Approximation to Performance Measures in Queuing Systems

Ehsan MoghimiHadji
Department of Industrial Engineering
Istanbul Aydin University
Florya, Istanbul, Turkey

Alagar Rangan
Department of Industrial Engineering
Eastern Mediterranean University
Famagusta, Via Mersin 10, Turkey

Abstract

Approximations to various performance measures in queuing systems have received considerable attention mainly because these measures have wide applicability. In this paper, we propose a method to approximate the queuing characteristics of a *GI/M/1* system, which is easy to evaluate and provide reasonably accurate results. This method is non-parametric in nature and utilizes only the first three moments of the arrival distribution. Numerical examples and optimality analysis of performance measures of *GI/M/1* queue are provided to illustrate the efficacy of the method and compared with benchmark approximations.

Keywords
Optimization; Non-parametric approximation; *GI/M/1* queuing system.

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 the methods of evaluation of the 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 and at best one might be in possession of the moments of the underlying distribution only. There are a number of cases where the moments of a distribution are easily obtained, but theoretical distributions are not available in closed forms (Lindsay et al. 2000). Alternatively, from the observed sample data, efficient estimators for the various moments of the underlying distributions could 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-I (1984) 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 (2011) 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. Li and Tian (2011), in their recent work, used the matrix analytical approach to obtain the steady state distributions of number of customers in the system for a GI/M/1 queue with single working vacation. Zhang and Hou (2011) analyzed the steady state of the GI/M/1/N queue with a variant of multiple working vacations by employing the supplementary variable and embedded Markov chain methods. Sohn and Lee (2004) conducted a Monte Carlo simulation in order to study the relation between various performance measures in a *G/G/1* queue. Recent works 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 (Li et al. 2008, Chae et al. 2009, and Baba 2005). In these applications the arrival epochs could be observed or at worst simulated. Thus, our motivation in this paper has 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 recourse to the arrival distribution explicitly.

In the following, we will discuss our problem with specific reference to a *GI/M/1* queuing system, even though our method works in similar situations for other non-Markovian queues as well. 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 commence 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 $\mu=1/E$(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* requires the prior knowledge of the inter-arrival distribution function *F* and not just the moments of *F*. As remarked in the beginning of this section, many queuing applications are likely to produce the moments of *F* and not the distribution itself. Thus, the problem is to find σ and *L* based on the first few moments of *F* only. Whitt-I (1984) showed that there is a considerable reduction in the range of possible values of σ and *L* when the third moment is also used as compared to two moments of *F*. Our endeavor in this note is to 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 on 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 examples 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), mixture of two exponentials, and exponential distribution. We also provide one optimization illustration to obtain economic performance measures in the application of *GI/M/1* queuing systems.

2. A non-parametric method

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

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.

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

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

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 $\Phi(s)$ is the Laplace Transform of the density function *f*.

Now $\Phi(0)=1$ and $\frac{d}{ds}\left(1-\Phi(s)\right)|_{s=0} = \acute{\Phi}(0) = m_1$ is non zero. This means that the denominator 1- $\Phi(s)$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 after 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 $s_0 \le 0$. Simple calculations show that this condition implies that $\Phi_2>2$ and $\Phi_3 \ge (3/2)\Phi_2^2$ or $\Phi_2<2$ and $\Phi_3 \le (3/2)\Phi_2^2$ (see Figure 1), where C^2 is the squared coefficient of variation of inter-arrival time, $\Phi_2= C^2+1$, and $\Phi_3=m_3/m_1^3$.

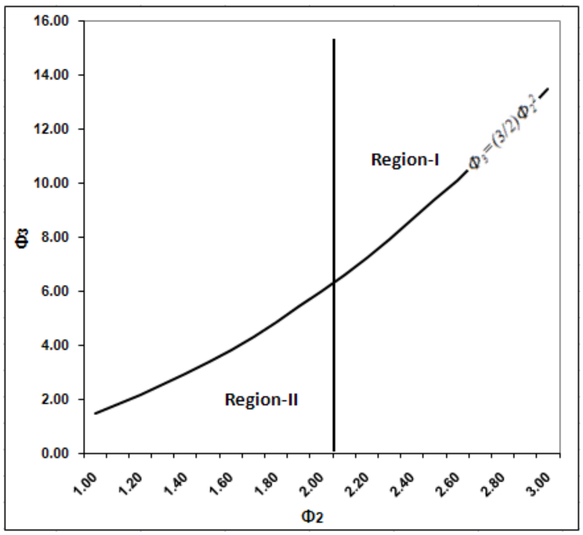

Figure 1: Region I and Region II are the feasible regions for the approximation to hold

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

Note3: The condition that s_0 is non-positive, which is necessary for the approximation to hold, is satisfied by many standard arrival distributions like 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* could be obtained. To illustrate the efficiency of the proposed method, we present in Tables 1 and 2 the values of σ and *L* computed using (4) for certain choice of the set $\{m_1, m_2, m_3\}$. The triplet (m_1, m_2, m_3) is chosen from the regions I and II (Figure 1). In order to compare the approximations with exact values we have considered the values of the moments of commonly used arrival distributions namely gamma and *PH4* (Phase type) distributions. The values of σ and *L* are calculated for various values of traffic intensity ρ. Eckberg (1977) specified upper and lower bound distributions that yield the maximum and minimum mean queue length in steady state among all inter-arrival time distributions with two and three moments specified. Using these distributions Whitt-I (1984) 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. Tables 1 and 2 also present 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.

