate; moments, fluid limit; Erlang-A queue; functional forward equation; moment generating function

2010 Mathematics Subject Classification: Primary 60K25
Secondary 90B22; 62L20

1. Introduction

Markov processes are important modeling tools that help researchers describe real-world phenomena. Thus, it comes as no surprise that the Erlang-A model, which is a Markovian and multiserver queueing model that incorporates customer abandonments, is an important modeling tool in a multitude of application settings. Some of the more prominent applications include telecommunications and call/contact centers, healthcare, urban mobility and transportation, and more recently cloud computing; see, for example, [28], [29], [45], [37], [3], [5], [13], [48], [44], [2], [41], [20], [42], [12], and [4]. A detailed overview of the model can be found in [26]. There is also a large body of work on non-Markovian abandonment models, such as [46], [25], [1], [39], [43], and [38], and the references therein. Despite its importance in many different applications and the well-established history of study regarding customer impatience [34], the Erlang-A queueing model has remained very difficult to analyze and understand. Even analysis of its moments beyond the fourth moment has remained an important topic for additional study.

It is well known that the stationary setting of the Erlang-A queue is much easier to analyze than its nonstationary counterpart [10]. Asymptotic methods, such as heavy traffic limit theory and strong approximations theory, are common approaches for analyzing nonstationary and

Received 10 January 2018; revision received 2 January 2019.
* Postal address: School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853-3801, USA.
** Email address: amd399@cornell.edu
*** Email address: jjp274@cornell.edu

state-dependent queueing models; see, for example, [16], [27], [40], [7], and [14]. Uniform acceleration is extremely useful for approximating the transition probabilities and moments, such as the mean and variance of Markov processes. Moreover, the strong approximation methods are useful for analyzing the sample path behavior of the Markov process by showing that the sample paths of properly rescaled queueing processes converge to deterministic dynamical systems and Gaussian process limits.

There are however two main drawbacks of these asymptotic methods. The first is that the method is asymptotic as a function of the model parameters and the results really only hold when the rates are large. Thus, the quality of the approximations significantly depend on the size of the model parameters, and these asymptotic methods have been shown to be quite inaccurate for moderate sized model parameter settings; see, for example, [30]. The second main drawback is that the asymptotic methods do not generate any important insights for the moments or cumulant moments beyond order two since the limits are based on Brownian motion. Since Brownian motion has symmetry, its cumulants are all zero beyond the second order. Thus, symmetric Brownian approximations are limited in their power to capture asymmetries in higher moments or even the dynamics of the moment generating function, cumulant generating function, or Fourier transform. Moreover, it has been recently shown in [9] and [35] that the Erlang-A queue and its variants have nontrivial amounts of skewness and excess kurtosis, which implies that the Erlang-A models are clearly not Gaussian for moderate sized queues. These results also demonstrate that it is important to capture the behavior of the Erlang-A model beyond its second moment as this information can be used in staffing decisions [31] and [47].

One common approximation method that is used in the stochastic networks, queueing, and chemical reactions literature is a *moment closure approximation*. Moment closure approximations are used to approximate the moments of the queueing process with a surrogate distribution. Often, the set of moment equations for a large number of queueing models is not closed; see, for example, [32]. Thus, the closure approximation helps approximate the moments with a closed system using the surrogate distribution. One such method used by Massey and Pender [30], [49] used Hermite polynomials for approximating the distribution of the queue length process. In fact, they showed that using a quadratic polynomial works quite well. Since the Hermite polynomials are orthogonal to the Gaussian distribution, which has support on the entire real line, these Hermite polynomial approximations do not take into account the discreteness of the queueing process nor do they account for the nonnegativity of the queueing process. However, they showed that Hermite polynomials are natural to analyze since they are orthogonal with respect to the Gaussian distribution and the heavy traffic limits of multiserver queues are Gaussian. In this paper we perform an in-depth analysis of the moments and the moment generating function of the nonstationary Erlang-A queue. As the Erlang-B and Erlang-C queueing models are special cases of the Erlang-A model, we obtain similar results for these models. In a manner similar to what was observed for peer-to-peer networks in [11], we show that this novel representation allows us to view our bounds and approximations in a new way.

1.1. Main contributions of the paper

The main contributions of this work can be summarized as follows.

- We provide new approximations for the moments, moment generating function, and cumulant generating function for the nonstationary Erlang-A queue exploiting the FKG and Jensen's inequalities.

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

- We derive a novel stochastic interpretation and representation of our approximations as shifted Poisson random variables or $M/M/\infty$ queues, depending on the context. This sheds new light on the complexity of queues in heavy traffic or critically loaded regimes.
- We prove precise error bounds for our approximations and we also prove new upper and lower bounds for the nonstationary Erlang-A queue that become exact in certain parameter settings.

1.2. Organization of the paper

The remainder of this paper is organized as follows. In Section 2 we introduce the nonstationary Erlang-A queueing model and its importance in stochastic network theory. In Section 3 we provide approximations for the moments of the Erlang-A system and bound the true values. In Section 4 we derive approximations for the moment generating function of the Erlang-A queue and find stochastic representations for these approximations. We demonstrate these results through several numerical illustrations. We conclude in Section 5.

2. The Erlang-A queueing model

The Erlang-A queueing model is a fundamental queueing model in the stochastic processes literature. The work of Mandelbaum *et al.* [27] shows that the queue length process for an $M(t)/M/c+M$ queueing system $Q \equiv \{Q(t) \mid t \geq 0\}$ is represented by the stochastic, time-changed integral equation

$$Q(t) = Q(0) + \Pi_1\left(\int_0^t \lambda(s)\,\mathrm{d}s\right) - \Pi_2\left(\int_0^t \mu(Q(s)\wedge c)\,\mathrm{d}s\right) - \Pi_3\left(\int_0^t \theta(Q(s)-c)^+\,\mathrm{d}s\right),$$

where $\Pi_i \equiv \{\Pi_i(t) \mid t \geq 0\}$, for $i = 1, 2, 3$, are independent and identically distributed (i.i.d.) standard (rate-1) Poisson processes and $(x \wedge y) = \min\{x, y\}$. Thus, we can write the sample path dynamics of the Erlang-A queueing process in terms of three independent unit-rate Poisson processes. A deterministic time change for Π_1 transforms it into a nonhomogeneous Poisson arrival process with rate $\lambda(t)$ that counts the customer arrivals that occurred in the time interval [0,t). A random time change for the Poisson process Π_2 gives a departure process that counts the number of serviced customers. We implicitly assume that the number of servers is $c \in \mathbb{Z}^+$ and that each server works at rate μ. Finally, the random time change of Π_3 gives a counting process for the number of customers that abandon the system before beginning service. We also assume that the abandonment distribution is exponential and the rate of abandonments is equal to θ.

One of the main reasons that the Erlang-A queueing model has been so extensively analyzed is that several important queueing models are special cases of it. One special case is the infinite server queue. The infinite-server queue can be derived from the Erlang-A queue in two ways. The first way is to set the number of servers to ∞. This precludes any abandonments since the abandonment rate $\theta(Q(t)-c)^+$ is always equal to 0 when the number of servers is infinite. The second way to derive the infinite-server queue is to set the service rate μ equal to the abandonment rate θ. When $\mu = \theta$, this implies that the sum of the service and abandonment departure processes is equal to a linear function, i.e. $\mu(Q(t)\wedge c) + \theta(Q(t)-c)^+ = \mu Q(t) = \theta Q(t)$. Thus, the Erlang-A queueing model becomes an infinite-server queue.

A key insight from [16] is that, for multiserver queueing systems, it is natural to scale up the arrival rate and the number of servers simultaneously. This scaling—known as the *Halfin–Whitt* scaling—has been an important technique for modeling call centers in the queueing literature.

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

Since the $M(t)/M/c+M$ queueing process is a special case of a single node *Markovian service network*, we can also construct an associated, *uniformly accelerated* queueing process where both the new arrival rate $\eta\lambda(t)$ and the new number of servers ηc are scaled by the same factor $\eta > 0$. Thus, using the Halfin-Whitt scaling for the Erlang-A model, we arrive at the following sample path representation for the queue length process:

$$\begin{aligned} Q^\eta(t) &= Q^\eta(0) + \Pi_1\Big(\int_0^t \eta\cdot\lambda(s)\,\mathrm{d}s\Big) - \Pi_2\Big(\int_0^t \mu\cdot(Q^\eta(s)\wedge\eta\cdot c)\,\mathrm{d}s\Big) \\ &\quad - \Pi_3\Big(\int_0^t \theta\cdot(Q^\eta(s)-\eta\cdot c)^+\,\mathrm{d}s\Big) \\ &= Q^\eta(0) + \Pi_1\Big(\int_0^t \eta\cdot\lambda(s)\,\mathrm{d}s\Big) - \Pi_2\Big(\int_0^t \eta\cdot\mu\cdot\Big(\frac{Q^\eta(s)}{\eta}\wedge c\Big)\,\mathrm{d}s\Big) \\ &\quad - \Pi_3\Big(\int_0^t \eta\cdot\theta\cdot\Big(\frac{Q^\eta(s)}{\eta}-c\Big)^+\,\mathrm{d}s\Big). \end{aligned}$$

The Halfin-Whitt scaling is defined by simultaneously scaling up the rate of customer demand (which is the arrival rate) with the number of servers. In the context of call centers this is scaling up the number of customers and scaling up the number of agents to answer the phones. In the context of hospitals or healthcare this might be scaling up the number of patients with the number of beds or nurses. Taking the following limits gives us the *fluid* models of [27], i.e.

$$\lim_{\eta\to\infty}\frac{1}{\eta}Q^\eta(t) = q(t) \quad \text{almost surely (a.s.)}$$

where the deterministic process $q(t)$, the *fluid mean*, is governed by the one-dimensional ordinary differential equation (ODE)

$$\dot{q}(t) = \lambda(t) - \mu(q(t)\wedge c) - \theta(q(t)-c)^+. \tag{1}$$

Moreover, if we take a diffusion limit, i.e.

$$\lim_{\eta\to\infty}\sqrt{\eta}\Big(\frac{1}{\eta}Q^\eta(t) - q(t)\Big) \Rightarrow \tilde{Q}(t),$$

we get a diffusion process where the variance of the diffusion is given by the following ODE:

$$\begin{aligned} \dot{\mathrm{var}}[\tilde{Q}(t)] &= \lambda(t) + \mu\cdot(q(t)\wedge c) + \theta\cdot(q(t)-c)^+ - 2\cdot\mathrm{var}[\tilde{Q}(t)]\cdot(\mu\cdot\{q(t)<c\} \\ &\quad + \theta\cdot\{q(t)\geq c\}). \end{aligned}$$

2.1. Mean-field approximation is identical to the fluid limit

In addition to using strong approximations to analyze the queue length process one can also use the functional Kolmogorov forward equations, as outlined in [30]. The functional forward equations for the Erlang-A model are derived as

$$\begin{aligned} \dot{\mathbb{E}}[f(Q(t))] &\equiv \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[f(Q(t))|Q(0)=q(0)] \\ &= \lambda(t)\mathbb{E}[f(Q(t)+1)-f(Q(t))] + \mathbb{E}[\delta(Q(t),c)(f(Q(t)-1)-f(Q(t)))] \end{aligned} \tag{2}$$

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

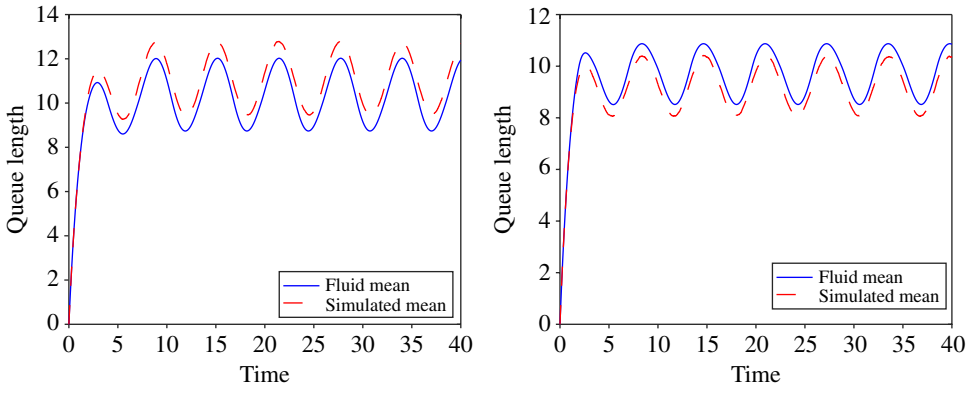

FIGURE 1: Comparison of a simulated Erlang-A queue and fluid model for $\lambda(t) = 10 + 2\sin(t)$, $\mu = 1$, $Q(0) = 0$, $c = 10$, and (a) $\theta = 0.5$ and (b) $\theta = 2$.

for all appropriate functions f and where $\delta(Q(t), c) = \mu(Q(t) \wedge c) + \theta(Q(t) - c)^+$. For the special case in which $f(x) = x$, we can derive an ODE for the mean queue length process as

$$\dot{\mathbb{E}}[Q(t)] = \lambda(t) - \mu\mathbb{E}[(Q(t) \wedge c)] - \theta\mathbb{E}[(Q(t) - c)^+]. \tag{3}$$

The first thing to note is that this equation is not autonomous and we need to know the distribution of $Q(t)$ *a priori* in order to compute the expectations on the right-hand side of (3). To know the distribution *a priori* is impossible except in some special cases like the infinite-server setting. However, it is easy to derive simple approximations for the mean queue length by making some assumptions on the queue length process. This is known as a closure approximation and one common closure approximation method is to simply take the expectations from outside the function to inside the function. This implies that the expectation $\mathbb{E}[f(X)]$ becomes $f(\mathbb{E}[X])$. This method is known as a mean-field approximation in physics and is also known as the deterministic mean approximation in [30]. By applying the mean-field approximation to (3), we can show that the resulting differential equation is given by the following autonomous ODE:

$$\dot{\mathbb{E}}[Q_{\mathrm{f}}(t)] = \lambda(t) - \mu(\mathbb{E}[Q_{\mathrm{f}}] \wedge c) - \theta(\mathbb{E}[Q_{\mathrm{f}}] - c)^+. \tag{4}$$

By careful inspection, we can observe that the ODE given by the mean-field approximation is identical to the fluid limit of (1). Moreover, if we simulate the queueing process and compare it to the mean-field limit, we notice an ordering property. For example, in Figure 1(a), we simulate the Erlang-A queue and compare it to the fluid model. We observe that, when $\theta < \mu$, the simulated mean is larger than the fluid mean. This is precisely what our results predict. Moreover, in Figure 1(b), we simulate the Erlang-A queue and compare it to the fluid model when $\theta > \mu$ and observe that the simulated queue length is smaller than the fluid limit.

Our goal in this work is to explain the behavior that we observe in Figure 1, which we will do in the following section. Before concluding our overview of the Erlang-A queueing model, we make a brief remark for notational clarity.

Remark 1. Throughout the remainder of this work, we use $Q(t)$ to represent the true queueing process and $Q_{\mathrm{f}}(t)$ to represent the fluid approximation of it. This fluid approximation is a

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

stochastic process that will be fully described in this work. In fact, in Section 4 we characterize the fluid approximations and use insight from these representations to bound the true queue length from above and below.

