Performance and Provisioning Analysis for the Dynamic Rate Erlang-A Queue

William A. Massey
Department of Operations Research and Financial Engineering
Princeton University
wmassey@princeton.edu

Jamol Pender
School of Operations Research and Information Engineering
Cornell University
228 Rhodes Hall, Ithaca, NY 14853
jjp274@cornell.edu

August 1, 2017

This work is dedicated to Ward Whitt, on the occasion of his 75th birthday. We are eternally grateful for his friendship, mentorship, guidance, kindness, and infinite knowledge about stochastic processes.


Abstract

The multi-server queue with non-homogeneous Poisson arrivals and customer abandonment is a fundamental dynamic rate queueing model for large scale service systems such as call centers and hospitals. Scaling the arrival rates and number of servers arises naturally when a manager updates a staffing schedule in response to a forecast of increased customer demand. Mathematically, this type of scaling ultimately gives us
the fluid and diffusion limits as found in Mandelbaum et al. [25] for Markovian service networks. The asymptotics used here reduce to the Halfin and Whitt [9] scaling for multi-server queues.

In this paper, we provide a review and an in-depth analysis of the Erlang-A queueing model. We prove new results about cumulant moments of the Erlang-A queue, the transient behavior of the Erlang-A limit cycle, new fluid limits for the delay time of a virtual customer, and optimal static staffing policies for healthcare systems. We combine tools from queueing theory, ordinary differential equations, complex analysis, cumulant moments, orthogonal polynomials, dynamic optimization to obtain new and important insights about this fundamental queueing model.

Keywords: Multi-Server Queues, Abandonment, Dynamical Systems, Asymptotics, Time-Varying Rates, Time Inhomogeneous Markov Processes, Hermite Polynomials, Fluid and Diffusion Limits, Skewness, Cumulant Moments.

1 Introduction

In operations research and applied probability, Markov processes play an important role in aiding researchers in modeling man-made phenomena. In fact this year marks the centennial of the Erlang blocking formula Erlang [6], which was derived by Agner Krarup Erlang and published in 1917. In the context of the constant rate setting, the Erlang blocking formula paved the way for queueing theory as we know it. In the stationary setting where rates are constant, the most common method of analysis is to study the behavior of the transition probabilities since a full understanding of the transition probabilities allows one to understand all of dynamics of the underlying Markov process. However, in real life, rates are not stationary and are often dynamic in nature. Furthermore, when rates are nonstationary or even state dependent, an explicit study of the transition probabilities is often intractable since explicit solutions are rare with exception in some very special cases.

Some common approaches used to analyze nonstationary and state dependent queueing models include asymptotic methods such as heavy traffic limit theory and strong approximations theory, see for example Halfin and Whitt [9], Mandelbaum et al. [25], Pender and Phung-Duc [11]. Uniform acceleration is extremely useful for approximating the transition probabilities and moments such as the mean and variance of the Markov process. 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 and Gaussian like limits. However, one drawback of these methods is that they are asymptotic as a function of the rates. This means that that the convergence of the methods depends on how large the rates are of the problem of interest. The inaccuracy for queueing processes with moderate to small rates has been noted by Massey and Pender [28].

Another method that is quite common in the queueing literature is a moment closure approximation. Moment closure approximations are used to approximate the queueing process distribution with a lower dimensional distribution and computes approximate moments of the queueing process. They can also be used to approximate unknown functions of the queue length to yield a set of equations that only depend on the moments that are being estimated.
One such method used by Massey and Pender [27, 28], Pender [38], Pender et al. [42], Pender and Ko [39], Engblom and Pender [5] 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, which were pioneered by Ward Whitt, are Gaussian.

In this paper, we perform an in-depth analysis of the Erlang-A queue. We are motivated to study the Erlang-A queue since it has served as a fundamental queueing model for analyzing the dynamic behavior of customer abandonments. It is also a canonical model in telecommunication systems and is emerging as a very useful model in various applications such as healthcare and cloud computing. Exact results for the Erlang-A queue are often hard to derive or are intractable, thus, we are interested in finding low dimensional dynamical system approximations and representations of the Erlang-A queue. To this end, our results in this work are inspired by two papers authored by Ward Whitt. The first paper Halfin and Whitt [9] proves many-server heavy traffic limits for a multi-server queueing model and the second paper Duffield et al. [2] exploits the cumulant generating function and cumulant moments to derive expressions for the mean and variance of packet networks. By combining new closure approximations, orthogonal polynomials, properties of cumulants, differential equations, and complex analysis we are able to derive several new results for the dynamic rate Erlang-A queue that highlight its complexity.

1.1 Main Contributions of Paper

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

- We analyze the cumulant moments of the Erlang-A queueing process and show that the all of the functional cumulants of the Erlang-A departure process are equal to the arrival rate in steady state.

- We show that any distribution that has a finite number of non-zero cumulant moments must be Gaussian.

- We review the Hermite polynomial closure approximation hierarchy that was developed in Massey and Pender [28] and relate it to cumulant moments of the Erlang-A queue.

- We derive a new fluid limit approximation for the mean delay of the Erlang-A queue and provide new algorithms for computing the mean delay using the Gaussian skewness methods for the queue length process.

- We show that the Gaussian skewness method can be used to stabilize delay probabilities effectively.
Using the Erlang-A fluid limit and dynamic optimization, we derive the static optimal number of beds to optimally design a nursing home facility over a finite time period [0,T].

1.2 Organization of Paper

The remainder of this paper is organized as follows. Section 2 describes the dynamic rate Erlang-A queue and its fluid and diffusion limits. Section 3 introduces cumulants, the new notion of a functional cumulant, and derives several properties of functional cumulants. Section 4 introduces the functional forward equations for the Erlang-A queue and shows that all cumulant moments for the departure process are equal to the arrival rate in steady state. We also show the limit cycle of an infinite server queue is an ellipse. Lastly, in Section 4, we review various Gaussian closure approximations for the Erlang-A queue. In Section 5, we show how to use the closure approximations to stabilize the probability of delay in the Erlang-A model. We also derive a new fluid limit result for the mean delay of the Erlang-A queue. In Section 6, we use dynamic optimization to find the static optimal number of beds to have in a nursing home. We derive new and simple bounds on the opportunity cost and optimal number of beds. Finally, we conclude in Section 7 where we offer suggestions for future research.

2 Queueing Model and Halfin-Whitt Scaling

The Erlang-A queueing model is a fundamental queueing model in the stochastic processes literature. The work of Mandelbaum et al. [23], shows that the $M(t)/M/c(t) + M$ queueing system process $Q \equiv \{Q(t)|t \geq 0\}$ is represented by the following stochastic, time changed integral equation:

$$Q(t) = Q(0) + \Pi_+ \left( \int_0^t \lambda(s)ds \right) - \Pi_- \left( \int_0^t \delta(Q(s), c(s)) ds \right)$$

(2.1)

where $\Pi_+ \equiv \{\Pi_+(t)|t \geq 0\}$ and $\Pi_- \equiv \{\Pi_-(t)|t \geq 0\}$ are independent and identically distributed standard (rate 1) Poisson processes and

$$\delta(Q, c) \equiv \mu \cdot (Q \wedge c) + \beta \cdot (Q - c)^+.$$  

(2.2)

Thus, we can write the sample path dynamics of the Erlang-A queueing process in terms of two independent unit rate Poisson processes. A deterministic time change for $\Pi_+$ transforms it into a non-homogeneous 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 $\Pi_-$, gives us a departure process that counts the number of both the serviced and abandoning customers. We implicitly assume that the number of servers is a deterministic function of time $c(t)$ and that each server works at rate $\mu$. We also assume that the abandonment distribution is exponential and the rate of abandonments is equal to $\beta$. When the mean number of the system $E Q(t)$ is less than the number of servers $c(t)$ or $E Q(t) < c(t)$, we say that the system is underloaded. Conversely, when $E Q(t) > c(t)$, we say that the system is overloaded. Finally, when $E Q(t) = c(t)$, we say that the system is critically loaded.
When the $M(t)/M/c(t) + M$ queueing system is underloaded for all $t$, then it behaves more like a nonstationary infinite server queue or when the number of servers is equal to $\infty$. Even with a time varying arrival rate, when initialized by a Poisson distribution, the nonstationary infinite server queue always has a Poisson transient distribution. Detailed explorations of infinite server queueing dynamics with non-homogeneous input can be found in the works of Palm [33], Khinchin et al. [16], and Eick et al. [4, 3].

Under general conditions, the Poisson distribution is uniquely characterized by having all its cumulant moments equal to its mean Khinchin et al. [16].

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 $\beta \cdot (Q(t) - c(t)) + M$ 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 $\mu$ equal to the abandonment rate $\beta$. When $\mu = \beta$, this implies that the sum of the service and abandonment departure processes is equal to a linear function i.e.

$$\mu = \beta \implies \delta(Q, c) = \mu \cdot Q = \beta \cdot Q.$$  

Thus, the Erlang-A queueing model becomes an infinite server queue.

One of the main and important insights of Halfin and Whitt [9] 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 and service systems in the queueing literature. Since the $M(t)/M/c(t) + M$ 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 $\eta \cdot c(t)$ 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

$$Q^\eta(t) = Q^\eta(0) + \Pi_+ \left( \int_0^t \eta \cdot \lambda(s) ds \right) - \Pi_- \left( \int_0^t \delta(Q^\eta(s), \eta \cdot c(s)) ds \right)$$

A call center interpretation of the Halfin-Whitt scaling is called resource pooling. We are scaling up simultaneously the customer demand (arrival rate) and the customer resource supply (number of service agents). Taking the following limits gives us the fluid and diffusion models of Mandelbaum et al. [25], i.e.

$$\lim_{\eta \to \infty} \frac{Q^\eta}{\eta}(t) = q(t) \text{ a.s. and } \lim_{\eta \to \infty} \sqrt{\eta} \cdot \left( \frac{Q^\eta}{\eta}(t) - q(t) \right) \overset{d}{=} \hat{Q}(t),$$

where the deterministic process $q(t)$, the fluid mean, is governed by the one dimensional dynamical system

$$\dot{q} = \lambda(t) - \delta(q, c).$$
Moreover, as pointed out in Mandelbaum et al. [25], if the set of time points \( \{ t \mid q(t) = c(t) \} \) has measure zero, then \( \hat{Q}(t) \) is a Gaussian diffusion process (with mean zero when \( Q^\eta(0) \) is only a constant scaled by \( \eta \)) whose variance combines with the fluid mean to form a two-dimensional dynamical system given by (2.5) and

\[
\frac{\dot{v} + \dot{q}}{2} = \lambda - \frac{\partial \delta}{\partial q}(q, c) \cdot v,
\]

where \( v \equiv \text{Var} \hat{Q} \),

\[
\frac{\partial \delta}{\partial q}(q, c) = \mu \cdot \{ q < c \} + \beta \cdot \{ q \geq c \}.
\]

and \( \{ q < c \} \) denotes an indicator function that equals one if the statement is true i.e, if \( q < c \), and zero if the statement is false. In fact, in the case where the arrival rate function is a constant for all time, we can recover the result of Halfin and Whitt [9].

