Poisson and Gaussian Approximations for Multi-Server Queues with Batch Arrivals and Batch Abandonment

Jamol Pender
School of Operations Research and Information Engineering
Cornell University
jjp274@cornell.edu

October 11, 2013

Abstract

In this paper, we consider nonstationary multi-server queues and multiserver loss queues with batch arrivals and where we allow customer abandonment to also occur as a Markovian batch process. To estimate the performance of the system, we develop two new approximations that take advantage of the functional Kolmogorov forward equations. The first approximation we develop uses the Poisson distribution, which is motivated from the infinite server queue offered load process. The second approximation uses a Gaussian distribution, which is motivated by the many server heavy traffic limit theorems of Mandelbaum et al [13]. We compare our results with Monte Carlo simulation and show that our approximations provide accurate estimates of the mean, variance, probability of delay, and the probability of blocking for the queueing process.

Keywords: Multi-Server Queues, Abandonment, Time-Varying Rates, Time Inhomogeneous Markov Processes, Gaussian, Poisson, Incomplete Gamma Function, Batch Arrivals, Batch Abandonment, Nonstationary Loss Queues.

1 Introduction

In this paper, we consider time varying multi-server queues and nonstationary loss queues with batch arrivals and where we allow customer abandonment to also occur as a Markovian batch process. We denote these queueing processes as the $M_t^X/M_t/c_t + M_t^X$ and $M_t^X/M_t/c_t/k_t + M_t^X$ for the multiserver and loss queue respectively. These types of queueing models have many applications in practice. For instance, in a healthcare setting, when accidents or emergency situations such as terrorist attacks, environmental disasters, or mass casualty events involve multiple people, it is often that the parties involved arrive to the hospital in batches, see for example Cohen et al [5]. Moreover, these people can also be rejected and rerouted to neighboring hospitals in batches if they exceed the hospital’s capacity.
A queueing model with batch arrivals also has several applications in communication networks. Often data packets often arrive in batches since they are often connected to one file, see for example Chan et al [1], Chaudhry et al [2, 3, 4], and Yao et al [23]. Most of these models consider stationary arrival rates and analyze the queueing dynamics in discrete-time. Queues with batch arrivals and time varying rates are much harder to study than single arrival queues with stationary rates. Thus, it is not surprising that the literature on the topic is quite scarce. A recent paper of Pang et al [18], considers the dynamics of the time varying queues with batch arrivals and dependent service times. Although they consider non-Markovian dynamics for the arrival and service processes, they do not consider the finite server setting, which is more realistic. Hence, it is still and open problem to study nonstationary and batch arrival systems with a finite number of servers and develop tractable approximations to estimate their dynamic behavior since this behavior is seen in real-world systems.

In addition, to the batch arrivals, our queueing model also includes a novel feature that we call batch abandonment. In a communications setting where many data packets arrive as a batch as seen in [2], it is also conceivable that jointly arriving data packets could also have joint processing deadlines as a batch, which would allow them to abandon the queue as a batch. More recently, many authors such as Mandelbaum and Zeltyn [24], Mandelbaum et al [13], Liu and Whitt [12] have analyzed the effects of customer abandonment and have found that abandonment affects the behavior of the queueing process substantially. Thus, it is natural to ask the question of how abandonment affects the behavior of the queueing process when abandonments can occur as a batch process.

As the model that we consider is Markovian, we exploit the functional Kolmogorov forward equations for the queueing process. Like Taaffe et al [22] or Rothkopf et al [17], we encounter difficulties from the forward equations since they are not a closed system unless the queue is an infinite server, a no-server queue, or the mean service rate is equal to the mean abandonment rate. Thus, it is necessary to prescribe an appropriate surrogate distribution for the queue length process in order to close the forward equations, which is non-trivial for systems with abandonment.

To overcome the fact that the forward equations are not closed, recently authors Massey and Pender [14] and Pender [19, 20] have developed novel approximation methods for time varying queueing processes using orthogonal polynomial expansions for the queue length process. For the multiserver case, [14] project the queue length process on a finite number of Hermite polynomial basis elements to close the forward equations for the mean, variance, and third cumulant moment. Furthermore, in the single server setting [20] projects the queue length process onto Laguerre polynomial basis elements since the heavy traffic density of the single server queue can be approximated by a exponential distribution. Moreover, the approximation methods of [19, 20] are shown to be quite accurate in a variety of parameter settings.

In the same spirit, we develop two new approximations that exploit the functional Kolmogorov forward equations of the queue length process. The first approximation uses a Poisson distribution, which is motivated by the time varying dynamics of the infinite server queue offered load process and reflects the discrete nature of the queueing process. We use the Poisson distribution to derive estimates for the mean and the variance of the queueing process as well as the probability of delay. These estimates are derived in terms of the tail
cdf of the Poisson distribution, which is intimately related to incomplete gamma functions. Our second approximation uses a Gaussian approximation, which is more motivated by the many server heavy traffic limit theorems of Mandelbaum et al [13]. This approximation is not discrete, but we will show that it is nonetheless quite accurate at approximating the dynamics of the queue length process.

