Applied Probability Trust (22 December 2017)

NEW PERSPECTIVES ON THE ERLANG-A QUEUE

ANDREW DAW,* *Cornell University*

JAMOL PENDER,** *Cornell University*

Abstract

The non-stationary Erlang-A queue is a fundamental queueing model that is used to describe the dynamic behavior of large scale multi-server 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 non-stationary Erlang-B and Erlang-C queueing models under certain stability conditions.

Keywords: Multi-Server Queues; Abandonment; Dynamical Systems; Asymptotics; Time-Varying Rates; Moments, Fluid Limits; Erlang-A Queue; Functional Forward Equations; 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

* Postal address: School of Operations Research and Information Engineering, Cornell University, Ithaca, NY

* Email address: amd399@cornell.edu

** Email address: jjp274@cornell.edu

Markovian and multi-server 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, healthcare, urban mobility and transportation, and more recently cloud computing. See for example the following work by [12, 13, 26, 24]. Despite its importance in many different applications, the Erlang-A queueing model has remained to be very difficult to analyze and understand. Even the analysis of the moments of Erlang-A queue beyond the fourth moment has remained an important topic for additional study.

It is well known that the stationary setting of the Erlang-A is much easier to analyze than its non-stationary counterpart. Some common approaches used to analyze non-stationary and state dependent queueing models including asymptotic methods such as heavy traffic limit theory and strong approximations theory, see for example [5, 11]. 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.

However, there are 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 and are nearly infinite. Thus, the quality of the approximations depends significantly on the size of the model parameters and these asymptotic methods have been shown to be quite inaccruate for moderate sized model parameter settings, see for example [14, 15]. 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 are based on Brownian motion. Since Brownian motion has symmetry, its cumulants are all zero beyond the second order. Thus, 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 shown recently by [19, 2] that the Erlang-A and its variants have non-trivial amounts of skewness and excess kurtosis, which implies that the Erlang-A are not nearly 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 [16].

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. It is often the case that the set of moment equations for a large number of queueing models are not closed, see for example [17, 20]. Thus, the closure approximation helps approximate the moments with a closed system using the surrogate distribution. One such method used by [21, 22] is to use Hermite polynomials for approximating the distribution of the queue length process. In fact, they show 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 and the fact the queueing process is non-negative. However, they show that Hermite polynomials are natural to analyze since they are orthogonal with respect to the Gaussian distribution and the heavy traffic limits of multi-server queues are Gaussian.

In this paper, we perform an in-depth analysis of the moments and the moment generating function of the non-stationary Erlang-A queue. As the Erlang-B and Erlang-C queueing models are special cases of the Erlang-A model, we are able to obtain similar results for those models. Our approach is to use convexity and exploit Jensen's and the FKG inequality to obtain bounds on the moments and moment generating function of the Erlang-A queue. What we find even more exciting is that we are able to provide a stochastic representation of our approximations and bounds as a Poisson random variables with a constant shift. This shifted Poisson was observed in peer to peer networks by [3], however, we will show in the sequel, this novel representation will allow 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 exploting

FKG and Jensen's inequalities.

- We derive a novel stochastic intepretation 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. Section 2 introduces 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 use these to bound the true values. In Section 4 we derive approximations for the moment generating function and cumulant moment generating function of the Erlang-A queue. We again bound the true values by these approximations, and we also find a representation for our approximations in terms of Poisson random variables or $M/M/\infty$ queues, depending on the context.

2. The Erlang-A Queueing Model

The Erlang-A queueing model is a fundamental queueing model in the stochastic processes literature. The work of [11], shows that the $M(t)/M/c+M$ queueing system process $Q \equiv \{Q(t)|t \geq 0\}$ is represented by the following stochastic, time changed integral equation:

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

where $\Pi_i \equiv \{\Pi_i(t)|t \geq 0\}$ for $i = 1, 2, 3$ are i.i.d. standard (rate 1) Poisson processes. 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 non-homogeneous Poisson arrival process with rate $\lambda(t)$ that counts the customer arrivals that occured in the time interval [0,t). A random time

change for the Poisson process Π_2 , gives us 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, a the random time change of Π_3 gives us a counting process for the number of customers that abandon 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 studied so extensively is because 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 infinity. This precludes any abandonments since the abandonment rate $\theta \cdot (Q(t) - c)^+$ is always equal to zero 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. $\mu \cdot (Q(t) \wedge c) + \theta \cdot (Q(t) - c)^+ = \mu \cdot Q(t) = \theta \cdot Q(t)$. Thus, the Erlang-A queueing model becomes an infinite server queue.

One of the main and important insights of [5] is that for multi-server 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 and been an important modeling technique for modeling call centers in the queueing literature. 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 $\eta \cdot \lambda(t)$ and the new number of servers $\eta \cdot c$ are both 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 as

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

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 [11], i.e.

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

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

$$\dot{q}(t) = \lambda(t) - \mu\cdot(q(t)\wedge c) - \theta\cdot(q(t)-c)^{+}. \tag{2}$$

Moreover, if one takes a diffusion limit i.e.

$$\lim_{\eta\to\infty}\sqrt{\eta}\left(\frac{1}{\eta}Q^{\eta}(t) - q(t)\right) \Rightarrow \tilde{Q}(t) \tag{3}$$

one gets a diffusion process where the variance of the diffusion is given by the following ODE

$$\begin{aligned}\dot{\text{Var}}\left[\tilde{Q}(t)\right] &= \lambda(t) + \mu\cdot(q(t)\wedge c) + \theta\cdot(q(t)-c)^{+} \\ &\quad -2\cdot\text{Var}\left[\tilde{Q}(t)\right]\cdot\left(\mu\cdot\{q(t)<c\} + \theta\cdot\{q(t)\geq c\}\right).\end{aligned} \tag{4}$$

2.1. Mean Field Approximation is Identical to the Fluid Limit

In addtion to using strong approximations to analyze the queue length process one can also use the functional Kolmogorov forward equations as outlined in [15]. The functional forward equations for the Erlang-A model are derived as,

$$\dot{\text{E}}[f(Q(t))] \equiv \frac{d}{dt}\text{E}[f(Q(t))|Q(0)=q(0)] \tag{5}$$

$$= \lambda\cdot\text{E}\left[f(Q(t)+1) - f(Q(t))\right] + \text{E}\left[\delta\left(Q(t),c\right)\cdot\left(f(Q(t)-1) - f(Q(t))\right)\right], \tag{6}$$

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

$$\dot{\text{E}}[Q(t)] = \lambda(t) - \mu\cdot\text{E}[(Q(t)\wedge c)] - \theta\cdot\text{E}[(Q(t)-c)^{+}]. \tag{7}$$

The first thing to note is that this equation is not autonomous and one needs to know the distribution of $Q(t)$ a priori in order to compute the expectations on the righthand side of Equation 7. 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 $E[f(X)]$ becomes $f(E[X])$. This method is known as a mean field approximation in physics and is also known as the deterministic mean approximation of [15]. By applying the mean field approximation to Equation 7, we can show that the resulting differential equation is given by the following autonomous ODE

$$\dot{\mathrm{E}}[Q_f(t)] \quad = \quad \lambda(t) - \mu \cdot (\mathrm{E}[Q_f] \wedge c) - \theta \cdot (\mathrm{E}[Q_f] - c)^+. \tag{8}$$

By careful inspection, one can observe that the ode given by the mean field approximation is identical to the fluid limit of Equation2. Moreover, if one simulates the queueing process and compare it to the mean field limit, one notices an ordering property. For example on the left of Figure 1, we simulate the Erlang-A queue and compare to the fluid model. We observe that when $\theta < \mu$, that the simulated mean is larger than the fluid mean. This is precisely what our results predict. Moreover, on the right of Figure 1, we simulate the Erlang-A queue and compare to the fluid model when $\theta > \mu$ and observe that simulated queue length is smaller than the fluid limit.

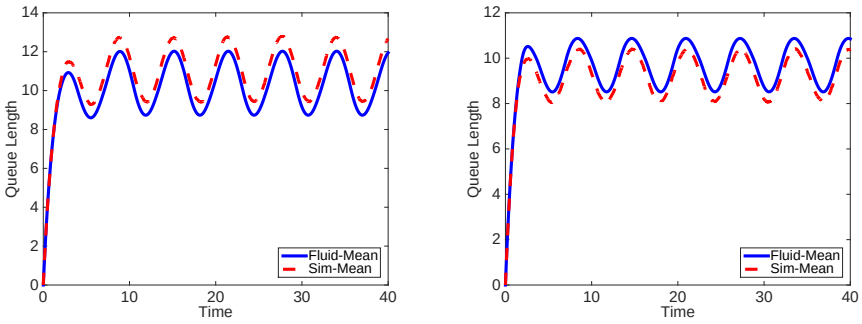


FIGURE 1: $\lambda(t) = 10 + 2 \cdot \sin(t)$, $\mu = 1$, $Q(0) = 0$, $c = 10$.
$\theta = 0.5$ (Left) and $\theta = 2$ (Right).

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 2.1. Throughout the remainder of this work, we use $Q(t)$ to represent the true queueing process and $Q_f(t)$ to represent the fluid approximation of it. This fluid approximation is a stochastic process that will be fully described in this work. In fact, in Section 4 we use 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 when the true moments of the Erlang-A queue are either dominated or dominates their corresponding fluid limit. We find that the relationship between the service rate and the abandonment rate determines whether or not the moment is dominated by the fluid limit. 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 m^{th} moment of the queueing system. Finally, in Subsection 3.3 we provide figures from numerical experiments that demonstrate these findings.

