APPROXIMATIONS TO PERFORMANCE MEASURES IN QUEUING SYSTEMS

N.S. Kambo[1], A. Rangan[2] & E. Moghimihadji[3]*

[1,2] Department of Industrial Engineering,
Eastern Mediterranean University, Famagusta, Turkey
[3] Department of Industrial Engineering,
Istanbul Aydin University, Istanbul, Turkey
ehsanhadji@aydin.edu.tr

ABSTRACT

Approximations to various performance measures in queuing systems have received considerable attention because these measures have wide applicability. In this paper we propose two methods to approximate the queuing characteristics of a *GI/M/1* system. The first method is non-parametric in nature, using only the first three moments of the arrival distribution. The second method treads the known path of approximating the arrival distribution – by a mixture of two exponential distributions – by matching the first three moments. Numerical examples and optimal analysis of performance measures of *GI/M/1* queues are provided to illustrate the efficacy of the methods, and are compared with benchmark approximations.

OPSOMMING

Benaderings tot verskeie prestasiemaatstawwe in toustaanstelsels ontvang beduidende aandag weens die wye toepasbaarheid daarvan. In hierdie artikel word twee metodes voorgestel om die toustaaneienskappe van die *GI/M/1*-stelsel te benader. Die eerste metode is nie-parametries van aard en gebruik slegs die eerste drie momente van die aankomsverdeling. Die tweede metode volg die bekende roete om die aankomsverdeling te benader deur die eerste drie momente te pas, deur middel van 'n kombinasie van twee eksponensiële verdelings. Die doelteffendheid van die metodes word aan die hand van numeriese voorbeelde en optimale ontleding van die prestasiemaatstawwe van *GI/M/1*–stelsels bewys, en word vergelyk met benaderings.

* Corresponding author

1. INTRODUCTION

The growth of queuing theoretic applications has been phenomenal, ranging from communication and multimedia systems to inventory and reliability theory. This has led to a sustained interest in methods to evaluate performance measures in queuing theory. In the case of non-Markovian queues, the computations of these measures involve the arrival and/or service distributions explicitly. However, in practical applications like management, optical, and communication networks, the specific forms of these distributions might not be known. At best, one might only be in possession of the moments of the underlying distribution. There are a number of cases where the moments of a distribution are easily obtained, but theoretical distributions are not available in closed forms [9]. Alternatively, from the sample data observed, efficient estimators for the various moments of the underlying distributions can be calculated. Thus, computation of performance measures based on the first two or three moments of the arrival and/or service distributions is very useful. Whitt [16], in a classic exposition, discussed approximations using extremal distributions giving the upper and lower bounds for the performance measures in a *GI/M/1* system. Smith [13] proposed a two-moment approximation for the probability distribution of *M/G/1/K* systems, and extended it to the analysis of *M/G/1/K* queuing networks. Sohn & Lee [14] conducted a Monte Carlo simulation in order to study the relation between various performance measures in a *G/G/1* queue. Recent work on such systems with working vacations for the server have immense applications in ATM machines and internet systems, such as optical nets, electric nets, and communication nets [8, 4, 2]. In these applications, the arrival epochs could be observed or, at worst, simulated. Our motivation in this paper has thus been to obtain approximations to the performance measures of a *GI/M/1* system using only the first three moments of the arrival distribution, without explicit recourse to the arrival distribution

We will discuss our problem with specific reference to a *GI/M/1* queuing system, even though our methods also work in a similar way for other non-Markovian queues. Consider a *GI/M/1* queue whose traffic intensity is ρ= *E* (service time)/*E* (arrival time), *L* is the expected equilibrium queue length, and σ is the steady state probability that a customer will have to wait to begin his service. It is well known that

$$L = \rho/(1-\sigma) \tag{1}$$

where σ is the unique root in the open interval (0, 1) of the equation

$$\Phi\big(\mu(1-\sigma)\big) = \sigma \tag{2}$$

with μ=1/*E*(service time) and $\Phi(s)$, the Laplace-Stieltjes Transform of the inter-arrival distribution function, say *F*, given by:

$$\Phi(s) = \int_0^\infty e^{-st} dF(t) \tag{3}$$

We note that the evaluation of the performance measures σ and *L* require prior knowledge of the inter-arrival distribution function *F*, and not just the moments of *F*. As mentioned in the beginning of this section, many queuing applications are likely to produce only the moments of *F* and not the distribution itself. Thus the problem is to find σ and *L* on the basis of the first few moments of *F* only. Whitt [16] showed that there is a considerable reduction in the range of possible values of σ and *L* when the third moment is also used, compared with using just two moments of *F*. We propose simple and accurate methods to evaluate σ and *L* based on the first three moments of the distribution function *F* in the absence of any knowledge of the form of *F*. In section 2, we propose a non-parametric method based only on the first three moments of *F* – without recourse to approximate *F* – by another distribution function. Numerical illustrations are provided to compare the values of σ and *L*, using the present method, with their exact values. The method provides exact results for certain important arrival distributions like Erlang of order 2, Coxian (K_2), a mixture of two exponentials, and exponential distribution. We also provide two optimisation illustrations to obtain economic performance measures in the application of *GI/M/1* queuing systems.