1.1 Contributions

In this work we make several key contributions to the literature of time varying queues with abandonment. Our first contribution is the development of a queueing model that includes both batch arrivals and batch abandonments, which serves to approximate the dynamics of many healthcare systems, emergency situations, and communication networks. To the best of our knowledge we believe that we are the first to analyze such a system with stationary or nonstationary parameters. Our second fundamental contribution is to identify that the nonstationary Poisson distribution can be used as a surrogate distribution to close the functional Kolmogorov forward equations. To the best of our knowledge, the Poisson approximation has not been considered in the literature and is new. Moreover, we also show that a Gaussian distribution can be quite useful to close the functional forward equations and develop accurate approximations for the mean and variance of the queueing process. Lastly, our approximation methods are very simple to implement and only involve the numerical integration of at most two differential equations, which is quite fast to solve numerically.

1.2 Organization of the Paper

The rest of the paper is organized as follows. We start with a description of the queueing model under consideration in Section 2. We also derive the functional forward equations for the model, which are important for the future analysis. We analyze a new Poisson approximation for the queueing model in Section 3. In Section 4, we develop a Gaussian approximation for the queueing model. Moreover, we give more numerical examples to support the use of our approximation methods. In Section 5, we show that our analysis is quite broad and apply it to a nonstationary loss queue with batch arrivals and batch abandonment. Lastly, concluding remarks are given in Section 6.

2 Stochastic Analysis of the Queueing Model

In this section, we give a overview of the queueing model that we will analyze in this paper. Using Kendall notation, the queueing model that we consider is the $M_t^X/M_t/c_t + M_t^X$ queue. Batches of customers arrive according to Poisson processes with rate $\lambda_j(t)$ for the batch of size $j$. We assume that the number of customers in a batch is a bounded positive integer-valued random variable with upper bound $n$. Moreover, the service distribution is exponential with rate $\mu(t)$. We also have $c(t)$ parallel and homogeneous servers. Lastly, batches of customers are allowed to abandon the queue if their time spent waiting is considered to be excessive. We also assume that the batch sizes of the abandonment process are also upper bounded by $n$, however, this assumption can be easily generalized to another bounds. The customers
abandon the queue according to Poisson processes with rate $\beta_j(t)$ for the batch of size $j$ if there are at least $j$ customers waiting to initiate service.

If we did not consider batch abandonment, then our queueing model would be in the class of queueing models called Markovian service networks that were analyzed in Mandelbaum et al [13]. However, the indicator function for the batch abandonment process is not a Lipschitz continuous function of the queue length and precludes us from applying the fluid and diffusion limit theorems of [13] to gain insight about the mean and variance of the queueing process. Thus, we must develop new ways of analyzing the queueing process, which can deal with the discontinuous behavior of the indicator function and other nonlinearities of the queueing process. To develop these new methods, we will need a thorough understanding of the functional Kolmogorov forward equations of the queue length process. The functional Kolmogorov forward equations for the $M^X_t/M_t/c_t + M^X_t$ have the following form

$$
\dot{E}[f(Q)] = \sum_{j=1}^{n} \lambda_j \cdot E[f(Q) + j] - \mu \cdot E[(f(Q) - f(Q)) \cdot (Q \wedge c)] \\
+ \sum_{j=1}^{n} \beta_j \cdot E[(f(Q) - f(Q)) \cdot (Q - c)^+ \cdot \{Q > c + j - 1\}], \quad (2.1)
$$

for all appropriate real-valued functions $f$ where the expectation is well defined. Although we consider the cases of when the arrival rate, service rate, and abandonment rates are functions of time, we suppress the time dependence to simplify the notation. Moreover, we use the ”dot” notation of physics to denote its time derivative i.e.

$$
\dot{E}[f(Q)] \equiv \frac{d}{dt} E[f(Q(t))]. \quad (2.2)
$$

For special cases such as the mean and variance of the queue length process, the functional Kolmogorov forward equations have the following expressions

$$
\dot{E}[Q] = \sum_{j=1}^{n} \lambda_j \cdot j - \mu \cdot E[(Q \wedge c)] - \sum_{j=1}^{n} \beta_j \cdot j \cdot E[(Q - c)^+ \cdot \{Q > c + j - 1\}] \\
\dot{\text{Var}[Q]} = \sum_{j=1}^{n} \lambda_j \cdot j^2 + \mu \cdot E[(Q \wedge c)] + \sum_{j=1}^{n} \beta_j \cdot j^2 \cdot E[(Q - c)^+ \cdot \{Q > c + j - 1\}] \\
- 2 \cdot \mu \cdot \text{Cov}[Q, (Q \wedge c)] - 2 \sum_{j=1}^{n} \beta_j \cdot j \cdot \text{Cov}[Q, (Q - c)^+ \cdot \{Q > c + j - 1\}].
$$

Although it may seem that we have a simple way to evaluate the mean and variance of the functional forward equations, it is important to notice that the right hand side of the functional equations depend nonlinearly on queue length. This implies that we need to know the distribution of the queue length in order to compute the mean and variance of the queue length process. For a further explanation of this, see for example Pender [20] where it is shown that the max and min functions depend on higher moments of the queue
length process. However, since we do not know the true distribution of the queueing process except in some special cases, we must develop methods that can accurately approximate the true distribution of the queue length process. Our first approximation, which exploits the dynamics of the infinite server queue is the subject of the next section.

