

# Waiting time probabilities in the $M/G/1 + M$ queue

Chihoon Lee\*  
Department of Statistics  
Colorado State University  
Fort Collins, CO, USA

Jianqiang C. Wang†  
Services Research Lab  
Hewlett-Packard Labs  
Palo Alto, CA, USA

November 29, 2010

## Abstract

We consider an  $M/G/1$  queueing system where the customers may leave the queue if their services do not commence before an exponentially distributed random time. The (conditional) offered waiting time distribution is approximated by a gamma distribution via matching the first and second moments of the actual waiting time. A simulation study is conducted to assess the accuracy of the approximation and it reveals that the approximation performs satisfactorily under general conditions on service time distributions.

*Keywords and Phrases: impatient customers, queueing system with customer abandonment, offered waiting time distribution, gamma distribution, moment matching.*

## 1 Introduction

In the  $M/G/1 + M$  queue, customers arrive at a single-server station according to a Poisson process with rate  $\lambda \in (0, \infty)$ , where successful/patient customers are served according to the First-Come-First-Served regime. Assume that each customer arrives to a service station with memoryless “patience” time clock, that is, if a service does not begin within an exponentially distributed amount of time, then the customer abandons the service system. Denote the independent and identically distributed service time random variable by  $B$  with continuous distribution function  $B(x) = \mathbb{P}(B \leq x)$  and  $B(0) = 0$ . Let  $R$  be an exponential random variable with mean parameter  $\gamma > 0$  denoting the customer’s “patience” time with distribution  $R(x) = \mathbb{P}(R \leq x)$ . We assume that  $B$  and  $R$  are independent of each other and

---

\*chihoon@stat.colostate.edu

†jianqiang.jay.wang@hp.com

the arrival process. Let  $W_o$  denote the time that a customer needs to wait before receiving a service, also known as *offered* waiting time or *virtual* waiting time. We further assume that  $R$  and  $W_o$  are independent each other (given the relevant covariates with respect to each individual customer). Knowledge of the waiting time probability plays an important role in the design and control of a manufacturing or service system in which customers may abandon based on their foreseen waiting time. Many researchers have addressed the problem of virtual offered waiting time for queueing systems with abandonment phenomenon, for single server cases, see DALEY (1965), STANFORD (1979), BACELLI and HÉBUTERNE (1981), BACELLI et al. (1984) and DE KOK and TIJMS (1985), among others. In Section 2 we review approaches taken in some of these papers and discuss their drawbacks from practical computational viewpoint. More detailed literature reviews will be presented along the way.

The goal of this paper is to provide a simple, easy-to-implement approximation scheme to the conditional waiting time distribution function

$$\mathbb{P}(W_o \leq x | W_o > 0) \quad \text{for } x > 0.$$

We assume the availability of the *actual* waiting times, that can be conveniently observed in many application scenarios. The main idea of our approximation is based on fitting a gamma distribution to the *virtual* waiting time distribution via matching the first and second moments of the actual waiting times. This approach is discussed in Section 2.1 and simulated examples of queueing models in Section 3 indicate that it performs satisfactorily on several different service time distributions. Our work is inspired by NOBEL and TIJMS (2006), which deals with waiting time probabilities for  $M/G/1$  retrial queue using a gamma distribution and the exact expressions for the first two moments of the conditional waiting time. In Section 4, we further discuss situations in which our method has limitations in providing accurate estimates and bring up some future research directions.

## 2 Approximation for waiting time probabilities

The waiting time probability distribution  $\mathbb{P}(W_o \leq x)$  is a mixture distribution having an atom at  $x = 0$  and a density for  $x > 0$ . We denote the point mass at 0 by  $P_0$ , which is also the probability of system idleness at steady state. Then the distribution function of virtual waiting time can be specified as

$$F(x) = P_0 + (1 - P_0)\mathbb{P}(W_o \leq x | W_o > 0), \quad x \geq 0.$$

It suffices to calculate the system idle probability  $P_0$  and approximate the conditional waiting time distribution  $\mathbb{P}(W_o \leq x | W_o > 0), x > 0$ . We note that in many practical cases the only available data are the *actual* waiting times that are right-censored by patience times, that is,  $W_a \equiv \min\{W_o, R\}$ . Hence we assume that the actual waiting times are conveniently given as observed data in our queueing model and, in addition, we assume the patience time distribution parameter  $\gamma$  is known *a priori*. The main purpose of our analysis is to fit a gamma distribution to the *virtual* waiting time distribution by moment matching, and

examine the goodness-of-fit through numerical studies. Details will be provided in Section 2.1.