3. Inequalities for the moments of the Erlang-A queue

In this section we prove conditions under which the true moments of the Erlang-A queue either dominate their corresponding fluid limits or are dominated by them. We find that the relationship between the service rate and the abandonment rate determines whether or not the moments of the Erlang-A queue are dominated by the associated fluid limits. The remainder of this section is organized as follows. In Subsection 3.1 we derive inequalities for the true mean of the Erlang-A and its fluid approximation. In Subsection 3.2 we extend these inequalities to analogous results for the mth moment of the queueing system. Finally, in Subsection 3.3 we provide figures from numerical experiments that validate our theoretical results.

3.1. Inequalities for the mean

We begin with analysis of the mean of the Erlang-A queue. While we will generalize to all higher moments in Subsection 3.2, we first provide the mean result separately to give the reader the essence and elegance of our result. Before we proceed, we first establish a lemma for comparisons of ordinary differential equations that will be fundamental for our subsequent analysis.

Lemma 1. (A comparison lemma.) *Let* $f : \mathbb{R}^2 \to \mathbb{R}$ *be a continuous function in both variables. If we assume that the initial value problem*

$$\dot{x}(t) = f(t, x(t)), \qquad x(0) = x_0,$$

has a unique solution for the time interval [0,T] and

$$\dot{y}(t) \leq f(t, y(t)) \quad \text{for } t \in [0, T] \text{ and } y(0) \leq x_0,$$

then $x(t) \geq y(t)$ *for all* $t \in [0, T]$.

Proof. The proof of this result is given in [15]. □

With this lemma in hand, we can now derive relationships for the fluid limit and the true mean. As seen in the proof of Theorem 1 below, this result follows from the application of Lemma 1 and the convexity seen in the fluid approximation.

Theorem 1. *For the Erlang-A queue, if* $Q(0) = Q_f(0)$ *then the true mean dominates the fluid limit when* $\theta < \mu$*, the fluid limit dominates the true mean when* $\theta > \mu$*, and the two means are equal when* $\theta = \mu$.

Proof. Recall from (3) that the true mean satisfies the differential equation

$$\dot{\mathbb{E}}[Q(t)] = \lambda(t) - \mu\mathbb{E}[(Q \wedge c)] - \theta\mathbb{E}[(Q - c)^+],$$

and from (4) the fluid limit satisfies the differential equation

$$\dot{\mathbb{E}}[Q_f(t)] = \lambda(t) - \mu(\mathbb{E}[Q_f] \wedge c) - \theta(\mathbb{E}[Q_f] - c)^+.$$

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

We can simplify both equations by observing that $(X \wedge c) + (X - c)^+ = X$ for any random variable X. Thus, we have the following two equations for the true mean and the fluid limit:

$$\dot{\mathbb{E}}[Q(t)] = \lambda(t) - \theta E[Q] + (\theta - \mu)\mathbb{E}[(Q \wedge c)],$$
$$\dot{\mathbb{E}}[Q_{\mathrm{f}}(t)] = \lambda(t) - \theta \mathbb{E}[Q_{\mathrm{f}}] + (\theta - \mu)(\mathbb{E}[Q_{\mathrm{f}}] \wedge c).$$

If we take the difference of the two equations, we obtain

$$\dot{\mathbb{E}}[Q(t)] - \dot{\mathbb{E}}[Q_{\mathrm{f}}(t)] = \theta(\mathbb{E}[Q_{\mathrm{f}}] - \mathbb{E}[Q]) + (\theta - \mu)(\mathbb{E}[(Q \wedge c)] - (\mathbb{E}[Q_{\mathrm{f}}] \wedge c)).$$

Now since the minimum function $(Q \wedge c)$ is a concave function, we have

$$(\mathbb{E}[(Q \wedge c)] - (\mathbb{E}[Q] \wedge c)) \leq 0$$

for any random variable Q. Thus, by Lemma 1, we have, for $\theta < \mu$,

$$\mathbb{E}[Q(t)] - \mathbb{E}[Q_{\mathrm{f}}(t)] \geq 0$$

and, for $\theta > \mu$,

$$\mathbb{E}[Q(t)] - \mathbb{E}[Q_{\mathrm{f}}(t)] \leq 0.$$

Finally, for $\theta = \mu$, we have

$$\mathbb{E}[Q(t)] - \mathbb{E}[Q_{\mathrm{f}}(t)] = 0$$

since both differential equations are initialized with the same value and the origin is an equilibrium point for the difference. □

As discussed in Section 2, the Erlang-A model is quite versatile in its relation to other queueing systems of practical interest. In the two following corollaries, we find that Theorem 1 can be applied to the Erlang-B and Erlang-C models.

Corollary 1. *For the Erlang-B queueing model, if* $Q(0) = Q_{\mathrm{f}}(0)$ *then* $\mathbb{E}[Q(t)] \leq \mathbb{E}[Q_{\mathrm{f}}(t)]$ *for all* $t \geq 0$.

Proof. This is obvious after noting that the Erlang-B queue is a limit of the Erlang-A queue by letting $\theta \to \infty$. □

Corollary 2. *For the Erlang-C queueing model, if* $Q(0) = Q_{\mathrm{f}}(0)$ *then* $\mathbb{E}[Q(t)] \geq \mathbb{E}[Q_{\mathrm{f}}(t)]$ *for all* $t \geq 0$.

Proof. This is obvious after noting that the Erlang-C queue is an Erlang-A queue with $\theta = 0$. Since μ is assumed to be positive, we fall into the case where $\theta < \mu$ and this completes the proof. □

Remark 2. Given that we use Jensen's inequality and the FKG inequality later in the paper, we consider itr important to differentiate them. Here we give an example that sets the two apart. If we have the following function Q^n then Jensen's inequality implies that $\mathbb{E}[Q^n] \geq \mathbb{E}[Q]^n$. However, the FKG inequality implies that $\mathbb{E}[Q^n] \geq \mathbb{E}[Q^{n-1}]\mathbb{E}[Q]$. We note that, by iterating the FKG inequality $n - 2$ more times, it yields Jensen's inequality for the moments of random variables.

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

3.2. Inequalities for the *m*th moment

In this subsection we extend the previous findings for the mean to higher moments of the queueing system. Like the result for the mean, this is again built through observation of the convexity in the differential equation of the fluid approximation.

Theorem 2. *For the Erlang-A queue, $m \in \mathbb{Z}^+$, and $t \geq 0$, if $Q(0) = Q_{\mathrm{f}}(0)$ then $\mathrm{E}[Q^m(t)] \geq \mathrm{E}[Q_{\mathrm{f}}^m(t)]$ when $\theta < \mu$, $\mathrm{E}[Q^m(t)] \leq \mathrm{E}[Q_{\mathrm{f}}^m(t)]$ when $\theta > \mu$, and $\mathrm{E}[Q^m(t)] = \mathrm{E}[Q_{\mathrm{f}}^m(t)]$ when $\theta = \mu$.*

Proof. We will use proof by induction. For the base case, we can apply Theorem 1. Now, suppose that the statement holds for $j \in \{1, 2, \ldots, m-1\}$. By (2) and expansion through the binomial theorem, we note that the mth moment satisfies the differential equation

$$
\begin{aligned}
&\dot{\mathbb{E}}[Q^m(t)] \\
&\quad = \lambda(t)\mathrm{E}\left[\sum_{j=0}^{m}\binom{m}{j}Q^j(t) - Q^m(t)\right] \\
&\qquad + \mathrm{E}\left[\left(\sum_{j=0}^{m}\binom{m}{j}(-1)^{m-j}Q^j(t) - Q^m(t)\right)(\theta Q(t) - (\theta-\mu)(Q(t)\wedge c))\right] \\
&\quad = \lambda(t)\sum_{j=0}^{m-1}\binom{m}{j}\mathrm{E}[Q^j(t)] + \theta\sum_{j=0}^{m-1}\binom{m}{j}(-1)^{m-j}\mathrm{E}[Q^{j+1}(t)] \\
&\qquad + (\theta-\mu)\mathbb{E}\left[\left(\sum_{j=0}^{m-1}\binom{m}{j}(-1)^{m-1-j}Q^{j+1}(t)\right)\wedge\left(c\sum_{j=0}^{m-1}\binom{m}{j}(-1)^{m-1-j}Q^j(t)\right)\right],
\end{aligned}
$$

whereas the approximate autonomous version satisfies

$$
\begin{aligned}
&\dot{\mathbb{E}}[Q_{\mathrm{f}}^m(t)] \\
&\quad = \lambda(t)\sum_{j=0}^{m-1}\binom{m}{j}\mathrm{E}[Q_{\mathrm{f}}^j(t)] + \theta\sum_{j=0}^{m-1}\binom{m}{j}(-1)^{m-j}\mathrm{E}[Q_{\mathrm{f}}^{j+1}(t)] \\
&\qquad + (\theta-\mu)\left(\mathbb{E}\left[\sum_{j=0}^{m-1}\binom{m}{j}(-1)^{m-1-j}Q_f^{j+1}(t)\right]\wedge\mathbb{E}\left[c\sum_{j=0}^{m-1}\binom{m}{j}(-1)^{m-1-j}Q_f^j(t)\right]\right),
\end{aligned}
$$

which is the closure approximation of the differential equation for $\mathrm{E}[Q^m(t)]$. Now by taking the difference of these ODEs, we have

$$
\begin{aligned}
&\dot{\mathbb{E}}[Q^m(t)] - \dot{\mathbb{E}}[Q_{\mathrm{f}}^m(t)] \\
&\quad = \lambda(t)\sum_{j=0}^{m-1}\binom{m}{j}\mathrm{E}[Q^j(t) - Q_{\mathrm{f}}^j(t)] + \theta\sum_{j=0}^{m-1}\binom{m}{j}(-1)^{m-j}\mathrm{E}[Q^{j+1}(t) - Q_{\mathrm{f}}^{j+1}(t)]
\end{aligned}
$$

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

$$+(\theta-\mu)\Bigg(\mathrm{E}\Bigg[\Bigg(\sum_{j=0}^{m-1}\binom{m}{j}(-1)^{m-1-j}Q^{j+1}(t)\Bigg)\wedge\Bigg(c\sum_{j=0}^{m-1}\binom{m}{j}(-1)^{m-1-j}Q^{j}(t)\Bigg)\Bigg]$$
$$-\mathrm{E}\Bigg[\sum_{j=0}^{m-1}\binom{m}{j}(-1)^{m-1-j}Q_{\mathrm{f}}^{j+1}(t)\Bigg]\wedge\mathrm{E}\Bigg[c\sum_{j=0}^{m-1}\binom{m}{j}(-1)^{m-1-j}Q_{\mathrm{f}}^{j}(t)\Bigg]\Bigg).$$

Because the minimum is a concave function, for any X and Y with real means, $\mathrm{E}[X\wedge Y]\leq \mathrm{E}[X]\wedge\mathrm{E}[Y]$. Thus, by Lemma 1, we have, for $\theta<\mu$,

$$\mathbb{E}[Q^m(t)]-\mathbb{E}[Q_{\mathrm{f}}^m(t)]\geq 0,$$

if $\theta>\mu$,

$$\mathbb{E}[Q^m(t)]-\mathbb{E}[Q_{\mathrm{f}}^m(t)]\leq 0,$$

and if $\theta=\mu$,

$$\mathbb{E}[Q^m(t)]-\mathbb{E}[Q_{\mathrm{f}}^m(t)]=0$$

since both differential equations are initialized with the same value, the origin is an equilibrium point for the difference, and all the lower-power terms in the differential equations follow this structure, which we know from the inductive hypothesis. Therefore, we see this holds for m, which completes the proof via Lemma 1. □

Again as we have seen for the mean, we can exploit the versatility of the Erlang-A queue to extend these insights to the Erlang-B and Erlang-C models as well.

Corollary 3. *For the Erlang-B queueing model, if* $Q(0)=Q_{\mathrm{f}}(0)$ *then* $\mathbb{E}[Q^m(t)]\leq\mathbb{E}[Q_{\mathrm{f}}^m(t)]$ *for all* $t\geq 0$ *and* $m\in\mathbb{Z}^+$.

Proof. This is obvious after noting that the Erlang-B queue is a limit of the Erlang-A queue by letting $\theta\to\infty$. □

Corollary 4. *For the Erlang-C queueing model, if* $Q(0)=Q_{\mathrm{f}}(0)$ *then* $\mathbb{E}[Q^m(t)]\geq\mathbb{E}[Q_{\mathrm{f}}^m(t)]$ *for all* $t\geq 0$ *and* $m\in\mathbb{Z}^+$.

Proof. This is obvious after noting that the Erlang-C queue is an Erlang-A queue with $\theta=0$. Since μ is assumed to be positive, we fall into the case where $\theta<\mu$ and this completes the proof. □

3.3. Numerical results

In this subsection we describe numerical results for approximating the moments of the Erlang-A queue and examine them relative to our findings. We do so through Figures 2, 3, 4, and 5. All four of these figures demonstrate the findings of Theorem 2. In Figures 2 and 3 we show the first four moments of the Erlang-A queue and their respective fluid approximations for the $\theta<\mu$ and $\theta>\mu$ cases, respectively. In these plots we take the arrival rate at time $t\geq 0$ to be $\lambda(t)=10+2\sin(t)$. We initialize the queue as empty, and we assume that the queueing system has $c=10$ servers each with an exponential service rate $\mu=1$. We test two different

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

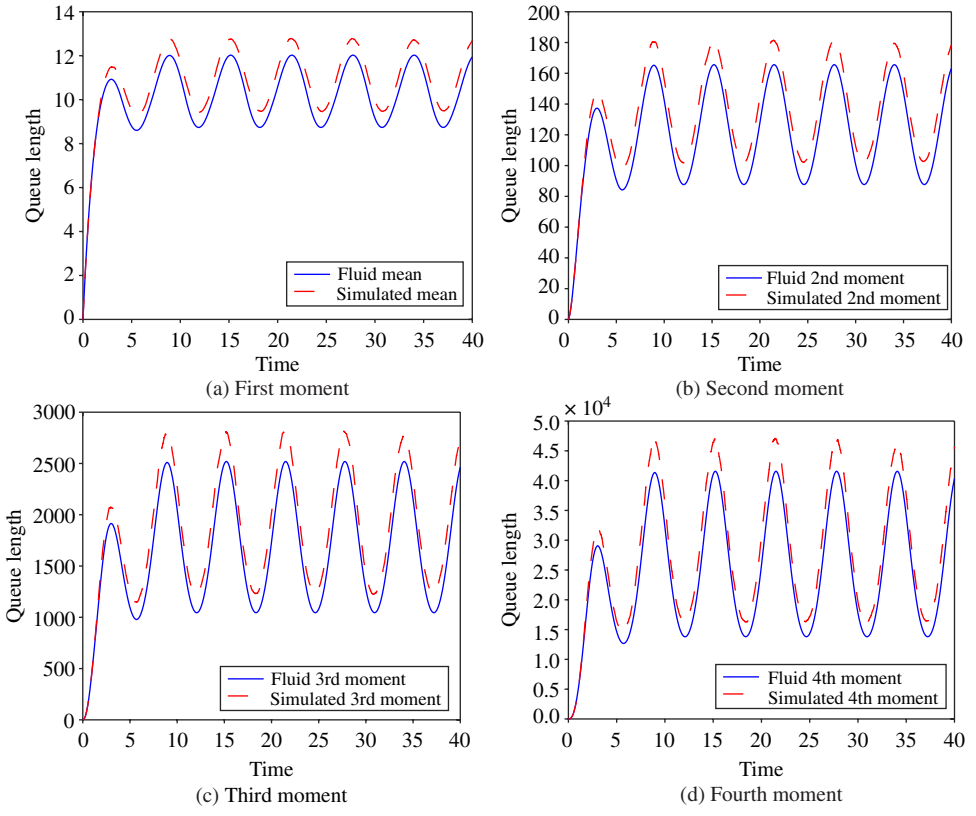

FIGURE 2: The first four moments of the Erlang-A queue and their respective fluid approximations for $\lambda(t) = 10 + 2\sin(t)$, $\mu = 1$, $\theta = 0.5$, $Q(0) = 0$, and $c = 10$.

cases for the abandonment rate, $\theta = 0.5$ and $\theta = 2$. In these settings we observe that, when $\theta < \mu$, the fluid approximations are below their corresponding simulated stochastic values and, when $\theta > \mu$, the fluid values are greater than the simulations, and this matches the statements of Theorems 1 and 2.