Approximations of probability distributions by phase type distributions (by matching moments up to a certain order) have attracted the attention of researchers because of necessity, and because of their wide applicability. Pioneering work on phase type distributions and their various applications was done by Neuts [11]. Among the various members of the family of phase type distributions that have been studied, mixtures of two exponential distributions (known as H_2 distributions) play a key role in many approximations used in queuing theory. These distributions are log convex in nature, and can approximate highly-skewed distributions quite accurately. In section 3, we suggest a simple nonlinear programming method in which the first two moments are matched exactly, while the third moment is matched as closely as possible. This method works in the entire region of possible values (of m_1, m_2, m_3), and provides exact three-moment match wherever possible. The approximated H_2 distribution with the given three moments of the inter-arrival distribution is then used to calculate σ and *L*. Numerical illustrations are provided to validate the approximation and computation of the performance measures.

2. A NON-PARAMETRIC METHOD

We observe from equation (2) that the computation of the performance measures *σ* and *L* in the *GI/M/1* system requires the use of *Φ(s)*, the Laplace Transform of the density function *f* corresponding to the distribution function *F*. However, without prior knowledge of *F,* and armed only with the first three moments of *F*, an approximation to *Φ(s)* is obtained from the proposition that follows.

Proposition

Suppose that the first three raw moments $m_1(\neq 0)$, m_2, and m_3 ($3m_2^2 \neq 2m_1m_3$) of the distribution function *F* exist and are known. Then the following approximation to the Laplace Transform of the distribution function *F* holds.

$$\Phi(s) \approx \frac{A(s-s_0)+Bs}{s(s-s_0)+A(s-s_0)+Bs} \tag{4}$$

where $s_0 = \frac{6m_1(m_2-2m_1^2)}{3m_2^2-2m_1m_3}$, $A = \frac{1}{m_1}$, and $B = \frac{-s_0(m_2-2m_1^2)}{2m_1^2}$ (5)

Proof

In the classical renewal theory, the renewal density *m(t)* of a renewal process with interval density *f(t)* satisfies the integral equation:

$$m(t) = f(t) + \int_0^t m(t-u)f(u)du \tag{6}$$

Applying Laplace Transform to both sides of (6), we obtain the Laplace Transform of *m(t)* as:

$$m^*(s) = \Phi(s) \;/\; (1-\Phi(s)) \tag{7}$$

where *Φ (s)* is the Laplace Transform of the density function *f*.

Now *Φ (0) =1* and $\frac{d}{ds}(1-\Phi(s))|_{s=0} = \acute{\Phi}(0) = m_1$ is non zero. This means that the denominator 1-*Φ (s)* of (7) has a simple zero at *s=0*. Thus we may approximate $m^*(s)$ by a rational function of the form

$$m^*(s) \approx A/s + B/(s - s_0) \tag{8}$$

where A, B, and s_0 are constants determined as follows. Assuming the existence of moments of the density function $f(t)$, we can express $\Phi\ (s)$ as

$$\Phi(s) = \sum_{n=0}^{\infty} \frac{(-1)^n s^n}{n!} m_n \tag{9}$$

where $m_0=1$ and m_n is the n^{th} order moment about the origin of f. Using (9) in (7) and (8), and comparing the coefficient of s^0, s^1 and s^2 on both sides, we obtain (using some algebra) the values of A, B, and s_0 as given in (5). This completes the proof.

Note1: For the approximation to hold, it is necessary that $s_0 \leq 0$. Simple calculations show that this condition implies that $\Phi_2 > 2$ and $\Phi_3 \geq (3/2)\Phi_2^2$ or $\Phi_2 < 2$ and $\Phi_3 \leq (3/2)\Phi_2^2$ (see Figure 1), where C^2 is the squared coefficient of variation of inter-arrival time, $\Phi_2 = C^2+1$, and $\Phi_3 = m_3/m_1^3$.