3 Cumulant Moments and Functional Generalizations

3.1 Cumulant Moments

Let \( X \) be a random variable whose moment generating function

\[
\mathbb{E} e^{zX} = \sum_{m=0}^{\infty} \frac{z^m}{m!} \cdot \mathbb{E} X^m
\]

is an analytic function for all complex \( z \) in a neighborhood of zero. It follows that the function \( \log \mathbb{E} e^{zX} \) is also analytic in a neighborhood of zero. We can then define the cumulant moments of \( X \) or \( \{ C^{(k)}X \mid k \geq 1 \} \) to be

\[
\sum_{k=1}^{\infty} \frac{z^k}{k!} \cdot C^{(k)}X \equiv \log \mathbb{E} e^{zX}.
\]

The first four moments are

\[
C^{(1)}X = \mathbb{E}X \quad (3.10)
\]
\[
C^{(2)}X = \text{Var}X \equiv \mathbb{E}(X - \mathbb{E}X)^2 \quad (3.11)
\]
\[
C^{(3)}X = \mathbb{E}(X - \mathbb{E}X)^3 \quad (3.12)
\]
\[
C^{(4)}X = \mathbb{E}(X - \mathbb{E}X)^4 - 3 \cdot (\text{Var}X)^2. \quad (3.13)
\]

Also, we define the \( \text{Skew}X \) and \( \text{Kurt}X \) to be the (generalized) skewness and kurtosis of the random variable \( X \), where

\[
\text{Skew}X \equiv \frac{C^{(3)}X}{\sqrt{(\text{Var}X)^3}} \quad \text{and} \quad \text{Kurt}X \equiv \frac{C^{(4)}X}{(\text{Var}X)^2}. \quad (3.14)
\]
These are the normalized versions of the third and fourth cumulant moments, in the same manner that the correlation of two random variables is the normalized version of their covariance.

Consider random variables with analytic moment generating functions near zero. We call such random variables analytic. The theorem below summarizes the mathematical appeal and utility of cumulant moments.

**Theorem 3.1.** For all analytic random variables $X$ and $Y$, we have

1. For all constants $a$, we have
   \[ C^{(k)}[aX] = a^k \cdot C^{(k)}X \quad (3.15) \]

2. The pair $X$ and $Y$ are independent if and only if
   \[ C^{(k)}[a \cdot X + b \cdot Y] = a^k \cdot C^{(k)}X + b^k \cdot C^{(k)}Y \quad (3.16) \]
   for all strictly positive integers $k$ as well as all constants $a$ and $b$.

3. All the cumulant moments of $X$ of degree 2 or better are zero if and only if $X$ has the distribution of a constant (point mass distribution).

4. All the cumulant moments of $X$ equal the mean for $X$ mean if and only if $X$ is Poisson.

Observe that skewness and kurtosis are invariant under constant translations and scalings of the underlying random variable. Since any Gaussian random variable is the translation and scaling of one that is standard (mean zero and variance one) Gaussian, the skewness and kurtosis of a random variable is an informal measure of how far a random variable is from being Gaussian.

Cumulant moments have a simple characterization of Gaussian random variables but they can also identify Gaussian random variables with the following stronger result.

**Theorem 3.2.** Analytic random variables only have a finite number of non-zero cumulant moments if and only if they are Gaussian.

**Proof.** For any complex number $z$, define the complex value $g(z)$ where
\[ g(z) \equiv \log Ee^{z \cdot X}. \quad (3.17) \]
Notice that $g(r)$ is a real number whenever $r$ is real. Moreover, we can apply Jensen’s inequality and obtain
\[ g(r) \geq r \cdot E X. \quad (3.18) \]
This gives us
\[ \log Ee^{r(X - EX)} = g(r) - r \cdot E X \geq 0, \quad (3.19) \]
so without any loss of generality, we can assume that $EX = 0$. Here, $g$ maps the real line to the positive reals.
For all complex \( z \), we have
\[
\text{Re} \, g(z) = \text{Re} \log E e^{z \cdot X} = |\log E e^{z \cdot X}| \leq \log E |e^{z \cdot X}| = \log E e^{\text{Re} z \cdot X}.
\] (3.20)

Moreover, if \( X \) has only a finite number of non-zero cumulant moments, then \( g \) must be a polynomial of some degree \( n \). The leading coefficient of this polynomial must be some strictly positive real number \( \alpha > 0 \) where
\[
\alpha = \lim_{r \to \infty} \frac{g(r)}{r^{n}}.
\] (3.21)

This gives us
\[
z^{n} = \lim_{r \to \infty} \frac{g(rz)}{\alpha \cdot r^{n}},
\] (3.22)

hence \( z^{n} \) must satisfy the same inequality for all complex \( z \) as \( g \), namely
\[
\text{Re} z^{n} \leq (\text{Re} z)^{n}.
\] (3.23)

If we confine ourselves to the unit circle or \( z = e^{i\theta} \), then we want to know which positive integers \( n \) have the property that
\[
\cos n\theta \leq (\cos \theta)^{n}
\] (3.24)

for all \( \theta \). This is true for the case of \( n = 1 \). Also for the case of \( n = 2 \) since \( \cos 2\theta = (\cos \theta)^{2} - (\sin \theta)^{2} \leq (\cos \theta)^{2} \). However for \( n \geq 3 \), a contradiction occurs when we set \( \theta \equiv 2\pi/n \). Since \( 0 < \theta < \pi \) here, it follows that \( (\cos \theta)^{n} < 1 = \cos n\theta \). This means that restricting \( g \) to be a polynomial forces its degree to be no greater than 2.

### 3.2 Functional Cumulant Moments

Consider a random variable \( X \), where there exists some positive real number \( r \) with \( E e^{r \cdot X} < \infty \). We can then construct an exponential change of measure with an expectation \( E_{r} \) such that
\[
E_{r} f(X) \equiv \frac{E \left[ f(X) \cdot e^{r \cdot X} \right]}{E e^{r \cdot X}}.
\] (3.25)

for all bounded continuous functions \( f \).

Now assume that the underlying distribution for \( X \) is sufficiently nice so that \( E e^{z \cdot X} \) is an analytic function when \( z \) is close to zero. Using analytic continuation, our change of measure expectation can be extended to an analytic function \( E_{z} \) when \( z \) is close to zero. We now have
\[
E_{z} X = \frac{E \left[ X \cdot e^{z \cdot X} \right]}{E e^{z \cdot X}} = \frac{d}{dz} \log E e^{z \cdot X} = \sum_{k=0}^{\infty} \frac{z^{k}}{k!} \cdot C^{(k+1)} X
\] (3.26)

and from this it follows that
\[
C^{(k+1)} X = \left. \frac{d^{k}}{dz^{k}} \right|_{z=0} E_{z} X.
\] (3.27)

Moreover, these cumulant moments uniquely characterize the underlying distribution for \( X \).
Using these results as motivation, we now define the functional cumulant of $X$ with respect to some bounded continuous function $f$ to be

$$C^{(k+1)}_z f [X] \equiv \frac{d^k}{dz^k} \bigg|_{z=0} E_z [f(X)].$$

The following result relates functional cumulants to exponential changes of measure

**Theorem 3.3.** Whenever $E_z f(X)$ is analytic in $z$

$$\frac{d^k}{dz^k} E_z f(X) = C^{(k+1)}_z f[X],$$

for all positive integers $k$

**Proof.** Observe that

$$E_{z+w} [f(X)] = \frac{E \left[ f(X) e^{(z+w)X} \right]}{E \left[ e^{(z+w)X} \right]} = \frac{E_z \left[ f(X) e^{wX} \right]}{E_z \left[ e^{wX} \right]} = (E_z)_w [f(X)].$$

From this it follows that

$$C^{(k+1)}_z f [X] = \frac{\partial^k}{\partial w^k} \bigg|_{w=0} E_{z+w} [f(X)] = \frac{d^k}{dz^k} E_z [f(X)]$$

and this completes the proof.

**Corollary 3.4.** The first 4 functional cumulant moment formulas are:

$$C^{(1)} f [X] = E[f(X)]$$
$$C^{(2)} f [X] = \text{Cov} [f(X), X]$$
$$C^{(3)} f [X] = \text{Cov} [f(X), (X - E X)^2]$$
$$C^{(4)} f [X] = \text{Cov} [f(X), (X - E X)^3] - 3 \cdot \text{Var}X \cdot \text{Cov} [f(X), X].$$

**Proof.** Since

$$E_z f(X) = \frac{E \left[ f(X) e^{zX} \right]}{E e^{zX}} = \frac{E \left[ f(X) e^{z(X - E X)} \right]}{E e^{z(X - E X)}},$$

we can expand the numerator of the final ratio and obtain

$$E \left[ f(X) e^{z(X - E X)} \right] = E f(X) + z \cdot E \left[ f(X) \cdot (X - E X) \right]$$
$$+ \frac{z^2}{2} \cdot E \left[ f(X) \cdot (X - E X)^2 \right] + \frac{z^3}{6} \cdot E \left[ f(X) \cdot (X - E X)^3 \right] + O(z^4).$$

Expanding the denominator of the final ratio yields

$$E \left[ f(X) e^{z(X - E X)} \right] = 1 + \frac{z^2}{2} \cdot \text{Var}X + \frac{z^3}{6} \cdot C^{(3)} X + O(z^4),$$

(3.33)
which gives us

$$\frac{1}{\mathbb{E}[f(X) \cdot e^{z(X-EX)}]} = 1 - \frac{z^2}{2} \cdot \text{Var}X - \frac{z^3}{6} \cdot C^{(3)}X + O(z^4). \quad (3.34)$$

Finally, by combining these expansions, we have

$$\mathbb{E}_z f(X) = \mathbb{E}[f(X)] + z \cdot \mathbb{E}[f(X) \cdot (X-EX)] + \frac{z^2}{2} \cdot (\mathbb{E}[f(X) \cdot (X-EX)^2] - \mathbb{E}[f(X)] \cdot \text{Var}X) + \frac{z^3}{6} \cdot \left( \mathbb{E}[f(X) \cdot (X-EX)^3] - \mathbb{E}[f(X)] \cdot C^{(3)}X - \frac{6}{2} \cdot \mathbb{E}[f(X) \cdot (X-EX)] \cdot \text{Var}X \right) + O(z^4).$$

When \( f \) is a scaling function, like \( f(x) = \delta(x,c) = \mu \cdot x \) when \( \mu = \beta \), then all the functional cumulant moments of \( X \) are equal to \( \mu \) times the corresponding cumulant moment of \( X \). In general, a functional moment of a random variable may *not* be equal or proportional to its cumulant moment or even a cumulant moment of the function applied to the random variable.