We observe the same relationships in Figures 4 and 5. For both of these figures we set $\lambda(t) = 100 + 20\sin(t)$ and $c = 100$, but otherwise use the same values as for Figures 2 and 3. With this increase in the arrival intensity and the number of servers, we see that the gaps between the fluid approximations and the simulations are again present, albeit proportionally smaller and may be most easily observed by inspection at the peaks and valleys of the trigonometric forms.

To gain further insight into the relationship between these moments and their approximations, we now plot the relative error for each moment displayed in the past four parameter settings. Specifically, we define the relative error as

$$\mathrm{RE}(t) = \frac{|\mathrm{E}[Q^m(t)] - \mathrm{E}[Q_f^m(t)]|}{\mathrm{E}[Q^m(t)]}.$$

Using empirically calculated moments via simulation and approximations calculated via differential equations, we plot this function across the four settings in Figures 6, 7, 8, and 9. We note that the error appears to become more localized to the valleys of the sine curve as

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

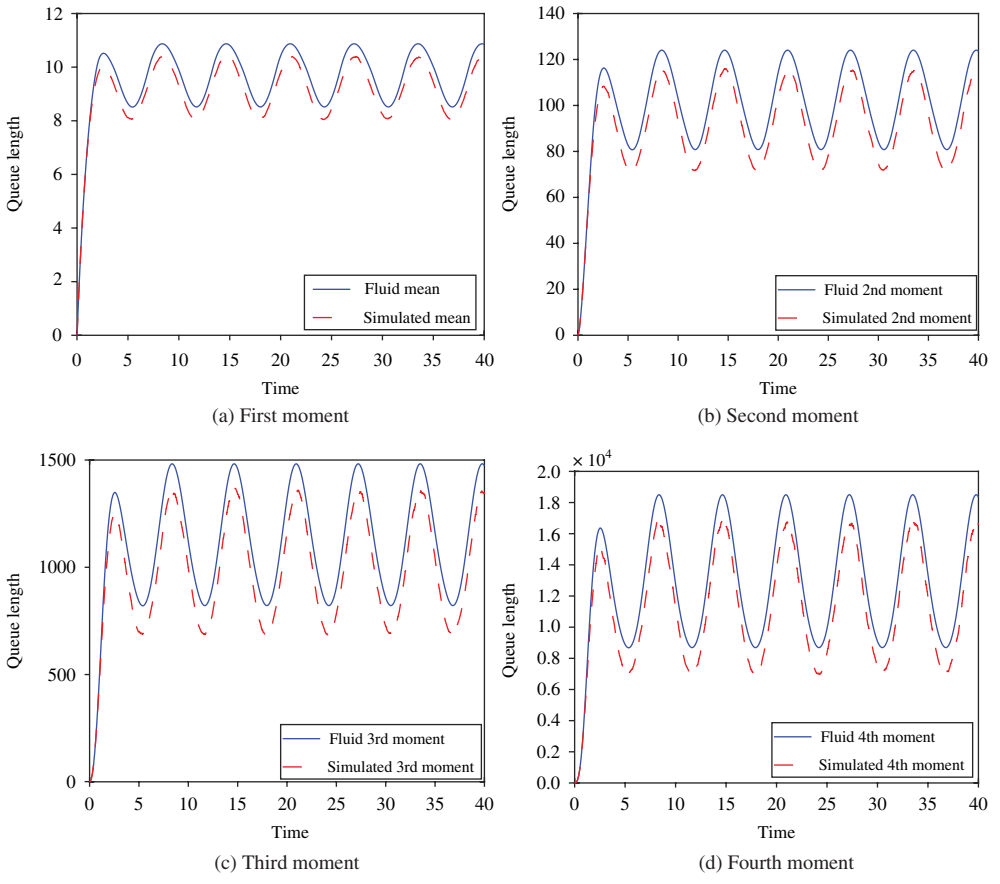

FIGURE 3: The first four moments of the Erlang-A queue and their respective fluid approximations for $\lambda(t) = 10 + 2\sin(t)$, $\mu = 1$, $\theta = 2$, $Q(0) = 0$, and $c = 10$.

the moments grow higher in power or for $\lambda(t)$ and c of greater magnitude, whereas the errors of the means when $\lambda(t)$ and c are not as large (i.e. Figures 6 and 7) oscillate more quickly with the maxima of its relative error occurring between the mean's extremes. This difference makes sense since at this point the denominator is at its smallest and this will have a more pronounced effect for the larger values at the higher moments. Furthermore, we can also infer that the highest relative error at the mean takes place when the value of the mean is equal to the number of servers, which creates a change in behavior. For more on these phenomena, see [30] and [35]. Additionally, we observe that as the rate and the number of servers grow large, the approximations appear to be more accurate.

4. Inequalities and characterizations for generating functions of the Erlang-A queue

Building on what we have proved for the moments of the Erlang-A, we can provide similar inequalities for the moment generating function and the cumulant generating function again through convexity in the ordinary or partial differential equations for the fluid approximations. We provide these inequalities in Subsections 4.1 and 4.2, respectively. As a result, we uncover new stochastic representations for our fluid approximations as shifted Poisson

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

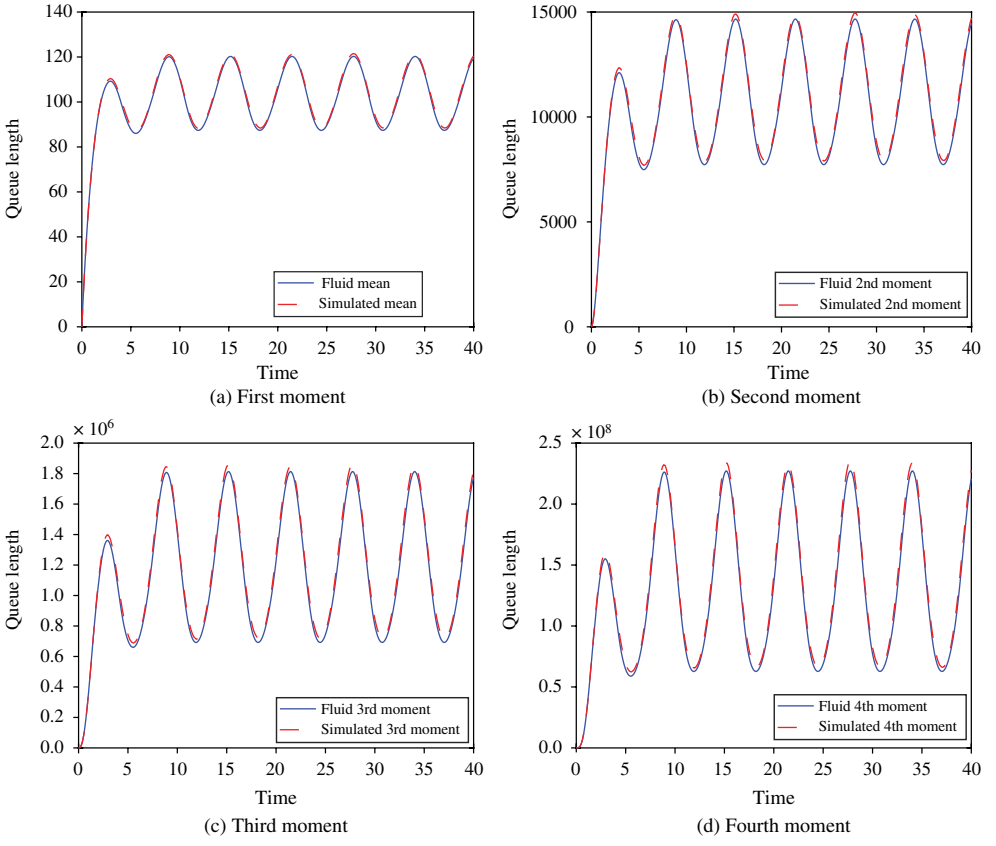

FIGURE 4: The first four moments of the Erlang-A queue and their respective fluid approximations for $\lambda(t) = 100 + 20\sin(t)$, $\mu = 1$, $\theta = 0.5$, $Q(0) = 0$, and $c = 100$.

random variables. We describe these stochastic representations for systems in steady-state in Subsection 4.3 and for nonstationary systems in Subsection 4.4. We conclude this section with a variety of demonstrations of these results through numerical experiments in Subsection 4.5.

4.1. An inequality for the moment generating function of the Erlang-A queue

Using the functional forward equations as given by (2), we can show that the moment generating function for the Erlang-A queue satisfies the following partial differential equation:

$$\dot{\mathbb{E}}[e^{\alpha Q(t)}] = \lambda(t)(e^{\alpha} - 1)\mathbb{E}[e^{\alpha \cdot Q(t)}] + \theta(e^{-\alpha} - 1) \cdot \mathbb{E}[Q(t)e^{\alpha \cdot Q(t)}] - (\theta - \mu)(e^{-\alpha} - 1)\mathbb{E}[(Q(t) \wedge c)e^{\alpha Q(t)}]. \quad (5)$$

As for the nonautonomous differential equation for the mean in (3), we also cannot directly compute the moment generating function since we do not know the distribution of the queue length *a priori*. This is also true for numerical purposes. Unless we can compute the expectation that includes the minimum function it is impossible to know the moment generating function, except in special cases such as the infinite-server queue and some cases of the stationary Erlang-B queue. Thus, it is useful to obtain approximations that are explicit upper

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

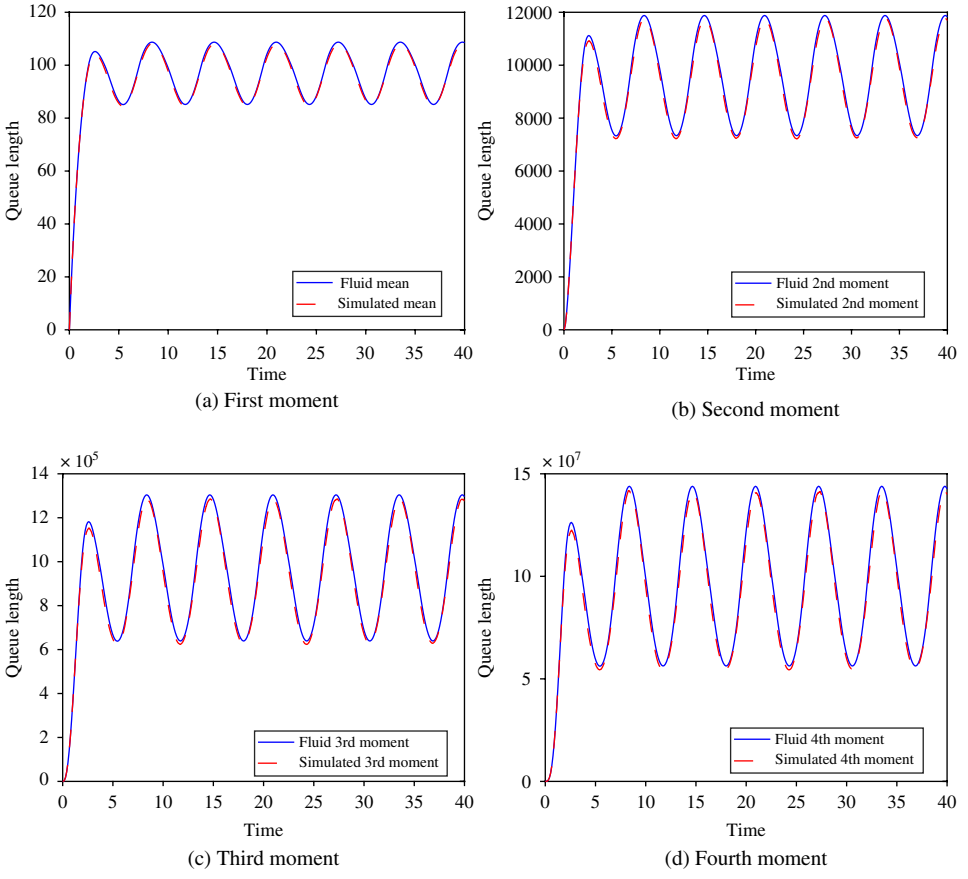

FIGURE 5: The first four moments of the Erlang-A queue and their respective fluid approximations for $\lambda(t) = 100 + 20\sin(t)$, $\mu = 1$, $\theta = 2$, $Q(0) = 0$, and $c = 100$.

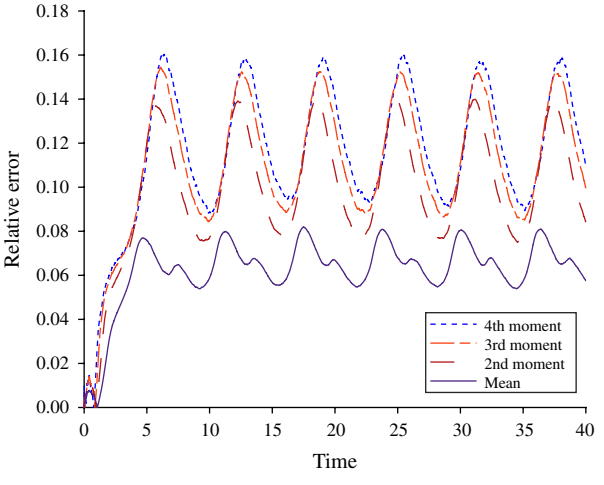

FIGURE 6: The relative error for $\lambda(t) = 10 + 2\sin(t)$, $\mu = 1$, $\theta = 0.5$, $Q(0) = 1$, and $c = 10$.