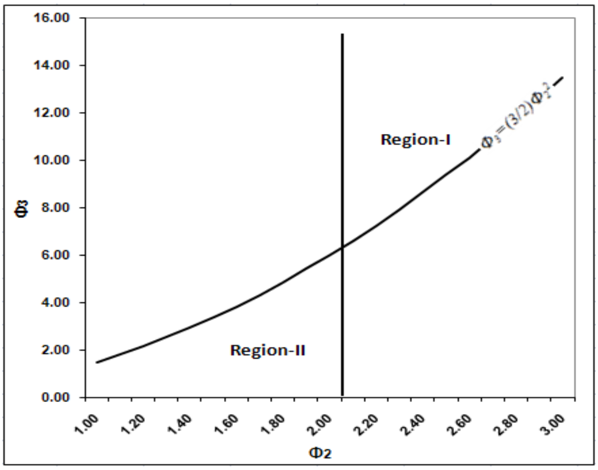

Figure1: Feasible regions for the approximation

Note2: It is worth mentioning that the error in the approximation (4) is of $o(1)$ as $s \rightarrow 0$.

Note3: The condition that s_0 is non-positive, which is necessary for the approximation to hold, is satisfied by many standard arrival distributions: uniform, gamma, mixed exponential, lognormal, Coxian (K_2), mixture of Erlangian, and Weibull. Also s_0 is non-positive for truncated normal and inverse Gaussian probability density functions under certain conditions.

Using (4) in equations (2) and (1) with the values of m_1, m_2, and m_3, the performance measures σ and L were obtained immediately. To illustrate the efficiency of the proposed method, we present (in Table 1) the values of σ and L computed using (4) for certain choices of the set $\{m_1, m_2, m_3\}$. In order to compare the approximations with exact values, we have considered the values of the moments of commonly used arrival distributions – namely, gamma (Table 1) and *PH4* distributions (Table 2). The values of σ and L are calculated for various values of traffic intensity ρ. Eckberg [5] specified upper and lower bound distributions that yield the maximum and minimum mean queue length in a steady state among all inter-arrival time distributions with two and three moments specified. Using these distributions, Whitt [16] calculated the maximum relative error for σ and L using the formula

$$MRE(in\ L) = (L_u - L_l)/L_l \tag{10}$$

where L_l and L_u are the minimum and maximum values of L using the lower and upper bound distributions. Table 1 also presents the upper and lower bounds for σ and L given two and three moments, and the corresponding maximum relative errors specified in (10). It can be seen that our method captures the values of σ and L with low relative errors.

Table1: Approximations for σ and L with gamma inter-arrival distribution

$f(t) = t^{k-1}\frac{e^{-\lambda t}\lambda^k}{\Gamma(k)}$		*ρ=0.3*		*ρ=0.7*		*ρ=0.9*	
Case-I	Exact	σ	0.27991	Σ	0.68748	σ	0.89541
		L	0.41661	L	2.23986	L	8.60503
	Non-parametric method	σ	0.31799	Σ	0.71160	σ	0.90422
		L	0.43988	L	2.42714	L	9.39653
	Relative error	5.58%		8.36%		9.20%	
C^2= 0.9091	Upper (2 moments)	σ	0.49760	σ	0.72080	σ	0.89890
		L	0.59713	L	2.50716	L	8.90208
m_1= 0.55	Lower (2 moments)	σ	0.04880	σ	0.46700	σ	0.80690
m_2= 0.5775		L	0.31279	L	1.31332	L	4.60798
m_3= 0.8951	MRE(in L)	90.91%		90.90%		93.19%	
Region-II	Upper (3 moments)	σ	0.49760	σ	0.72080	σ	0.89890
		L	0.59713	L	2.50716	L	8.90208
	Lower (3 moments)	σ	0.16180	σ	0.68076	σ	0.89520
		L	0.35791	L	2.19271	L	8.58779
	MRE(in L)	66.84%		14.34%		3.66%	
Case-II	Exact	σ	0.60233	σ	0.85242	σ	0.95297
		L	0.75439	L	4.74319	L	19.13469
	Non-parametric method	σ	0.55157	σ	0.85082	σ	0.95292
		L	0.66900	L	4.69232	L	19.11640
C^2= 3.3333	Relative error	11.31%		1.05%		0.03%	
m_1= 0.45	Upper (2 moments)	σ	0.77867	σ	0.87700	σ	0.95540
m_2= 0.8775		L	1.35542	L	5.69106	L	20.17937
m_3= 3.0273	Lower (2 moments)	σ	0.04088	σ	0.46700	σ	0.80690
Region-I		L	0.31279	L	1.31332	L	4.66080
	MRE(in L)	333.34%		333.33%		332.96%	
	Upper (3 moments)	σ	0.77867	σ	0.87700	σ	0.95540
		L	1.35544	L	5.69106	L	20.17937
	Lower (3 moments)	σ	0.24600	σ	0.84200	σ	0.95270
		L	0.39788	L	4.43038	L	19.02748
	MRE(in L)	240.67%		28.46%		6.05%	

3. INTER-ARRIVAL DISTRIBUTION APPROXIMATION BY MATCHING MOMENTS

Approximating general distributions by phase type distributions is important in queuing theory because their structure leads to Markovian state description and consequently analytical tractability. Although several phase type distributions have been used in the literature, two distributions that are prime candidates for such approximations (because of their simplicity and suitability) are mixtures of two exponentials (H_2) and Coxian (K_2) distributions. On this point, we use the former for analysis, as these distributions provide a fairly accurate match when the C^2 is large - which is true for arrival distribution in a queuing system. Furthermore, in some of the examples discussed by Whitt [17], the maximum relative error (MRE) when two moments are fitted was found to be 200 percent, while working with mixtures of exponentials reduced the MRE to 50 percent. Specifying the third moment reduced the MRE to 5 percent. Thus, a three-moment match using H_2 distributions for inter-arrival distributions seems to provide useful results.