3 Poisson Approximation

The first approximation method that we provide is called the Poisson Infinite Server Approximation (PISA), which exploits the results of Eick et al [6] for the time varying infinite server queue. In the paper of [6], they use the properties of the Poisson arrival process and use Poisson random measure arguments to show that the $M_t/G/\infty$ queue $Q^\infty(t)$, has a Poisson distribution with time varying rate $q^\infty(t)$. The exact analysis of the infinite server queue is often useful since it represents the dynamics of the queueing process if there were an unlimited amount of resources to satisfy the demand process. As observed in [6], $q^\infty(t)$ has the following representation

$$q^\infty(t) \approx E[Q^\infty(t)]$$

$$= \int_{-\infty}^{t} G(t-u)\lambda(u)du$$

$$= E\left[\int_{t-S}^{t} \lambda(u)du\right]$$

$$= E[\lambda(t-S_e)] \cdot E[S]$$

where $S$ represents a service time with distribution $G$, $\overline{G} = 1 - G(t) = P(S > t)$, and $S_e$ is a random variable with distribution that follows the stationary excess of residual-lifetime cdf $G_e$, defined by

$$G_e(t) \equiv P(S_e < t) = \frac{1}{E[S]} \int_{0}^{t} \overline{G}(u)du, \quad t \geq 0.$$  

Using the knowledge that the queue length in the infinite server setting has a Poisson distribution gives us the idea that it might be appropriate for us to assume that our queueing process with batch arrivals and batch abandonment might be close to the Poisson distribution, but with a different mean. Thus, the infinite server analysis motivates our approximation of the queue length as

$$Q(t) \approx \text{Poisson}(q(t)).$$

for all $t \geq 0$, where $\{q(t)|t \geq 0\}$ is some one-dimensional dynamical system. This Poisson approximation for the queue length process leads us to our first theorem.

**Theorem 3.1.** The functional forward equations for the mean of the queue length process when the queue length has the distribution given by 3.8 has the form

$$\dot{E}[Q] = \sum_{j=1}^{n} \lambda_j \cdot j - \mu \cdot E[(Q \wedge c)] - \sum_{j=1}^{n} \beta_j \cdot j \cdot E[(Q - c)^+ \cdot \{Q > c + j - 1\}].$$
Moreover, the rate functions of the functional forward equations have the following closed form expressions

\[
E [(Q \wedge c)] = \Gamma(c, q(t)) + c \cdot \Gamma(c + 1, q(t)) \tag{3.9}
\]
\[
E [(Q - c)^+ \cdot \{Q \geq c + j\}] = q \cdot \Gamma(c - 1 + j, q(t)) - c \cdot \Gamma(c + j, q(t)) \tag{3.10}
\]

and where

\[
\Gamma(c, x) = \frac{1}{\Gamma(c)} \int_0^x e^{-y} y^{c-1} dy \quad \text{and} \quad \Gamma(c, x) = \frac{1}{\Gamma(c)} \int_x^\infty e^{-y} y^{c-1} dy \tag{3.11}
\]

are the lower and upper incomplete gamma functions respectively.

**Proof.** To prove this result, we must show that under the Poisson(q(t)) assumption, that the rate functions of the mean satisfy the closed form expressions of Equations 3.9 and 3.10. Before we prove the results, we provide a lemma that shows that the tail distribution of a Poisson distribution can be expressed in terms of the incomplete gamma function.

**Lemma 3.2.**

\[
\Gamma(c, x) = \sum_{m=c}^{\infty} e^{-x} \cdot \frac{x^m}{m!} = \frac{1}{\Gamma(c)} \int_0^x e^{-y} y^{c-1} dy \tag{3.12}
\]
\[
\Gamma(c, x) = \sum_{m=0}^{c-1} e^{-x} \cdot \frac{x^m}{m!} = \frac{1}{\Gamma(c)} \int_x^\infty e^{-y} y^{c-1} dy. \tag{3.13}
\]