3.1. Inequalities for the Mean

We begin with analysis of the mean of the Erlang-A queue. Before we proceed, we first establish a lemma for comparisons of ordinary differential equations that will be fundamental to our approach to the results.

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

$$\dot{x}(t) = f(t, x(t)),\ x(0) = x_0 \tag{9}$$

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

$$\dot{y}(t) \leq f(t, y(t)) \quad \text{for } t \in [0, T] \text{ and } y(0) \leq x_0 \tag{10}$$

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

Proof. The the proof of this result is given in [4]. □

With this lemma in hand, we can now derive relationships for the fluid limit and the true mean. As seen in the proof, these results follow from the application of this differential equation comparison lemma 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 that the true mean satisfies the following differential equation

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

and the fluid limit satisfies the following differential equation

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

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

$$\begin{aligned} \dot{\mathrm{E}}[Q(t)] &= \lambda(t) - \theta \cdot E[Q] + (\theta - \mu) \cdot \mathrm{E}[(Q \wedge c)] \\ \dot{\mathrm{E}}[Q_f(t)] &= \lambda(t) - \theta \cdot \mathrm{E}[Q_f] + (\theta - \mu) \cdot (\mathrm{E}[Q_f] \wedge c). \end{aligned}$$

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

$$\begin{aligned}\dot{\mathrm{E}}[Q(t)] - \dot{\mathrm{E}}[Q_f(t)] &= \lambda(t) - \theta \cdot E[Q] + (\theta - \mu) \cdot \mathrm{E}[(Q \wedge c)] \\ &- \lambda(t) + \theta \cdot \mathrm{E}[Q_f] - (\theta - \mu) \cdot (\mathrm{E}[Q_f] \wedge c) \\ &= \theta \cdot (\mathrm{E}[Q_f] - E[Q]) + (\theta - \mu) \cdot (\mathrm{E}[(Q \wedge c)] - (\mathrm{E}[Q_f] \wedge c))\end{aligned}$$

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

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

for any random variable Q. Thus, we have that for $\theta < \mu$

$$\dot{\mathrm{E}}[Q(t)] - \dot{\mathrm{E}}[Q_f(t)] \geq 0,$$

and for $\theta > \mu$

$$\dot{\mathrm{E}}[Q(t)] - \dot{\mathrm{E}}[Q_f(t)] \leq 0.$$

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

$$\dot{\mathrm{E}}[Q(t)] - \dot{\mathrm{E}}[Q_f(t)] = 0$$

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

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* $Q(0) = Q_f(0)$*, then* $\mathrm{E}[Q(t)] \leq \mathrm{E}[Q_f(t)]$ *for all* $t \geq 0$.

Proof. This is obvious after noticing 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* $Q(0) = Q_f(0)$*, then* $\mathrm{E}[Q(t)] \geq \mathrm{E}[Q_f(t)]$ *for all* $t \geq 0$.

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

Remark 3.1. Given that we use Jensen's inequality and the FKG inequality later on in the paper, we find it 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, FKG implies that $\mathbb{E}[Q^n] \geq \mathbb{E}[Q^{n-1}] \cdot \mathbb{E}[Q]$. We find it interesting that by iterating the FKG inequality $n-2$ more times, it yields Jensen's inequality for the moments of random variables.

3.2. Inequalities for the m^{th} Moment

In this subsection we will now 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 and $m \in \mathbb{Z}^+$, if $Q(0) = Q_f(0)$, then $\mathrm{E}\left[Q^m(t)\right] \geq \mathrm{E}\left[Q_f^m(t)\right]$ when $\theta < \mu$, $\mathrm{E}\left[Q^m(t)\right] \leq \mathrm{E}\left[Q_f^m(t)\right]$ when $\theta > \mu$, and $\mathrm{E}\left[Q^m(t)\right] = \mathrm{E}\left[Q_f^m(t)\right]$ 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 $j \in \{1, 2, \ldots, m-1\}$. Recall that the m^{th} moment satisfies

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

and the approximate autonomous version satisfies

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

Now by taking the difference, we have that

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

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

$$\dot{\mathrm{E}}\left[Q^m(t)\right] - \dot{\mathrm{E}}\left[Q_f^m(t)\right] \geq 0,$$

if $\theta<\mu$,

$$\dot{\mathrm{E}}\left[Q^m(t)\right] - \dot{\mathrm{E}}\left[Q_f^m(t)\right] \leq 0,$$

and if $\theta=\mu$,

$$\dot{\mathrm{E}}\left[Q^m(t)\right] = \dot{\mathrm{E}}\left[Q_f^m(t)\right] = 0$$

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. □

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* $Q(0) = Q_f(0)$*, then* $\mathrm{E}[Q^m(t)] \leq \mathrm{E}\left[Q_f^m(t)\right]$ *for all* $t \geq 0$ *and* $m \in \mathbb{Z}^+$*.*

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

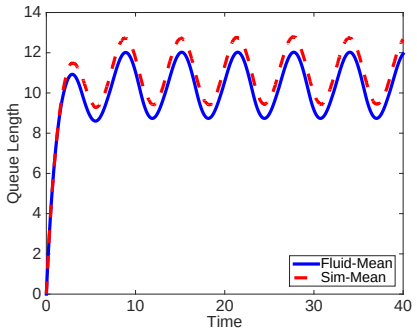
Corollary 4. *For the Erlang-C queueing model, if* $Q(0) = Q_f(0)$*, then* $\mathrm{E}[Q^m(t)] \geq \mathrm{E}\left[Q_f^m(t)\right]$ *for all* $t \geq 0$ *and* $m \in \mathbb{Z}^+$*.*

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

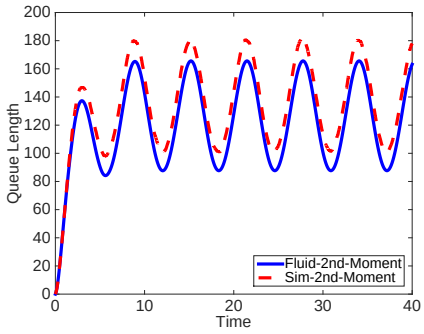
3.3. Numerical Results

In this section we describe numerical results for approximating the moments of the Erlang-A queue and examine them relative to our findings. In Figures 2 and 3, we show the first four moments of the Erlang-A queue and their respective fluid approximations for cases of $\theta < \mu$ and $\theta > \mu$, 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 exponential service rate $\mu = 1$. We test two different 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 that when $\theta > \mu$ the fluid values are greater than the simulations, and this matches the statements of Theorems 3 and 2.

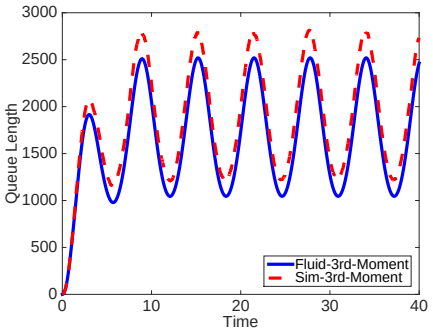
We observe the same relationships in Figures 4 and 5. For these plots we instead set $\lambda(t) = 100 + 20\sin(t)$ and $c = 100$ and otherwise use the same values as for Figures 2 and 3. With this increase in the arrival intensity and the number of servers, we see that the gaps between the fluid approximations and the simulations are again present, albeit proportionally smaller.



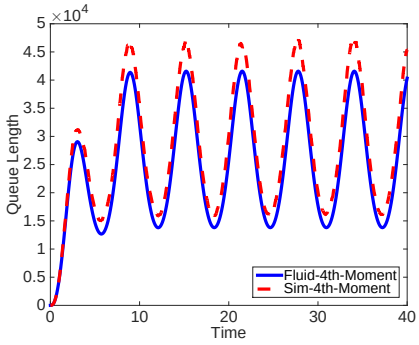
(a) First Moment



(b) Second Moment

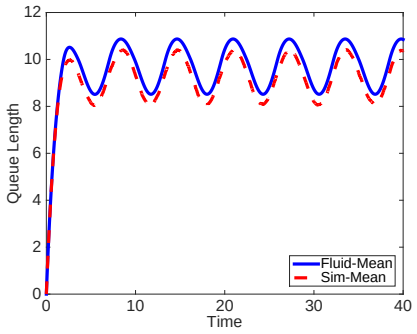


(c) Third Moment

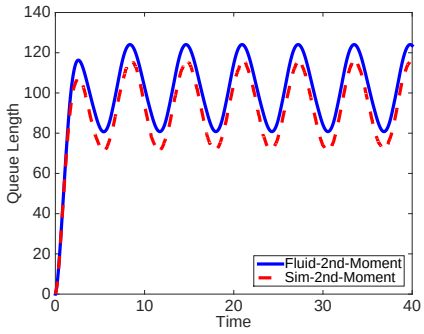


(d) Fourth Moment

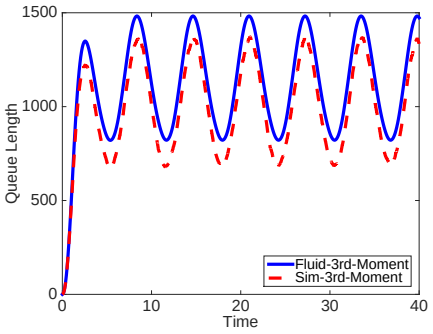
FIGURE 2: $\lambda(t) = 10 + 2 \cdot \sin(t)$, $\mu = 1$, $\theta = 0.5$, $Q(0) = 0$, $c = 10$.