## 4 Closure Approximations for the Erlang-A Queue

In this section, we derive the functional forward equations for the Erlang-A queueing model and approximate the cumulant moments of the Erlang-A queue using Gaussian based closure approximations. We show that these closure approximations are effective at approximating the cumulant moments of the Erlang-A queue and are effective tools for approximating other important performance measures like the probability of delay.

### 4.1 Functional Forward Equations

To gain a better understanding of the dynamics of the mean, variance, and third cumulant moment of the Erlang-A queueing process, we need to study their rates of change over time. This leads us to analyze the *functional version* of the Kolmogorov forward equations for the \( M(t)/M/c(t) + M \) queue, which is of the form

$$\dot{\mathbb{E}}[f(Q)] = \lambda \cdot \mathbb{E}[f(Q + 1) - f(Q)] + \mathbb{E}[\delta(Q,c) \cdot (f(Q - 1) - f(Q))], \quad (4.35)$$

for all appropriate functions \( f \). We always assume, for the remainder of this paper, that quantities such as \( \beta \) and \( \mu \) are constant. To simplify our notation, time dependent quantities such as \( Q(t) \), \( \lambda(t) \), and \( c(t) \) are denoted in this paper as \( Q \), \( \lambda \), and \( c \), with their time dependence suppressed. For an expression like \( \mathbb{E}[f(Q)] \) we use the “dot” notation of physics to denote its time derivative when we do not make time explicit or

$$\dot{\mathbb{E}}[f(Q)] \equiv \frac{d}{dt} \mathbb{E}[f(Q(t)) \mid Q(0) = q(0)]. \quad (4.36)$$
Theorem 4.1. The dynamics for the cumulant moment generating function of an Erlang-A queueing process are
\[
\log (E^{z \cdot Q}) = (1 - e^{-z}) \cdot (\lambda \cdot e^{z} - E_{z} \delta(Q, c)) .
\] (4.37)

Proof. For any given function \(f\), the functional forward equation for the Erlang-A queue is
\[
\dot{E} [f(Q)] = \lambda \cdot E [f(Q + 1) - f(Q)] + E [\delta(Q, c) \cdot (f(Q - 1) - f(Q))].
\] (4.38)
For the special case of \(f(Q) \equiv e^{zQ}\), this reduces to
\[
\dot{E} e^{zQ} = \lambda \cdot (e^{z} - 1) \cdot E e^{zQ} + (e^{-z} - 1) \cdot E [\delta(Q, c) \cdot e^{zQ}]
\] (4.39)
Dividing both sides by \(E e^{zQ}\) gives us
\[
\log (E^{z \cdot Q}) = \lambda \cdot (e^{z} - 1) + (e^{-z} - 1) \cdot E [\delta(Q, c) \cdot e^{zQ}] / E e^{zQ}
\] (4.40)
When \(\lambda\) is a constant and we are in steady state, then the time derivative of the logarithm term is zero. We now have
\[
\lambda \cdot e^{z} = E_{z} \delta(Q, c).
\] (4.42)
The rest follows by differentiation with respect to \(z\) and setting \(z = 0\).

Now we have the terminology to express a new result for the steady state distribution of the Erlang-A queue in steady state.

Corollary 4.2. For an Erlang-A queueing process with a constant arrival rate \(\lambda\), we have in steady state
\[
\lambda = C^{(k)} \delta_{c}[Q]
\] (4.43)
for all strictly positive integers \(k\), where \(\delta_{c}(x) \equiv \delta(x, c)\).

Proof. When \(\lambda\) is a constant and we are in steady state, then the time derivative of the logarithm term is zero. The rest follows by differentiation with respect to \(z\) and setting \(z = 0\).

Corollary 4.3. The dynamics of the cumulant moments for the Erlang-A queue are
\[
\dot{C}^{(k)} Q = \lambda - \sum_{j=1}^{k} \binom{k}{j-1} (-1)^{k-j} \cdot C^{(j)} \delta_{c}[Q].
\] (4.44)
for all strictly positive integers \(k\).

Proof. If we differentiate this formula \(k\) times, with respect to \(z\), then we have for all strictly positive integers \(k\)
\[
\dot{C}^{(k)} z \cdot Q = \lambda \cdot e^{z} + (e^{-z} - 1) \cdot C^{(k+1)} \delta_{c}[Q] + e^{-z} \cdot \sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j} \cdot C^{(j+1)} \delta_{c}[Q].
\] (4.45)
Setting \(z = 0\) gives us the rest.
Corollary 4.4. We have the following dynamics for the first 6 cumulant moments:

\[ C^{(1)} Q = \lambda - C^{(1)} \delta_c [Q] \] (4.46)

\[ \frac{C^{(1)}}{2} Q + \frac{C^{(2)}}{2} Q = \lambda - C^{(2)} \delta_c [Q] \] (4.47)

\[ \frac{C^{(1)}}{6} Q + \frac{C^{(2)}}{3} Q + \frac{C^{(3)}}{4} Q = \lambda - C^{(3)} \delta_c [Q] \] (4.48)

\[ \frac{C^{(2)}}{4} Q + \frac{C^{(3)}}{4} Q + \frac{C^{(4)}}{5} Q = \lambda - C^{(4)} \delta_c [Q] \] (4.49)

\[ -\frac{C^{(1)}}{30} Q + \frac{C^{(2)}}{3} Q + \frac{C^{(3)}}{2} Q + \frac{C^{(4)}}{5} Q = \lambda - C^{(5)} \delta_c [Q] \] (4.50)

\[ -\frac{C^{(2)}}{12} + 5\frac{C^{(4)}}{12} Q + \frac{C^{(5)}}{2} Q + \frac{C^{(6)}}{6} Q = \lambda - C^{(6)} \delta_c [Q] \] (4.51)

Proof. Rewriting cumulant moments dynamics in infinite matrix form gives us

\[
\begin{bmatrix}
  C^{(1)} Q \\
  C^{(2)} Q \\
  C^{(3)} Q \\
  C^{(4)} Q \\
  \vdots
\end{bmatrix} = \begin{bmatrix}
  \lambda \\
  \lambda \\
  \lambda \\
  \lambda \\
  \vdots
\end{bmatrix} - \begin{bmatrix}
  1 & 0 & 0 & 0 & \cdots \\
  -1 & 2 & 0 & 0 & \cdots \\
  1 & -3 & 3 & 0 & \cdots \\
  -1 & 4 & -6 & 4 & \cdots \\
  \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix} \cdot \begin{bmatrix}
  C^{(1)} \delta_c [Q] \\
  C^{(2)} \delta_c [Q] \\
  C^{(3)} \delta_c [Q] \\
  C^{(4)} \delta_c [Q] \\
  \vdots
\end{bmatrix}.
\] (4.52)

The column vector of all \( \lambda \)s is an eigenvector for this matrix or

\[
\begin{bmatrix}
  \lambda \\
  \lambda \\
  \lambda \\
  \lambda \\
  \vdots
\end{bmatrix} = \begin{bmatrix}
  1 & 0 & 0 & 0 & \cdots \\
  -1 & 2 & 0 & 0 & \cdots \\
  1 & -3 & 3 & 0 & \cdots \\
  -1 & 4 & -6 & 4 & \cdots \\
  \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix} \cdot \begin{bmatrix}
  \lambda \\
  \lambda \\
  \lambda \\
  \lambda \\
  \vdots
\end{bmatrix}.
\] (4.53)

Applying the inverse of the matrix to () then gives us

\[
\begin{bmatrix}
  1 & 0 & 0 & 0 & \cdots \\
  -1 & 2 & 0 & 0 & \cdots \\
  1 & -3 & 3 & 0 & \cdots \\
  -1 & 4 & -6 & 4 & \cdots \\
  \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}^{-1} \cdot \begin{bmatrix}
  C^{(1)} Q \\
  C^{(2)} Q \\
  C^{(3)} Q \\
  C^{(4)} Q \\
  \vdots
\end{bmatrix} = \begin{bmatrix}
  \lambda \\
  \lambda \\
  \lambda \\
  \lambda \\
  \vdots
\end{bmatrix} - \begin{bmatrix}
  C^{(1)} \delta_c [Q] \\
  C^{(2)} \delta_c [Q] \\
  C^{(3)} \delta_c [Q] \\
  C^{(4)} \delta_c [Q] \\
  \vdots
\end{bmatrix}.
\] (4.54)
Observing that
\[
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & \cdots \\
-1 & 2 & 0 & 0 & 0 & 0 & \cdots \\
1 & -3 & 3 & 0 & 0 & 0 & \cdots \\
-1 & 4 & -6 & 4 & 0 & 0 & \cdots \\
1 & -5 & 10 & -10 & 5 & 0 & \cdots \\
-1 & 6 & -15 & 20 & -15 & 6 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & \cdots \\
1/2 & 1/2 & 0 & 0 & 0 & 0 & \cdots \\
1/6 & 1/2 & 1/3 & 0 & 0 & 0 & \cdots \\
0 & 1/4 & 1/2 & 1/4 & 0 & 0 & \cdots \\
-1/30 & 0 & 1/3 & 1/2 & 1/5 & 0 & \cdots \\
0 & -1/12 & 0 & 5/12 & 1/2 & 1/6 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix},
\]
gives us the equations.