**Proof.** See Janssen et al [10].

Now to complete the proof of the theorem, we have for the service rate function that

\[
E [(Q \wedge c)] = E[Q] - E [(Q - c)^+] \\
= q - E [(Q - c) \cdot \{Q \geq c\}] \\
= q - E [Q \cdot \{Q > c\}] + c \cdot E [\{Q > c\}] \\
= q - \sum_{m=c+1}^{\infty} m \cdot e^{-q} \cdot \frac{q^m}{m!} + c \cdot \sum_{m=c+1}^{\infty} e^{-q} \cdot \frac{q^m}{m!} \\
= q - \sum_{m=c+1}^{\infty} m \cdot e^{-q} \cdot \frac{q^m}{m!} + c \cdot \Gamma(c + 1, q(t)) \\
= q - q \cdot \sum_{m=c}^{\infty} e^{-q} \cdot \frac{q^m}{m!} + c \cdot \Gamma(c + 1, q(t)) \\
= q \cdot \Gamma(c, q(t)) + c \cdot \Gamma(c + 1, q(t)) \\
= q \cdot \Gamma(c, q(t)) + c \cdot \Gamma(c + 1, q(t)).
\]
Moreover, for the batch abandonment rate function, we have that

\[ E \left[ (Q - c)^+ \cdot \{Q > c + j - 1\} \right] = E \left[ (Q - c) \cdot \{Q \geq c\} \cdot \{Q > c + j - 1\} \right] \\
= E \left[ (Q - c) \cdot \{Q > c + j - 1\} \right] \\
= E \left[ Q \cdot \{Q > c + j - 1\} - c \cdot E \left[ \{Q > c + j - 1\} \right] \right] \\
= \sum_{m=c+j}^{\infty} m \cdot e^{-q} \frac{q^m}{m!} - c \cdot \sum_{m=c+j}^{\infty} e^{-q} \frac{q^m}{m!} \\
= q \cdot \Gamma(c + j - 1, q(t)) - c \cdot \Gamma(c + j, q(t)). \]

In order to visualize and assess the quality of our approximations methods, we simulate the queueing process (10,000 sample paths) with the parameters given in Table 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\lambda_1)</td>
<td>(50 + 20 \sin(t))</td>
</tr>
<tr>
<td>(\lambda_2)</td>
<td>(50/2 + 10 \sin(t))</td>
</tr>
<tr>
<td>(\lambda_3)</td>
<td>(50/3 + 20/3 \sin(t))</td>
</tr>
<tr>
<td>(\beta_1)</td>
<td>1</td>
</tr>
<tr>
<td>(\beta_2)</td>
<td>.5</td>
</tr>
<tr>
<td>(\beta_3)</td>
<td>1/3</td>
</tr>
<tr>
<td>(\mu)</td>
<td>1</td>
</tr>
<tr>
<td>(c)</td>
<td>200</td>
</tr>
<tr>
<td>(q(0))</td>
<td>0</td>
</tr>
</tbody>
</table>

On the left of Figure 1 we simulate the mean queue length and compare it to PISA mean approximation. We see that the PISA is good at estimating the dynamic behavior of the mean of the queue length process. On the right of Figure 1, we also simulate the variance of the queue and compare it to the variance of the PISA method. To compute the variance using the PISA method, we use the fact that the cumulant moments of the Poisson distribution are all determined by the mean. Thus, we are actually plotting the mean for the variance. The approximation for the variance is not great, but not terrible either. Moreover, on the left of Figure 2, we compare the skewness of the queueing process with the prediction from the Poisson distribution. We see that the Poisson assumption is decent at approximating the skewness of the queueing process, especially its sign since it is mostly positively skewed.

### 3.1 Probability of Delay (PISA)

In addition to providing estimates for the mean and variance using the Poisson approximation, we can also derive an approximation for the probability of delay. Using the PISA method, the probability of delay is
\[ P(Delay) = P(Q \geq c) \]  
\[ = \sum_{m=c}^{\infty} e^{-q} \cdot \frac{q^m}{m!} \]  
\[ = \Gamma(c, q). \]  

On the right of Figure 2, we compare the simulated probability of delay with our estimate from the PISA method. We see that the PISA method is quite accurate at estimating the probability of delay for our queueing model. Thus, we see that the PISA method seems to be quite useful in estimating queueing performance. However, the main drawback of the PISA method is that it is a one dimensional projection onto the rate functions for the
Poisson distribution. Perhaps a higher dimensional projection might provide more flexibility in approximating the performance measures. This is the subject of the next section.

4 Gaussian Approximation

In this section, we are not motivated by the infinite server queueing process as in the Poisson setting, but instead the many server heavy traffic limit theorems of [13]. We also would like to extend the one dimensional projection to the two dimensional setting. Thus, we now assume that the queueing process has a Gaussian distribution. This approximation, which was coined the Gaussian Variance Approximation (GVA) by [14] where it was shown that this method was equivalent to the Gaussian mollifier approach used by Ko and Gautam [11]. The Gaussian distribution assumption on the queue length is equivalent to

\[ Q(t) \overset{D}{=} q(t) + X \cdot \sqrt{v(t)}. \]  

(4.17)

for all \( t \geq 0 \), where \( \{q(t), v(t)\mid t \geq 0\} \) is some two-dimensional dynamical system where the \( v \) process is always positive and \( X \) is a standard Gaussian random variable. This approximation for the queue length process leads us to our second theorem.