In practice, the gamma distribution is a flexible life distribution model that offers a good fit to numerous sets of failure data. A random variable  $X$  with gamma distribution  $\Gamma(\alpha, \beta)$  is given by the probability density function

$$f_{\Gamma(\alpha, \beta)}(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} 1_{\{x \geq 0\}},$$

where  $\alpha, \beta \in (0, \infty)$  are shape and scale parameters respectively. Then the mean and variance of  $X$  are given as  $\mathbb{E}[X] = \alpha\beta$ ,  $\text{Var}[X] = \alpha\beta^2$ . For instance, the gamma distribution arises in situations where one is concerned about the waiting time for a finite number ( $\alpha \in \mathbb{N}$ ) of independent events to occur, assuming that events occur at a constant rate  $1/\beta > 0$ . The distribution  $\Gamma(k, 1/\lambda)$  with integer values of  $\alpha = k$  equals to the well-known Erlang distribution denoted by *Erlang*( $k, \lambda$ ).

Denote the complementary distribution function of service time and patience time as  $\bar{B}(x)$  and  $\bar{R}(x)$ , respectively. We assume that density function of virtual waiting time distribution exists for  $x > 0$  and denote it by  $f(x)$ . We use a level-crossing argument to derive an integral equation satisfied by the density function  $f(x)$ . Suppose the process  $\{W_o(t) : t \geq 0\}$  is stationary. The expected number of down-crossings at level  $x > 0$  during interval  $(t, t+h)$  is  $[F(x+h) - F(x)](1 - \lambda h)$ . The expected number of up-crossings at level  $x$  during  $(t, t+h)$  consists of two parts: (i) if customers whose service time requirements are greater than  $x$  arrive at an empty system then it yields the up-crossings  $\lambda h P_0 \bar{B}(x)$  and similarly, (ii) if customers whose service requirements are in excess of  $x - t$  find, on their arrival,  $t$  amount of workload and are willing to stay at least  $t$  then the expected number of up-crossings is equal to  $\lambda h \int_0^x \bar{B}(x-t) \bar{R}(t) f(t) dt$ . Note here that these arguments are implicitly due to the PASTA (Poisson arrivals see time averages) property.

The level-crossing method as described by COHEN (1977) and DOSHI (1992) implies that the conservation law holds in the long run, that is, we have

$$[F(x+h) - F(x)](1 - \lambda h) = \lambda h P_0 \bar{B}(x) + \lambda h \int_0^x \bar{B}(x-t) \bar{R}(t) f(t) dt. \quad (1)$$

It is through this observation that the distribution functions for the virtual waiting time, service time, and patience time can be related. Dividing both sides in (1) by  $h$  and letting  $h \rightarrow 0$  we get

$$f(x) = \lambda P_0 \bar{B}(x) + \lambda \int_0^x \bar{B}(x-t) \bar{R}(t) f(t) dt, \quad (2)$$

and the normalizing equation

$$P_0 + \int_0^\infty f(t) dt = 1. \quad (3)$$

Exact computation of waiting time density  $f(x)$  and system idle probability  $P_0$  is a formidable task for general service time distribution  $B(x)$ . For the case of exponential service

time and patience time distributions, i.e.  $M/M/1 + M$  queueing system, equation (81) on page 142 of ANCKER and GAFARIAN (1963) (see also the relation (49) of STANFORD (1979)) provides an explicit formula for  $f(x)$ , that is,  $f(x) = c_1 \exp(-c_2x - c_3e^{-c_4x})$ , where  $c_i > 0$  ( $i = 1, \dots, 4$ ) are appropriate constants. Hence for large values of  $x$ , the density  $f(x)$  can be well approximated by that of exponential random variables. See also BAE et al. (2001) and DE KOK and TIJMS (1985) for the case of the  $M/G/1 + D$  queue with deterministic patience time. DALEY (1965) provided the Laplace-Stieltjes transform of the density  $f(x)$  in the form of an infinite series