Using an empirical study, Bere [3] showed that when the service time distribution is approximated using the first two of its moments, the third moment has a considerable effect on the probability distribution of the number of customers in an *M/G/1* queue if

C^2>*1*. He also showed that the probability distribution of the number of customers and the average number of customers in *λ(n)/G/1/N* system are highly sensitive to the third moment of the service time distribution if C^2>*1*. Altiok [1], in justifying the inclusion of the third moment in matching, refers to Bere's empirical work. Whitt [15] empirically showed that the effect of the third moment on the average number in the system in a *GI/G/1* queue becomes considerable as C^2 increases. If the service time distribution has C^2<*1*, the impact of the third moment is not significant [3, 15]. Since the present work deals with matching an H_2 distribution with three moments, we confine our attention to the range C^2>*1* only.

It is well known [15] that three numbers m_1, m_2, and m_3 can be the first three raw moments of a distribution function F, provided that $m_1 \geq 0$, $m_2/m_1^2 \geq 1$, and $m_3/m_1^3 \geq m_2^2/m_1^4$. Furthermore, if the first three moments exist for a distribution *F*, then an H_2 distribution exists with these three moments if, and only if, the first three moments of *F* satisfy the conditions $m_1 \geq 0$, $\Phi_2 = m_2/m_1^2 = C^2 + 1 \geq 2$, and $\Phi_3 = m_3/m_1^3 \geq (3/2)\ \Phi_2^2$ [1]. However, Karlin & Studden [7] have shown that m_1, m_2, and m_3 are the moments of some probability distributions on the positive real line if, and only if, $m_1 \geq 0$, $\Phi_2 \geq 1$, and $\Phi_3 \geq \Phi_2^2$. Thus, in the region where exact three-moment matching is not possible, researchers have used adhoc methods to find the approximate H_2 distribution. These regions are clearly shown in Figure 2.

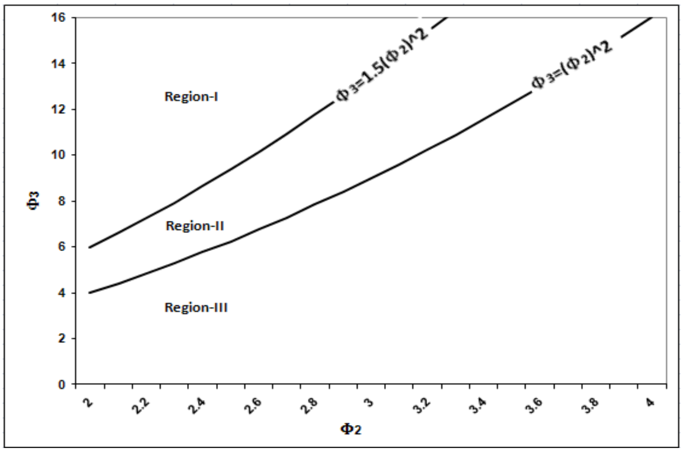

Figure 2: Region of three-moment matching (Region-I: Exact three-moment match possible; Region-II: Exact three-moment match not possible; Region-III: Infeasible region for three moments of a distribution function)

The parameters *p*, μ_1, and μ_2 of H_2 (see (11) below) when (m_1, m_2, m_3) falls in Region-I, and where the three moments can be matched exactly, are well documented. However, when these moments fall in Region-II, such that the three moments cannot be matched exactly, methods suggested in the literature are recipes in nature. Lopez-Herrero [10], in the absence of information on service distribution, used the maximum entropy principle approach to estimate the true distribution of the number of customers served during the busy period in an *M/G/1* retrial system. Whitt [15] suggests that "if m_3 turns out to be too small when attempting an H_2-fit, one procedure is to replace m_3 by something slightly larger than $(3/2)\ m_2^2/m_1$". Altiok [1] suggests the use of $m_3 = 3m_2^2/2m_1 + \varepsilon m_1^3$ but does not indicate how to calculate the perturbation parameter *ε*. In the following algorithm, we propose a goal programming procedure of matching the first two moments exactly, and matching the third moment as closely as possible in Region-II. This procedure subsumes Region-I as a particular case. We note that Region-III is an infeasible region for the existence of (m_1, m_2, m_3).