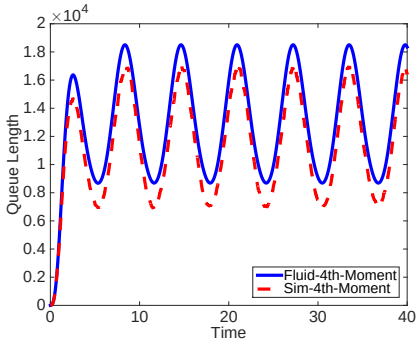
(a) First Moment



(b) Second Moment

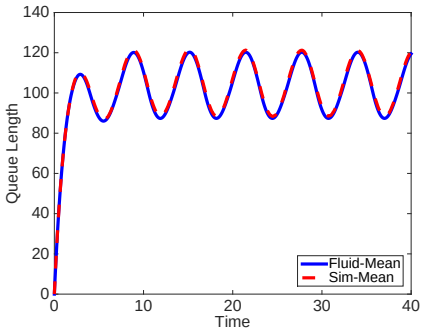


(c) Third Moment

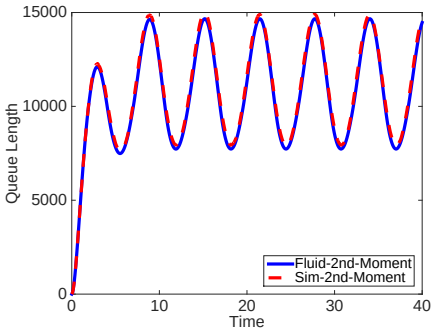


(d) Fourth Moment

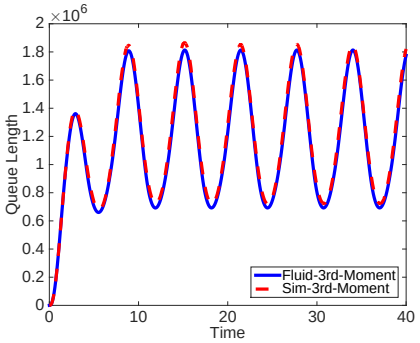
FIGURE 3: $\lambda(t) = 10 + 2 \cdot \sin(t)$, $\mu = 1$, $\theta = 2$, $Q(0) = 0$, $c = 10$.



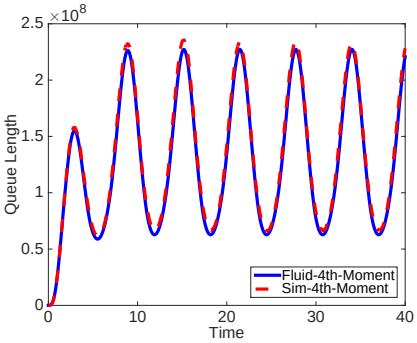
(a) First Moment



(b) Second Moment

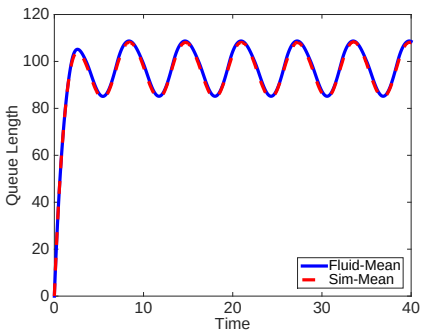


(c) Third Moment

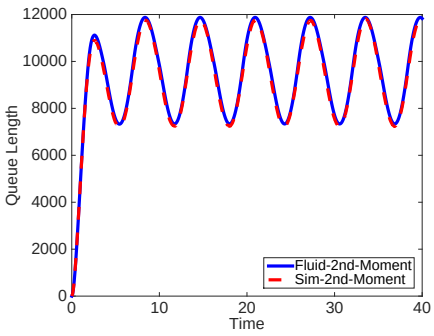


(d) Fourth Moment

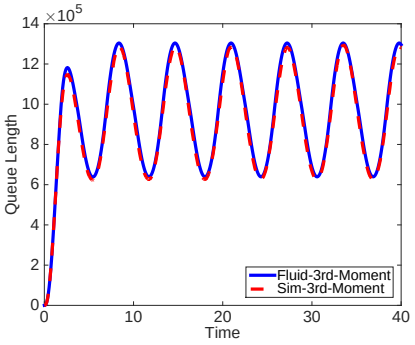
FIGURE 4: $\lambda(t) = 100 + 20 \cdot \sin(t)$, $\mu = 1$, $\theta = 0.5$, $Q(0) = 0$, $c = 100$.



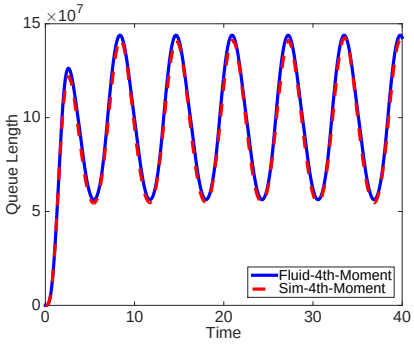
(a) First Moment



(b) Second Moment



(c) Third Moment



(d) Fourth Moment

FIGURE 5: $\lambda(t) = 100 + 20 \cdot \sin(t)$, $\mu = 1$, $\theta = 2$, $Q(0) = 0$, $c = 100$.

4. Inequalities and Characterizations for Generating Functions of the Erlang-A Queue

Building on what we have found 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 differential equations for the fluid approximations. We provide these inequalities in Subsections 4.1 and 4.2, respectively. In doing so, we find forms for the fluid approximations that we can interpret in terms of expectations of other random quantities. Through these recognitions, we characterize the fluid approximations. We describe these 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 empirical experiments in Subsection 4.5.

4.1. An Inequality for the Moment Generating Function of the Erlang-A Queue

Using the functional forward equations [15], we can show that the moment generating function for the Erlang-A queue satisfies the following partial differential equation

$$\begin{aligned}
\dot{\mathrm{E}}\left[e^{\alpha\cdot Q(t)}\right] &= \lambda(t)\cdot(e^{\alpha}-1)\cdot \mathrm{E}\left[e^{\alpha\cdot Q(t)}\right] + \theta\cdot(e^{-\alpha}-1)\cdot \mathrm{E}\left[Q(t)\cdot e^{\alpha\cdot Q(t)}\right] && (11)\\
&\quad -(\theta-\mu)\cdot(e^{-\alpha}-1)\cdot \mathrm{E}\left[(Q(t)\wedge c)\cdot e^{\alpha\cdot Q(t)}\right] && (12)\\
&= \lambda(t)\cdot(e^{\alpha}-1)\cdot \mathrm{E}\left[e^{\alpha\cdot Q(t)}\right] + \theta\cdot(e^{-\alpha}-1)\cdot \frac{\partial M(t,\alpha)}{\partial \alpha} && (13)\\
&\quad -(\theta-\mu)\cdot(e^{-\alpha}-1)\cdot \mathrm{E}\left[(Q(t)\wedge c)\cdot e^{\alpha\cdot Q(t)}\right]. && (14)
\end{aligned}$$

Just like the non-autonomous differential equation for the mean in Equation 7, 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 Erlang-B queue. Thus, it is useful to obtain approximations that are explicit upper 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 following partial differential equation

$$\begin{aligned}
\dot{\mathrm{E}}\left[e^{\alpha\cdot Q_f(t)}\right] &= \lambda(t)\cdot(e^{\alpha}-1)\cdot \mathrm{E}\left[e^{\alpha\cdot Q_f(t)}\right] + \theta\cdot(e^{-\alpha}-1)\cdot \frac{\partial M_f(t,\alpha)}{\partial \alpha}\\
&\quad -(\theta-\mu)\cdot(e^{-\alpha}-1)\cdot \left(\mathrm{E}\left[Q_f(t)\cdot e^{\alpha\cdot Q_f(t)}\right]\wedge \mathrm{E}\left[c\cdot e^{\alpha\cdot Q_f(t)}\right]\right) && (15)\\
\frac{\partial M_f(t,\alpha)}{\partial t} &= \lambda(t)\cdot(e^{\alpha}-1)\cdot M_f(t,\alpha) + \theta\cdot(e^{-\alpha}-1)\cdot \frac{\partial M_f(t,\alpha)}{\partial \alpha}\\
&\quad -(\theta-\mu)\cdot(e^{-\alpha}-1)\cdot \left(\frac{\partial M_f(t,\alpha)}{\partial \alpha}\wedge c\cdot M_f(t,\alpha)\right). && (16)
\end{aligned}$$

The following theorem determines exactly when $\mathrm{E}\left[e^{\alpha\cdot Q_f(t)}\right]$ is a lower or upper bound for the exact moment generating function of the Erlang-A queue.