We can write the first three of these equations more explicitly as

\[
\dot{E}Q = \lambda - E\delta(Q, c), \quad (4.55)
\]
\[
\frac{\dot{E}Q + \dot{\text{Var}}Q}{2} = \lambda - \text{Cov}[Q, \delta(Q, c)], \quad (4.56)
\]
and
\[
\frac{\dot{E}Q}{6} + \frac{\dot{\text{Var}}Q}{2} + \frac{C^{(3)}Q}{3} = \lambda - \text{Cov}\left[(Q - E Q)^2, \delta(Q, c)\right]. \quad (4.57)
\]

From a computational perspective, we want the ensemble of formulas for the time derivatives of the mean, variance, and third-cumulant moment, as summarized in (4.55)-(4.57), to be an autonomous set of differential equations. This means that their current behavior should be some integral functional of their past behavior. We can achieve this by making a closure approximation in the same spirit as Rothkopf and Oren Rothkopf and Oren [44]. The philosophy that they give for this technique is as follows (see page 524 of Rothkopf and Oren [44]):

... The basic strategy of a closure technique is to reduce an infinite system of equations to a finite system by making a “closure assumption” in the form of a functional relationship between the variables of the system.

Similar techniques for approximating non-stationary (dynamic rate) queueing models are also used in Taaffe and Ong [49], Clark [1], Ingolfsson et al. [13], Taaffe and Ong [48], Taaffe and Clark [47], Nasr and Taaffe [31], Schwarz et al. [45], Pender [38].

In general, we start by assuming that our underlying closure distribution for the queueing process is uniquely defined by a finite set of parameters. We assume that these parameters are uniquely defined by the same number of expectations of some distinct functions of the queueing process. The forward equations for these functional expectations then form a finite dimensional, dynamical system for these parameters. Whereas Rothkopf and Oren [44] and Taaffe and Ong [49] assume an underlying discrete distribution for their closure assumptions, our underlying distribution is continuous, is based on polynomials of Gaussian random variables, and was inspired from the Halfin-Whitt asymptotics.
4.2 Deterministic Mean Approximation is the Fluid Limit

For example, we can define a deterministic mean approximation for our queueing model by assuming that some underlying deterministic process \( q \equiv \{q(t) | t \geq 0 \} \) approximates our Markovian queueing process or \( Q \approx q \). If we replace \( Q \) by \( q \) in the Kolmogorov forward equation for the mean of \( Q \) as given by (4.55), then \( q \) solves the resulting one-dimensional, dynamical system.

\[
\dot{q} = \lambda - \mu \cdot (q \wedge c) - \beta \cdot (q - c)^+,
\]

where we set \( Q(0) = q(0) \). This method, however, takes us right back to the fluid limit given by (2.5). We can also write this fluid equation as

\[
\begin{cases}
\lambda - \mu \cdot q & \text{when } q < c, \\
\lambda + (\beta - \mu) \cdot c - \beta \cdot q & \text{when } q \geq c.
\end{cases}
\]

\[(4.59)\]

We plot the deterministic mean and the fluid limit against a 10,000 sample paths of a stochastic simulation on the left of Figure 2. It is observed that the fluid limit does a decent job of approximating the dynamics, however, it is not perfect. We find that the fluid limit underestimates the queue length when compared to the simulation. This can be explained by Jensen’s inequality in combination with the fact that the minimum function is concave and the service rate is larger than the abandonment rate. On the right of Figure 2 we also observe that the variance of the diffusion limit is not a good estimator of the variance of the queue length process. This inaccuracy of the fluid and diffusion limits motivated the GVA method Ko and Gautam [17], Massey and Pender [28], which we will review in the sequel.

However, before we review the GVA method, we now want to visualize the fluid limit or deterministic mean approximation dynamical system for the special case of \( \lambda(t) = a + \hat{a} \cdot \cos bt \) and \( \beta = \mu \). Our Erlang-A system now reduces to the case of an \( M/M/\infty \) queue, which was extensively analyzed by Eick et al. [33, 34], Massey and Whitt [29], McCalla and Whitt [30].
Here, the mean behavior is the dynamical system that equals the fluid limit. Notice that the transient behavior of this system captures the dynamics of the corresponding Erlang-A system for the distinct cases of \( q < c \) and \( q \geq c \).

**Theorem 4.5.** If \( q(t) \) is the mean for an \( M(t)/M/\infty \) queue, then the phase space pair \((q(t), \dot{q}(t))\) belongs to an ellipse of extremal radii \( r^* \) and \( r_* \) where

\[
r^* = \frac{\hat{a} b}{\sqrt{\mu^2 + b^2}} \quad \text{and} \quad r_* = \frac{\hat{a}}{\sqrt{\mu^2 + b^2}}. \tag{4.60}
\]

Moreover, the ellipse is centered at \((a/\mu + x \cdot e^{-\mu t}, y \cdot e^{-\mu t})\), where

\[
x = q(0) - \frac{a}{\mu} - \frac{\hat{a} \mu}{\mu^2 + b^2} \quad \text{and} \quad y = \dot{q}(0) - \frac{\hat{a} b^2}{\mu^2 + b^2}. \tag{4.61}
\]

**Proof.** First the solution for \( \dot{q} = \lambda - \mu \cdot q \) is

\[
q(t) = q(0) \cdot e^{-\mu t} + \int_0^t \lambda(s) \cdot e^{-\mu(t-s)} \, ds. \tag{4.62}
\]

Using complex numbers, we can rewrite the arrival rate function as

\[
\lambda(t) = a + \hat{a} \cdot \cos bt = a + \hat{a} \cdot \Re(e^{ibt}). \tag{4.63}
\]

Now we have

\[
q(t) = q(0) \cdot e^{-\mu t} + \int_0^t \lambda(s) \cdot e^{-\mu(t-s)} \, ds \tag{4.64}
\]

\[
= q(0) \cdot e^{-\mu t} + \int_0^t (a + \hat{a} \cdot \Re(e^{ibs})) \cdot e^{-\mu(t-s)} \, ds \tag{4.65}
\]

\[
= q(0) \cdot e^{-\mu t} + a \cdot \int_0^t e^{-\mu(t-s)} \, ds + \hat{a} \cdot e^{-\mu t} \cdot \Re \left( \int_0^t e^{(\mu+ib)s} \, ds \right) \tag{4.66}
\]

\[
= q(0) \cdot e^{-\mu t} + a \cdot \left( 1 - e^{-\mu t} \right) + \hat{a} \cdot e^{-\mu t} \cdot \Re \left( e^{(\mu+ib)t} - 1 \right) \tag{4.67}
\]

\[
= q(0) \cdot e^{-\mu t} + \frac{a}{\mu} \cdot \left( 1 - e^{-\mu t} \right) + \hat{a} \cdot \Re \left( \frac{e^{ibt} - e^{-\mu t}}{\mu + ib} \right) \tag{4.68}
\]

\[
= q(0) \cdot e^{-\mu t} + \frac{a}{\mu} \cdot \left( 1 - e^{-\mu t} \right) + \frac{\hat{a}}{\mu^2 + b^2} \cdot \Re(e^{ibt} - e^{-\mu t}, \mu + ib) \tag{4.69}
\]

\[
= \frac{a}{\mu} + \left( q(0) - \frac{a}{\mu} - \frac{\hat{a} \mu}{\mu^2 + b^2} \right) \cdot e^{-\mu t} + \frac{\hat{a}}{\mu^2 + b^2} \cdot \Re(e^{ibt}, \mu + ib) \tag{4.70}
\]

\[
= \frac{a}{\mu} + x \cdot e^{-\mu t} + \frac{\hat{a}}{\mu^2 + b^2} \cdot \Re(e^{ibt}, \mu + ib), \tag{4.71}
\]

where \( \langle z, w \rangle = \Re(z \cdot \overline{w}) \) is the dot product when we think of the two complex numbers as two-vectors over the real numbers. Moreover, we can also rewrite the solution for \( q(t) \) as

\[
q(t) = q(0) \cdot e^{-\mu t} + \int_0^t \lambda(t-s) \cdot e^{-\mu s} \, ds. \tag{4.72}
\]
Now if we take the time derivative of \( q(t) \), then we have

\[
\dot{q}(t) = -\mu \cdot q(t) \cdot e^{-\mu t} + \lambda(0) \cdot e^{-\mu t} + \int_0^t \dot{\lambda}(t-s) \cdot e^{-\mu s} \, ds \tag{4.73}
\]

\[
= \dot{q}(0) \cdot e^{-\mu t} + \int_0^t (\dot{a} \cdot \Re(i \cdot e^{ib(t-s)}) \cdot e^{-\mu s} \, ds \tag{4.74}
\]

\[
= \dot{q}(0) \cdot e^{-\mu t} + \int_0^t (\dot{a} \cdot \Re(i \cdot e^{ibt}) \cdot e^{-\mu s} \, ds \tag{4.75}
\]

\[
= y \cdot e^{-\mu t} + \frac{\dot{a}b^2}{\mu^2 + b^2} \cdot \langle i \cdot e^{ibt}, \mu + ib \rangle \tag{4.76}
\]

Since \( \langle e^{i\theta}, z \rangle^2 + \langle i \cdot e^{i\theta}, z \rangle^2 = |z|^2 \), we then have

\[
b^2 \cdot \left( q(t) - \frac{a}{\mu} - x \cdot e^{-\mu t} \right)^2 + \left( \dot{q}(t) - y \cdot e^{-\mu t} \right)^2 = \frac{\dot{a}^2b^2}{\mu^2 + b^2}. \tag{4.79}
\]