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

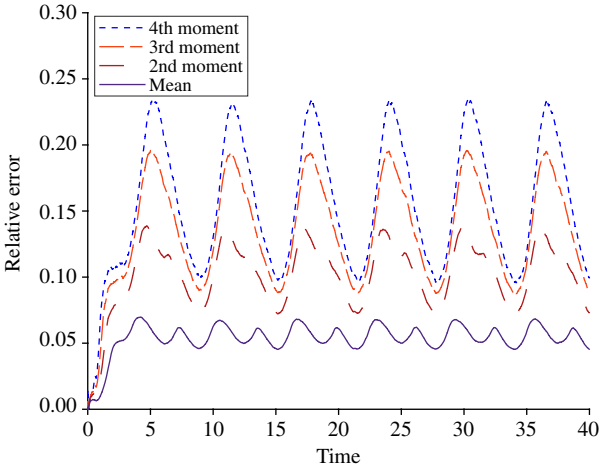

FIGURE 7: The relative error for $\lambda(t) = 10 + 2\sin(t)$, $\mu = 1$, $\theta = 2$, $Q(0) = 1$, and $c = 10$.

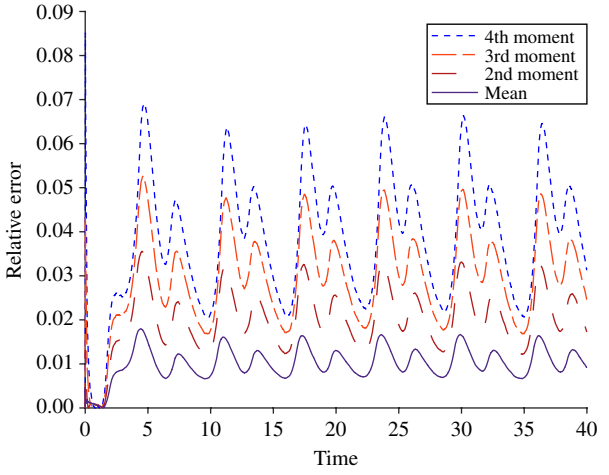

FIGURE 8: The relative error for $\lambda(t) = 100 + 20\sin(t)$, $\mu = 1$, $\theta = 0.5$, $Q(0) = 1$, and $c = 100$.

or lower bounds for the moment generating function. By using Jensen's inequality for concave functions, we can approximate the moment generating function with the partial differential equation

$$\begin{aligned}\dot{\mathbb{E}}[\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}] &= \lambda(t)(\mathrm{e}^{\alpha} - 1)\mathbb{E}[\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}] + \theta(\mathrm{e}^{-\alpha} - 1)\mathbb{E}[Q_{\mathrm{f}}(t)\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}] \\ &\quad -(\theta - \mu)(\mathrm{e}^{-\alpha} - 1)(\mathbb{E}[Q_{\mathrm{f}}(t)\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}] \wedge \mathbb{E}[c\mathrm{e}^{\alpha \cdot Q_{\mathrm{f}}(t)}]), \end{aligned} \tag{6}$$

which, for $M_f(t, \alpha) \equiv \mathrm{E}\,[\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}]$, can be expressed as

$$\begin{aligned}\frac{\partial M_f(t,\alpha)}{\partial t} &= \lambda(t)(\mathrm{e}^{\alpha} - 1) \cdot M_f(t,\alpha) + \theta(\mathrm{e}^{-\alpha} - 1)\frac{\partial M_f(t,\alpha)}{\partial \alpha} \\ &\quad -(\theta - \mu)(\mathrm{e}^{-\alpha} - 1)\Big(\frac{\partial M_f(t,\alpha)}{\partial \alpha} \wedge cM_f(t,\alpha)\Big). \end{aligned} \tag{7}$$

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

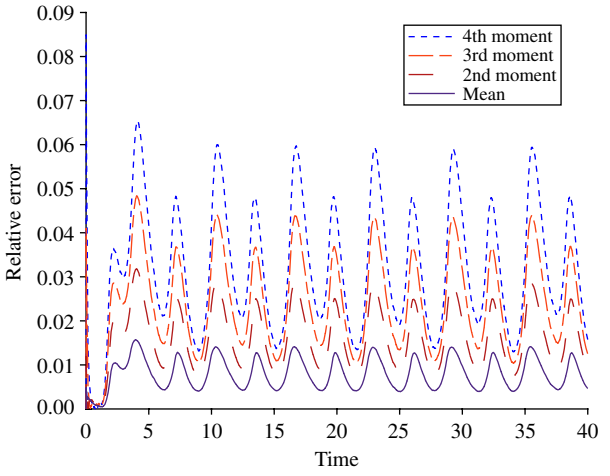

FIGURE 9: The relative error for $\lambda(t) = 100 + 20\sin(t)$, $\mu = 1$, $\theta = 2$, $Q(0) = 1$, and $c = 100$.

The following theorem specifies when $M_f(t, \alpha) = \mathbb{E}[e^{\alpha Q_f(t)}]$ is a lower or upper bound for the exact moment generating function of the Erlang-A queue.

Theorem 3. *Let $t \geq 0$ and $\alpha \in \mathbb{R}$. For the Erlang-A queue, if $Q(0) = Q_f(0)$ then $\mathbb{E}[e^{\alpha Q(t)}] \geq \mathbb{E}[e^{\alpha Q_f(t)}]$ when $(\theta - \mu)(e^{-\alpha} - 1) > 0$, $\mathbb{E}[e^{\alpha Q(t)}] \leq \mathbb{E}[e^{\alpha Q_f(t)}]$ when $(\theta - \mu)(e^{-\alpha} - 1) < 0$, and $\mathbb{E}[e^{\alpha Q(t)}] = \mathbb{E}[e^{\alpha Q_f(t)}]$ when $(\theta - \mu)(e^{-\alpha} - 1) = 0$.*

Proof. Let $M(t, \alpha) = \mathrm{E}\,[e^{\alpha Q(t)}]$. From (5), we note that $M(t, \alpha)$ is given by the partial differential equation

$$\dot{M}(t, \alpha) = \lambda(t)(e^{\alpha} - 1)M(t, \alpha) + \theta(e^{-\alpha} - 1)\frac{\partial M(t, \alpha)}{\partial \alpha} - (\theta - \mu)(e^{-\alpha} - 1)\mathbb{E}[(Q(t) \wedge c) \cdot e^{\alpha Q(t)}].$$

If $(\theta - \mu)(e^{-\alpha} - 1) > 0$ then, by application of Jensen's inequality on the minimum function, we note further that

$$\dot{M}(t, \alpha) \leq \lambda(t)(e^{\alpha} - 1)M(t, \alpha) + \theta(e^{-\alpha} - 1)\frac{\partial M(t, \alpha)}{\partial \alpha} - (\theta - \mu)(e^{-\alpha} - 1)\Big(\frac{\partial M(t, \alpha)}{\partial \alpha} \wedge cM(t, \alpha)\Big).$$

We can now observe by comparison to (7) that the right-hand side of this inequality is equivalent to the partial differential equation for $M_f(t, \alpha)$, and, thus, by application of Lemma 1, $M(t, \alpha) \geq M_f(t, \alpha)$. By symmetric arguments for $(\theta - \mu)(e^{-\alpha} - 1) < 0$ and $(\theta - \mu)(e^{-\alpha} - 1) = 0$, we complete the proof. □

Remark 3. If we restrict to $\alpha > 0$, we can note that $(\theta - \mu)(e^{-\alpha} - 1) > 0$ if and only if $\theta < \mu$, forming a connection to the conditions seen in Section 3 for the moment bounds.

As with the moments, we can observe these relationships in our numerical experiments. Moreover, we provide figures demonstrating our analytical results in Subsection 4.5.

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

4.2. An inequality for the cumulant moment generating function of the Erlang-A queue

As a consequence of the findings for the moment generating function, we can also provide similar inequalities for the cumulant moment generating function. Using (5), we have

$$\begin{aligned}
\log\left(\dot{\mathbb{E}}[\mathrm{e}^{\alpha Q(t)}]\right) &\equiv \frac{\partial}{\partial t}\log\left(\mathbb{E}[\mathrm{e}^{\alpha Q(t)}]\right)\\
&= \frac{\dot{\mathbb{E}}[\mathrm{e}^{\alpha Q(t)}]}{\mathbb{E}[\mathrm{e}^{\alpha Q(t)}]}\\
&= \lambda(t)(\mathrm{e}^{\alpha}-1)+\theta(\mathrm{e}^{-\alpha}-1)\frac{\mathbb{E}[Q(t)\mathrm{e}^{\alpha Q(t)}]}{\mathbb{E}[\mathrm{e}^{\alpha Q(t)}]}\\
&\quad -(\theta-\mu)(\mathrm{e}^{-\alpha}-1)\frac{\mathbb{E}[(Q(t)\wedge c)\mathrm{e}^{\alpha Q(t)}]}{\mathbb{E}[\mathrm{e}^{\alpha Q(t)}]},\\
\frac{\partial}{\partial t}\log\left(\mathbb{E}[\mathrm{e}^{\alpha Q(t)}]\right) &= \lambda(t)(\mathrm{e}^{\alpha}-1)+\theta(\mathrm{e}^{-\alpha}-1)\frac{\partial}{\partial \alpha}\log\left(\mathbb{E}[\mathrm{e}^{\alpha Q(t)}]\right)\\
&\quad -(\theta-\mu)(\mathrm{e}^{-\alpha}-1)\frac{\mathbb{E}[(Q(t)\wedge c)\mathrm{e}^{\alpha Q(t)}]}{\mathbb{E}[\mathrm{e}^{\alpha Q(t)}]}.
\end{aligned}$$

As for the moment generating function, we note that we cannot compute the cumulant moment generating function directly without knowing the distribution of the queue length. By applying Jensen's inequality again, we can describe the fluid approximation as follows, where we let $G(t,\alpha)=\log\left(\mathrm{E}\left[\mathrm{e}^{\alpha Q(t)}\right]\right)$ and $G_f(t,\alpha)=\log\left(\mathrm{E}\left[\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}\right]\right)$:

$$\begin{aligned}
\log\left(\dot{\mathbb{E}}[\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}]\right) &= \lambda(t)(\mathrm{e}^{\alpha}-1)+\theta(\mathrm{e}^{-\alpha}-1)\frac{\partial}{\partial \alpha}\log\left(\mathbb{E}[\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}]\right)\\
&\quad -(\theta-\mu)(\mathrm{e}^{-\alpha}-1)\left(\frac{\mathbb{E}[Q_{\mathrm{f}}(t)\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}]\wedge\mathbb{E}[c\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}]}{\mathbb{E}[\mathrm{e}^{\alpha Q(t)}]}\right),\\
\frac{\partial G_f(t,\alpha)}{\partial t} &= \lambda(t)(\mathrm{e}^{\alpha}-1)+\theta(\mathrm{e}^{-\alpha}-1)\frac{\partial G_f(t,\alpha)}{\partial \alpha}\\
&\quad -(\theta-\mu)(\mathrm{e}^{-\alpha}-1)\left(\frac{\partial G_f(t,\alpha)}{\partial \alpha}\wedge c\right).
\end{aligned}\tag{8}$$

Using this observation and our approach in finding the inequalities for the moment generating function, we find the equivalent inequalities for the cumulant moment generating function in the following corollary.

Corollary 5. *Let* $t\geq 0$ *and* $\alpha\in\mathbb{R}$. *For the Erlang-A queue, if* $Q(0)=Q_{\mathrm{f}}(0)$ *then* $\log\left(\mathbb{E}[\mathrm{e}^{\alpha Q(t)}]\right)\geq\log\left(\mathbb{E}[\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}]\right)$ *when* $(\theta-\mu)(\mathrm{e}^{-\alpha}-1)>0$, $\log\left(\mathbb{E}[\mathrm{e}^{\alpha Q(t)}]\right)\leq\log\left(\mathbb{E}[\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}]\right)$ *when* $(\theta-\mu)(\mathrm{e}^{-\alpha}-1)<0$, *and* $\log\left(\mathbb{E}[\mathrm{e}^{\alpha Q(t)}]\right)=\log\left(\mathbb{E}[\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}]\right)$ *when* $(\theta-\mu)(\mathrm{e}^{-\alpha}-1)=0$.

Proof. The proof follows from the same argument that was given in Theorem 3 applied to (8) and the fact that the log function is strictly increasing. □

4.3. Characterization of the moment generating function in steady state

From what we have observed for the moment generating function, we can derive an exact representation for the fluid approximation of the moment generating function in steady state. We assume a stationary arrival rate $\lambda>0$. We will investigate the differential equations for the stationary fluid approximation in a casewise manner based on the relationship of λ

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

and the system's service parameters. To do so, we begin with a lemma bounding the fluid approximation of the mean.

Lemma 2. *Suppose that λ is constant. If $\lambda < c\mu$ then $\mathrm{E}[Q_{\mathrm{f}}(\infty)] < c$. Moreover, if $\lambda \geq c\mu$ then $\mathrm{E}[Q_{\mathrm{f}}(\infty)] \geq c$.*

Proof. We will prove this by contradiction. For the first part, we assume that $\mathbb{E}[Q_{\mathrm{f}}(\infty)] \geq c$. Now by using the differential equation for the mean in steady state, we have

$$0 \quad = \quad \lambda - \mu(\mathbb{E}[Q_{\mathrm{f}}(\infty)] \wedge c) - \theta(\mathbb{E}[Q_{\mathrm{f}}(\infty)] - c)^{+} = \lambda - \mu c - \theta(\mathbb{E}[Q_{\mathrm{f}}(\infty)] - c)^{+}.$$

Since we assumed that $\mathbb{E}[Q_{\mathrm{f}}(\infty)] \geq c$, then this yields the inequality

$$\lambda \quad \geq \quad c\mu,$$

which yields a contradiction. For the second case, where we assume that $\lambda \geq c\mu$ and $\mathbb{E}[Q_{\mathrm{f}}(\infty)] < c$, then by the same differential equation we have

$$\lambda = \mu(\mathbb{E}[Q_{\mathrm{f}}(\infty)] \wedge c) + \theta(\mathbb{E}[Q_{\mathrm{f}}(\infty)] - c)^{+} = \mu(\mathbb{E}[Q_{\mathrm{f}}(\infty)] \wedge c) < c\mu,$$

which yields another contradiction. □

We now begin characterizing the fluid approximations with our second case, $\lambda \geq c\mu$, in the following proposition.