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

Proof. If we take the difference of the two partial differential equations, we obtain the following

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

Now by exploiting the positive scalability property and the concavity of the minimum function, we have by Jensen's inequality that

$$\begin{aligned}
\mathrm{E}\left[(Q(t) \wedge c) \cdot e^{\alpha \cdot Q(t)}\right] &= \mathrm{E}\left[\left(Q(t) \cdot e^{\alpha \cdot Q(t)} \wedge c \cdot e^{\alpha \cdot Q(t)}\right)\right] \\
&\leq \left(\mathrm{E}\left[Q_f(t) \cdot e^{\alpha \cdot Q_f(t)}\right] \wedge \mathrm{E}\left[c \cdot e^{\alpha \cdot Q_f(t)}\right]\right).
\end{aligned}$$

Thus, we have when $\theta < \mu$ that

$$\dot{\mathrm{E}}\left[e^{\alpha \cdot Q(t)}\right] - \dot{\mathrm{E}}\left[e^{\alpha \cdot Q_f(t)}\right] \geq 0, \tag{17}$$

when $\theta > \mu$

$$\dot{\mathrm{E}}\left[e^{\alpha \cdot Q(t)}\right] - \dot{\mathrm{E}}\left[e^{\alpha \cdot Q_f(t)}\right] \leq 0, \tag{18}$$

and finally when $\theta = \mu$,

$$\dot{\mathrm{E}}\left[e^{\alpha \cdot Q(t)}\right] - \dot{\mathrm{E}}\left[e^{\alpha \cdot Q_f(t)}\right] = 0 \tag{19}$$

since they solve the same partial differential equation. This completes our proof. □

As with the moments, we can observe these relationships occurring in numerical experiments. We provide figures demonstrating this in Subsection 4.5.

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 Equation 11, we have

$$\log\left(\mathrm{E}\left[\overset{\bullet}{e^{\alpha\cdot Q(t)}}\right]\right) \equiv \frac{\partial}{\partial t}\log\left(\mathrm{E}\left[e^{\alpha\cdot Q(t)}\right]\right) = \frac{\overset{\bullet}{\mathrm{E}}\left[e^{\alpha\cdot Q(t)}\right]}{\mathrm{E}\left[e^{\alpha\cdot Q(t)}\right]} \tag{20}$$

$$= \lambda(t)\cdot(e^{\alpha}-1)+\theta\cdot(e^{-\alpha}-1)\cdot\frac{\mathrm{E}\left[Q(t)\cdot e^{\alpha\cdot Q(t)}\right]}{\mathrm{E}\left[e^{\alpha\cdot Q(t)}\right]}$$

$$-(\theta-\mu)\cdot(e^{-\alpha}-1)\cdot\frac{\mathrm{E}\left[(Q(t)\wedge c)\cdot e^{\alpha\cdot Q(t)}\right]}{\mathrm{E}\left[e^{\alpha\cdot Q(t)}\right]} \tag{21}$$

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

$$-(\theta-\mu)\cdot(e^{-\alpha}-1)\cdot\frac{\mathrm{E}\left[(Q(t)\wedge c)\cdot e^{\alpha\cdot Q(t)}\right]}{\mathrm{E}\left[e^{\alpha\cdot Q(t)}\right]}. \tag{22}$$

Like for the MGF, we note that we cannot compute the cumulant moment generating function directly without knowing the distribution of the queue length. So, by again applying Jensen's inequality, we can describe the fluid approximation as follows.

$$\log\left(\mathrm{E}\left[\overset{\bullet}{e^{\alpha\cdot Q_f(t)}}\right]\right) = \lambda(t)\cdot(e^{\alpha}-1)+\theta\cdot(e^{-\alpha}-1)\cdot\frac{\partial G_f(t,\alpha)}{\partial\alpha}$$

$$-(\theta-\mu)\cdot(e^{-\alpha}-1)\cdot\left(\frac{\mathrm{E}\left[Q_f(t)\cdot e^{\alpha\cdot Q_f(t)}\right]\wedge\mathrm{E}\left[c\cdot e^{\alpha\cdot Q_f(t)}\right]}{\mathrm{E}\left[e^{\alpha\cdot Q(t)}\right]}\right) \tag{23}$$

$$\frac{\partial G_f(t,\alpha)}{\partial t} = \lambda(t)\cdot(e^{\alpha}-1)+\theta\cdot(e^{-\alpha}-1)\cdot\frac{\partial G_f(t,\alpha)}{\partial\alpha}$$

$$-(\theta-\mu)\cdot(e^{-\alpha}-1)\cdot\left(\frac{\partial G(t,\alpha)}{\partial\alpha}\wedge c\right). \tag{24}$$

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. *For the Erlang-A queue, if* $Q(0)=Q_f(0)$*, then* $\log\left(\mathrm{E}\left[e^{\alpha\cdot Q(t)}\right]\right)\geq\log\left(\mathrm{E}\left[e^{\alpha\cdot Q_f(t)}\right]\right)$ *when* $\theta<\mu$*,* $\log\left(\mathrm{E}\left[e^{\alpha\cdot Q(t)}\right]\right)\leq\log\left(\mathrm{E}\left[e^{\alpha\cdot Q_f(t)}\right]\right)$ *when* $\theta>\mu$*, and* $\log\left(\mathrm{E}\left[e^{\alpha\cdot Q(t)}\right]\right)=\log\left(\mathrm{E}\left[e^{\alpha\cdot Q_f(t)}\right]\right)$ *when* $\theta=\mu$*.*

Proof. The proof follows from the same argument that was given in Theorem 3 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 stationary fluid approximation differential equations in a casewise manner based on the relationship of λ 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 $E[Q_f(\infty)] < c$. Moreover, if $\lambda \geq c\mu$, then $E[Q_f(\infty)] \geq c$.*

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

$$\begin{aligned} 0 &= \lambda - \mu \cdot (E[Q_f(\infty)] \wedge c) - \theta \cdot (E[Q_f(\infty)] - c)^+ \\ &= \lambda - \mu \cdot c - \theta (E[Q_f(\infty)] - c)^+. \end{aligned}$$

Since we assumed that $E[Q_f(\infty)] \geq c$, then this yields the following inequality

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

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

$$\begin{aligned} \lambda &= \mu \cdot (E[Q_f(\infty)] \wedge c) + \theta \cdot (E[Q_f(\infty)] - c)^+ \\ &= \mu \cdot (E[Q_f(\infty)] \wedge c) \\ &= c\mu + \mu \cdot (E[Q_f(\infty)] - c) \\ &< c\mu, \end{aligned}$$

which yields another contradiction. □

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

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

$$\frac{\partial M_f(\infty, \alpha)}{\partial \alpha} = \frac{\lambda \cdot (e^{\alpha} - 1) + (\theta - \mu) \cdot (1 - e^{-\alpha}) \cdot c}{\theta \cdot (1 - e^{-\alpha})} \cdot M_f(\infty, \alpha) \tag{25}$$

which yields a solution of

$$M_f(\infty, \alpha) \quad = \quad e^{\frac{\alpha\cdot(\theta-\mu)\cdot c+\lambda\cdot(e^{\alpha}-1)}{\theta}} \tag{26}$$

for $\alpha \in \mathbb{R}$.

Proof. To find the partial differential equation, we use functional cumulant bound for any non-decreasing function $h(\cdot)$ (which can be seen as a form of the FKG inequality),

$$\frac{\mathrm{E}[h(X)\cdot e^{\alpha\cdot X}]}{\mathrm{E}[e^{\alpha\cdot X}]} \geq \mathrm{E}[h(X)]. \tag{27}$$

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

$$\frac{\mathrm{d}G_f(\alpha)}{\mathrm{d}\alpha} = \frac{\lambda(e^{\alpha}-1)+c(\theta-\mu)(1-e^{-\alpha})}{\theta(1-e^{-\alpha})} = \frac{\lambda e^{\alpha}}{\theta} + \frac{c(\theta-\mu)}{\theta} \tag{28}$$

where here we have used the identity $e^x = \frac{e^x-1}{1-e^{-x}}$, which can be observed by multiplying each side of the equation by $1-e^{-x}$. Because the MGF is equal to 1 when $\alpha = 0$, we also have that $G_f(0) = 0$. Using this initial condition and integrating left and right sides of Equation 28 with respect to α, we find that

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

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

We can now observe that the fluid approximation is equivalent in distribution to a Poisson random variable shifted by $\gamma \equiv \frac{c(\theta-\mu)}{\theta}$, as the moment generation 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 MGF. This gives rise to the following.

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

$$\mathrm{E}\left[(Q_f(\infty)-\gamma)^m\right] \leq \mathrm{E}\left[(Q(\infty))^m\right] \leq \mathrm{E}\left[(Q_f(\infty))^m\right]$$

and if $\theta < \mu$

$$\mathrm{E}\left[(Q_f(\infty))^m\right] \leq \mathrm{E}\left[(Q(\infty))^m\right] \leq \mathrm{E}\left[(Q_f(\infty)-\gamma)^m\right]$$

where $\gamma = \frac{c(\theta-\mu)}{\theta}$.

Proof. From Proposition 1, we have that the fluid approximation of the MGF in steady-state is

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

where $\Gamma \sim \text{Pois}\left(\frac{\lambda}{\theta}\right)$ and $\gamma = \frac{c(\theta-\mu)}{\theta}$. From the uniqueness of MGF's, we have that

$$\mathrm{E}\left[(Q_f(\infty))^m\right] = \mathrm{E}\left[(\Gamma + \gamma)^m\right]$$

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 $\frac{\lambda}{\theta}$. 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 that $\mathrm{E}\left[(Q(\infty))^m\right] \leq \mathrm{E}\left[(\Gamma + \gamma)^m\right]$. Additionally, by comparing the steady-state infinite server queue representation of Γ to $Q(\infty)$, we can further observe that $\mathrm{E}\left[(Q(\infty))^m\right] \geq \mathrm{E}\left[\Gamma^m\right]$, 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 that

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

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

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

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

Remark 4.1. Note that in Theorem 2, we require that $Q(0) = Q_f(0)$ 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 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.