### 4.3 Gaussian Variance Approximation

Let us now extend this method to the two-dimensional, dynamical system case. Inspired by our diffusion limit being Gaussian, suppose that we approximate the dynamics of the mean and variance of \( Q \) by a random process \( Q \equiv \{ Q(t) | t \geq 0 \} \) such that

\[
Q \overset{d}{=} q + G \cdot \sqrt{v} \tag{4.80}
\]

or more formally

\[
P \{ Q \geq c \} \approx P \{ G \geq \chi \} \tag{4.81}
\]

for all \( t \geq 0 \), where \( \{ q(t), v(t) | t \geq 0 \} \) is some two-dimensional, deterministic, dynamical system, with the \( v \) process always being positive. In this paper, we always define \( G \) to be a standard Gaussian random variable or Gaussian(0, 1). Either one is shorthand for a Gaussian distribution with zero mean and unit variance. We define \( \varphi \) and \( \Phi \) to be the density and the cumulative distribution functions, for \( G \) respectively, where

\[
\varphi(x) \equiv \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \Phi(x) \equiv \int_{-\infty}^{x} \varphi(y) \, dy, \quad \text{and} \quad \Phi(x) \equiv 1 - \Phi(x) = \int_{x}^{\infty} \varphi(y) \, dy. \tag{4.82}
\]

Now we make this substitution of \( q, v, \) and \( G \) into the forward equations for the mean and the variance of \( Q \), i.e. (4.55) and (4.56). We can call the resulting two dimensional
dynamical system the \textit{Gaussian variance approximation} (GVA). The dynamical for $q$ and $v$ are the differential equations

\begin{equation}
\dot{q} = \lambda - \mu \cdot q - E\delta(G,\chi) \cdot \sqrt{v} \tag{4.83}
\end{equation}

and

\begin{equation}
\frac{\dot{q} + \dot{v}}{2} = \lambda - \text{Cov}[G,\delta(G,\chi)] \cdot v, \tag{4.84}
\end{equation}

where

\begin{equation}
EQ \approx q, \ VarQ \approx v, \text{ but } \chi \equiv \frac{c - q}{\sqrt{v}}. \tag{4.85}
\end{equation}

To solve these equations numerically, we need to compute the expectation and covariance terms involving functions of the standard Gaussian random variable $G$. The final results yield generic functions of $\chi$, which is a simple function of $q$ and $v$. We compute these Gaussian terms by using the following lemma:

\textbf{Lemma 4.6} (Stein \cite{46}). \textit{The random variable $G$ is Gaussian(0, 1) if and only if}

\begin{equation}
E[G \cdot f(G)] = E\left[\frac{d}{dG}f(G)\right], \tag{4.86}
\end{equation}

\textit{for all generalized functions $f$.}
For example, since \((G - \chi)^+ = (G - \chi) \cdot \{G \geq \chi\}\), then Stein’s lemma can be used to obtain
\[
E(G - \chi)^+ = E[G \cdot \{G \geq \chi\}] - \chi \cdot P\{G \geq \chi\} = \varphi(\chi) - \chi \cdot \Phi(\chi) \tag{4.87}
\]
Observe that as a function of \(G\), \(\{G \geq \chi\}\) is an increasing unit, single step function. Moreover, the density \(\varphi\) is an infinitely differentiable function. Finally, the derivative of \(\{G \geq \chi\}\), which is a unit step function at the value \(\chi\), is a unit point mass measure at \(\chi\). This is a special case of a generalized function. As a result, Stein’s lemma gives us
\[
E[ G \cdot \{G \geq \chi\}] = \varphi(\chi) \tag{4.88}
\]
Moreover, since \((G - \chi)^+ = G - G \wedge \chi\), we have
\[
E[G \wedge \chi] = -E(G - \chi)^+ = \chi \cdot \Phi(\chi) - \varphi(\chi) \tag{4.89}
\]
Similar arguments give us
\[
\text{Cov}[G, G \wedge \chi] = E[G \cdot (G \wedge \chi)] = E[\{G \leq \chi\}] = P\{G \leq \chi\} = \Phi(\chi) \tag{4.90}
\]
and
\[
\text{Cov}[G, (G - \chi)^+] = 1 - \text{Cov}[G, G \wedge \chi] = \Phi(\chi). \tag{4.91}
\]
These positive covariances are in keeping with the FKG inequality by Fortuin et al. \[8\]. This theorem states that increasing functions of the same random variable are always positively correlated.

When we make the substitutions into (4.83) and (4.84), the GVA dynamical system reduces to
\[
\dot{q} = \lambda - \mu \cdot q + (\beta - \mu) \cdot (\chi \cdot \Phi(\chi) - \varphi(\chi)) \cdot \sqrt{v} \tag{4.92}
\]
and
\[
\frac{\dot{q} + v}{2} = \lambda - (\mu \cdot \Phi(\chi) + \beta \cdot \Phi(\chi)) \cdot v. \tag{4.93}
\]
These equations are the same as the ones with the “\(g\) functions” as used in Ko and Gautam \[17\]. On the left of Figure 3, we observe that GVA yields a better approximation to the simulated mean than the fluid model. In fact, it almost matches the simulation perfectly so you can only see one line. One reason for the better accuracy is that the GVA method includes stochastic fluctuations that the fluid limit or the deterministic mean approximation does not. We see for the variance on the right of Figure 3, however, the GVA plot does not match the simulation as well the mean, but it is a significant improvement over the diffusion variance approximation. This inaccuracy in the variance motivated us to construct a new approximation called the Gaussian Skewness Approximation (GSA), which will review in the sequel. Just like GVA was a refinement to the fluid limit, GSA will serve as a refinement to the diffusion limit and GVA.

### 4.4 Gaussian Skewness Approximation

Skewness captures an intrinsic property of the distribution that is invariant with respect to deterministic translations and positive scalings of the underlying random variable. It follows
from these invariance properties that any Gaussian distribution has the same skewness as the standard Gaussian distribution for $G$, which is zero. Moreover, a Gaussian distribution is uniquely characterized by having its third and all higher degree cumulant moments all equal to zero. Thus we can use skewness informally as a metric for how “close” a random variable is to being Gaussian.

We now introduce a new approximation method and call it the Gaussian skewness approximation (GSA). For some three-dimensional, dynamical system \( \{q(t), v(t), \sigma_\theta(t) | t \geq 0\} \), we assume that

\[
Q^d \approx q + H_\theta \cdot \sqrt{v},
\]  

where

\[
H_\theta \equiv G \cos \theta(t) + \frac{G^2 - 1}{\sqrt{2}} \sin \theta(t)
\]  

or more formally

\[
P\{Q(t) \geq c\} \approx P\{H_{\theta(t)} \geq \chi(t)\} = \Phi_\theta(\chi(t))
\]  

for all \( t \geq 0 \), where

\[
H_\theta = G \cos \theta + \frac{G^2 - 1}{\sqrt{2}} \sin \theta
\]  

for all \( t \geq 0 \). Since 1, \( G \), and \( G^2 - 1 \) are orthogonal vectors, they are all uncorrelated as random variables. This means that \( \mathbb{E}Q = q \) and \( \text{Var}Q = v \). We also define \( \sigma_\theta \) to be the skewness of \( Q \) or

\[
\sigma_\theta \equiv \text{Skew}Q \approx \text{Skew}H_\theta = \sqrt{2} \cdot (3 - \sin \theta) \cdot \sin \theta.
\]

This approximation is a natural extension of the GVA method since we are using the next orthogonal Hermite polynomial of \( G \). Notice that neither the mean nor the variance are a function of \( \theta \). Moreover, the second and third orthonormal components of \( Q \), i.e. \( G \) and \( (G^2 - 1)/\sqrt{2} \), are multiplied by \( \sqrt{v} \cos \theta \) and \( \sqrt{v} \sin \theta \) respectively. This is the general representation for any two dimensional rectangular coordinate system in terms of its polar
coordinates, i.e. the “radius” \( \sqrt{v} \) and the “angle” \( \theta \). Any two dimensional vector in this subspace with mean squared norm \( \sqrt{v} \) would have this form for some value of \( \theta \).

Now observe that the one-dimensional distributions of the process \( Q \) are uniquely parameterized by the deterministic processes given by \( q, v, \) and \( \sin \theta \) where \( v \geq 0 \). This is due to the fact that the squares of either \( \cos \theta \) or \( -\cos \theta \) plus the square of \( \sin \theta \) equals 1. Moreover, \( G \cos \theta \) and \( -G \cos \theta \) are identically distributed since \( G \) is a standard Gaussian random variable. Hence \( G \cos \theta \) and \( -G \cos \theta \) are identically distributed with \( G^2 = (-G)^2 \).

The next theorem shows that \( \sigma_\theta \) is an invertible function of \( \sin \theta \). This means that the one-dimensional distributions of \( Q \) are also uniquely parameterized by the deterministic processes \( q, v, \) and \( \sigma_\theta \). Now we state and prove our main result.

**Proposition 4.7.** Suppose that the one-dimensional distributions of the Erlang-A process \( Q \) are given by

\[
Q \approx q + H_\theta \cdot \sqrt{v}, \quad \text{where} \quad H_\theta \equiv G \cos \theta + \frac{1}{\sqrt{2}} (G^2 - 1) \sin \theta \tag{4.99}
\]

and the time dependent parameters \( q \) and \( v \), combined with \( \sigma_\theta \) or \( \sin \theta \), form a three-dimensional dynamical system.

If \( Q \) solves the same moment forward equations as our queueing process \( Q \) for its mean, variance, and third cumulant moment, then we must have

\[
\text{E}Q \approx q, \text{Var}Q \approx v, \quad \text{and} \quad \text{Skew}Q \approx C^{(3)}H_\theta = \sigma_\theta, \tag{4.100}
\]

where

\[
\sigma_\theta \equiv \sqrt{2} \cdot (3 - \sin^2 \theta) \cdot \sin \theta, \tag{4.101}
\]

which holds if and only if

\[
\sin \theta = 2 \cdot \sin \left( \frac{1}{3} \sin^{-1} \left( \frac{\sigma_\theta}{2\sqrt{2}} \right) \right). \tag{4.102}
\]

Finally, our dynamical system is

\[
\dot{q} = \lambda - \mu \cdot q - \text{E}\delta(H_\theta, \chi) \cdot \sqrt{v}, \tag{4.103}
\]

\[
\frac{\dot{q} + \dot{v}}{2} = \lambda - \text{Cov} \left[ H_\theta, \delta(H_\theta, \chi) \right] \cdot v, \tag{4.104}
\]

and

\[
\frac{\dot{q}}{6} + \frac{\dot{v}}{2} + \frac{(\sigma_\theta \cdot \sqrt{v^3})}{3} = \lambda - \text{Cov} \left[ H_\theta^2, \delta(H_\theta, \chi) \right] \cdot \sqrt{v^3}, \tag{4.105}
\]

with \( \chi \equiv (c - q)/\sqrt{v} \) or equivalently \( c = q + \chi \cdot \sqrt{v} \).