Theorem 4.1. The functional forward equations for the mean and variance of the queue length process when the queue length has the distribution given by 4.17 have the form

\[
\begin{align*}
\dot{E}[Q] &= \sum_{j=1}^{n} \lambda_j \cdot j - \mu \cdot E[(Q \wedge c)] - \sum_{j=1}^{n} \beta_j \cdot j \cdot E[(Q - c)^+ \cdot \{Q > c + j - 1\}] \\
\dot{\text{Var}}[Q] &= \sum_{j=1}^{n} \lambda_j \cdot j^2 + \mu \cdot E[(Q \wedge c)] + \sum_{j=1}^{n} \beta_j \cdot j^2 \cdot E[(Q - c)^+ \cdot \{Q > c + j - 1\}] \\
& \quad - 2 \cdot \mu \cdot \text{Cov}[Q, (Q \wedge c)] - 2 \sum_{j=1}^{n} \beta_j \cdot j \cdot \text{Cov}[Q, (Q - c)^+ \cdot \{Q > c + j - 1\}].
\end{align*}
\]

Moreover, the rate functions have closed form expressions

\[
\begin{align*}
E[(Q - c)^+ \cdot \{Q \geq c + j\}] &= \sqrt{v} \cdot (\varphi(\psi_j) - \chi \cdot \Phi(\psi_j)) \\
E[(Q \wedge c)] &= q - \sqrt{v} \cdot (\varphi(\chi) - \chi \cdot \Phi(\chi)) \\
\text{Cov}[Q, (Q - c)^+ \cdot \{Q \geq c + j\}] &= v \cdot \psi_j \cdot \varphi(\psi_j) + v \cdot \Phi(\psi_j) - v \cdot \chi \cdot \varphi(\psi_j) \\
\text{Cov}[Q, (Q \wedge c)] &= v \cdot \Phi(\chi)
\end{align*}
\]

where we define \( \varphi \) and \( \Phi \) to be the Gaussian density and the cumulative distribution functions

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

(4.22)

and

\[
\chi = \frac{c - q}{\sqrt{v}} \quad \text{and} \quad \psi_j = \frac{c - q + j - 1}{\sqrt{v}}.
\]

(4.23)
Proof. Like in the Poisson case, we must make sure that the rate functions have the closed form expressions given in 4.18. In order to prove that the rate functions we provide a lemma known as Stein’s lemma [21], which is crucial to the proof.

Lemma 4.2 (Stein [21]). Suppose that $X$ is standard Gaussian random variable and $E[f'(X)] < \infty$, then
\[ E[X \cdot f(X)] = E[f'(X)]. \] (4.24)

Moreover,
\[ E[h_n(X) \cdot f(X)] = E[f^{(n)}(X)] \] (4.25)

where $h_n(X)$ is the $n^{th}$ Hermite polynomial.

Now the rate functions expressions for the mean dynamics are
\[
E \left[ (Q \wedge c) \right] = E[Q] - E[(Q - c)^+] \\
= E[Q] - \sqrt{v} \cdot E[(X - \chi)^+] \\
= q - \sqrt{v} \cdot E[(X - \chi) \cdot \{X \geq \chi\}] \\
= q - \sqrt{v} \cdot (\varphi(\chi) - \chi \cdot \Phi(\chi))
\]

and
\[
E \left[ (Q - c)^+ \cdot \{Q > c + j - 1\} \right] = \sqrt{v} \cdot E \left[ (X - \chi)^+ \cdot \{X \geq \psi_j\} \right] \\
= \sqrt{v} \cdot E \left[ (X - \chi) \cdot \{X \geq \chi\} \cdot \{X \geq \psi_j\} \right] \\
= \sqrt{v} \cdot E \left[ (X - \chi) \cdot \{X \geq \psi_j\} \right] \\
= \sqrt{v} \cdot (\varphi(\psi_j) - \chi \cdot \Phi(\psi_j)).
\]