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

$$\mathrm{E}[Q_f^n(\infty)] = \sum_{j=0}^{n} \binom{n}{j} \cdot \left(\frac{c(\theta-\mu)}{\theta}\right)^j \cdot \mathcal{P}_{n-j}\left(\frac{\lambda}{\theta}\right) \tag{29}$$

where $\mathcal{P}_m\left(\frac{\lambda}{\theta}\right)$ is the m^{th} Touchard polynomial with parameter $\frac{\lambda}{\theta}$.

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

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

□

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

$$\frac{\mathrm{d}G_f(\infty,\alpha)}{\mathrm{d}\alpha}\Big|_{\alpha=0} = \frac{\lambda}{\theta} + \frac{c(\theta-\mu)}{\theta} = \mathrm{E}[Q_f(\infty)] \tag{30}$$

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

$$\frac{\mathrm{d}^n G_f(\infty,\alpha)}{\mathrm{d}^n\alpha}\Big|_{\alpha=0} = \frac{\lambda}{\theta} = \mathrm{C}^{(n)}[Q_f(\infty)] \tag{31}$$

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

We now consider the second case, which is $\lambda < c\mu e^{-\alpha}$. 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 $\alpha \geq 0$,*

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

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

Proof. To begin, suppose that $\frac{\partial M_f(\infty,\alpha)}{\partial \alpha} < cM_f(\infty,\alpha)$. Using this information in conjunction with the steady-state form of the partial differential equation for the fluid MGF given in Equation 16, we have that

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

which simplifies to

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

Using our assumption, we see that

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

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

We now move to showing the opposite direction and instead assume that $\frac{\partial M_f(\infty,\alpha)}{\partial \alpha} \geq cM_f(\infty,\alpha)$. In this case, Equation 16 is given by

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

and this simplifies to

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

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

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

and this now yields

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

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 $\lambda < c\mu e^{-\alpha}$.

Proposition 2. *For $\alpha \geq 0$, if $\lambda < c\mu e^{-\alpha}$, then in steady-state we have that*

$$\frac{\partial M_f(\infty,\alpha)}{\partial \alpha} = \frac{\lambda \cdot e^{\alpha}}{\mu} \cdot M_f(\infty,\alpha) \tag{32}$$

which yields a solution of

$$M_f(\infty,\alpha) = e^{\frac{\lambda\cdot(e^{\alpha}-1)}{\mu}} \tag{33}$$

for $\alpha \in \mathbb{R}$.

Proof. By Lemma 3 and our assumption that $\lambda < c\mu e^{-\alpha}$, we know that $\frac{\partial M_f(\infty,\alpha)}{\partial \alpha} < cM_f(\infty,\alpha)$. Thus, by observing this in the steady-state MGF equation, we easily obtain the result in Equation 32. Moreover, the solution to Equation 32 can be easily seen by inserting our proposed solution in and noting that it satisfies our differential equation. Moreover, the solution is unique by the properties of linear ordinary differential equation theory. □

Remark 4.2. We now pause to note that the $\lambda \geq c\mu e^{-\alpha}$ case of Lemma 3 implies Proposition 1 (and its following consequences) with a weaker assumption. However, because the condition $\lambda \geq c\mu$ does not depend on the choice of α it is more general, and thus we leave those results as stated with that assumption instead of $\lambda \geq c\mu e^{-\alpha}$.

Here we observe that Equation 33 is equivalent to the moment generating function of a Poisson random variable with parameter $\frac{\lambda}{\mu}$. Now, by recalling again that the steady-state distribution of a $M/M/\infty$ queue is a Poisson distribution with parameter equal to the arrival rate over the service rate, we find the following inequalities.

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

$$\mathrm{E}\left[\Gamma_\theta^m\right] \leq \mathrm{E}\left[Q(\infty)^m\right] \leq \mathrm{E}\left[\Gamma_\mu^m\right], \tag{34}$$

and if $\mu > \theta$

$$\mathrm{E}\left[\Gamma_\mu^m\right] \leq \mathrm{E}\left[Q(\infty)^m\right] \leq \mathrm{E}\left[\Gamma_\theta^m\right] \tag{35}$$

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

Proof. In each case, the inequality involving $\Gamma_\mu \sim \mathrm{Pois}\left(\frac{\lambda}{mu}\right)$ 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. Here we are using Proposition 2 with $\alpha = 0$, and by continuity we know this holds for some ball around 0. This validates the use of the derivatives of the steady-state MGF with respect to α evaluated at $\alpha = 0$ in finding the moments for the fluid approximation. Thus, we are left to prove the inequalities for $\Gamma_\theta \sim \mathrm{Pois}\left(\frac{\lambda}{\theta}\right)$.

To do so, let's first note that the stationary distribution of a $M/M/\infty$ queue with service rate θ is equivalent to that of Γ_θ. Suppose now that $\theta > \mu$. Then, any state

of such a $M/M/\infty$ queue has a larger rate of departure than the same state in the Erlang-A system. Thus, we have that

$$\mathrm{E}\left[\Gamma_\theta^m\right] \leq \mathrm{E}\left[Q(\infty)^m\right] \leq \mathrm{E}\left[\Gamma_\mu^m\right]$$

for all $m \in \mathbb{Z}^+$. By symmetric arguments in the $\theta < \mu$ case, we complete the proof. □

As we did for the case when $\lambda \geq c\mu$, 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 that*

$$\frac{\mathrm{d}G_f(\infty,\alpha)}{\mathrm{d}\alpha}\Big|_{\alpha=0} = \frac{\lambda}{\mu} = \mathrm{E}[Q_f(\infty)] \tag{36}$$

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

$$\frac{\mathrm{d}^n G_f(\infty,\alpha)}{\mathrm{d}^n\alpha}\Big|_{\alpha=0} = \frac{\lambda}{\mu} = \mathrm{C}^{(n)}[Q_f(\infty)] \tag{37}$$

$$\frac{\mathrm{d}^n M_f(\infty,\alpha)}{\mathrm{d}^n\alpha}\Big|_{\alpha=0} = \mathcal{P}_n\left(\frac{\lambda}{\mu}\right) = \mathrm{E}\left[Q_f(\infty)^n\right] \tag{38}$$

where $\mathrm{C}^{(n)}[Q_f(\infty)]$ is defined as the n^{th} cumulant moment of $Q_f(\infty)$ and $\mathcal{P}_m\left(\frac{\lambda}{\mu}\right)$ is the m^{th} Touchard polynomial with parameter $\frac{\lambda}{\mu}$.

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 incorporate this, we now incorporate a point process that can be used to approximate any periodic mean arrival pattern, as discussed in [1]. Specifically, we define $\lambda(t)$ by a Fourier series: let λ_0 and $\{(a_k, b_k), k \in \mathbb{Z}^+\}$ be such that

$$\lambda(t) = \lambda_0 + \sum_{k=1}^{\infty} a_k \sin(kt) + b_k \cos(kt). \tag{39}$$

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 sufficiently large. 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* $\underline{\lambda} \equiv \inf_{t\geq 0} \lambda(t) > c\mu$ *and that* $\mathrm{E}\left[Q_f(0)\right] > c$. *Then,*

$$\mathrm{E}\left[Q_f(t)\right] > c$$

for all time $t \geq 0$.

Proof. We have seen that $\mathrm{E}\left[Q_f(t)\right]$ evolves according to

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

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

$$\dot{\mathrm{E}}\left[Q_f(\hat{t})\right] = \lambda(\hat{t}) - c\mu - \theta\epsilon \geq \underline{\lambda} - c\mu - \theta\epsilon > 0.$$

By the continuity of the fluid mean and the fact that $\mathrm{E}\left[Q_f(0)\right] = q(0) > c$, we see that $\mathrm{E}\left[Q_f(t)\right] > c$ 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 an infinite server queue with nonstationary Poisson arrival rate* $\lambda(t)$ *and exponential service rate* μ *and initial value* $Q_\infty(t) = q_0$. *Then,*

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

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

Proof. To start, we have that time derivative of the MGF is

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

where $\lambda(t)$ is as defined previously:

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

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

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

where $M(\alpha, t)$ 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 MGF, which is $G(\alpha, t) = \log(M(\alpha, t))$. This PDE is

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

with the initial condition that

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

Using the notation that $G_x = \frac{\partial G}{\partial x}$, we seek to solve the system

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

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

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

with the initial conditions

$$\begin{aligned} \alpha(r, 0) &= r, \\ t(r, 0) &= 0, \\ g(r, 0) &= r q_0. \end{aligned}$$

We can first see that the ODE's for α and t solve to

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

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

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

which gives a solution of

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

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

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

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

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

Theorem 6. *If* $\inf_{t\leq\infty} \lambda(t) \equiv \underline{\lambda} > c\mu$ *and* $q(0) > c$*, then in for all* $t \geq 0$ *we have that*

$$\frac{\partial G_f(t,\alpha)}{\partial t} = \lambda(t)\cdot(e^{\alpha} - 1) + \theta\cdot(e^{-\alpha} - 1)\cdot\frac{\partial G_f(t,\alpha)}{\partial \alpha} - c\cdot(\theta - \mu)\cdot(e^{-\alpha} - 1) \quad (40)$$

which gives a solution of

$$\begin{aligned} G_f(t,\alpha) &= (e^{\alpha} - 1)\left(\frac{\lambda_0}{\theta}(1 - e^{-\theta t}) + \sum_{k=1}^{\infty} \frac{(a_k\theta + b_k k)\sin(kt) + (b_k\theta - a_k k)(\cos(kt) - e^{-\theta t})}{\theta^2 + k^2}\right) \\ &\quad + \frac{c(\theta - \mu)}{\theta}\alpha + \log((e^{\alpha} - 1)e^{-\theta t} + 1)\left(q(0) - \frac{c(\theta - \mu)}{\theta}\right) \end{aligned} \quad (41)$$