Table 1: Comparison of the values of the performance measures σ and L computed using the approximation with the exact values when the arrival distribution is *PH4*

$f(t) = \sum_{i=1}^{4} p_i \lambda_i e^{-\lambda_i t}$		**ρ=0.3**		**ρ=0.7**		**ρ=0.9**	
	Exact	**σ**	**0.33210**	**σ**	**0.73870**	**σ**	**0.91790**
		L	**0.44917**	**L**	**2.67891**	**L**	**10.96224**
	Proposed approximation	**σ**	**0.33143**	**σ**	**0.73583**	**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**
C^2= 1.517		**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%**	

Table 2: Comparison of the approximations values of the performance measures σ and L computed using the approximation with the exact values when the inter-arrival distribution is gamma

$f(t) = t^{k-1} \frac{e^{-\lambda t} \lambda^k}{\Gamma(k)}$		**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**
	Proposed approximation	**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**
	Proposed approximation	**σ**	**0.55157**	**σ**	**0.85082**	**σ**	**0.95292**
		L	**0.66900**	**L**	**4.69232**	**L**	**19.11640**
	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**
		L	**0.31279**	**L**	**1.31332**	**L**	**4.66080**
Region - I	**MRE(in L)**	**333.34%**		**333.33%**		**332.96%**	
	Upper (3 moments)	**σ**	**0.77867**	**σ**	**0.87700**	**σ**	**0.95540**
C^2= 3.3333		**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. Optimization Illustration

In practice, queuing managers are generally interested in optimizing the model parameters under their control by minimizing the operating cost or maximizing the business profit. In this illustration, we will be interested in obtaining the optimal service rate in a cost minimization problem for a *GI/M/1* queuing system. The objective cost function consists of two components, which are the cost due to customers waiting in line known as the delay cost, and the service cost rate. Thus, the cost function to be minimized is given by:

$$C(\mu) = c_1(\lambda W) + c_2\mu \tag{11}$$

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, (11) reduces to

$$C(\mu) = c_1 L + c_2\mu \tag{12}$$

The optimal μ^* of the above objective function was computed using our non-parametric method introduced in section 2 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 so as 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^* whereas 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 exactly match)

$f(t)=\left(\frac{p\lambda_1-\lambda_2}{\lambda_1-\lambda_2}\right)\lambda_1\exp(-\lambda_1 t)+\left(1-\frac{p\lambda_1-\lambda_2}{\lambda_1-\lambda_2}\right)\lambda_2\exp(-\lambda_2 t)$								p=0.8, λ_1=2, λ_2=0.2 (m_1=1.5, m_2=11.5, m_3=167.25 and s_0= -0.6, A=0.667, B=0.933)
C(μ)								
Approximation	**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 s_0=-0.273, A=0.5, B=0.136)
C(μ)								
	μ=1.6667	**μ=1.25**	**μ=1.00**	**μ=0.8333**	**μ=0.7143**	**μ=0.625**	**μ=0.5556**	
Exact Values	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	17.139	13.265	11.198	10.213	***10.206***	11.75	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	6.75	14.061	c_1=1, c_2=2
	Relative Error =		0.16%	0.03%	0.07%			

4. Conclusions

This note introduces a method of approximation for the evaluation of 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. This method is non-parametric in the sense that it does not use the explicit form of the underlying arrival distribution function. Such situations arise frequently in queuing models wherein only the instants of customer arrivals are known. This information will obviously lead us to the sample moments and thus our present approximation comes in handy. This approximation is simple to execute and generates exact results for some important arrival distributions such as exponential, mixture of two exponentials, Coxian, and gamma distribution function. It is worth mentioning that it is also possible to employ moments matching procedures as an approximation for the unknown arrival distribution. Typical distributions used in the matching procedure are phase type distributions. This is because of the fact that any distribution can be approximated by a phase type distribution although the number of phases needed might be large rendering the approximation ineffective. Work in this direction is in progress. The usefulness of the methods in optimization procedures has been illustrated with examples.

References

Baba, Y., Analysis of a GI/M/1 queue with multiple working vacations, *Operations Research Letters*, vol. 33, no. 2, pp. 201-209, 2005.

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

Eckberg, A.E., Sharp bounds on Laplace-Stieltjes transforms, with applications to various queuing problems, *Mathematics of Operations Research*, vol. 2, no. 2, pp. 135-142, 1977.

Li, J., and Tian, N., Performance analysis of a GI/M/1 queue with single working vacation, *Journal of Applied Mathematics and Computation*, vol. 217, no. 10, pp. 4960-4971, 2011.

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

Lindsay, B.G., Pilla, R.S., and Basak, P., Moment-based approximations of distributions using mixtures: Theory and applications, *Ann. Inst. Statist. Math*, vol. 52, no. 2, pp. 215-230, 2000.

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

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

Whitt, W., On approximations for queues, I: Extremal distributions, *AT&T Bell laboratory technical journal*, vol. 63, no. 1, pp. 115-138, 1984.

Zhang, M., and Hou, Z., Steady state analysis of the GI/M/1/N queue with a variant of multiple working vacation, *Journal of Computers and Industrial Engineering*, vol. 61, no. 4, pp. 1296-1301, 2011.