Moreover, for the variance rate functions we have that
\[
Cov \left[ Q, (Q \wedge c) \right] = Cov \left[ Q, Q \right] - Cov \left[ Q, (Q - c)^+ \right] \\
= v - v \cdot Cov \left[ X, (X - \chi)^+ \right] \\
= v - v \cdot E \left[ X \cdot (X - \chi)^+ \right] \\
= v - v \cdot \Phi(\chi) \\
= v \cdot \Phi(\chi)
\]

\[
Cov \left[ Q, (Q - c)^+ \cdot \{Q > c + j - 1\} \right] = v \cdot Cov \left[ X, (X - \chi) \cdot \{X \geq \chi\} \cdot \{X \geq \psi_j\} \right] \\
= v \cdot Cov \left[ X, (X - \chi) \cdot \{X \geq \psi_j\} \right] \\
= v \cdot E \left[ X \cdot (X - \chi) \cdot \{X \geq \psi_j\} \right] \\
= v \cdot E \left[ X^2 \cdot \{X \geq \psi_j\} \right] - v \cdot \chi \cdot E \left[ X \cdot \{X \geq \psi_j\} \right] \\
= v \cdot E \left[ h_2(X) \cdot \{X \geq \psi_j\} \right] + v \cdot E \left[ \{X \geq \psi_j\} \right] \\
- v \cdot \chi \cdot E \left[ X \cdot \{X \geq \psi_j\} \right] \\
= v \cdot \psi_j \cdot \varphi(\psi_j) + v \cdot \Phi(\psi_j) - v \cdot \chi \cdot \varphi(\psi_j).
\]
We compare the GVA method with Monte Carlo simulation and use the same parameters of Table 1. On the left of Figure 3 we simulate the mean queue length and compare it to GVA mean approximation. We see that the GVA is good at estimating the mean dynamics. On the right of Figure 3, we also simulate the variance of the queueing system and compare it to the variance of the GVA method. In this case, we see that the Gaussian approximation is quite accurate at estimating the variance as well. Moreover, on the left of Figure 4, we compare the skewness of the queueing process with the prediction from the Gaussian distribution. Since the Gaussian is defined by all of its cumulant moments of order greater than 2 are equal to zero, it is obvious the Gaussian will not estimate the non-zero value of the skewness accurately.

![Figure 3: Comparison of Simulated and Gaussian Means (Left). Comparison of Simulated and Gaussian Variances (Right).](image)

### 4.1 Probability of Delay (GVA)

In addition to providing estimates for the mean and variance, we can also derive an approximation for the probability of delay using the GVA method. On the right of Figure 4, we compare the simulated probability of delay with our estimate from the GVA method. We see that the GVA method is quite accurate at estimating the probability of delay for our queueing model. Thus, we see that the GVA method seems to be quite useful in estimating queueing performance.

\[
\mathbb{P}(\text{Delay}) = \mathbb{P}(Q \geq c) \tag{4.26}
\]

\[
= \mathbb{E}[\{Q \geq c\}] \tag{4.27}
\]

\[
= \mathbb{E}[\{X \geq \chi\}] \tag{4.28}
\]

\[
= \Phi(\chi) \tag{4.29}
\]
4.2 An Additional Numerical Example

In addition, to the numerical example given in Table 1, we provide another numerical example where the customers are considered to be patient relative to the mean service time in Table 2. This implies that they are willing to wait longer than the mean service time to initiate service with a server.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_1$</td>
<td>$50 + 20 \sin(t)$</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>$50/2 + 10 \sin(t)$</td>
</tr>
<tr>
<td>$\lambda_3$</td>
<td>$50/3 + 20/3 \sin(t)$</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>.5</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>.25</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>.25</td>
</tr>
<tr>
<td>$\mu$</td>
<td>1</td>
</tr>
<tr>
<td>$c$</td>
<td>150</td>
</tr>
<tr>
<td>$q(0)$</td>
<td>0</td>
</tr>
</tbody>
</table>

In this new numerical example, we compare the Poisson and Gaussian approximations against the Monte Carlo simulation of the queueing process. On the left of Figure 5 we see that the mean of the queue length process is well approximated by both the Poisson and Gaussian approximations. However, on the right of Figure 5 we see that only the Gaussian method is able to approximate the dynamics of the variance. Since the mean and variance are different, the Poisson approximation is not able to capture this difference. On the left of Figure 6, we simulate the skewness of the process and it appears that the Gaussian and Poisson approximations are good at realizing that it is small. However, the Gaussian approximation is better since the skewness oscillates around zero. On the right of Figure 6 we simulate the probability of delay and compare it to the Poisson and Gaussian
approximations. We see that the both Poisson and Gaussian approximations work well, but the Gaussian approximation is slightly better. This dominance of the Gaussian distribution can be explained by the fact that the Poisson approximation only has one degree of freedom for projecting the mean queue length process, however, the Gaussian has two for the mean and variance. This gives the Gaussian approximation a little more flexibility than the Poisson approximation.