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

Proof. From Equation 24, we have that the PDE for the fluid approximation's cumulant moment generating function is

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

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

$$\mathrm{E}\left[Q_f(t)e^{\alpha Q_f(t)}\right] \geq \mathrm{E}\left[Q_f(t)\right]\mathrm{E}\left[e^{\alpha Q_f(t)}\right] > c\mathrm{E}\left[e^{\alpha Q_f(t)}\right]$$

and so $\left(\frac{\partial G_f(t,\alpha)}{\partial \alpha} \wedge c\right) = c$. Thus, we have the PDE given in Equation 40 and so now we seek to find it's solution. We approach this via the method of characteristics. Because $G_f(0,\alpha) = \log(\mathrm{E}\left[e^{\alpha Q_f(0)}\right]) = \alpha q(0)$, we see that we seek to solve the following system

$$\begin{cases} \theta(1-e^{-\alpha})G_{(\alpha)} + G_{(t)} = \lambda(t)(e^{\alpha}-1) + c(\theta-\mu)(1-e^{-\alpha}) \\ G_f(0,\alpha) = \alpha q(0) \end{cases}$$

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

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

with initial conditions $\alpha(r,0) = r$, $t(r,0) = t$, and $g(r,0) = rq(0)$. Then, we can solve the first two ODE's to see that

$$\begin{aligned} \alpha(r,s) &= \log((e^r-1)e^{\theta s}+1) \\ t(r,s) &= s \end{aligned}$$

and so we can use these to solve the remaining equation. Substituting in, we have the ODE as

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

and this now solves to

$$\begin{aligned} g(r,s) = (e^r-1)&\left(\frac{\lambda_0}{\theta}(e^{\theta s}-1) + \sum_{k=1}^{\infty}\frac{(a_k\theta+b_k k)\sin(ks)e^{\theta s} + (b_k\theta - a_k k)(\cos(ks)e^{\theta s}-1)}{\theta^2+k^2}\right) \\ &+ \frac{c(\theta-\mu)}{\theta}\left(\log\left((e^r-1)e^{\theta s}+1\right) - r\right) + rq(0). \end{aligned}$$

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

$$\begin{aligned} G_f(t,\alpha) &= g(\log((e^{\alpha}-1)e^{-\theta t}+1), t) \\ &= (e^{\alpha}-1)e^{-\theta t}\left(\frac{\lambda_0}{\theta}(e^{\theta t}-1) + \sum_{k=1}^{\infty}\frac{(a_k\theta+b_k k)\sin(kt)e^{\theta t} + (b_k\theta - a_k k)(\cos(kt)e^{\theta t}-1)}{\theta^2+k^2}\right) \\ &\quad + \frac{c(\theta-\mu)}{\theta}\left(\alpha - \log((e^{\alpha}-1)e^{-\theta t}+1)\right) + \log((e^{\alpha}-1)e^{-\theta t}+1)q(0) \end{aligned}$$

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 a infinite server queue shifted by $\gamma \equiv \frac{c(\theta-\mu)}{\theta}$. This gives rise to the following.

Theorem 7. *For the Erlang-A queue with nonstationary arrival rate $\lambda(t)$ such that $\underline{\lambda} \equiv \inf_{t\geq 0} \lambda(t) > c\mu$ and initial value $q(0) > c$, the fluid approximation of the MGF is equivalent to that of a shifted $M/M/\infty$ queue with arrival rate $\lambda(t)$, service rate θ, initial value $q(0) - \frac{c(\theta-\mu)}{\theta}$, and linear shift $\frac{c(\theta-\mu)}{\theta}$.*

Proof. Observe from Theorem 6 that the fluid MGF for the Erlang-A under these conditions is

$$\begin{aligned} M_f(t,\alpha) &= e^{G_f(t,\alpha)} \\ &= e^{(e^\alpha-1)\left(\frac{\lambda_0}{\theta}(1-e^{-\theta t})+\sum_{k=1}^{\infty}\frac{(a_k\theta+b_kk)\sin(kt)+(b_k\theta-a_kk)(\cos(kt)-e^{-\theta t})}{\theta^2+k^2}\right)+\frac{c(\theta-\mu)}{\theta}\alpha}\left((e^\alpha-1)e^{-\theta t}+1\right)^{q(0)-\frac{c(\theta-\mu)}{\theta}} \end{aligned}$$

which is of a form that we can recognize. Comparing it to Lemma 5, we can see that Q_f is of the form of a shifted $M/M/\infty$ queue with arrival rate $\lambda(t)$, service rate θ, initial value $q(0) - \frac{c(\theta-\mu)}{\theta}$, and linear shift $\frac{c(\theta-\mu)}{\theta}$, thus enforcing that the fluid model does start at $q(0)$. □

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

Corollary 9. *Let $Q(t)$ represent the Erlang-A queue with nonstationary arrival rate $\lambda(t)$ 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 $\theta > \mu$*

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

and if $\theta < \mu$

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

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

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 describe various numerical experiments demonstrating these findings. We first have Figures 6, 7, 8, and 9, which compare simulated value of the moment generating function to their fluid approximations. In the first two figures, the arrival intensity is $\lambda(t) = 5 + \sin(t)$, the service rate is $\mu = 1$, and the number of servers is $c = 5$. The abandonment rates are the differing component of these plots, with $\theta = 0.5$ and $\theta = 2$ as the two respective values. These same comparisons are made in the latter two figures, however in this case the arrival rate is instead $\lambda(t) = 10 + 2\sin(t)$ and the number of servers is $c = 10$.

Through these plots one can observe that the true MGF dominates the fluid approximation when $\theta < \mu$ and that the fluid dominates the stochastic value when $\theta > \mu$. This is of course stated with the understanding that for small values of α or for times near 0 the values of the MGF and the approximation are quite close and so with numerical error the surfaces may overlap.

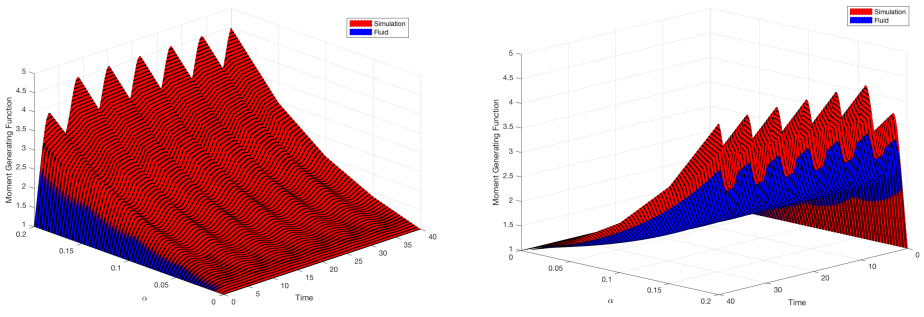


FIGURE 6: $\lambda(t) = 5 + \sin(t)$, $\mu = 1$, $\theta = 0.5$, $Q(0) = 0$, $c = 5$.
Front view (left) and rear view (right).

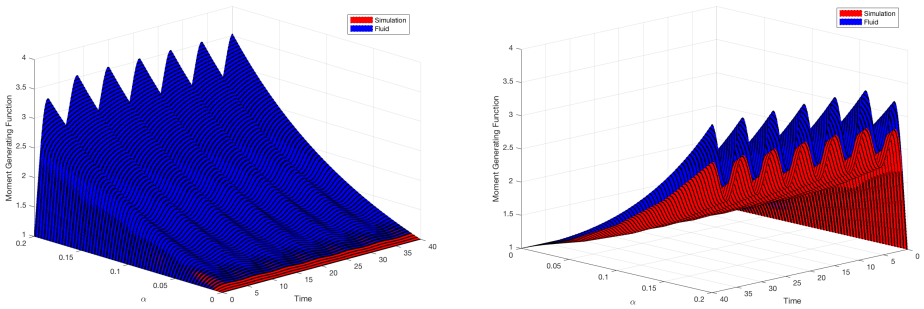


FIGURE 7: $\lambda(t) = 5 + \sin(t)$, $\mu = 1$, $\theta = 2$, $Q(0) = 0$, $c = 5$.
Front view (left) and rear view (right).

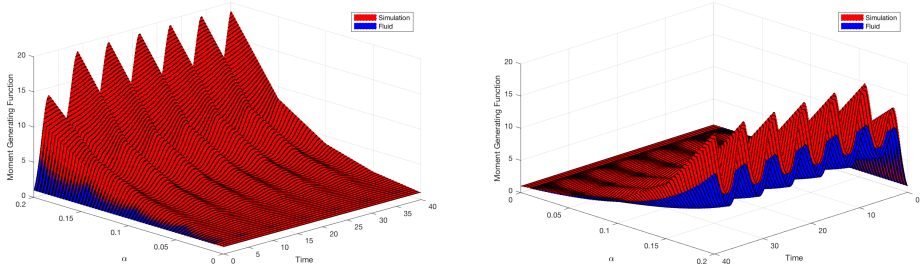


FIGURE 8: $\lambda(t) = 10 + 2 \cdot \sin(t)$, $\mu = 1$, $\theta = 0.5$, $Q(0) = 0$, $c = 10$.
Front view (left) and rear view (right).