Proposition 1. *If $\lambda \geq c\mu$ then in steady state we have*

$$\frac{\partial M_f(\infty, \alpha)}{\partial \alpha} = \frac{\lambda(\mathrm{e}^{\alpha} - 1) + (\theta - \mu)(1 - \mathrm{e}^{-\alpha})c}{\theta(1 - \mathrm{e}^{-\alpha})} M_f(\infty, \alpha)$$

with $M_f(\infty, 0) = 1$, which yields a solution of

$$M_f(\infty, \alpha) = \mathrm{e}^{(\alpha(\theta-\mu)c + \lambda(\mathrm{e}^{\alpha}-1))/\theta} \quad \textit{for } \alpha \geq 0.$$

Proof. To find the partial differential equation, we use a functional cumulant bound for any nondecreasing function $h(\cdot)$ (which can be seen as a form of the FKG inequality as $\mathrm{e}^{\alpha x}$ is also nondecreasing in x for $\alpha \geq 0$):

$$\frac{\mathbb{E}[h(X)\mathrm{e}^{\alpha X}]}{\mathbb{E}[\mathrm{e}^{\alpha X}]} \geq \mathbb{E}[h(X)].$$

In the case that $\lambda \geq c\mu$ we have $\mathrm{E}[Q_{\mathrm{f}}(t)] \geq c$ in steady state by Lemma 2, and so we know how to evaluate the minimum in the fluid equation. Thus, the derivative of $G_f(\infty, \alpha) = \log(M_f(\infty, \alpha))$ with respect to α is

$$\frac{\mathrm{d}G_f(\infty, \alpha)}{\mathrm{d}\alpha} = \frac{\lambda(\mathrm{e}^{\alpha} - 1) + c(\theta - \mu)(1 - \mathrm{e}^{-\alpha})}{\theta(1 - \mathrm{e}^{-\alpha})} = \frac{\lambda \mathrm{e}^{\alpha}}{\theta} + \frac{c(\theta - \mu)}{\theta}, \tag{9}$$

where we have used the identity $\mathrm{e}^{x} = (\mathrm{e}^{x} - 1)/(1 - \mathrm{e}^{-x})$. Because the moment generating function is equal to 1 when $\alpha = 0$, we also have $G_f(0) = 0$. Using this initial condition and integrating the left- and right-hand sides of (9) with respect to α, we find that

$$G_f(\infty, \alpha) = \frac{\lambda(\mathrm{e}^{\alpha} - 1) + c\alpha(\theta - \mu)}{\theta},$$

and since $M_f(\infty, \alpha) = \mathrm{e}^{G_f(\infty, \alpha)}$, we attain the stated result. □

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

We can now observe that the fluid approximation is equivalent in distribution to a Poisson random variable shifted by $\gamma \equiv c(\theta - \mu)/\theta$, as the moment generating function for the Poisson distribution is $e^{\beta(e^{\alpha}-1)}$, where β is the rate of arrival and α is the space parameter of the moment generating function. This gives rise to the following result.

Theorem 4. *For the Erlang-A queue with $\lambda \geq c\mu$ and $m \in \mathbb{Z}^+$, if $\theta > \mu$,*

$$\mathrm{E}[(Q_{\mathrm{f}}(\infty) - \gamma)^m] \leq \mathrm{E}[(Q(\infty))^m] \leq \mathrm{E}[(Q_{\mathrm{f}}(\infty))^m],$$

and, if $\theta < \mu$,

$$\mathrm{E}[(Q_{\mathrm{f}}(\infty))^m] \leq \mathrm{E}[(Q(\infty))^m] \leq \mathrm{E}[(Q_{\mathrm{f}}(\infty) - \gamma)^m],$$

where $\gamma = c(\theta - \mu)/\theta$.

Proof. From Proposition 1, the fluid approximation of the moment generating function in steady-state is

$$M_f(\infty, \alpha) = e^{(\lambda(e^{\alpha}-1)+c\alpha(\theta-\mu))/\theta} = \mathrm{E}\,[e^{\alpha(\Gamma+\gamma)}]$$

where $\Gamma \sim \mathrm{Pois}(\lambda/\theta)$ and $\gamma = c(\theta - \mu)/\theta$. From the uniqueness of moment generating functions, we have

$$\mathrm{E}[(Q_{\mathrm{f}}(\infty))^m] = \mathrm{E}[(\Gamma + \gamma)^m] \quad \text{for all } m \in \mathbb{Z}^+.$$

Now, recall that for an $M/M/\infty$ queue with arrival rate λ and service rate θ, the stationary distribution is that of a Poisson random variable with rate parameter λ/θ. So, we can think of Γ as representing the steady-state distribution of an infinite-server queue with Poisson arrival rate λ and exponential service rate θ.

Suppose now that $\theta > \mu$. Then, by Theorem 2 and our preceding observation, we have $\mathrm{E}[(Q(\infty))^m] \leq \mathrm{E}[(\Gamma + \gamma)^m]$. Additionally, by comparing the steady-state infinite server queue representation of Γ to $Q(\infty)$, we can further observe that $\mathrm{E}[(Q(\infty))^m] \geq \mathrm{E}[\Gamma^m]$, as, for any state j, the service rate in $Q(\infty)$ is no more than the service rate in the same state in the Γ queueing system. Thus, we have

$$\mathrm{E}[(Q_{\mathrm{f}}(\infty) - \gamma)^m] = \mathrm{E}[\Gamma^m] \leq \mathrm{E}[(Q(\infty))^m] \leq \mathrm{E}[(\Gamma + \gamma)^m] = \mathrm{E}[(Q_{\mathrm{f}}(\infty))^m]$$

for all $m \in \mathbb{Z}^+$ whenever $\theta > \mu$. By symmetric arguments, we also find that if $\mu > \theta$ then

$$\mathrm{E}[(Q_{\mathrm{f}}(\infty))^m] = \mathrm{E}[(\Gamma + \gamma)^m] \leq \mathrm{E}[(Q(\infty))^m] \leq \mathrm{E}[\Gamma^m] = \mathrm{E}[(Q_{\mathrm{f}}(\infty) - \gamma)^m]$$

for all $m \in \mathbb{Z}^+$, as in this case $\gamma = c(\theta - \mu)/\theta < 0$. □

Remark 4. Note that in Theorem 2 we require that $Q(0) = Q_{\mathrm{f}}(0)$, but in this case we have not assumed such a condition. This is because the inequalities in Theorem 2 hold for all time, and we simply need the relationship to hold in steady state, which can be seen to occur regardless of initial conditions.

By knowing the fluid form of the moment generating function explicitly as a Poisson distribution, we can also provide exact expressions for the fluid moments and the fluid cumulant moments. These are given in the two following corollaries.

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

Corollary 6. *If $\lambda \geq c\mu$ then in steady state we find that the first n moments have the steady state expressions*

$$\mathbb{E}[Q_{\mathrm{f}}^n(\infty)] = \sum_{j=0}^{n} \binom{n}{j} \left(\frac{c(\theta-\mu)}{\theta}\right)^j \mathcal{P}_{n-j}\left(\frac{\lambda}{\theta}\right),$$

where $\mathcal{P}_m(\lambda/\theta)$ is the mth Touchard polynomial with parameter λ/θ.

Proof. This can be seen by direct use of the Poisson form of the fluid moment generating function. Let $\Gamma \sim \mathrm{Pois}(\frac{\lambda}{\theta})$ and let $\gamma = \frac{c(\theta-\mu)}{\theta}$. Then,

$$\begin{aligned}
\mathbb{E}[Q_{\mathrm{f}}^n(\infty)] &= \mathbb{E}[(\Gamma+\gamma)^n] \\
&= \sum_{j=0}^{n} \binom{n}{j} \gamma^j \mathbb{E}[\Gamma^{n-j}] \\
&= \sum_{j=0}^{n} \binom{n}{j} \gamma^j \mathcal{P}_{n-j}\left(\frac{\lambda}{\theta}\right) \\
&= \sum_{j=0}^{n} \binom{n}{j} \left(\frac{c(\theta-\mu)}{\theta}\right)^j \mathcal{P}_{n-j}\left(\frac{\lambda}{\theta}\right).
\end{aligned}$$

□

Corollary 7. *If $\lambda \geq c\mu$ then in steady state we have*

$$\left.\frac{\mathrm{d}G_f(\infty,\alpha)}{\mathrm{d}\alpha}\right|_{\alpha=0} = \frac{\lambda}{\theta} + \frac{c(\theta-\mu)}{\theta} = \mathbb{E}[Q_{\mathrm{f}}(\infty)]$$

and, for $n \geq 2$,

$$\left.\frac{\mathrm{d}^n G_f(\infty,\alpha)}{\mathrm{d}^n\alpha}\right|_{\alpha=0} = \frac{\lambda}{\theta} = \mathrm{C}^{(n)}[Q_{\mathrm{f}}(\infty)],$$

where $\mathrm{C}^{(n)}[Q_{\mathrm{f}}(\infty)]$ is defined as the nth cumulant moment of $Q_{\mathrm{f}}(\infty)$.

We now consider the second case in which $\lambda < c\mu\mathrm{e}^{-\alpha}$. Note that this now also requires a relationship involving the space parameter of the moment generating function, α. This is less general than the first case, but it allows us to derive Lemma 3.

Lemma 3. *For $\alpha \in \mathbb{R}$,*

$$\frac{\partial M_f(\infty,\alpha)}{\partial\alpha} < cM_f(\infty,\alpha)$$

if and only if $\lambda < c\mu\mathrm{e}^{-\alpha}$.

Proof. To begin, suppose that $\partial M_f(\infty,\alpha)/\partial\alpha < cM_f(\infty,\alpha)$. Using this information in conjunction with the steady-state form of the partial differential equation for the fluid moment generating function given in (7), we have

$$0 = \lambda(\mathrm{e}^{\alpha}-1)M_f(\infty,\alpha) + \theta(\mathrm{e}^{-\alpha}-1)\frac{\partial M_f(\infty,\alpha)}{\partial\alpha} - (\theta-\mu)(\mathrm{e}^{-\alpha}-1)\frac{\partial M_f(\infty,\alpha)}{\partial\alpha},$$

which simplifies to

$$\frac{\partial M_f(\infty,\alpha)}{\partial\alpha} = \frac{\lambda}{\mu}\mathrm{e}^{\alpha}M_f(\infty,\alpha).$$

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

Using our assumption, we see that

$$\frac{\lambda}{\mu}\mathrm{e}^{\alpha}M_f(\infty,\alpha) < cM_f(\infty,\alpha)$$

and this yields $\lambda < c\mu\mathrm{e}^{-\alpha}$, which shows one direction.

We now move to showing the opposite direction and instead assume that $\partial M_f(\infty,\alpha)/\partial\alpha \geq cM_f(\infty,\alpha)$. In this case, (7) is equivalently stated as

$$0 = \lambda(\mathrm{e}^{\alpha}-1)M_f(\infty,\alpha) + \theta(\mathrm{e}^{-\alpha}-1)\frac{\partial M_f(\infty,\alpha)}{\partial\alpha} - c(\theta-\mu)(\mathrm{e}^{-\alpha}-1)M_f(\infty,\alpha),$$

and this simplifies to

$$\frac{\partial M_f(\infty,\alpha)}{\partial\alpha} = \frac{\lambda(\mathrm{e}^{\alpha}-1)+c(\theta-\mu)(1-\mathrm{e}^{-\alpha})}{\theta(1-\mathrm{e}^{-\alpha})}M_f(\infty,\alpha) = \frac{\lambda\mathrm{e}^{\alpha}+c(\theta-\mu)}{\theta}M_f(\infty,\alpha).$$

Again by use of this case's assumption, we have

$$\frac{\lambda\mathrm{e}^{\alpha}+c(\theta-\mu)}{\theta}M_f(\infty,\alpha) \geq cM_f(\infty,\alpha),$$

and this now yields

$$\lambda \geq \mathrm{e}^{-\alpha}(c\theta - c(\theta-\mu)) = c\mu\mathrm{e}^{-\alpha},$$

thus completing the proof. □

We can now use this lemma to find an explicit form for the fluid approximation of the steady-state moment generating function when $\lambda < c\mu\mathrm{e}^{-\alpha}$.

Proposition 2. *For $\alpha \in \mathbb{R}$, if $\lambda < c\mu\mathrm{e}^{-\alpha}$ then in steady state we have*

$$\frac{\partial M_f(\infty,\alpha)}{\partial\alpha} = \frac{\lambda\mathrm{e}^{\alpha}}{\mu}M_f(\infty,\alpha), \tag{10}$$

which yields the solution

$$M_f(\infty,\alpha) = \mathrm{e}^{\lambda(\mathrm{e}^{\alpha}-1)/\mu}. \tag{11}$$

Proof. By Lemma 3 and our assumption that $\lambda < c\mu\mathrm{e}^{-\alpha}$, we know that $\partial M_f(\infty,\alpha)/\partial\alpha < cM_f(\infty,\alpha)$. Thus, by observing this in the steady-state moment generating function equation, we easily obtain the result in (10). Moreover, the solution to (10) can readily be verified by substitution and it is unique by the properties of linear ODEtheory. □

Here we observe that the right-hand side of (11) is equivalent to the moment generating function of a Poisson random variable with parameter λ/μ. Now, by recalling again that the steady-state distribution of an $M/M/\infty$ queue is a Poisson distribution with parameter equal to the arrival rate divided by the service rate, we find the following inequalities.

Theorem 5. *Let $\lambda < c\mu$ and $m \in \mathbb{Z}^+$. Then, if $\theta > \mu$,*

$$\mathrm{E}[\Gamma_\theta^m] \leq \mathrm{E}[Q(\infty)^m] \leq \mathrm{E}[\Gamma_\mu^m] = \mathrm{E}[Q_{\mathrm{f}}(\infty)^m],$$

and, if $\mu > \theta$,

$$\mathrm{E}[Q_{\mathrm{f}}(\infty)^m] = \mathrm{E}[\Gamma_\mu^m] \leq \mathrm{E}[Q(\infty)^m] \leq \mathrm{E}[\Gamma_\theta^m]$$

where $\Gamma_x \sim \mathrm{Pois}(\lambda/x)$ for $x > 0$.

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

Proof. In each case, the inequality involving $\Gamma_\mu \sim \text{Pois}(\lambda/\mu)$ follows directly from Proposition 2 and Theorem 2 via the observation that the fluid form of the moment generating function is equivalent in distribution to that of Γ_μ. Thus, we are left to prove the inequalities for $\Gamma_\theta \sim \text{Pois}(\lambda/\theta)$.

To this end, let us first note that the stationary distribution of an $M/M/\infty$ queue with service rate θ is equivalent to that of Γ_θ. Suppose now that $\theta > \mu$. Then, any state of such an $M/M/\infty$ queue has a larger rate of departure than the same state in the Erlang-A system. Thus, we have

$$\mathrm{E}[\Gamma_\theta^m] \le \mathrm{E}[Q(\infty)^m] \le \mathrm{E}[\Gamma_\mu^m] \quad \text{for all } m \in \mathbb{Z}^+.$$

By symmetric arguments in the $\theta < \mu$ case, we complete the proof. □

As we did for the case in which $\lambda \ge c\mu$, we can use these findings to give explicit expressions for the fluid approximations of the moments and the cumulant moments.

Corollary 8. *If $\lambda < c\mu$ then in steady state we have*

$$\frac{\mathrm{d}G_f(\infty,\alpha)}{\mathrm{d}\alpha}\bigg|_{\alpha=0} = \frac{\lambda}{\mu} = \mathbb{E}[Q_{\mathrm{f}}(\infty)],$$

and, for $n \in \mathbb{Z}^+$,

$$\frac{\mathrm{d}^n G_f(\infty,\alpha)}{\mathrm{d}^n\alpha}\bigg|_{\alpha=0} = \frac{\lambda}{\mu} = \mathrm{C}^{(n)}[Q_{\mathrm{f}}(\infty)],$$
$$\frac{\mathrm{d}^n M_f(\infty,\alpha)}{\mathrm{d}^n\alpha}\bigg|_{\alpha=0} = \mathcal{P}_n\left(\frac{\lambda}{\mu}\right) = \mathrm{E}[Q_{\mathrm{f}}(\infty)^n],$$

where $\mathrm{C}^{(n)}[Q_{\mathrm{f}}(\infty)]$ is defined as the nth cumulant moment of $Q_{\mathrm{f}}(\infty)$ and $\mathcal{P}_m(\lambda/\mu)$ is the mth Touchard polynomial with parameter λ/μ.