$$\tilde{f}(s) = P_0 \sum_{j=0}^{\infty} \prod_{k=0}^j \lambda \tilde{B} \left( s + \frac{k}{\gamma} \right), \quad P_0^{-1} = 1 + \sum_{j=0}^{\infty} \prod_{k=0}^j \lambda \tilde{B} \left( \frac{k}{\gamma} \right), \quad (4)$$

where for any function  $g(x)$ ,  $\tilde{g}(s) \equiv \int_0^{\infty} e^{-sx} dg(x)$ . To obtain  $\mathbb{P}(W_0 \leq x | W_0 > 0)$ ,  $x > 0$  one needs to evaluate the infinite series in (4) and invert  $\tilde{f}(s)$  at required  $x$  values, which in practice is quite difficult procedure to perform both analytically and numerically. The moments of  $W_0$  can be expressed in terms of derivatives of  $\tilde{f}$  in (4). However, this analytic method seems practically infeasible to obtain computable formulas for the moments, since it frequently involves numerical Laplace transform and evaluation of infinite series.

On the other hand, one can approximate (2) and (3) using numerical integration, e.g. the trapezoidal rule, in solving equation (2), which is in the form of the Volterra integral equation of the second kind. The numerical approximation procedures are described in IRAVANI and BALCIOGLU (2008), NETRAVALI (1973), DEN ISEGER et al. (1997), and references therein.

We point out that both Laplace transform (4) and numerical approximation approaches aforementioned are excessively complex and computationally expensive, and must be adapted to specific service time distributions. As a practical alternative, we propose a simple approximation scheme of the waiting time distribution based on the gamma distribution.

## 2.1 Gamma approximation

We consider a family of service time distributions  $B(x)$  that has the following tail behavior: As  $x \rightarrow \infty$ ,  $\overline{B}(x) \sim x^{-\theta_1} e^{-\theta_2 x}$  for some  $\theta_1, \theta_2 \in (0, \infty)$ , where  $f(x) \sim g(x)$  means  $\lim_{x \rightarrow \infty} f(x)/g(x) = c$  for some finite constant  $c$ . Then choosing an appropriate set of parameters  $\theta_1$  and  $\theta_2$ ,  $B(x)$  would yield a very flexible class of service time distributions in practice. For instance, decaying rates of  $\overline{B}(x)$  will be ranging from polynomial to exponential. Also, notice that  $x^{-\theta_1} e^{-\theta_2 x}$  is the form of probability density function of a gamma random variable up to a multiplicative constant. Then by formerly replacing  $f(x)$  in (2) by the density of a gamma random variable, the second term on the right side of (2) becomes the (lower) incomplete gamma function. An asymptotic behavior of incomplete gamma function is known (see, e.g., Chapter 8 of OLVER et al. (2010)); for large  $x$ , its tail decay rate is of order  $x^{-c_1} e^{-c_2 x}$  for appropriate constants  $c_1, c_2 \in (0, \infty)$ . Hence for large values of  $x > 0$ , the solution  $f(x)$  of (2) can be well approximated by the density of gamma random variables, for the aforementioned class of service time distributions.

Recall the *actual* waiting time  $W_a \equiv \min\{W_o, R\}$ . Let  $\gamma = \mathbb{E}[R]$  so that  $R$  is  $\Gamma(1, \gamma)$  random variable. Then one can derive by assuming that  $W_o$  follows a  $\Gamma(\alpha, \beta)$  distribution that

$$\mathbb{E}[W_a] = \gamma \left( 1 - \left( \frac{\gamma}{\gamma + \beta} \right)^\alpha \right), \quad (5)$$

$$\mathbb{E}[W_a^2] = 2\gamma^2 \left( 1 - \left( \frac{\gamma}{\gamma + \beta} \right)^\alpha \right) - 2\gamma\alpha\beta \left( \frac{\gamma}{\gamma + \beta} \right)^{\alpha+1}. \quad (6)$$