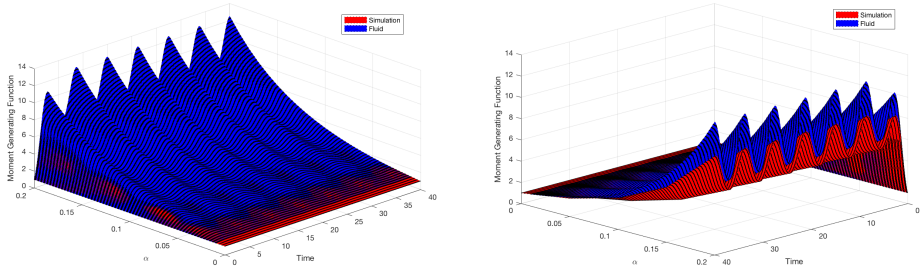
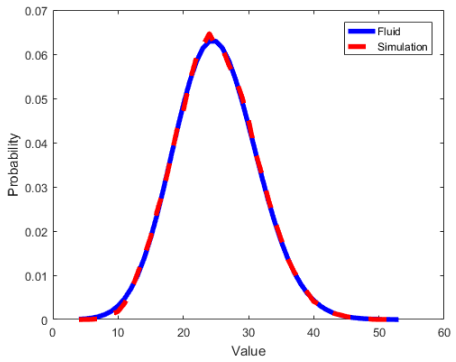
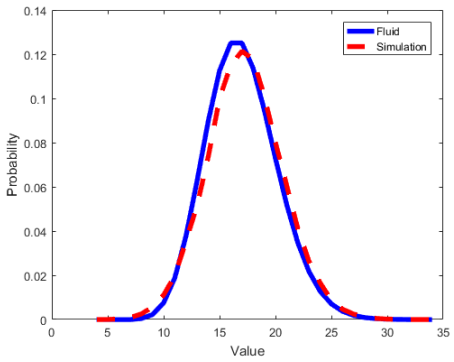


FIGURE 9: $\lambda(t) = 10 + 2 \cdot \sin(t)$, $\mu = 1$, $\theta = 2$, $Q(0) = 0$, $c = 10$.
Front view (left) and rear view (right).

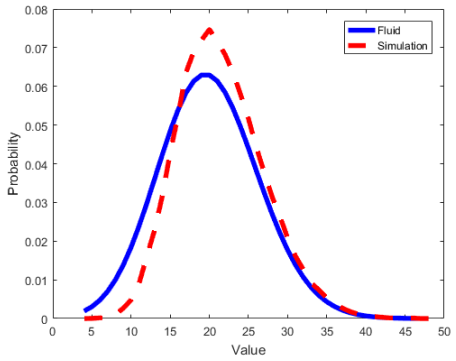
In Figure 10 we plot the limiting distribution for the steady-state Erlang-A. For these plots we take $\lambda = 20$ and $\mu = 1$, and then vary θ and c. For the three plots on the left we take the abandonment rate to be $\theta = 0.5$ and for those on the right we set $\theta = 2$. For the top two 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 [7, 6, 8, 22, 18, 25]. 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 $\frac{\lambda}{\theta}$, but when $\lambda < c\mu$ it is equivalent to a Poisson distribution with parameter $\frac{\lambda}{\mu}$.



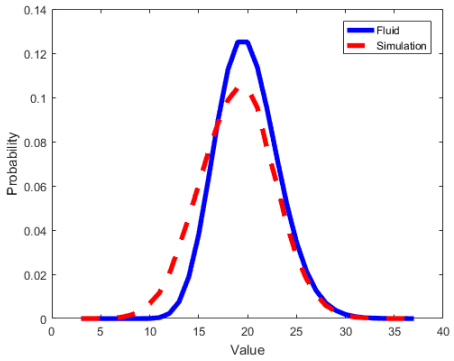
(a) $\theta = 0.5$, $c = 15$



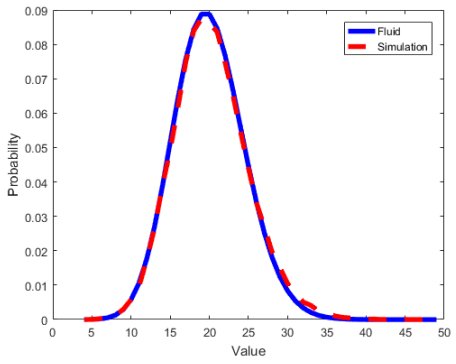
(b) $\theta = 2$, $c = 15$



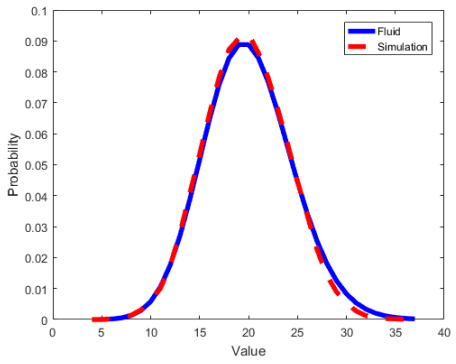
(c) $\theta = 0.5$, $c = 20$



(d) $\theta = 2$, $c = 20$



(e) $\theta = 0.5$, $c = 25$

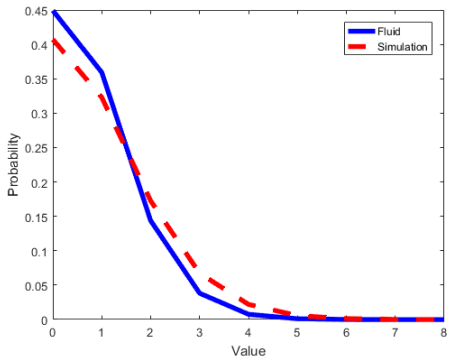


(f) $\theta = 2$, $c = 25$

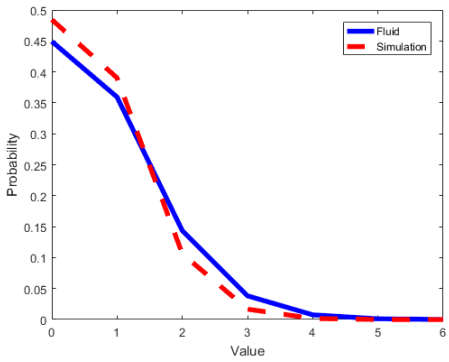
FIGURE 10: Empirical and Fluid Limiting Distributions for $\lambda = 20$ and $\mu = 1$.

In Figure 11 we examine the limiting distributions for the single server case. In

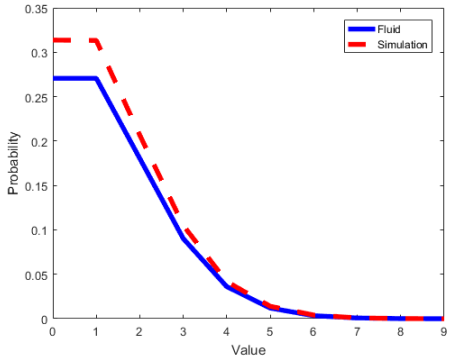
these plots we set $\mu = 1$ and then vary the arrival rate and the abandonment rate. On all plots on the left we set $\theta = 0.5$ and on the right $\theta = 2$. Further, in the top pair we make $\lambda = 0.8$, in the middle we let $\lambda = 1$, and in the bottom pair $\lambda = 1.2$. As in Figure 10, Figure 11 shows that our approximations are quite good. Thus, we are able to capture single server dynamics as well as large-scale multi-server dynamics even though they are quite different. This is even more useful as our approximations are non-asymptotic and don't rely on scaling the number of servers.



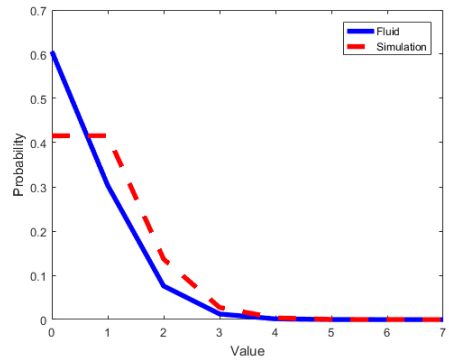
(a) $\theta = 0.5$, $\lambda = 0.8$



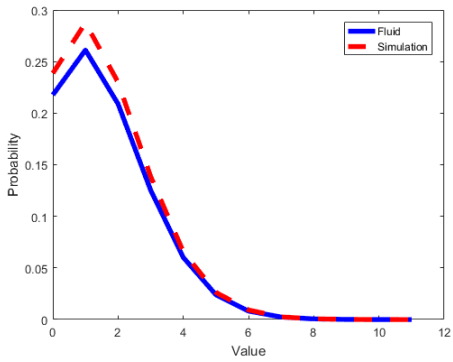
(b) $\theta = 2$, $\lambda = 0.8$



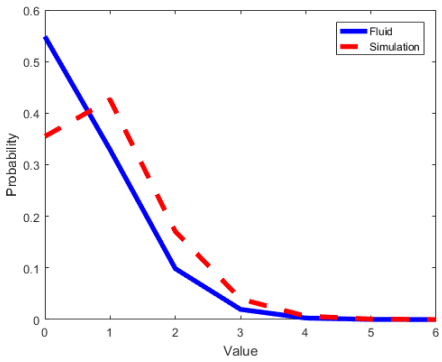
(c) $\theta = 0.5$, $\lambda = 1$



(d) $\theta = 2$, $\lambda = 1$



(e) $\theta = 0.5$, $\lambda = 1.2$



(f) $\theta = 2$, $\lambda = 1.2$

FIGURE 11: Empirical and Fluid Limiting Distributions for $c = 1$ and $\mu = 1$.