Algorithm

Given the first three moments of a distribution function *F* (say, m_1, m_2, and m_3, which are assumed to be finite), the parameters *p* $(0<p<1)$, μ_1, and μ_2 of the H_2 distribution given by (11) – whose first three moments either match m_1, m_2, and m_3 exactly, or match the first two moments exactly, and match the third moment as closely as possible in the sense of squared differences – are given by the following algorithm:

Step 1: Find optimal *p* using the Golden section method [12, 6] to solve the following one dimension optimisation problem:

$$\underset{\frac{C^2-1}{C^2+1}<p<1}{MinZ} = \left\{ m_3 - 6m_1^3\left[1+\frac{3}{2}\left(C^2-1\right)+\left(\frac{C^2-1}{2}\right)^{3/2}\frac{2p-1}{\sqrt{p(1-p)}}\right]\right\}^2$$

Step 2: Using the value of *p* found in step 1, compute:

$$\mu_1 = m_1\left[1-\sqrt{\frac{1-p}{2p}\left(C^2-1\right)}\right]$$

and

$$\mu_2 = m_1\left[1+\sqrt{\frac{p}{2(1-p)}\left(C^2-1\right)}\right]$$

The steps of the algorithm are justified using the following arguments. Consider the following probability density function of a mixture of two exponentials (H_2 distribution):

$$f(x) = p\frac{\exp\left(-x/\mu_1\right)}{\mu_1}+(1-p)\frac{\exp\left(-x/\mu_2\right)}{\mu_2}, x \ge 0, \qquad 0 \le p \le 1 \tag{11}$$

In the following, we do not consider the trivial cases of *p=0, 1* as they lead us to exponential density.

The first three moments of the density function above are:

$$m_1 = p\mu_1+(1-p)\mu_2 \tag{12}$$

$$m_2 = 2p\mu_1^2+2(1-p)\mu_2^2 \tag{13}$$

$$m_3 = 6\left[p\mu_1^3+(1-p)\mu_2^3\right] \tag{14}$$

From (12) we have

$$\mu_2 = \frac{m_1-p\mu_1}{1-p} \tag{15}$$

Substituting (15) in (13), and after simple algebra

$$\mu_1 = m_1\left[1\pm\sqrt{\frac{1-p}{2p}\left(C^2-1\right)}\right] \tag{16}$$

we have two cases. First we consider

$$\mu_1 = m_1\left[1-\sqrt{\frac{1-p}{2p}\left(C^2-1\right)}\right] \tag{17}$$

We know that

$\frac{1-p}{2p}\left(C^2-1\right)\geq 0$, which implies that $C^2 \geq 1$

Also, $\mu_1 > 0$ implies that $p > \frac{C^2-1}{C^2+1}$ (18)

Substituting (17) in (15) results in

$$\mu_2 = m_1\left[1+\sqrt{\frac{p}{2(1-p)}\left(C^2-1\right)}\right] \tag{19}$$

The third parameter *p* is obtained by matching the third moment as closely as possible. Thus, we minimise

$$\left\{m_3 - 6m_1^3\left[1+\frac{3}{2}\left(C^2-1\right)+\left(\frac{C^2-1}{2}\right)^{3/2}\frac{2p-1}{\sqrt{p(1-p)}}\right]\right]\right\}^2 \tag{20}$$

subject to: $1 > p > \frac{C^2-1}{C^2+1}$

In the second case, we set

$$\mu_1 = m_1\left[1+\sqrt{\frac{1-p}{2p}\left(C^2-1\right)}\right] \tag{21}$$

Using algebra similar to the first case results in

$$\mu_2 = m_1\left[1-\sqrt{\frac{p}{2(1-p)}\left(C^2-1\right)}\right] \tag{22}$$

and the third parameter *p* is determined by solving

$$\underset{\frac{2}{C^2+1} > p > 0}{MinZ =} \quad \left\{m_3 - 6m_1^3\left[1+\frac{3}{2}\left(C^2-1\right)+\left(\frac{C^2-1}{2}\right)^{3/2}\frac{2p-1}{\sqrt{p(1-p)}}\right]\right]\right\}^2 \tag{23}$$

It is easily seen that both cases lead to the same result, but with the roles of *p* and *1-p* interchanged.

In Table 2 we continue with *PH4* inter-arrival distribution. To illustrate this, we fit H_2 distribution for this distribution by matching the moments specified by the algorithm. The performance measures σ and L obtained by using the fitted H_2 distribution in steps 1 and 2 are given in Table 2. Attention is drawn to the MRE for these measures when three moments are matched [16] and to the actual relative error in using the approximated H_2 distributions. Also note that these two methods improve in accuracy for increasing ρ as expected.