**Proof.** See Massey and Pender \[28\].
To explore the effectiveness of the GSA method, we compare the method to simulation, GVA, and the fluid and diffusion limits. On the left of Figure 4, we observe that GSA is as good an estimator of the mean as GVA. Moreover, we observe on the right of Figure 4 that GSA is a better estimator of the simulated variance of the queue length process. Thus, by incorporating the skewness of the queue length distribution, we are better able to capture the non-Gaussian dynamics of the queue length process. On the left of Figure 5, we also observe that the GSA method is able to capture the simulated skewness of the queue length distribution. This is an important insight since it is observed that the skewness is not zero and is positive. Positive skewness implies that extreme values of the queue length are more likely to be positive and the mean is larger than the median. We observe that the mean queue length is larger than the median queue length on the right of Figure 5.

Figure 4: $\lambda(t) = 10 + 2 \cdot \sin t$, $\mu = 1$, $\beta = 0.5$, $Q(0) = 100$, $c = 10$.

Figure 5: $\lambda(t) = 10 + 2 \cdot \sin t$, $\mu = 1$, $\beta = 0.5$, $Q(0) = 100$, $c = 10$. 
4.5 Related Closure Approximations

Although we describe a hierarchy of closure approximations for the Erlang-A queue that successively approximate the dynamics better and better, it is important to highlight other closure approximation methods that exist. The first method that we mention here is that of Pender [34]. The major insight of this method was to approximate the queue length density with a Gram-Charlier series expansion and a Gaussian surrogate distribution. Using this method, Pender [34] was able to demonstrate that one can view the Gram Charlier series as a perturbation of the GVA method that incorporates higher cumulant moments linearly. This method was also applied successfully to nonstationary Erlang-loss queues with abandonment in Pender [36]. Although the Gram-Charlier method works well in practice, it was not satisfying theoretically since one could not prove error bounds for the methodology. One reason was that the Gram Charlier series approach or any continuous distribution approach could not approximate the discreteness of the queue length process itself. However, an initial idea of Pender [35], inspired by the infinite server queue, was to approximate the queue length process with a Poisson distribution. Since the Poisson distribution was discrete, it was better at approximating the discreteness of the queue length process. However, rigorous error bounds of the method in Pender [35] were very hard to obtain since it used a process approximation idea instead of a density approach. However, Engblom and Pender [5], using Poisson-Charlier polynomial expansions and weighted Sobolev spaces, were able to derive a new discrete closure approximation method based on a density approximation with rigorous error bounds. This was the first closure approximation method that guaranteed error bounds for all moments as a function of the number of terms that were used in the approximation.

5 Performance Analysis

In this section, we analyze various performance measures of the Erlang-A queue using the closure approximations we have developed. However, before we begin, we provide a 3D plot of the Erlang-A queue to understand its behavior in a variety of settings.

5.1 3D Dynamic Analytics

In Figure 6, we provide a 3 dimensional plot of the Erlang-A queue in three different settings. The x-axis represents the queue length, the y-axis represents the time derivative of the queue length, and the z-axis represents time. We plot three different settings of the Erlang-A queue. The first setting in blue is the impatient case where the abandonment rate $\beta$ is twice that of the service rate $\mu$ and this represents the setting where customers are more impatient. The second setting is when the abandonment rate is equal to the service rate, thus yielding an infinite server queue. The third and last setting is where the abandonment rate is half that of the service rate and this represents the setting where customers are more patient. In Figure 6, we observe an ordering of the queue length processes, where we see that when customers are more patient, the higher the queue length is going to be since more of them do not leave. Thus, we see that the Erlang-A queue with more patient customers has a larger amplitude of periodic behavior.
Figure 6: Erlang-A Queue 3D Cylinder \((t, \frac{d}{dt}q(t), q(t))\).
\[
\lambda(t) = 100 + 20 \cdot \sin t, \mu = 1, Q(0) = 80, \beta \in \{0.5, 1, 2\}.
\]

In Figure 6, we also observe that the maximum and minimum queue lengths are obtained when \(q(t)\) are equal to zero. Thus, it is obvious that the times of the peak arrival rate do not correspond to the peak queue length and times of the minimum arrival rate do not correspond to the smallest queue length. Moreover, we observe a bit of symmetry in the case where the abandonment rate and the service rate are equal. In this case, the queue length in steady state is an ellipse. However, when the abandonment and the service rate are not equal, the dynamics are equal to an ellipse since there are different dynamics above and below the number of servers.

### 5.2 Delay Analysis

The dynamics of the Erlang-A queue is important, but it is also important to understand how much time a customer should expect to wait in line. To this end, we describe a new way to estimate the delay experienced by a random customer. Given a time \(\tau\), let \(\{Q_\tau(t)|t \geq 0\}\) be an \(M(t)/M/c + M\) queueing process where the only change to the original model is that the arrival process after time \(\tau\) has been turned off. More precisely, our Poisson arrival rate process \(\{\lambda_\tau(t)|t \geq \tau\}\) equals

\[
\lambda_\tau(t) \equiv \lambda(t) \cdot \{t \leq \tau\}
\]
where the second factor denotes an indicator function that equals 1 if $t \leq \tau$ and 0 otherwise. The sample path construction of this process reduces to

$$Q_{\tau}(t) = Q_{\tau}(\tau) - II \left( \int_{\tau}^{t} \delta \left( Q_{\tau}(s), c \right) \, ds \right). \quad (5.107)$$

The process $Q_{\tau}(t)$ has decreasing sample paths and so we can define the virtual delay process to be

$$D_{\tau}[c] \equiv \min \{ t \mid Q_{\tau}[c](t) < c \} - \tau \quad (5.108)$$

Assuming the FCFS discipline, $D_{\tau}[c]$ is the time that an incoming customer arriving at time $\tau$ has to wait to get service from an agent. Note that for all $t \geq \tau$, we have

$$\{ Q_{\tau}(t) \geq c \} = \{ D_{\tau} > t - \tau \}. \quad (5.109)$$

Note that the indicator event for the delay is a strict inequality while on the left of Equation 5.109 it is not. There is a non-zero probability that a customer arriving a time $\tau$ experiences no delay at all. Hence the delay distribution will always have a point mass at zero.

By taking expectations on both sides of Equation 5.109 gives us the following probabilities

$$P \{ Q_{\tau}(t) \geq c \} = P \{ D_{\tau} > t - \tau \}. \quad (5.110)$$

Moreover, the uniformly accelerated virtual delay time process is defined as

$$D_{\tau}[\eta c] = \min \{ t \mid Q_{\tau}[\eta c](t) < \eta c \} - \tau = \min \left\{ t \mid \frac{Q_{\tau}[\eta c](t)}{\eta} < c \right\} - \tau, \quad (5.111)$$

and has the following limit

$$\lim_{\eta \to \infty} D_{\tau}[\eta c] = d_{\tau}, \quad (5.112)$$

where

$$d_{\tau} \equiv \min \{ t \mid q_{\tau}(t) \leq c \} - \tau. \quad (5.113)$$

Using the fluid limit for the queue length process we can approximate the delay experience by a customer when the number of customers in the queue is given by $q_{\tau}(\tau)$. This result is given by the following theorem.

**Theorem 5.1.** The mean delay can be approximated by the fluid limit and the following expression

$$d_{\tau} = \frac{1}{\beta} \log \left( 1 + \frac{\beta \cdot (q_{\tau}(\tau) - c)}{\mu \cdot c} \right). \quad (5.114)$$

**Proof.** If $q_{\tau}(\tau)$ is smaller than the number of servers $c$, then the delay time is equal to zero. This agrees with our formula given in Equation 5.114. Thus, it only remains to show the proof for the case when $q_{\tau}(\tau)$ is greater than the number of servers $c$. The queue length satisfies the following differential equation

$$\dot{q}_{\tau}(t) = -\mu \cdot c - \beta \cdot (q_{\tau}(t) - c) = (\beta - \mu) \cdot c - \beta \cdot q_{\tau}(t). \quad (5.115)$$
Since the differential equation is linear, we can solve the differential equation explicitly. The solution to the differential equation is given by

\[ q_\tau(t) = \left( \frac{\beta - \mu}{\beta} \right) \cdot c \cdot (1 - e^{-\beta \cdot (t-\tau)}) + q_\tau(\tau) \cdot e^{-\beta \cdot (t-\tau)}. \] (5.116)

Now we set the solution of the differential equation equal to \( c \), where the customer has the opportunity to receive service

\[ c = \left( \frac{\beta - \mu}{\beta} \right) \cdot c \cdot (1 - e^{-\beta \cdot d_\tau}) + q_\tau(\tau) \cdot e^{-\beta \cdot d_\tau}. \] (5.117)

Solving for \( d_\tau \), we finally have the following solution to the delay time

\[ d_\tau = \frac{1}{\beta} \log \left( 1 + \frac{\beta \cdot (q_\tau(\tau) - c)}{\mu \cdot c} \right), \] (5.118)

which completes the proof.

Similar to the fluid delay approximation, we can also make a similar approximation for the GVA and GSA methods. In the case of GVA we have that the delay time can be approximated with

\[ d_{GVA}^\tau \equiv \min \{ t \mid q_{GVA}^\tau(t) \leq c \} - \tau \] (5.119)

and for GSA, we have that

\[ d_{GSA}^\tau \equiv \min \{ t \mid q_{GSA}^\tau(t) \leq c \} - \tau. \] (5.120)

Our mean fluid delay approximation is useful because it is exact and can be compared with many other estimators of the mean delay quite easily. Furthermore, since we have illustrated via numerical examples that GVA and GSA are better approximations for the mean dynamics of the mean and variance, one would expect that a delay approximation based on the GVA and GSA methods should outperform the fluid approximation. In Table 1, we see that GVA and GSA are better estimators of the mean delay time. In fact we see in Table 1 that the fluid delay approximation consistently underestimates the mean delay time. This is consistent with the fact that in this example, the fluid mean underestimates the true mean queue length process. However, in Table 2, we see that the fluid delay approximation is doing much better than in Table 1 since the demand is 10 times as much and the number of servers is also increased ten fold. The example in Table 2 is much closer to the limiting dynamics, so we expect this improvement. Thus, when the demand is high enough we can use the fluid delay time approximation and have confidence that it is close to the true mean delay time.