4.4. Characterization of the nonstationary moment generating function

Many scenarios that feature customer abandonments may also feature an arrival process that is nonstationary. To address this, we now incorporate a point process that can be used to approximate any generalized periodic, nonstationary arrival rate function, as discussed in [8]. Specifically, we define $\lambda(t)$ by a Fourier series: let λ_0 and $\{(a_k, b_k), k \in \mathbb{Z}^+\}$ be such that

$$\lambda(t) = \lambda_0 + \sum_{k=1}^{\infty} (a_k \sin(kt) + b_k \cos(kt)). \tag{12}$$

We now take $\lambda(t)$ as the rate of arrivals at time t in the Erlang-A model. Under this setting, we derive the following expression for the cumulant moment generating function of the fluid approximation and its corresponding partial differential equation whenever the arrival rate is greater than a certain threshold. We do so through a series of technical lemmas. First, we bound the fluid mean when the arrival rate and initial value are sufficiently large.

Lemma 4. *Suppose that $\underline{\lambda} \equiv \inf_{t\ge 0} \lambda(t) > c\mu$ and that $\mathrm{E}[Q_{\mathrm{f}}(0)] > c$. Then*

$$\mathrm{E}[Q_{\mathrm{f}}(t)] > c$$

for all time $t \ge 0$.

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

Proof. We have seen in (4) that $\mathrm{E}[Q_{\mathrm{f}}(t)]$ evolves according to

$$\dot{\mathbb{E}}[Q_{\mathrm{f}}(t)] = \lambda(t) - \mu(\mathrm{E}[Q_{\mathrm{f}}(t)] \wedge c) - \theta(\mathrm{E}[Q_{\mathrm{f}}(t)] - c)^{+}$$

at all times t. Now, suppose that $\hat{t} > 0$ is a time such that $\mathrm{E}[Q_{\mathrm{f}}(\hat{t})] = c + \varepsilon$ for some $\varepsilon > 0$. Then, if $\varepsilon < (\underline{\lambda} - c\mu)/\theta$, we have

$$\dot{\mathbb{E}}[Q_{\mathrm{f}}(\hat{t})] = \lambda(\hat{t}) - c\mu - \theta\varepsilon \geq \underline{\lambda} - c\mu - \theta\varepsilon > 0.$$

By the continuity of the fluid mean and the fact that $\mathrm{E}[Q_{\mathrm{f}}(0)] > c$, we see that $\mathrm{E}[Q_{\mathrm{f}}(t)] > c$ for all time $t \geq 0$. □

With this in hand, we now also provide the moment generating function for an $M/M/\infty$ queue with nonstationary arrival rate $\lambda(t)$, which we will use for comparison later in this section.

Lemma 5. *Let $Q_\infty(t)$ be the number in system for an infinite-server queue with periodic Poisson arrival rate $\lambda(t)$ as defined in (12), exponential service rate μ, and initial value $Q_\infty(0) = q_0$. Then*

$$\begin{aligned}\mathrm{E}\left[\mathrm{e}^{\alpha Q_\infty(0)}\right] = \exp\Bigg((\mathrm{e}^{\alpha} - 1)\Bigg(&\frac{\lambda_0}{\mu}(1 - \mathrm{e}^{-\mu t}) \\ &+ \sum_{k=1}^{\infty} \frac{(a_k\mu + b_k k)\sin(kt) + (b_k\mu - a_k k)(\cos(kt) - \mathrm{e}^{-\mu t})}{\mu^2 + k^2}\Bigg)\Bigg) \\ \times (\mathrm{e}^{-\mu t}(\mathrm{e}^{\alpha} - 1) + 1)^{q_0}\end{aligned}$$

for all $t \geq 0$ and $\alpha \in \mathbb{R}$.

Proof. To start, the time derivative of the moment generating function is

$$\frac{\mathrm{dE}\left[\mathrm{e}^{\alpha Q_\infty(0)}\right]}{\mathrm{d}t} = \lambda(t)(\mathrm{e}^{\alpha} - 1)\mathrm{E}\left[\mathrm{e}^{\alpha Q_\infty(0)}\right] + \mu(\mathrm{e}^{-\alpha} - 1)\mathrm{E}\left[Q_\infty(t)\mathrm{e}^{\alpha Q_\infty(t)}\right].$$

This differential equation can be viewed as a partial differential equation when expressed as

$$\mu(1 - \mathrm{e}^{-\alpha})\frac{\partial M(t,\alpha)}{\partial \alpha} + \frac{\partial M(t,\alpha)}{\partial t} = \lambda(t)(\mathrm{e}^{\alpha} - 1)M(t,\alpha),$$

where $M(t,\alpha)$ is the moment generating function at time t and space parameter α. To simplify our effort, we instead consider the differential equation for the cumulant moment generating function, which is $G(\alpha, t) = \log(M(t,\alpha))$. This PDE is

$$\mu(1 - \mathrm{e}^{-\alpha})\frac{\partial G(t,\alpha)}{\partial \alpha} + \frac{\partial G(t,\alpha)}{\partial t} = \lambda(t)(\mathrm{e}^{\alpha} - 1)$$

with the initial condition that

$$G(0,\alpha) = \log(\mathrm{E}[\mathrm{e}^{\alpha Q_\infty(0)}]) = \log(\mathrm{e}^{\alpha q_0}) = \alpha q_0.$$

Using the compressed notation $G_x = \partial G/\partial x$, we seek to solve the system

$$\mu(1 - \mathrm{e}^{-\alpha})G_\alpha + G_t = \lambda(t)(\mathrm{e}^{\alpha} - 1), \qquad G(0,\alpha) = \alpha q_0,$$

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

and we do so via the method of characteristics. For this approach, we introduce the characteristic variables r and s and establish the characteristic equations, which are ODEs, as

$$\frac{\mathrm{d}\alpha}{\mathrm{d}s}(r,s)=\mu(1-\mathrm{e}^{-\alpha}),\qquad \frac{\mathrm{d}t}{\mathrm{d}s}(r,s)=1,\qquad \frac{\mathrm{d}g}{\mathrm{d}s}(r,s)=\lambda(t)(\mathrm{e}^{\alpha}-1)$$

with the initial conditions

$$\alpha(r,0)=r,\qquad t(r,0)=0,\qquad g(r,0)=rq_0.$$

We can first see that the ODEs for α and t solve to

$$\begin{aligned}\alpha(r,s)=\log(\mathrm{e}^{c_1(r)+\mu s}+1) \quad &\longrightarrow \quad \alpha(r,s)=\log((\mathrm{e}^{r}-1)\mathrm{e}^{\mu s}+1),\\ t(r,s)=s+c_2(r) \quad &\longrightarrow \quad t(r,s)=s,\end{aligned}$$

and so we can now use these to solve the remaining ODE. After substituting we have

$$\frac{\mathrm{d}g}{\mathrm{d}s}(r,s)=\lambda(s)(\mathrm{e}^{r}-1)\mathrm{e}^{\mu s},$$

which gives the solution

$$\begin{aligned}g(r,s)=(\mathrm{e}^{r}-1)&\\ \times\left(\frac{\lambda_0}{\mu}(\mathrm{e}^{\mu s}-1)+\sum_{k=1}^{\infty}\frac{(a_k\mu+b_k k)\sin(ks)\mathrm{e}^{\mu s}+(b_k\mu-a_k k)(\cos(ks)\mathrm{e}^{\mu s}-1)}{\mu^2+k^2}\right)&\\ +rq_0.&\end{aligned}$$

So, using $s=t$ and $r=\log(\mathrm{e}^{-\mu t}(\mathrm{e}^{\alpha}-1)+1)$, we have

$$\begin{aligned}G(t,\alpha)&=g(\log(\mathrm{e}^{-\mu t}(\mathrm{e}^{\alpha}-1)+1),t)\\ &=(\mathrm{e}^{\alpha}-1)\left(\frac{\lambda_0}{\mu}(1-\mathrm{e}^{-\mu t})\right.\\ &\qquad\left.+\sum_{k=1}^{\infty}\frac{(a_k\mu+b_k k)\sin(kt)+(b_k\mu-a_k k)(\cos(kt)-\mathrm{e}^{-\mu t})}{\mu^2+k^2}\right)\\ &\quad+\log(\mathrm{e}^{-\mu t}(\mathrm{e}^{\alpha}-1)+1)q_0\end{aligned}$$

and, therefore, by solving for $M(t,\alpha)=\mathrm{e}^{G(t,\alpha)}$ we attain the stated result. □

Now that we have established these lemmas we proceed with the analysis of the nonstationary Erlang-A model. In the next theorem we give explicit forms for the fluid form of the cumulant moment generating function and its corresponding partial differential equation.

Theorem 6. *If* $\inf_{t\leq\infty}\lambda(t)\equiv\underline{\lambda}>c\mu$, *where* $\lambda(t)$ *is given by (12) and* $Q_{\mathrm{f}}(0)=q_0>c$, *then, for all* $t\geq 0$, *we have*

$$\frac{\partial G_f(t,\alpha)}{\partial t}=\lambda(t)(\mathrm{e}^{\alpha}-1)+\theta(\mathrm{e}^{-\alpha}-1)\frac{\partial G_f(t,\alpha)}{\partial\alpha}-c(\theta-\mu)(\mathrm{e}^{-\alpha}-1),\tag{13}$$

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

which gives the solution

$$G_f(t,\alpha) = (\mathrm{e}^{\alpha}-1) \times \left(\frac{\lambda_0}{\theta}(1-\mathrm{e}^{-\theta t}) + \sum_{k=1}^{\infty} \frac{(a_k\theta + b_k k)\sin(kt) + (b_k\theta - a_k k)(\cos(kt) - \mathrm{e}^{-\theta t})}{\theta^2 + k^2}\right) + \frac{c(\theta-\mu)}{\theta}\alpha + \log\left((\mathrm{e}^{\alpha}-1)\mathrm{e}^{-\theta t}+1\right)\left(q_0 - \frac{c(\theta-\mu)}{\theta}\right) \tag{14}$$

for all $t \geq 0$ and all $\alpha \geq 0$.

Proof. From (8), the PDE for the fluid approximation's cumulant moment generating function is

$$\frac{\partial G_f(t,\alpha)}{\partial t} = \lambda(t)(\mathrm{e}^{\alpha}-1) + \theta(\mathrm{e}^{-\alpha}-1)\frac{\partial G_f(t,\alpha)}{\partial \alpha} - (\theta-\mu)(\mathrm{e}^{-\alpha}-1)\left(\frac{\partial G_f(t,\alpha)}{\partial \alpha} \wedge c\right).$$

Now, recall that $\partial G_f(t,\alpha)/\partial\alpha = \mathrm{E}[Q_{\mathrm{f}}(t)\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}]/\mathrm{E}[\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}]$. Using the FKG inequality and our observation from Lemma 4 that $\mathrm{E}[Q_{\mathrm{f}}(t)] > c$, we have

$$\mathrm{E}[Q_{\mathrm{f}}(t)\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}] \geq \mathrm{E}[Q_{\mathrm{f}}(t)]\mathrm{E}[\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}] > c\mathrm{E}[\mathrm{e}^{\alpha Q_{\mathrm{f}}(t)}],$$

and so $(\partial G_f(t,\alpha)/\partial\alpha \wedge c) = c$. Thus, we have the PDE given in (13) and so now we seek to find its solution. We approach this via the method of characteristics. Because $G_f(0,\alpha) = \log(\mathrm{E}[\mathrm{e}^{\alpha Q_{\mathrm{f}}(0)}]) = \alpha q_0$, we see that we seek to solve the system

$$\theta(1-\mathrm{e}^{-\alpha})G_{(\alpha)} + G_{(t)} = \lambda(t)(\mathrm{e}^{\alpha}-1) + c(\theta-\mu)(1-\mathrm{e}^{-\alpha}), \qquad G_f(0,\alpha) = \alpha q_0,$$

where $G_{(x)} = \partial G_f/\partial x$. Introducing characteristic variables r and s, we have the characteristic ODEs as

$$\frac{\mathrm{d}\alpha}{\mathrm{d}s}(r,s) = \theta(1-\mathrm{e}^{-\alpha}), \qquad \frac{\mathrm{d}t}{\mathrm{d}s}(r,s) = 1,$$
$$\frac{\mathrm{d}g}{\mathrm{d}s}(r,s) = \lambda(t)(\mathrm{e}^{\alpha}-1) + c(\theta-\mu)(1-\mathrm{e}^{-\alpha})$$

with initial conditions $\alpha(r,0) = r$, $t(r,0) = t$, and $g(r,0) = rq_0$. Then we can solve the first two ODEs to show that

$$\alpha(r,s) = \log((\mathrm{e}^{r}-1)\mathrm{e}^{\theta s}+1), \qquad t(r,s) = s.$$

Substituting these solutions into the remaining ODE, we have

$$\frac{\mathrm{d}g}{\mathrm{d}s}(r,s) = \lambda(s)\mathrm{e}^{\theta s}(\mathrm{e}^{r}-1) + c(\theta-\mu)\frac{\mathrm{e}^{\theta s}(\mathrm{e}^{r}-1)}{\mathrm{e}^{\theta s}(\mathrm{e}^{r}-1)+1},$$

and this now solves to

$$g(r,s) = (\mathrm{e}^{r}-1) \times \left(\frac{\lambda_0}{\theta}(\mathrm{e}^{\theta s}-1) + \sum_{k=1}^{\infty} \frac{(a_k\theta + b_k k)\sin(ks)\mathrm{e}^{\theta s} + (b_k\theta - a_k k)(\cos(ks)\mathrm{e}^{\theta s} - 1)}{\theta^2 + k^2}\right) + \frac{c(\theta-\mu)}{\theta}(\log((\mathrm{e}^{r}-1)\mathrm{e}^{\theta s}+1) - r) + rq_0.$$

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

Now, we can rearrange our solutions to find $s = t$ and $r = \log((e^{\alpha} - 1)e^{-\theta t} + 1)$. Then, we have

$$\begin{aligned} G_f(t, \alpha) &= g(\log((e^{\alpha} - 1)e^{-\theta t} + 1), t) \\ &= (e^{\alpha} - 1)e^{-\theta t} \\ &\quad \times \left(\frac{\lambda_0}{\theta}(e^{\theta t} - 1) + \sum_{k=1}^{\infty} \frac{(a_k\theta + b_k k)\sin(kt)e^{\theta t} + (b_k\theta - a_k k)(\cos(kt)e^{\theta t} - 1)}{\theta^2 + k^2} \right) \\ &\quad + \frac{c(\theta - \mu)}{\theta}(\alpha - \log((e^{\alpha} - 1)e^{-\theta t} + 1)) + \log((e^{\alpha} - 1)e^{-\theta t} + 1)q_0, \end{aligned}$$

and this simplifies to the stated result. □

Like the approach in our investigation of the steady-state scenario, we can now observe that the fluid approximation is equivalent in distribution to the number in system for an infinite-server queue shifted by $\gamma \equiv c(\theta - \mu)/\theta$. This gives rise to the following result.