In Figures 12, 13, and 14, 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 $M/M/\infty$. In the first figure we consider the mean for $\theta = 1.1$ and 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 find the same for the latter two figures, in which we plot the moment generating function for $\theta = 1.1$ and $\theta = 0.9$, respectively.

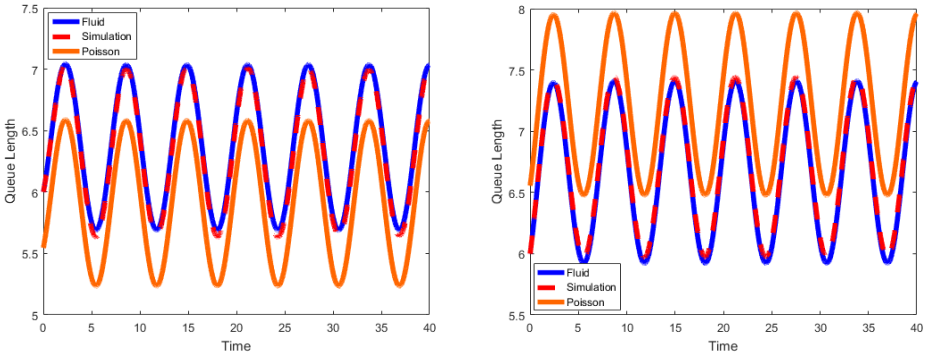


FIGURE 12: Queue Mean for $\lambda(t) = 6.5 + \sin(t)$, $\mu = 1$, $Q(0) = 6$, $c = 5$. $\theta = 1.1$ (left) and $\theta = 0.9$ (right).

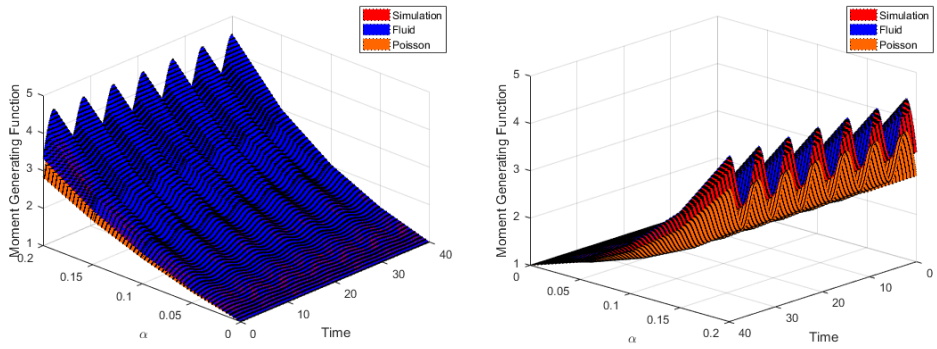


FIGURE 13: MGF for $\lambda(t) = 6.5 + \sin(t)$, $\mu = 1$, $\theta = 1.1$, $Q(0) = 6$, $c = 5$. Front view (left) and rear view (right).

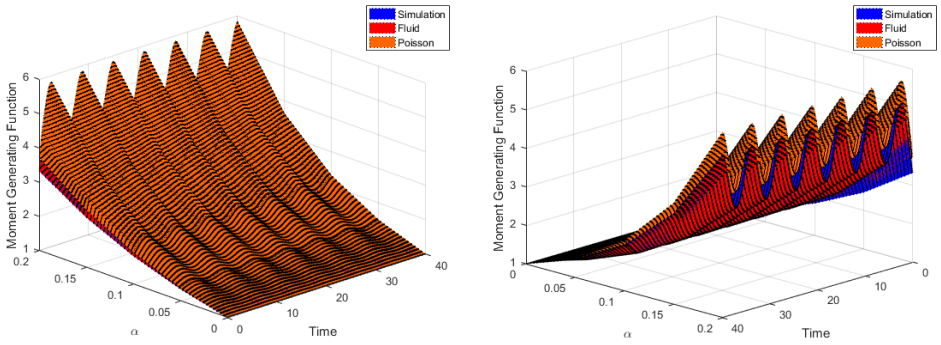


FIGURE 14: MGF for $\lambda(t) = 6.5 + \sin(t)$, $\mu = 1$, $\theta = 0.9$, $Q(0) = 6$, $c = 5$. Front view (left) and rear view (right).

5. Conclusion

In this paper we have investigated the Erlang-A queueing system through comparison to the fluid approximations of its moments and moment generating function as well as of its cumulants and cumulant moment generating function. Through recognizing the convexity in the differential equations describing these approximations, we have 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 larger the approximation dominates the true value, and when the rates of abandonment and service are equal, the two are equivalent.

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 the stationary and non-stationary cases, respectively. 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 of 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 non-trivial endeavor, which stems from the non-differentiablility and non-closure 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 [22, 9, 10].

Additionally, it would be even more 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 distributions that we have found as characterizations of our fluid 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 appears to be a $M/M/\infty$ 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. Finally, it would be interesting to extend this to networks of Erlang-A queues like in [23], however, we would have to keep track of the routing probabilities carefully to keep track of the convexity/concavity of the rate functions.

Acknowledgements

References

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

[2] Engblom, S. and Pender, J. (2014). Approximations for the moments of nonstationary and state dependent birth-death queues. *arXiv preprint*

arXiv:1406.6164.

[3] Ferragut, A. and Paganini, F. (2012). Content dynamics in P2P networks from queueing and fluid perspectives. In *Proceedings of the 24th International Teletraffic Congress*. International Teletraffic Congress. p. 11.

[4] Hale, J. K. and Lunel, S. M. V. (2013). *Introduction to functional differential equations* vol. 99. Springer Science & Business Media.

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

[6] Hampshire, R. C., Jennings, O. B. and Massey, W. A. (2009). A time-varying call center design via Lagrangian mechanics. *Probability in the Engineering and Informational Sciences* **23,** 231–259.

[7] Hampshire, R. C. and Massey, W. A. (2010). Dynamic optimization with applications to dynamic rate queues. In *Risk and Optimization in an Uncertain World*. INFORMS pp. 208–247.

[8] Hampshire, R. C., Massey, W. A. and Wang, Q. (2009). Dynamic pricing to control loss systems with quality of service targets. *Probability in the Engineering and Informational Sciences* **23,** 357–383.

[9] Ko, Y. M. and Pender, J. (2016). Strong approximations for time-varying infinite-server queues with non-renewal arrival and service processes. *Stochastic Models*.

[10] Ko, Y. M. and Pender, J. (2017). Diffusion limits for the $(MAP_t/Ph_t/\infty)$ N queueing network. *Operations Research Letters* **45,** 248–253.

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

[12] Mandelbaum, A., Massey, W. A., Reiman, M. I., Stolyar, A. and Rider, B. (2002). Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems* **21,** 149–171.

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

[14] Massey, W. A. and Pender, J. (2011). Poster: skewness variance approximation for dynamic rate multiserver queues with abandonment. *ACM SIGMETRICS Performance Evaluation Review* **39,** 74–74.

[15] Massey, W. A. and Pender, J. (2013). Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems* **75,** 243–277.

[16] Massey, W. A. and Pender, J. (2017). Performance and provisioning analysis for the dynamic rate Erlang-A queue.

[17] Matis, T. I. and Feldman, R. M. (2001). Transient analysis of state-dependent queueing networks via cumulant functions. *Journal of Applied Probability* **38,** 841–859.

[18] Niyirora, J. and Pender, J. (2016). Optimal staffing in nonstationary service centers with constraints. *Naval Research Logistics (NRL)* **63,** 615–630.

[19] Pender, J. (2014). Gram-Charlier expansion for time varying multiserver queues with abandonment. *SIAM Journal on Applied Mathematics* **74,** 1238–1265.

[20] Pender, J. Laguerre polynomial expansions for time varying multiserver queues with abandonment 2014.

[21] Pender, J. (2016). Sampling the functional Kolmogorov forward equations for nonstationary queueing networks. *INFORMS Journal on Computing* **29,** 1–17.

[22] Pender, J. and Ko, Y. M. (2017). Approximations for the queue length distributions of time-varying many-server queues. *INFORMS Journal on Computing* **29,** 688–704.

[23] Pender, J. and Massey, W. A. (2017). Approximating and stabilizing dynamic rate Jackson networks with abandonment. *Probability in the Engineering and Informational Sciences* **31,** 1–42.

[24] Pender, J. and Phung-Duc, T. (2016). A law of large numbers for $M/M/c/Delay$ off-setup queues with nonstationary arrivals. In *International Conference on Analytical and Stochastic Modeling Techniques and Applications*. Springer. pp. 253–268.

[25] Qin, Z. and Pender, J. (2017). Dynamic control for nonstationary queueing networks.

[26] Yom-Tov, G. B. and Mandelbaum, A. (2014). Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management* **16,** 283–299.