Table 2: Approximations for *σ* and *L* with *PH4* inter-arrival distribution

$f(t) = \sum_{i=1}^{4} p_i \lambda_i \, e^{-\lambda_i t}$, $0 < p_i < 1$		***ρ=0.3***		***ρ=0.7***		***ρ=0.9***	
Case-I	**Exact**	σ	0.33210	σ	0.73870	σ	0.91790
		L	0.44917	L	2.67891	L	10.96224
	Non-parametric method	σ	0.33143	σ	0.73583	σ	0.91792
		L	0.44872	L	2.67717	L	10.96491
	Relative error	0.10%		0.06%		0.02%	
C^2= 1.517	**H2**	σ	0.33143	σ	0.73853	σ	0.91792
		L	0.44872	L	2.67717	L	10.96491
m_1= 19.997	**Relative error**	0.10%		0.06%		0.02%	
m_2= 1006.48	**Upper (2 moments)**	σ	0.61894	σ	0.78823	σ	0.92330
m_3= 101021.195		L	0.78728	L	3.30547	L	11.73403
	Lower (2 moments)	σ	0.04088	σ	0.46700	σ	0.80690
Region-I		L	0.31279	L	1.31332	L	4.66080
	MRE(in L)	151.70%		151.69%		151.76%	
	Upper (3 moments)	σ	0.61894	σ	0.78823	σ	0.92330
		L	0.78728	L	3.30547	L	11.73400
	Lower (3 moments)	σ	0.12374	σ	0.71530	σ	0.91730
		L	0.34236	L	2.45873	L	10.88270
	MRE(in L)	129.96%		34.44%		7.82%	
Case-II	**Exact**	σ	0.59490	σ	0.89250	σ	0.96900
		L	0.74056	L	6.51163	L	29.03226
	Non-parametric method	σ	0.53737	σ	0.89113	σ	0.96901
		L	0.64847	L	6.42699	L	29.04163
	Relative error	12.44%		1.30%		0.03%	
C^2= 5.821	**H2**	σ	0.53740	σ	0.89110	σ	0.96900
		L	0.64851	L	6.42792	L	29.03226
m_1= 20.004	**Relative error**	12.43%		1.29%		0.00%	
m_2= 2729.4164	**Upper (2 moments)**	σ	0.85938	σ	0.92186	σ	0.97170
m_3= 836216.8692		L	2.13341	L	8.95828	L	31.80212
	Lower (2 moments)	σ	0.04088	σ	0.46700	σ	0.80690
Region-I		L	0.31279	L	1.31332	L	4.66080
	MRE(in L)	582.06%		582.11%		582.33%	
	Upper (3 moments)	σ	0.85938	σ	0.19286	σ	0.97170
		L	2.13341	L	8.95828	L	31.80212
	Lower (3 moments)	σ	0.18173	σ	0.87540	σ	0.96880
		L	0.36663	L	5.61798	L	28.84615
	MRE(in L)	481.90%		59.46%		10.25%	

4. TWO OPTIMISATION ILLUSTRATIONS

4.1 Illustration 1

In practice, queuing managers are generally interested in optimising the model parameters under their control by minimising the operating cost or maximising the business profit. In the first illustration, we will be interested in obtaining the optimal service rate in a cost minimisation problem for a *GI/M/1* queuing system. The objective cost function consists of two components: the cost due to customers waiting in line (known as the delay cost), and the service cost rate. Thus the cost function to be minimised is given by:

$$C(\mu) = c_1(\lambda W) + c_2\mu \qquad (24)$$

where λ and μ are the arrival and service rates respectively, W is the expected waiting time of a customer in the system, c_1 is the expected cost per unit time of a customer's wait, and c_2 is the service cost rate. Using Little's formula, (24) reduces to

$$C(\mu) = c_1 L + c_2 \mu \tag{25}$$

The optimal μ* of the objective function above was computed using our moments matching method (introduced in section 3) by assuming the first three moments of the arrival distribution only, and the cost rates. However, in order to compare our results with the exact values, the moments were chosen to correspond to Coxian (K_2) and Inverse Gaussian distributions commonly used in queuing theory. The results are presented in Tables 3 and 4. When the Coxian arrival distribution was used, our method provided the exact values of μ*. In the case of Inverse Gaussian distribution, the relative errors were significantly small.