5.3 Computing the Probability of Delay with Diffusion Limits, GVA and GSA

In this section, we explore the additional usefulness of using the GVA and GSA methods. In addition to estimating the mean and variance of the Erlang-A queue with good accuracy,
Table 1: Delay Analysis Comparisons
\[ \lambda(t) = 10 + 2 \cdot \sin t, \mu = 1, \beta = 0.5, Q(0) = 0, c = 10. \]

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>Fluid Approx</th>
<th>GVA</th>
<th>GSA</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>0</td>
<td>0.0590</td>
<td>0.0620</td>
<td>0.1298</td>
</tr>
<tr>
<td>8</td>
<td>0.1336</td>
<td>0.2140</td>
<td>0.2140</td>
<td>0.2102</td>
</tr>
<tr>
<td>9</td>
<td>0.1910</td>
<td>0.2680</td>
<td>0.2670</td>
<td>0.2456</td>
</tr>
<tr>
<td>10</td>
<td>0.1013</td>
<td>0.1790</td>
<td>0.1770</td>
<td>0.1956</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0.0180</td>
<td>0.0170</td>
<td>0.1252</td>
</tr>
</tbody>
</table>

Table 2: Delay Analysis Comparisons
\[ \lambda(t) = 100 + 20 \cdot \sin t, \mu = 1, \beta = 0.5, Q(0) = 0, c = 100. \]

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>Fluid Approx</th>
<th>GVA</th>
<th>GSA</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0.0010</td>
<td>0.0426</td>
</tr>
<tr>
<td>8</td>
<td>0.1336</td>
<td>0.1480</td>
<td>0.1480</td>
<td>0.1377</td>
</tr>
<tr>
<td>9</td>
<td>0.1910</td>
<td>0.2020</td>
<td>0.2020</td>
<td>0.1938</td>
</tr>
<tr>
<td>10</td>
<td>0.1013</td>
<td>0.1120</td>
<td>0.1120</td>
<td>0.1133</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0248</td>
</tr>
</tbody>
</table>

we can also use the GVA and GSA methods to approximate the probability of delay for the queueing process. We can approximate the probability of delay using the fluid and diffusion limits, GVA or GSA for the Erlang-A queue. The probability of delay is defined as the probability that the queue length exceeds or is equal to the number of servers i.e

\[ P\{D_t > 0\} = P\{Q(t) \geq c\}. \] (5.121)

For the fluid and diffusion limits and GVA, we can approximate the probability of delay with the Gaussian tail cdf function. This implies that

\[ P\{Q_{FD}(t) \geq c\} \approx P\{G \geq \chi_{FD}\} = \Phi(\chi_{FD}) \] (5.122)

and

\[ P\{Q_{GVA}(t) \geq c\} \approx P\{G \geq \chi_{GVA}\} = \Phi(\chi_{GVA}). \] (5.123)

However, for GSA, we have that

\[ P\{Q_{GSA}(t) \geq c\} \approx P\{H_\theta \geq \chi_{GSA}\} = \Psi(\chi_{GSA}). \] (5.124)

Thus, by using the Gaussian tail cdf function, we can approximate the probability of delay for each node of the network. Moreover, we can show that by inverting our probability of delay approximations, we can construct delay stabilizing staffing schedules for the queue length process.
5.4 Stabilizing the Probability of Delay

The closed form expression for the probability of delay for each node of the network is an important object to exploit for constructing a staffing algorithm that stabilizes the probability of delay. This idea of stabilizing performance measures was introduced by Ward Whitt and his colleagues in the work of Jennings et al. [15] where they produced stable staffing schedules for the multi-server queue with general service and arrival distributions. A follow-up paper by Ward Whitt and his collaborators Feldman et al. [7] constructed a simulation staffing algorithm to stabilize the probability of delay for arbitrary queueing networks. More work on the topic has been also pursued by Ward Whitt, his students and collaborators in Liu and Whitt [21, 22, 23], Li et al. [20], Pender [37], He et al. [12], Liu and Whitt [24], Whitt and Zhao [50], Pender and Massey [40].

Recall that in the case of the diffusion limits and GVA the probability of delay approximation uses the tail cdf of the Gaussian distribution. Thus, using the properties of the cdf, one can invert the tail cdf to get
\[
P\{Q_{FD}(t) \geq c\} = \Phi(\chi_{FD}) = \epsilon \quad \text{(5.125)}
\]
\[
\Phi^{-1}(\Phi(\chi_{FD})) = \Phi^{-1}(\epsilon) \quad \text{(5.126)}
\]
\[
\chi_{FD} = \Phi^{-1}(\epsilon) \quad \text{(5.127)}
\]
\[
c_{FD} = q_{FD} + \sqrt{v_{FD}} \cdot \Phi^{-1}(\epsilon). \quad \text{(5.128)}
\]

Thus, by inverting the probability of delay approximation using the GVA method we see that we should use the following staffing function
\[
c_{GVA} = \left[ q_{GVA} + \sqrt{v_{GVA}} \cdot \Phi^{-1}(\epsilon) \right]. \quad \text{(5.129)}
\]
in order to stabilize the queueing process with approximately $\epsilon$ probability of delay. For GSA, we have the number of servers to stabilize the delay probability at a level of $\epsilon$ is
\[
c_{GSA} = \left[ q_{GSA} + \sqrt{v_{GSA}} \cdot \Phi^{-1}(\theta(\epsilon)) \right]. \quad \text{(5.130)}
\]

Although the staffing procedures appear to be the same, it is the case that the actual staffing functions are different because the dynamics for the mean and variance of the GVA and GSA methods are better estimates than the fluid and diffusion limits of the true mean and variance dynamics. Since GVA and GSA do a better job of estimating the mean and variance dynamics than the fluid and diffusion limits, we expect them to produce better staffing schedules for stabilizing the delay probabilities at their target values.

Our method of stabilization is deterministic and, thus is very powerful since it does not require the use of simulation, which is computationally expensive. Unlike Jennings et al. [15], Feldman et al. [7] there is no need to actually simulate the queueing system in order to update the staffing schedule. Our method simply requires the numerical solution of 2 odes for GVA and 3 odes for GSA, which is computationally fast and does not require much computational effort especially in the large scale setting or when there are a large number of servers. To demonstrate the effectiveness of our algorithms, we plot in Figures 7 - 9 the delay probabilities produced by our algorithms on the left and the number of servers on the
right. We see that all three algorithms do a good job of producing stable delay probabilities. However, we see that GSA is the best at producing the most stable delay probabilities for each of the target values, especially near the value $\epsilon = .5$. This value should produce the median queue length and one can observe in Figure 9 that the GSA does a better job of stabilization near $\epsilon = .5$ (the magenta colored curve). $\epsilon = .5$ is also an important value since this is when the fluid limit is near the number of servers and it is well known that the queue length is non-Gaussian near this region. Thus, GSA is able to model this value quite well despite being non-Gaussian.

Figure 7: Stabilizing the Probability of Delay with (FD) Target Levels:
$\epsilon = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$
Stabilized Delay Probabilities (Left) The Number of Servers (Right),
$\lambda(t) = 100 + 20 \cdot \sin t$, $\mu = 1$, $\beta = 0.5$, $Q(0) = 100$.

6 Application of the Erlang-A to Healthcare

In this section, we illustrate that the Erlang-A queueing model is important for emerging healthcare applications. In the spirit of Jennings et al. [14], McCalla and Whitt [30], Hampshire and Massey [10], we develop a novel staffing algorithm for static planning for a nursing home during a time interval $[0,T]$. This is in contrast to the work of Niyirora and Pender [32], Qin and Pender [43], where the optimal staffing algorithms were dynamic and varied over time. In the context of capacity planning in a nursing home, the time scale is larger and it is not practical to change the number of beds or rooms dynamically with time. Thus, we exploit dynamic optimization to construct a static optimal algorithm for determining the number of beds needed in a nursing home for achieving profitability.
Figure 8: Stabilizing the Probability of Delay with (GVA) Target Levels: 
\( \epsilon = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\} \)
Stabilized Delay Probabilities (Left) The Number of Servers (Right), 
\( \lambda(t) = 100 + 20 \cdot \sin t, \mu = 1, \beta = 0.5, Q(0) = 100. \)

6.1 Lagrangian Algorithm to Find the Optimal Number of Beds

The fluid model for the mean behavior of Erlang-A model is defined by the dynamical system 
\( \{q(t) \mid t \geq 0\} \)
\[ \dot{q} = \lambda - \mu \cdot (q \wedge c) - \beta \cdot (q - c)^+ \]  
(6.131)
where \( q \) and \( \lambda \) are implicitly time dependent. We use \( q[c] \) with square and not circular brackets to stress the dependency of \( q \) on the constant number of beds \( c \). Notice that when \( c \) equals zero or infinity, we have
\[ \dot{q}[0] = \lambda - \beta \cdot q[0] \quad \text{and} \quad \dot{q}[\infty] = \lambda - \mu \cdot q[\infty]. \]  
(6.132)

Now let \( r \) equal the revenue obtained from each successfully served customer, i.e. one that departs as a service completion and not as a customer abandonment. If \( w \) equals the cost rate of each agent, then we define \( P[c] \) to equal the mean nursing home profit over the time interval \((0, T]\) or
\[ P_T[c] \equiv \int_0^T (r\mu \cdot (q \wedge c) - wc) \, dt. \]  
(6.133)
Since \( P_T[0] = 0 \) and \( P_T[c] < 0 \) as \( c \) becomes sufficiently large, then there exists some number of beds \( c_* \) such that
\[ P_T[c_*] = \max_{c \geq 0} P_T[c]. \]  
(6.134)

The way that we construct an algorithm to find \( c_* \) is to superimpose a Lagrangian structure onto the profit function for fixed \( c \). Observe that
\[ P_T[c] = \max_{q, \dot{q} = \lambda - \mu \cdot (q \wedge c) - \beta \cdot (q - c)^+} \int_0^T (r\mu \cdot (q \wedge c) - wc) \, dt. \]  
(6.135)
We can rewrite this as a constrained optimization problem, i.e.