5 Nonstationary Loss Queues

In addition to the time varying multiserver queue, we can also analyze the nonstationary loss queue with extra waiting spaces, batch arrivals, and batch abandonment using the same approximation methods. We denote the nonstationary loss queue with $k_t$ extra waiting
spaces, batch arrivals, and batch abandonment as the $M_t^X/M_t/c_t/k_t + M_t^X$ queue. The functional Kolmogorov forward equations for the mean and variance have the following representation for the $M_t^X/M_t/c_t/k_t + M_t^X$ queue

$$
\dot{E}[Q] = \sum_{j=1}^{n} \lambda_j \cdot j \cdot E \left[ \{ Q < c + k - j + 1 \} \right] - \mu \cdot E \left[ (Q \wedge c) \right] \\
- \sum_{j=1}^{n} \beta_j \cdot j \cdot E \left[ (Q - c)^+ \cdot \{ Q > c + j - 1 \} \right]
$$

$$
\dot{\text{Var}}[Q] = \sum_{j=1}^{n} \lambda_j \cdot j^2 \cdot E \left[ \{ Q < c + k - j + 1 \} \right] + \mu \cdot E \left[ (Q \wedge c) \right] \\
+ \sum_{j=1}^{n} \beta_j \cdot j^2 \cdot E \left[ (Q - c)^+ \cdot \{ Q > c + j - 1 \} \right] \\
+ 2 \sum_{j=1}^{n} \lambda_j \cdot j \cdot \text{Cov}[Q, \{ Q < c + k - j + 1 \}] - 2 \cdot \mu \cdot \text{Cov}[Q, (Q \wedge c)] \\
- 2 \sum_{j=1}^{n} \beta_j \cdot j \cdot \text{Cov}[Q, (Q - c)^+ \cdot \{ Q > c + j - 1 \}] .
$$

One should notice that the equations are similar to the multiserver queue functional equations, however, the main difference is that the arrival rate process is a state dependent function of the queue length process that serves to block customers when the capacity is reached. Thus, in order to derive the Poisson and Gaussian approximations for the $M_t^X/M_t/c_t/k_t + M_t^X$ queue, we need to only derive expressions for the rate functions that pertain to the arrival rate since the service and abandonment terms are identical to the case of the multiserver queue.

5.1 Poisson Approximation

**Theorem 5.1.** The functional forward equations for the mean of the queue length process when the queue length has the distribution given by 3.8 has the form

$$
\dot{E}[Q] = \sum_{j=1}^{n} \lambda_j \cdot j \cdot E \left[ \{ Q < c + k - j + 1 \} \right] - \mu \cdot E \left[ (Q \wedge c) \right] \\
- \sum_{j=1}^{n} \beta_j \cdot j \cdot E \left[ (Q - c)^+ \cdot \{ Q > c + j - 1 \} \right] .
$$

Moreover, the rate functions have closed form expressions

$$
E \left[ \{ Q < c + k - j + 1 \} \right] = \Gamma(c + k - j, q(t)) \tag{5.30}
$$

$$
E \left[ (Q \wedge c) \right] = \Gamma(c, q(t)) + c \cdot \Gamma(c + 1, q(t)) \tag{5.31}
$$

$$
E \left[ (Q - c)^+ \cdot \{ Q > c + j - 1 \} \right] = q \cdot \Gamma(c - 1 + j, q(t)) - c \cdot \Gamma(c + j, q(t)) \tag{5.32}
$$
Proof. In order to prove this result, it only remains to give the expression for the rate function corresponding to the arrival rate since the others are identical to the multi-server case.

\[
E\left[\{Q < c + k - j + 1\}\right] = \sum_{m=0}^{c+k-j} e^{-q} \cdot \frac{q^m}{m!} = \Gamma(c + k - j, q(t))
\]

\[\square\]

5.1.1 Probability of Blocking (PISA)

One can also get the probability of blocking for each batch as

\[
E\left[\{Q > c + k - j\}\right] = \sum_{m=c+k-j}^{\infty} e^{-q} \cdot \frac{q^m}{m!} = \Gamma(c + k - j, q(t)). \quad (5.33)
\]

5.2 Gaussian Approximation

Theorem 5.2. The functional forward equations for the mean and variance of the \(\text{M}_t\text{X} / \text{M}_t / c_t / k_t + \text{M}_t\text{X}\) queue length process when the queue length has the distribution given by 4.17 have the form

\[
\dot{E}[Q] = \sum_{j=1}^{n} \lambda_j \cdot j \cdot E\left[\{Q < c + k - j + 1\}\right] - \mu \cdot E\left[(Q \wedge c)\right]
- \sum_{j=1}^{n} \beta_j \cdot j \cdot E\left[(Q - c)^+ \cdot \{Q > c + j - 1\}\right]
\]

\[
\dot{\text{Var}}[Q] = \sum_{j=1}^{n} \lambda_j \cdot j^2 \cdot E\left[\{Q < c + k - j + 1\}\right] + \mu \cdot E\left[(Q \wedge c)\right]
+ \sum_{j=1}^{n} \beta_j \cdot j^2 \cdot E\left[(Q - c)^+ \cdot \{Q > c + j - 1\}\right]
+ 2 \sum_{j=1}^{n} \lambda_j \cdot j \cdot \text{Cov}[Q,\{Q < c + k - j + 1\}] - 2 \cdot \mu \cdot \text{Cov}[Q, (Q \wedge c)]
- 2 \sum_{j=1}^{n} \beta_j \cdot j \cdot \text{Cov}[Q, (Q - c)^+ \cdot \{Q > c + j - 1\}].
\]