Table 3. The optimal service rate *μ with Coxian inter-arrival distribution. (The optimal *μ** computed, using our method and using the Coxian distribution, match exactly)**

$f(t) = \left(\frac{\rho\lambda_1 - \lambda_2}{\lambda_1 - \lambda_2}\right)\lambda_1 \exp(-\lambda_1 t) + \left(1 - \frac{\rho\lambda_1 - \lambda_2}{\lambda_1 - \lambda_2}\right)\lambda_2 exp(-\lambda_2 t)$								ρ=0.8, λ_1=2, λ_2=0.2 (m_1=1.5, m_2=11.5, m_3=167.25 and estimated p=0.77778, μ_1= 0.500006, and μ_2=5.000023)
	C(μ*)							
Approximation (moments matching)	**μ*=2.22**	**μ*=1.667**	**μ*=1.333**	**μ*=1.111**	**μ*=0.952**	**μ*=0.833**	**μ*=0.741**	
	22.9	17.87	15.33	14.35	14.86	17.91	29.74	c_1=1, c_2=10
	11.79	9.54	8.667	8.793	10.1	13.74	26.04	c_1=1, c_2=5
	5.127	4.54	4.667	5.459	7.242	11.24	23.81	c_1=1, c_2=2

Table 4. The optimal service rate *μ with Inverse Gaussian inter-arrival distribution**

$f(t) = \left[\frac{K}{2\pi t^3}\right]^{1/2} \exp\left(\frac{-K(t-M)^2}{2M^2 t}\right)$								K=1, M=2 (m_1=2, m_2=12, m_3=152 and estimated p=0.8430, μ_1= 1.3897, and μ_2=5.27770)
	C(μ*)							
Exact values	**μ*=1.667**	**μ*=1.25**	**μ*=1.00**	**μ*=0.833**	**μ*=0.714**	**μ*=0.625**	**μ*=0.556**	
	17.126	13.255	11.193	10.215	***10.213***	11.76	18.505	c_1=1, c_2=10
	8.793	7.005	6.193	***6.048***	6.642	8.635	15.727	c_1=1, c_2=5
	3.793	3.255	***3.193***	3.548	4.499	6.76	14.061	c_1=1, c_2=2
Approximation (moments matching)								
	17.139	13.265	11.198	10.213	**10.206**	11.750	18.505	c_1=1, c_2=10
	8.806	7.015	6.198	**6.046**	6.635	8.625	15.727	c_1=1, c_2=5
	3.806	3.265	**3.198**	3.546	4.492	6.750	14.061	c_1=1, c_2=2
	Relative error =		0.14%	0.03%	0.07%			

4.2 Illustration 2

In the second illustration, we consider an optimisation problem in a *GI/M/1* queue with working vacation for the server. These problems have wide application in Internet systems such as optical, electrical, and communication nets [8]. We consider a single server queuing system that has the general arrival process. The working vacation and vacation interruption are connected, and the server enters into vacation when there are no customers, and it can take service at the lower rate during the vacation period. If there are customers in the system at the instant of a service completion during the vacation period, the server will return to the normal working level regardless of whether the vacation ends. Otherwise, it continues the vacation. The performance measure *L*, the mean queue length, and *P(J=0)* and *P(J=1)* – which are the state probabilities of a server in the steady state - have been derived by Li et al. [8]. We refer the reader to their paper for the relevant expressions. Li et al. [8] considered the problem of optimising the service rate η during the server's vacation period for a given cost structure. Let c_w represent the unit time cost of every waiting customer, and c_1 and c_2 the service costs per unit time during the normal working level and vacation period respectively. The expected net cost function to be optimised can be seen to be

$$\min: Z = c_w L + c_1 \mu P(J=1) + c_2 \eta P(J=0) \qquad (26)$$

where μ is the service rate during the service period. The optimal service rate η^* was computed using our non-parametric method of section 2 for certain values of the model parameters and cost parameters. We have used the Coxian arrival distribution and its moments (as used in Illustration 1) to obtain the optimal service rate η^*. However, the Inverse Gaussian distribution could not be used, as the objective function (26) loses its convexity and becomes monotonic. We have used Erlangian of order 2 (used by Li et al. [8]) in its place. In Figures 3 and 4, we present the values of η versus the associated cost. The optimal η^* and the corresponding cost obtained using our method are very close to the exact values.

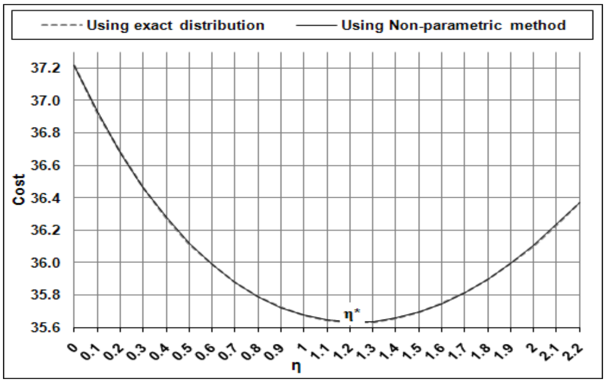

Figure 3: (*c_w=4, c_1=15, c_2=10, Θ=1, ρ=0.65*, and Coxian distribution parameters are *p=0.8, λ_1=2, λ_2=0.2*) Optimal service rate *η during servers vacation period with Coxian inter-arrival distribution**