Theorem 7. *For the Erlang-A queue with periodic arrival rate $\lambda(t)$ given by (12) such that $\underline{\lambda} \equiv \inf_{t \geq 0} \lambda(t) > c\mu$ and initial value $q_0 > c$, the fluid approximation of the moment generating function is equal to the moment generating function of a shifted $M/M/\infty$ queue length process with arrival rate $\lambda(t)$, service rate θ, initial value $q_0 - c(\theta - \mu)/\theta$, and linear shift $c(\theta - \mu)/\theta$.*

Proof. Observe from Theorem 6 that the fluid moment generating function for the Erlang-A queue under these conditions is

$$\begin{aligned} M_f(t, \alpha) &= e^{G_f(t,\alpha)} \\ &= \exp\Bigg((e^{\alpha} - 1) \\ &\quad \times \left(\frac{\lambda_0}{\theta}(1 - e^{-\theta t}) \sum_{k=1}^{\infty} \frac{(a_k\theta + b_k k)\sin(kt) + (b_k\theta - a_k k)(\cos(kt) - e^{-\theta t})}{\theta^2 + k^2} \right) \\ &\quad + \frac{c(\theta - \mu)}{\theta}\alpha \Bigg)((e^{\alpha} - 1)e^{-\theta t} + 1)^{q_0 - c(\theta - \mu)/\theta}, \end{aligned}$$

which is of a form that we can recognize. Comparing it to Lemma 5, we can see that Q_f is equivalent to the number in system of a shifted $M/M/\infty$ queue with arrival rate $\lambda(t)$, service rate θ, initial value $q_0 - c(\theta - \mu)/\theta$, and linear shift $c(\theta - \mu)/\theta$, thus enforcing that the fluid model does start at q_0. □

This representation of the fluid approximation now allows us to provide upper and lower bounds for the moments of the Erlang-A system.

Corollary 9. *Let $Q(t)$ represent the Erlang-A queue with periodic arrival rate $\lambda(t)$ given by (12) such that $\underline{\lambda} \equiv \inf_{t \geq 0} \lambda(t) > c\mu$ and initial value $q_0 > c$, and let $Q_f(t)$ represent the corresponding fluid approximation. Then, if $\theta > \mu$,*

$$\mathrm{E}[(Q_f(t) - \gamma)^m] \leq \mathrm{E}[Q(t)^m] \leq \mathrm{E}[Q_f(t)^m],$$

and, if $\theta < \mu$,

$$\mathrm{E}[Q_f(t)^m] \leq \mathrm{E}[Q(t)^m] \leq \mathrm{E}[(Q_f(t) - \gamma)^m],$$

for all time $t > 0$ and all $m \in \mathbb{Z}^+$, where $\gamma = c(\theta - \mu)/\theta$.

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

Proof. In each case, the bound involving the fluid approximation of the moment is a direct consequence of Theorem 2 and so only the other two bounds remain to be shown. We now note that since we have characterized the fluid approximation as a shifted $M/M/\infty$ queue, the remaining bounds are from the unshifted version of this system and, by following the same arguments as in Theorems 4 and 5 regarding the rates of departure in the corresponding states of the Erlang-A queue and the $M/M/\infty$ queue, this completes the proof. □

4.5. Numerical results

In this subsection we provide numerical experiments that demonstrate the findings of the previous subsections. For the interested reader, we note that an extended version of this paper containing additional plots and figures is available on arxiv.

In Figure 10 we plot the limiting distribution for the steady-state Erlang-A model. For these plots we take $\lambda = 20$ and $\mu = 1$, and then vary θ and c. For the three plots on the left, we take the abandonment rate to be $\theta = 0.5$ and, for those on the right, we set $\theta = 2$. For the top two plots, we set the number of servers as $c = 15$, in the middle two $c = 20$, and in the bottom two we make $c = 25$. We observe that the approximate distribution is quite close when λ is not near $c\mu$, but the approximation is less accurate when $\lambda = c\mu$. This finding is consistent with much of the literature that focuses on finding novel approximations for queueing networks and optimal control of these networks; see, for example, [17], [18], [19], and [33]. We note here that these approximations are not all of the same form: recall that when $\lambda \geq c\mu$ the fluid approximation is equivalent in distribution to a shifted Poisson random variable with parameter λ/θ, but when $\lambda < c\mu$ it is equivalent to a Poisson distribution with parameter λ/μ.

In Figure 11 we examine the limiting distributions for the single server case. In these plots we set $\mu = 1$ and then vary the arrival rate and the abandonment rate. For all plots on the left, we set $\theta = 0.5$ and on the right $\theta = 2$. Furthermore, for the top, middle, and bottom pairs of plots, we set λ to 0.8, 1, and 1.2, respectively. As in Figure 10, Figure 11 shows that our approximations are quite good. Thus, we are able to capture single server dynamics as well as large-scale multiserver dynamics even though they are quite different. This is even more useful as our approximations are nonasymptotic and do not rely on scaling the number of servers.

In Figure 12 we take the arrival rate as $\lambda(t) = 6.5 + \sin(t)$, the service rate as $\mu = 1$, and the number of servers as $c = 5$. Because $\inf_{t\geq 0} \lambda(t) > c\mu$, we use the characterization of the fluid approximation as a shifted $M/M/\infty$ queue and compare the simulated system, the fluid approximation, and the unshifted $M/M/\infty$, as stated in Theorem 7. We consider the mean for $\theta = 1.1$ and $\theta = 0.9$ and find that while the fluid approximation is quite close the unshifted system is not near to the Erlang-A system, even for these relatively similar rates of service and abandonment. We note that the simulated values in Figure 12 are plotted using a dashed line, and the differences between these points and the fluid approximations can be most readily observed at the extremes of the trigonometric function.

As a final comparison, in Table 1 we test the bounds in Theorems 4 and 5 by comparison to simulations of the first three moments. We perform this experiment with $c = 10$ and $\mu = 1$ while taking $\lambda \in \{8, 10, 12\}$ and $\theta \in \{0.5, 2\}$ so that we have considered all different conditions contained in these two results. As can also be observed in Figure 12, the empirical values are much closer to the fluid approximation than they are to the opposite bounds from Theorems 4 and 5. This is to be expected since the fluid values are the only intended approximation. Still, the pursuit of closer bounds in conjunction with the fluid approximation may be a future direction of practical importance.

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

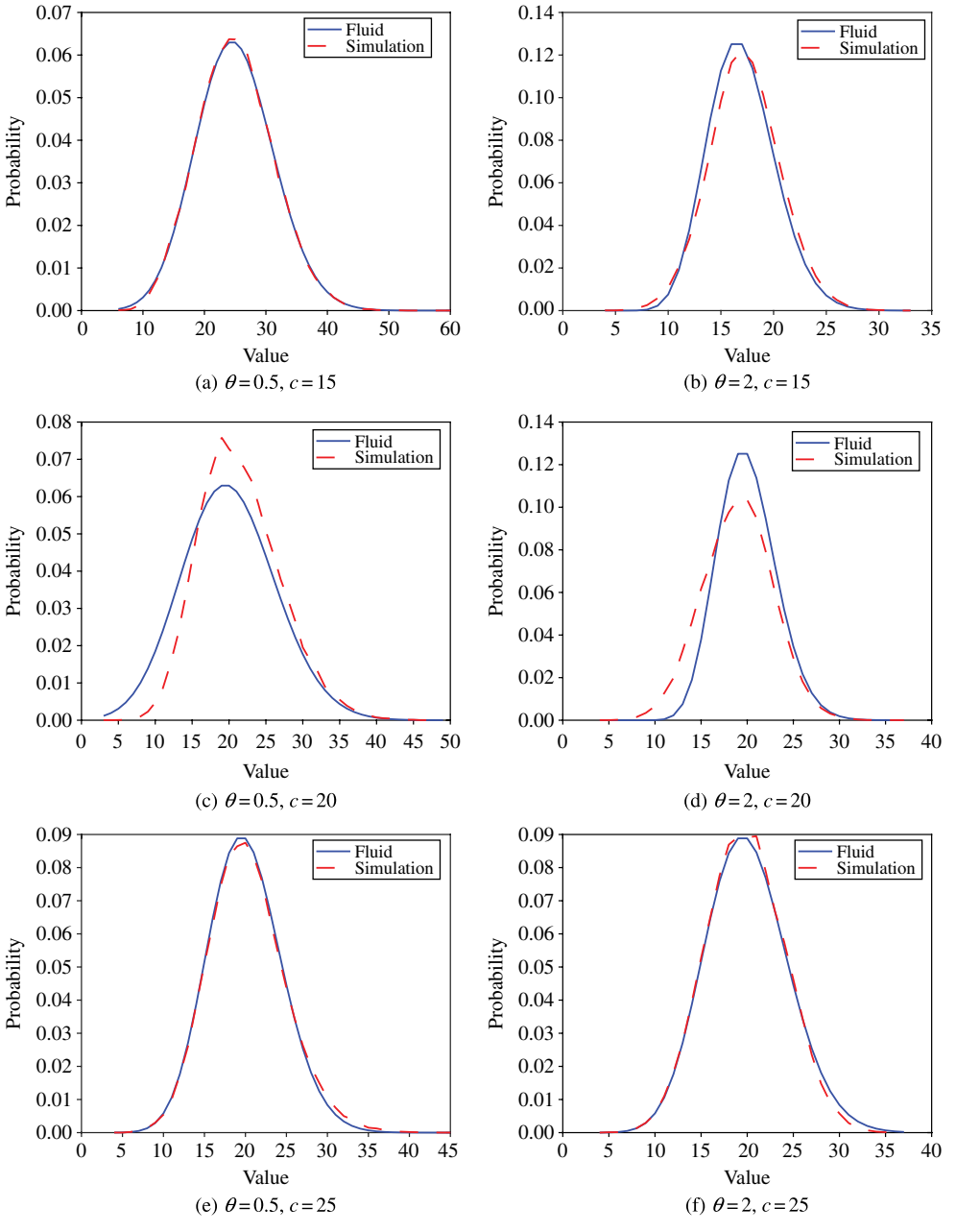

FIGURE 10: Empirical and fluid limiting distributions for $\lambda = 20$ and $\mu = 1$.

5. Conclusion

In this paper we investigated the Erlang-A queueing system through comparison to the fluid approximations of its moments, moment generating function, and cumulant moment generating function. Through recognizing the convexity in the differential equations describing

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

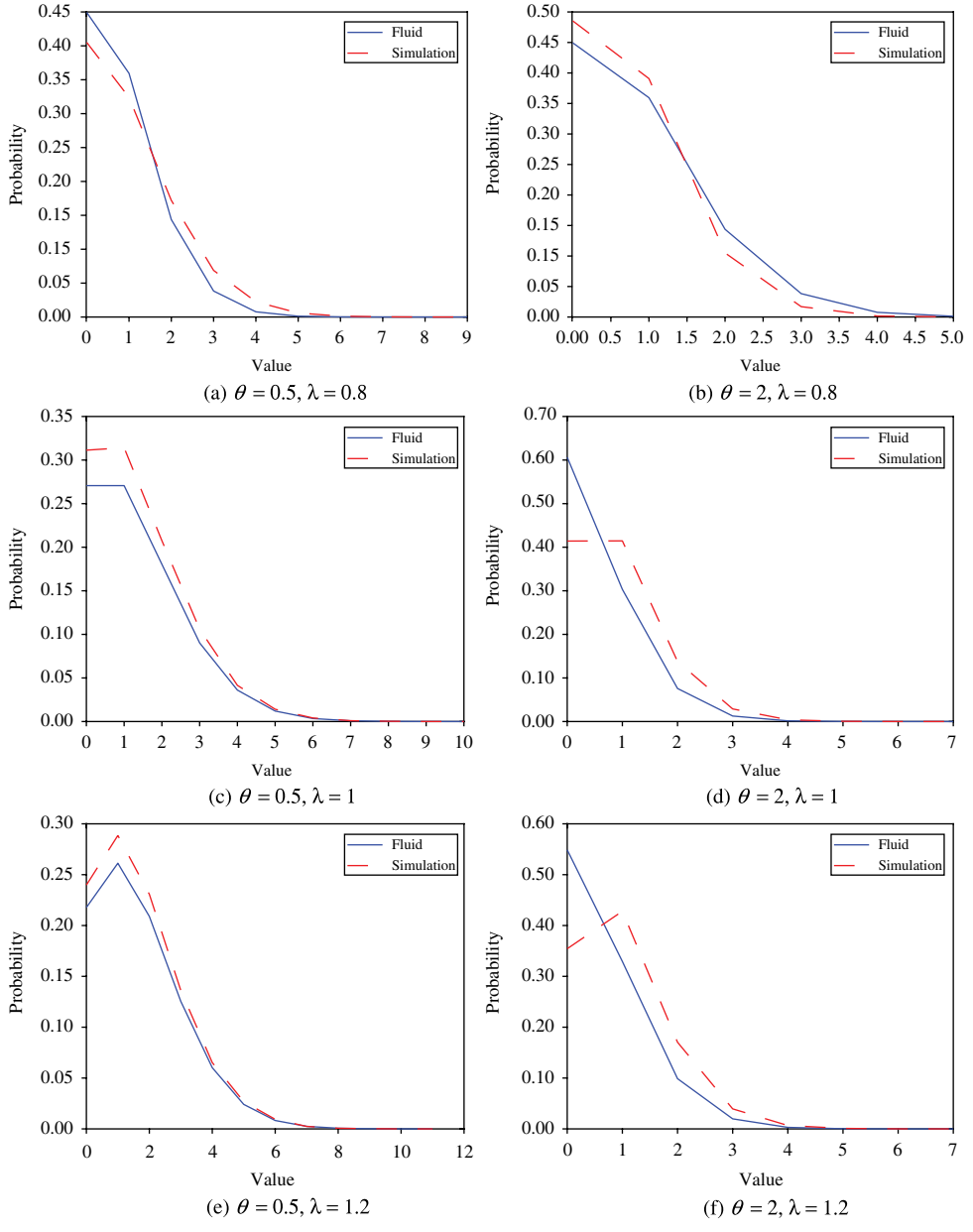

FIGURE 11: Empirical and fluid limiting distributions for $c = 1$ and $\mu = 1$.

these approximations, we found fundamental relationships between the values of these quantities and their fluid counterparts: when the rate of abandonment is less than the rate of service the true value dominates the approximation, when the service rate is smaller the approximation dominates the true value, and when the rates of abandonment and service are equal, the two are equivalent. That is, for any (possibly nonstationary) arrival rate, and for θ

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

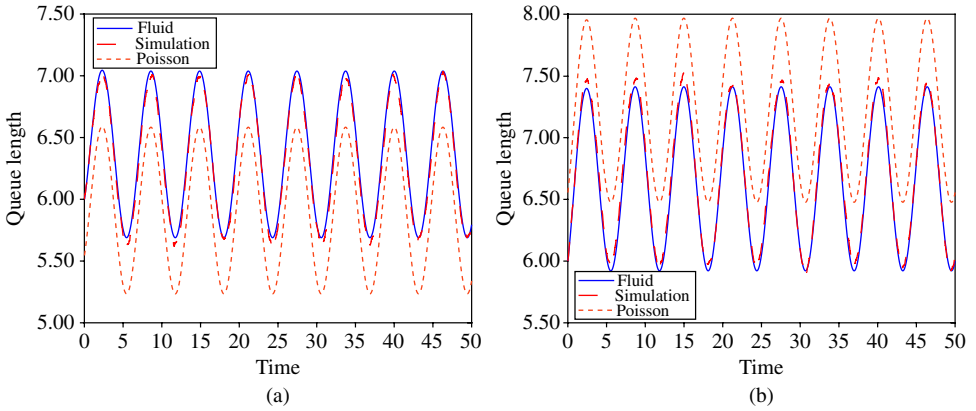

FIGURE 12: The queue mean for $\lambda(t) = 6.5 + \sin(t)$, $\mu = 1$, $Q(0) = 6$, $c = 5$, and (a) $\theta = 1.1$ and (b) $\theta = 0.9$.