Moreover, the rate functions have closed form expressions.
where

\[
\chi = \frac{c-q}{\sqrt{v}}, \quad \psi_j = \frac{c-q+j-1}{\sqrt{v}}, \quad \text{and} \quad \psi_{k,j} = \frac{c+k-q-j+1}{\sqrt{v}}.
\]

**Proof.** In order to prove this it only remains to give the expressions for the arrival rate functions. For the mean we have that

\[
E \{Q < c + k - j + 1\} = E \{X < \psi_{k,j}\} \quad (5.41)
\]

\[
= \Phi(\psi_{k,j}) \quad (5.42)
\]

and for the variance, we have that

\[
\text{Cov} \{Q, \{Q < c + k - j + 1\}\} = \text{Cov} \{q + \sqrt{v} \cdot X, \{X < \psi_{k,j}\}\} \quad (5.43)
\]

\[
= \sqrt{v} \cdot \text{Cov} \{X, \{X < \psi_{k,j}\}\} \quad (5.44)
\]

\[
= -\sqrt{v} \cdot \phi(\psi_{k,j}) \quad (5.45)
\]

\[
= -\sqrt{v} \cdot \phi(\psi_{k,j}) \quad (5.46)
\]

5.2.1 Probability of Blocking (GVA)

Like in the Poisson case, we can also derive an approximation for the probability of blocking for each batch as

\[
E \{Q > c + k - j + 1\} = E \{X > \psi_{k,j}\} \quad (5.41)
\]

\[
= \Phi(\psi_{k,j}) \quad (5.42)
\]

For the nonstationary loss queue we also compare our approximations against Monte Carlo simulation with the parameters given in Table 3. On the left of Figure 7 we see that the Poisson and Gaussian approximations do quite well at approximating the mean when the queue length is not near the boundary. However, when the queue length is near capacity, then the Poisson approximation is not as good as the Gaussian approximation. On the right of Figure 7, we see a similar picture for the variance of the queue length process. On the left of Figure 8 we compare the probability of delay and on the right of Figure 8 we compare
Table 3: Batch Queueing Model Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_1$</td>
<td>$50 + 10 \sin(t)$</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>$50/2 + 5\sin(t)$</td>
</tr>
<tr>
<td>$\lambda_3$</td>
<td>$50/3 + 10/3 \sin(t)$</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>.5</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>.25</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>.25</td>
</tr>
<tr>
<td>$\mu$</td>
<td>1</td>
</tr>
<tr>
<td>$c$</td>
<td>150</td>
</tr>
<tr>
<td>$k$</td>
<td>25</td>
</tr>
<tr>
<td>$q(0)$</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 7: Comparison of Simulated, Poisson, and Gaussian Means (Left). Comparison of Simulated, Poisson, and Gaussian Variances (Right).

the blocking probability with our Poisson and Gaussian approximations. In both figures we see that the Gaussian approximation is better and the Poisson is not accurate, especially when the mean queue length process is near capacity. Thus, we once again conclude that the Gaussian approximation is much better than the Poisson since it is more flexible.

6 Conclusions and Future Work

We have shown in this paper that the Poisson and Gaussian approximations can be quite useful for estimating the dynamic behavior of queueing systems with time varying arrival rates, batch arrivals, and batch abandonment. Using the Poisson and Gaussian approximations we can mimic very accurately the real behavior of the mean, variance, probability of delay, and the probability of blocking for our queueing system by numerically integrating at most two differential equations. Thus, to estimate these performance measures, it is no longer necessary to simulate the stochastic system. Moreover, since we have accurate estimates
of the mean and variance, we can construct stabilizing staffing schedules for probability of delay and blocking by inverting the approximations as in [19]. Thus, this analysis yields insight and information for managers who want to optimally staff their service systems.

With respect to queueing, it is also interesting to consider when the number of servers is near one as in [20]. In the single server setting, it is more appropriate to use a generalized exponential distribution in order to approximate the dynamics of the queue length process. Furthermore, it is interesting to consider the stationary dynamics of the model and to possibly understand the transition probabilities of the queueing process using generating function techniques.

In this paper we have only analyzed the one dimensional queueing model. Thus, we find it interesting to consider a multi-dimensional network version of our model. This would enable us to model multiple hospitals or multiple communication networks and their interactions. Based on our one dimensional approximations, it seems natural to use a multi-dimensional Poisson or multivariate Gaussian distribution to generate accurate estimates of the mean, variance, and several other performance measures in the network setting. We plan to pursue these extensions in future work.

We should also mention that this approach of analyzing the queueing process is not limited to the study of queues. This approach applies to any Markov process with batch arrivals and batch departures. Thus, these methods can be applied to and can be applied to such processes like stochastic epidemic processes, which also have the difficulty that the system of forward equations is not closed. We also plan to pursue these extensions in future work.

References