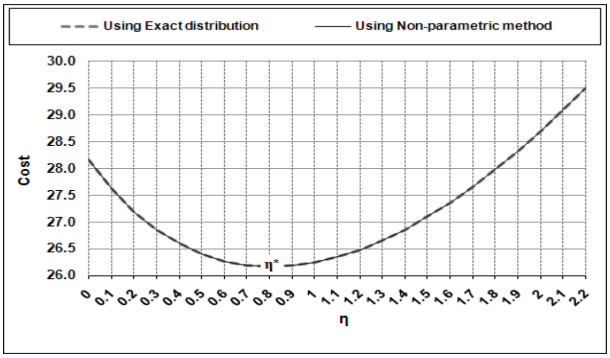

Figure 4: (*c_w=4, c_1=15, c_2=10, Θ=1, ρ=0.65*, and Erlangian distribution parameters are *K=2, M=2.5*) Optimal service rate *η during servers vacation period with Erlangian of order 2 inter-arrival distribution**

5. CONCLUSION

This study introduces two methods for evaluating performance measures in a *GI/M/1* queuing system in the absence of information on the arrival distribution, and when only the first three moments are known. The first method is non-parametric as it does not use the distribution function, whereas the second method uses an H_2 distribution obtained by moment matching procedure. This procedure involves the computationally economical Golden section method. Note that a Coxian (K_2) distribution is also a good phase type distribution to consider. It is worth pursuing the regions ($\Phi 2$, $\Phi 3$) in which each of these approximations scores higher than the others in terms of relative errors. The usefulness of the methods in optimisation procedures has been illustrated with examples.

REFERENCES

[1] **Altiok, T.** 1985. On the phase-type approximations of general distributions. *IIE Transactions*, 17(2), pp 110-116.

[2] **Baba, Y.** 2005. Analysis of a GI/M/1 queue with multiple working vacations. *Operations Research Letters*, 33(2), pp 201-209.

[3] **Bere, B.** 1981. *Influence du moment d'ordre 3 sur les files d'attente a lois de services generales.* Thesis (MS). Universite de Rennes, France.

[4] **Chae, K.C., Lim, D.E. & Yang, W.S.** 2009. The GI/M/1 queue and the GI/Geo/1 queue both with single working vacation. *Performance Evaluation*, 66(7), pp 356-367.

[5] **Eckberg, A.E.** 1977. Sharp bounds on Laplace-Stieltjes transforms, with applications to various queuing problems. *Mathematics of Operations Research*, 2(2), pp 135-142.

[6] **Kambo, N.S**. 1991. *Mathematical programming techniques*, revised edition, East-West Press Private Limited.

[7] **Karlin, S. & Studden, W.J.** 1966. *Tchebycheff systems: With applications in analysis and statistics*. New York: John Wiley & Sons.

[8] **Li, J., Tian, N. & Ma, Z.** 2008. Performance analysis of GI/M/1 queue with working vacations and vacation interruption. *Applied Mathematical Modeling*, 32(12), pp 2715-2730.

[9] **Lindsay, B.G., Pilla, R.S. & Basak, P.** 2000. Moment-based approximations of distributions using mixtures: Theory and applications. *Ann. Inst. Statist. Math*, 52(2), pp 215-230.

[10] **Lopez-Herrero, M.J.** 2002. On the number of customers served in the M/G/1 retrial queue: First moments and maximum entropy approach. *Journal of Computers and Operations Research*, 29(12), pp 1739-1757.

[11] **Neuts, M.F.** 1981. *Matrix geometric solutions in stochastic models: An algorithmic approach*. John Hopkins University Press.

[12] **Rao, S.S.** 2009. *Engineering optimization: Theory and practice*. 4th edition, John Wiley & Sons.

[13] **Smith, J.M.** 2011. Properties and performance modeling of finite buffer M/G/1/K networks. *Journal of Computers and Operations Research*, 38(4), pp 740-754.

[14] **Sohn, S.Y. & Lee, S.H.** 2004. Sensitivity analysis for output performance measures in long-range dependent queuing system. *Journal of Computers and Operations Research*, 31(9), pp 1527-1536.

[15] **Whitt, W.** 1982. Approximate a point process by a renewal process, I: Two basic methods. *Operations Research*, 30(1), pp 125-147.

[16] **Whitt, W.** 1984. On approximations for queues, I: Extremal distributions. *AT&T Bell Laboratory Technical Journal*, 63(1), pp 115-138.

[17] **Whitt, W.** 1984. On approximations for queues, III: Mixtures of exponential distributions. *AT&T Bell Laboratory Technical Journal*, 63(1), pp 163-175.