TABLE 1: Comparison of the opposite (opp.) bounds from Theorems 4 and 5 with simulations (sim.) for the first three moments for various λ and θ, where $\mu = 1$ and $c = 10$.

	1st			2nd			3rd		
Scenario	Sim	Fluid	Opp.	Sim	Fluid	Opp.	Sim	Fluid	Opp.
$\lambda = 8, \theta = .5$	8.4	8	16	80.4	72	272	864.4	712	4880
$\lambda = 8, \theta = 2$	7.8	8	4	67	72	20	626.8	712	116
$\lambda = 10, \theta = .5$	11.1	10	20	137.4	120	420	1883.5	1620	9220
$\lambda = 10, \theta = 2$	9.2	10	5	92.5	105	30	986.8	1155	205
$\lambda = 12, \theta = .5$	14.3	14	24	225.8	220	600	3873.7	3776	15576
$\lambda = 12, \theta = 2$	10.6	11	6	119.4	127	42	1422.3	1535	330

TABLE 2: Approximation characterization overview.

Parameter condition	Approximation and model relationship
$\theta > \mu$	Approximation $\geq$ true value
$\theta < \mu$	Approximation $\leq$ true value
$\theta = \mu$	The approximation and the true value are equal

as the abandonment rate and μ as the service rate, we can summarize these relationships as follows.

Note that the parameter conditions in Table 2 have actually been subjected to some specification in the case of the generating functions. As used in Subsections 4.1 and 4.2, the general conditions are actually $(\theta - \mu)(e^{-\alpha} - 1) < 0$, $(\theta - \mu)(e^{-\alpha} - 1) > 0$, and $(\theta - \mu)(e^{-\alpha} - 1) > 0$. However, as we remarked previously, these conditions are equivalent when the space parameter of the generating functions is restricted to positive values, i.e. $\alpha > 0$.

In forming these inequalities we have found explicit representations of the fluid approximations through equivalences in distribution with Poisson random variables and infinite-server queues, in cases of stationary and nonstationary arrival rates, respectively. The respective subsection references for these are Subsections 4.3 and 4.4. For convenience, in Table 3 we give an overview of the bounds and characterizations shown in this paper.

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

TABLE 3: Approximation characterization overview.

Model type	Fluid characterization	Opposite bound
Stationary with $\lambda \geq c\mu$	$\text{Pois}(\frac{\lambda}{\theta}) + \frac{c(\theta-\mu)}{\theta}$	$\text{Pois}(\frac{\lambda}{\theta})$
Stationary with $\lambda < c\mu$	$\text{Pois}(\frac{\lambda}{\mu})$	$\text{Pois}(\frac{\lambda}{\theta})$
Periodic with $\underline{\lambda} > c\mu$	$M_t/M(\theta)/\infty$ shifted by $\frac{c(\theta-\mu)}{\theta}$	$M_t/M(\theta)/\infty$

We note here that we use the notation $M_t/M(\theta)/\infty$ to represent an infinite-server queue with arrivals according to the nonstationary Poisson process with rate $\lambda(t)$ as given in (12) and exponential service with rate $\theta > 0$. We should also note that the periodic case also requires the initial number in system to be greater than c. These characterizations both give insight into the approximations themselves and yield natural inequalities that complement those from the approximations. We have demonstrated the performance of these bounds through simulations. Through consideration of both these findings and the empirical experiments, we can identify interesting directions for future work. For example, it would be of great interest to gain more explicit insights into the gap between the fluid approximations and the true values. This is a nontrivial endeavor, which stems from the nondifferentiability and nonclosure in the differential equations for the true expectations. The numerical experiments in this work indicate that the fluid approximations may often be quite close but not exact, and additional understanding would be useful in practice. Moreover, extending our results to more complicated queueing systems where the arrival and service processes follow phase type distributions is of interest given the new work of Ko and Pender [36], [22], [21].

It would be additionally useful to gain a better understanding of the limiting distribution of the Erlang-A queue. As we discuss in the paper, the empirical experiments in Subsection 4.5 indicate that the true limiting distributions closely resemble the shifted Poisson approximations. In particular, the approximations seem quite close when λ is not near $c\mu$. As a simple extension of this work, it can be observed that some sort of combination of the approximation when $\lambda < c\mu$ and of the approximation when $\lambda > c\mu$ could make a nice choice for approximation of the distribution when $\lambda = c\mu$. In some sense, it is not surprising that these approximations are similar to the true limiting distribution, as the Erlang-A queue appears to be an $M/M/\infty$ queue with service rate μ (the approximation when $\lambda < c\mu$), when only considering the states up to c, and it also resembles some sort of shifted $M/M/\infty$ queue with service rate θ (which also describes the approximation when $\lambda \geq c\mu$) for states $c+1$ and beyond.

Because the relationships between the true quantities and their approximations are conditioned only on the service and abandonment rate, it may be possible for this to be extended to stochastic-intensity, non-Poisson arrival processes, such as the Hawkes process or shot noise driven queues studied in [23], [24], and [6]. Finally, in a similar manner it would be interesting to extend this to networks of Erlang-A queues; however, we would have to keep track of the routeing probabilities carefully to keep track of the convexity/concavity of the rate functions. We plan to consider these extensions in future work.

Acknowledgements

The authors are grateful for the generous support of the National Science Foundation through J. Pender's Civil, Mechanical and Manufacturing Innovation (CMMI) Career Award (1751975), and A. Daw's Graduate Research Fellowship (grant number DGE-1650441).

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

References

[1] BACCELLI, F. AND HEBUTERNE, G. (1981). On queues with impatient customers. In *Performance*, ed. F. J. Kylstra, North-Holland, pp. 159–179.

[2] BORST, S., MANDELBAUM, A. AND REIMAN, M. I. (2004). Dimensioning large call centers. *Operat. Res.* **52,** 17–34.

[3] BOXMA, O. J. AND DE WAAL, P. R. (1994). Multiserver queues with impatient customers. *Teletraffic Sci. Eng.* **1,** 743–756.

[4] BRAVERMAN, A., DAI, J. G. AND FENG, J. (2016). Stein's method for steady-state diffusion approximations: an introduction through the Erlang-A and Erlang-C models. *Stoch. Systems* **6,** 301–366.

[5] BROWN, L. *et al.* (2005). Statistical analysis of a telephone call center: a queueing-science perspective. *J. Amer. Statist. Assoc.* **100,** 36–50.

[6] DAW, A. AND PENDER, J. (2018). Queues driven by Hawkes processes. *Stoch. Systems* **8,** 192–229.

[7] DONG, J., FELDMAN, P. AND YOM-TOV, G. B. (2015). Service systems with slowdowns: potential failures and proposed solutions. *Operat. Res.* **63,** 305–324.

[8] EICK, S. G., MASSEY, W. A. AND WHITT, W. (1993). $M_t/G/\infty$ queues with sinusoidal arrival rates. *Manag. Sci.* **39,** 241–252.

[9] ENGBLOM, S. AND PENDER, J. (2014). Approximations for the moments of nonstationary and state dependent birth-death queues. Preprint. Available at https://arxiv.org/abs/1406.6164.

[10] FELDMAN, Z., MANDELBAUM, A., MASSEY, W. A. AND WHITT, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Manag. Sci.* **54,** 324–338.

[11] FERRAGUT, A. AND PAGANINI, F. (2012). Content dynamics in P2P networks from queueing and fluid perspectives. In *Proc. 24th Internat. Teletraffic Congress*, IEEE, p. 11.

[12] FRALIX, B. H. (2013). On the time-dependent moments of Markovian queues with reneging. *Queueing Systems* **75,** 149–168.

[13] GARNETT, O., MANDELBAUM, A. AND REIMAN, M. (2002). Designing a call center with impatient customers. *Manufacturing Service Operat. Manag.* **4,** 208–227.

[14] GURVICH, I., HUANG, J. AND MANDELBAUM, A. (2014). Excursion-based universal approximations for the Erlang-A queue in steady-state. *Math. Operat. Res.* **39,** 325–373.

[15] HALE, J. K. AND VERDUYN LUNEL, S. M. (2013). *Introduction to Functional Differential Equations*, Vol. 99. Springer.

[16] HALFIN, S. AND WHITT, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operat. Res.* **29,** 567–588.

[17] HAMPSHIRE, R. C. AND MASSEY, W. A. (2010). Dynamic optimization with applications to dynamic rate queues. *INFORMS Tutorials Operat. Res.* **2010,** 208–247.

[18] HAMPSHIRE, R. C., JENNINGS, O. B. AND MASSEY, W. A. (2009). A time-varying call center design via Lagrangian mechanics. *Prob. Eng. Inf. Sci.* **23,** 231–259.

[19] HAMPSHIRE, R. C., MASSEY, W. A. AND WANG, Q. (2009). Dynamic pricing to control loss systems with quality of service targets. *Prob. Eng. Inf. Sci.* **23,** 357–383.

[20] KNESSL, C. AND VAN LEEUWAARDEN, J. S. H. (2015). Transient analysis of the Erlang-A model. *Math. Meth. Operat. Res.* **82,** 143–173.

[21] KO, Y. M. AND PENDER, J. (2017). Diffusion limits for the $(MAP_t/Ph_t/\infty)^N$ queueing network. *Operat. Res. Lett.* **45,** 248–253.

[22] KO, Y. M. AND PENDER, J. (2018). Strong approximations for time-varying infinite-server queues with non-renewal arrival and service processes. *Stoch. Models* **34,** 186–206.

[23] KOOPS, D. T., BOXMA, O. J. AND MANDJES, M. R. H. (2017). Networks of ·/G/∞ queues with shot-noise-driven arrival intensities. *Queueing Systems* **86,** 301–325.

[24] KOOPS, D. T., SAXENA, M., BOXMA, O. J. AND MANDJES, M. (2018). Infinite-server queues with Hawkes input. *J. Appl. Prob.* **55,** 920–943.

[25] MANDELBAUM, A. AND ZELTYN, S. (2004). The impact of customers' patience on delay and abandonment: Some empirically-driven experiments with the M/M/n+G queue. *OR Spektrum* **26,** 377–411.

[26] MANDELBAUM, A. AND ZELTYN, S. (2007). Service engineering in action: The Palm/Erlang-A queue, with applications to call centers. In *Advances in Services Innovations*. Springer, Berlin, pp. 17–45.

[27] MANDELBAUM, A., MASSEY, W. A. AND REIMAN, M. I. (1998). Strong approximations for Markovian service networks. *Queueing Systems Theory Appl.* **30,** 149–201.

[28] MANDELBAUM, A. *et al.* (2002). Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommun. Systems* **21,** 149–171.

[29] MASSEY, W. A. (2002). The analysis of queues with time-varying rates for telecommunication models. *Telecommun. Systems* **21,** 173–204.

[30] MASSEY, W. A. AND PENDER, J. (2013). Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems* **75,** 243–277.

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press

[31] MASSEY, W. A. AND PENDER, J. (2018). Dynamic rate Erlang-A queues. *Queueing Systems* **89,** 127–164.
[32] MATIS, T. I. AND FELDMAN, R. M. (2001). Transient analysis of state-dependent queueing networks via cumulant functions. *J. Appl. Prob.* **38,** 841–859.
[33] NIYIRORA, J. AND PENDER, J. (2016). Optimal staffing in nonstationary service centers with constraints. *Naval Res. Logistics* **63,** 615–630.
[34] PALM, C. (1953). Methods of judging the annoyance caused by congestion. *Tele* **4,** 189–208.
[35] PENDER, J. (2014). Gram Charlier expansion for time varying multiserver queues with abandonment. *SIAM J. Appl. Math.* **74,** 1238–1265.
[36] PENDER, J. AND KO, Y. M. (2017). Approximations for the queue length distributions of time-varying many-server queues. *INFORMS J. Computing* **29,** 688–704.
[37] PENDER, J. AND PHUNG-DUC, T. (2016). A law of large numbers for M/M/c/Delayoff-setup queues with nonstationary arrivals. In *Lecture Notes in Computer Science*, Springer, pp. 253–268.
[38] REED, J. E. AND WARD, A. R. (2008). Approximating the GI/GI/1+GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Math. Operat. Res.* **33,** 606–644.
[39] SHIMKIN, N. AND MANDELBAUM, A. (2004). Rational abandonment from tele-queues: nonlinear waiting costs with heterogeneous preferences. *Queueing Systems* **47,** 117–146.
[40] TALREJA, R. AND WHITT, W. (2009). Heavy-traffic limits for waiting times in many-server queues with abandonment. *Ann. Appl. Prob.* **19,** 2137–2175.
[41] VAN LEEUWAARDEN, J. S. H. AND KNESSL, C. (2012). Spectral gap of the Erlang A model in the Halfin-Whitt regime. *Stoch. Systems* **2,** 149–207.
[42] WARD, A. R. AND GLYNN, P. W. (2003). A diffusion approximation for a Markovian queue with reneging. *Queueing Systems* **43,** 103–128.
[43] WARD, A. R. AND GLYNN, P. W. (2005). A diffusion approximation for a GI/GI/1 queue with balking or reneging. *Queueing Systems* **50,** 371–400.
[44] WHITT, W. (2006). Sensitivity of performance in the Erlang-A queueing model to changes in the model parameters. *Operat. Res.* **54,** 247–260.
[45] YOM-TOV, G. B. AND MANDELBAUM, A. (2014). Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing Service Operat. Manag.* **16,** 283–299.
[46] ZELTYN, S. AND MANDELBAUM, A. (2005). Call centers with impatient customers: Many-server asymptotics of the M/M/n+G queue. *Queueing Systems* **51,** 361–402.
[47] ZHANG, B., VAN LEEUWAARDEN, J. S. H. AND ZWART, B. (2012). Staffing call centers with impatient customers: Refinements to many-server asymptotics. *Operat. Res.* **60,** 461–474.
[48] ZOHAR, E., MANDELBAUM, A. AND SHIMKIN, N. (2002). Adaptive behavior of impatient customers in tele-queues: theory and empirical support. *Manag. Sci.* **48,** 566–583.
[49] PENDER J. (2016). Sampling the functional kolmogorov forward equations for nonstationary queueing networks. *INFORMS J. Computing* **29**.

https://doi.org/10.1017/apr.2019.11 Published online by Cambridge University Press