\[ P_T[c] = \max_{p,q: q = \lambda - \mu \cdot (q \wedge c) - \beta \cdot (q - c)^+} \int_0^T L(c,p,q,\dot{q}) \, dt, \]  

(6.136)

where our Lagrangian is

\[ L(c,p,q,\dot{q}) = r \mu \cdot (q \wedge c) - w \cdot c + p \cdot (\dot{q} - \lambda + \mu \cdot (q \wedge c) + \beta \cdot (q - c)^+) \]  

(6.137)

The resulting Euler-Lagrange equations are

\[ \frac{d}{dt} \frac{\partial L}{\partial \dot{q}}(c,p,q,\dot{q}) = \frac{\partial L}{\partial p}(c,p,q,\dot{q}) \quad \Leftrightarrow \quad \dot{p} = \mu \cdot (p + r) \cdot \{q < c\} + \beta \cdot p \cdot \{q \geq c\} \]

\[ \frac{\partial L}{\partial p}(c,p,q,\dot{q}) = 0 \quad \Leftrightarrow \quad \dot{q} = \lambda - \mu \cdot (q \wedge c) - \beta \cdot (q - c)^+. \]

We can now rewrite \( P_T[c] \) as

\[ P_T[c] = \int_0^T L(p,q,\dot{q})[c] \, dt, \]  

(6.138)

where \( L(p,q,\dot{q})[c] \equiv L(c,p[c],q[c],\dot{q}[c]) \) and the Lagrange multiplier \( p[c] \) solves the dynamical system

\[ \dot{p} = \mu \cdot (p + r) \cdot \{q < c\} + \beta \cdot p \cdot \{q \geq c\}. \]  

(6.139)
Theorem 6.1. If $p$ is the opportunity cost process for the Erlang-A fluid profit function, then we have

$$-r < p(t) \leq 0$$  \hspace{1cm} (6.140)

for all $0 \leq t \leq T$, with $p(T) = 0$. Moreover, if $p(t) = 0$ for some $0 \leq t < T$ then $p(s) = 0$ and $q(s) \geq c$ for all $t \leq s \leq T$.

Proof. If $p(t) \leq -r$ for some $0 \leq t \leq T$, then $p(t) \leq 0$. This would force $p$ to be a decreasing function on the interval $[0,T]$ and bounded above by $-r$. Since this would rule out the possibility of $p(T) = 0$, then the premise cannot be true.

By similar reasoning, we cannot have $p(t) > 0$ for some $0 \leq t \leq T$. This is also why $p(t) = 0$ for such a $t$ holds only if $p = 0$ and $q \geq c$ on the entire interval $[t,T]$. This completes the proof.

Given some real-valued function $f$ on $(0,T]$, we define the decreasing rearrangement of $f$ to be the unique right continuous, decreasing function $f^\downarrow$ on $(0,T]$ where

$$\int_0^T \{ f(t) > x \} \, dt = \int_0^T \{ f^\downarrow(t) > x \} \, dt$$  \hspace{1cm} (6.141)

for all real values $x$. For all decreasing functions $g$, we define its generalized inverse to be $g^{-1}$ where

$$g^{-1}(x) \equiv \inf \{ y \mid g(y) \geq x \}.$$  \hspace{1cm} (6.142)

For all $x$ less than some number in the range of $f$, we have

$$\int_0^T \{ f(t) > x \} \, dt = (f^\downarrow)^{-1}(x).$$  \hspace{1cm} (6.143)

Theorem 6.2. Assume that $w < r\mu$, $q^\downarrow[c]$ is the decreasing rearrangement of $q$ over the time interval $(0,T]$, and

$$\bar{c} \equiv q^\downarrow[c] \left( \frac{w}{r\mu} T + \frac{\beta - \mu}{r\mu} \int_0^{q^\downarrow[c]-1(c)} p^\uparrow[c] \, dt \right),$$  \hspace{1cm} (6.144)

where $p^\uparrow[c]$ is defined to be the rearrangement of $p$ that transforms $q$ into its decreasing rearrangement $q^\downarrow$. We then have

$$P_T^\uparrow[c] \geq 0 \Rightarrow c \leq \bar{c} \text{ and } c \leq q^\downarrow[c] \left( \frac{w}{r\beta} T \right).$$  \hspace{1cm} (6.145)

Similarly, we also have

$$P_T^\downarrow[c] \leq 0 \Rightarrow c \geq \bar{c} \geq q^\downarrow[c] \left( \frac{w}{r\mu} T \right).$$  \hspace{1cm} (6.146)
Proof. Using the sensitivity results for Lagrangian optimality, we have

\[
\mathcal{P}_T'[c] = \frac{d}{dc} \int_0^T \mathcal{L} \left( p, q, \dot{q} \right) [c] \, dt
\]

\[
= \int_0^T \frac{\partial \mathcal{L}}{\partial c} \left( c, p[c], q[c], \dot{q}[c] \right) [c] \, dt
\]

\[
= \int_0^T r\mu \cdot \{ q[c] > c \} - w + p \cdot (\mu \cdot \{ q[c] > c \} - \beta \cdot \{ q[c] > c \} ) \, dt
\]

\[
= r\mu \cdot \int_0^T \{ q[c] > c \} \, dt - \left( wT + (\beta - \mu) \cdot \int_0^T p[c] \cdot \{ q[c] > c \} \, dt \right)
\]

\[
= r\mu \cdot \int_0^T \{ q^+[c] > c \} \, dt - \left( wT + (\beta - \mu) \cdot \int_0^T p[c] \cdot \{ q^+[c] > c \} \, dt \right)
\]

\[
= r\mu \cdot q^+[c]^{-1}(c) - \left( wT + (\beta - \mu) \cdot \int_0^{q^+[c]^{-1}(c)} p^+[c] \, dt \right).
\]

This completes the proof. \(\square\)

Corollary 6.3. For any Erlang-A system, the number of servers \(c_*\) needed for profit optimality has the following properties:

1. We have \(c_* = \tau_*\) or

\[
c_* = q^+[c_*] \left( \frac{w}{r\mu} T + \frac{\beta - \mu}{r\mu} \int_0^{q^+[c_*]^{-1}(c_*)} p^+[c_*] \, dt \right).
\] (6.147)

2. We have the following upper and lower bounds for \(c_*\),

\[
q^+[c_*] \left( \frac{w}{r\mu} T \right) \leq c_* \leq q^+[c_*] \left( \frac{w}{r\beta} T \right).
\] (6.148)

3. Given our rearrangement \(p^+[c_*]\) of \(p[c_*]\), there exists a unique \(\tau\) that solves the equation

\[
\tau = \frac{w}{r\mu} T + \frac{\beta - \mu}{r\mu} \int_0^\tau p^+[c_*] \, dt.
\] (6.149)

We then have \(c_* = q^+[c_*](\tau)\).

Proof. Part 1 and Part 2 are obvious from the previous theorem. However, for part 3, we have

\[
0 \leq \frac{w}{r\mu} \cdot T + \frac{\beta - \mu}{r\mu} \cdot \int_0^0 p^+ \, dt = \frac{w}{r\mu} \cdot T.
\] (6.150)

but since \(-r < p[c_*] \leq 0\) on \((0, T]\), we have

\[
T \geq \frac{w}{r\mu} \cdot T + \frac{\beta - \mu}{r\mu} \cdot \int_0^T p^+ \, dt \geq \frac{w}{r\mu} \cdot T + \frac{\mu - \beta}{r\mu} \cdot rT = \left( 1 + \frac{w - \beta r}{r\mu} \right) T \geq T.
\] (6.151)

This gives us the unique fixed point value \(\tau\) where \(0 \leq \tau \leq T\). This completes the proof. \(\square\)

32
6.2 Opportunity Costs (Rebates)

We can write the differential equation for the dual variable \( p \) as

\[
\dot{p} = \begin{cases} 
\mu \cdot (p + r) & \text{when } q < c, \\
\beta \cdot p & \text{when } q \geq c.
\end{cases}
\]

This gives our dual variable a piecewise log linearity property over time as we now state.

\textbf{Theorem 6.4.} The following two statements hold:

1. When \( q \geq c \) over any open subinterval of \((0, T]\), the process \( \log(-p) \) is a linear function with slope \( \beta \).

2. When \( q < c \) over any closed subinterval of \((0, T]\), the process \( \log(p + r) \) is a linear function with slope \( \mu \).

This means that we can easily compute \( p \) over an interval \([0, T]\) provided we know the times that \( q \) equals \( c \) and which intervals of these time points correspond to \( q \) being over or below \( c \). In fact Figure 10 illustrates how to transform \( p \) logarithmically into a piecewise linear plot.

7 Conclusion and Final Remarks

In this paper, we prove and review several important results for the dynamic rate Erlang-A queue. Although we provide many new results, there are many areas that are still ripe for new research. One interesting area of research would be to explore the possibilities for extending and replicating our closure approximations and cumulant moment results for non-Markovian queueing systems. For example, the many-server limit theorems in the spirit of
Halfin and Whitt have been extended to phase-type and general distributions in the work of Ko and Pender [19], Pender and Ko [39], Ko and Pender [18] and it would be great to see if the closure approximations work well for these non-Markovian systems. Moreover, there is new research that explores the new aspect of adding delayed information to drive the arrival processes of queueing networks, see for example Pender et al. [42]. It would be interesting to extend this work to the dynamic rate Erlang-A queueing model and generalizations of it. We plan to pursue some of this work in the future.

References


[40] Jamol Pender and William A Massey. Approximating and stabilizing dynamic rate
jackson networks with abandonment. *Probability in the Engineering and Informational

[41] Jamol Pender and Tuan Phung-Duc. A law of large numbers for m/m/c/delayoff-
setup queues with nonstationary arrivals. In *International Conference on Analytical

[42] Jamol Pender, Richard H Rand, and Elizabeth Wesson. Queues with choice via delay
2017.

2017.


[45] Justus Arne Schwarz, Gregor Selinka, and Raik Stolletz. Performance analysis of time-


[47] Michael R Taaffe and Gordon M Clark. Approximating nonstationary two-priority non-


[49] Michael R Taaffe and Kim L Ong. Approximating nonstationaryph (t)/m (t)/s/c queue-

[50] Ward Whitt and Jingtong Zhao. Many-server loss models with non-poisson time-varying
arrivals. *Naval Research Logistics (NRL), 2017.*