We use (5) and (6) to set the parameters  $\alpha$  and  $\beta$  of the gamma distribution. More specifically, we replace  $\mathbb{E}[W_a]$  and  $\mathbb{E}[W_a^2]$  by sample moments  $\bar{W}_a$  and  $\sum_{n=1}^N W_{a,n}^2/N$ , respectively, and seek to solve equations (5) and (6). It is often difficult to get the solution to the estimating equations (5) and (6) to converge. The reason for this is that the shape parameter  $\alpha$  is in the exponent and a small change in  $\alpha$  may result in magnified change in the exponentiation. A reparameterization fixes the convergence problem. We define  $\eta \equiv \log(\gamma/(\gamma + \beta))$ , then one can show that the aforementioned estimating equations are equivalent to the following:

$$\log \left( 1 - \frac{\bar{W}_a}{\gamma} \right) = \alpha\eta \quad (7)$$

$$C + \alpha\gamma(1 - \exp(\eta)) = 0, \quad (8)$$

where  $C = (\sum_{n=1}^N W_{a,n}^2/N - 2\gamma\bar{W}_a)/(2\gamma - 2\bar{W}_a)$ . The two equations (7) and (8) in  $\alpha$  and  $\eta$  are always solvable and have a unique solution in a neighborhood of true parameter values, provided  $(\alpha, \beta)$  are isolated roots to the population estimating equations (5) and (6). See Chapter 7 of SERFLING (1980) for an overview of statistical theory on estimators defined by estimating equations. Then we can transform back to estimated  $\beta$  after solving for  $\alpha$  and  $\eta$ . We denote the estimated parameters as  $(\hat{\alpha}, \hat{\beta})$ . Any statistical package will provide the value of  $F_{\hat{\alpha}, \hat{\beta}}(x)$  for arbitrary  $x > 0$ , where  $F_{\alpha, \beta}(\cdot)$  is the cumulative distribution function of a gamma random variable with parameter values  $\alpha, \beta \in (0, \infty)$ .

The true tail probability of the offered waiting time  $\mathbb{P}(W_o > x | W_o > 0)$  can then be estimated by taking the proportion of customers with actual waiting time greater than  $x$ , and then using the formula

$$\mathbb{P}(W_o > x | W_o > 0) = \mathbb{P}(W_a > x | W_a > 0) / \mathbb{P}(R > x). \quad (9)$$

This computing procedure has been used in the next section on numerical studies. The idea of moment matching gamma approximation is not restricted to  $M/G/1 + M$  queueing systems, but can be readily generalized to arbitrary known patience time distribution, i.e.,  $M/G/1 + G$ . In the general case, one can analytically derive the first and second moment of  $W_a$  as functions of gamma parameters  $\alpha, \beta$  and patience time distribution, and then solve the sample moment matching equations to obtain gamma parameters. But one complication for  $M/G/1 + G$  systems is that the estimating equations often need to be evaluated numerically, as the form of (5) and (6) may not be obtainable for general patience time distribution.

We close this section with one side result, that may be of independent interest. Suppose we are interested in estimating customer abandonment rate  $\gamma^{-1} \in (0, \infty)$ . We propose a natural estimator of  $\gamma^{-1}$  given data up to time  $t \in (0, \infty)$  by

$$r(t) = \frac{\text{Total number of abandonments by time } t}{\text{Total waiting time by time } t}. \quad (10)$$

In practice, one observes the  $M/G/1 + M$  system up to a finite time horizon limit  $T$ , and we propose to use  $r(T)$  as an estimate of abandon rate parameter  $\gamma^{-1}$ . We note that the proposed estimator is still obtainable even if only queue length and abandonment data are at hand, as the aggregated waiting time by time  $t$  equals to the integral of queue length process up to time  $t$ . The following theorem establishes consistency and asymptotic unbiasedness results, which justify the use of an estimator in (10).

**Theorem 1.** *For  $t \in (0, \infty)$ , let  $r(t)$  be as in (10) and suppose  $W_o$  follows a gamma distribution. Then we have*

- (i) *with probability one,  $r(t) \rightarrow \gamma^{-1}$ , as  $t \rightarrow \infty$ ,*
- (ii)  *$\mathbb{E}[r(t)] \rightarrow \gamma^{-1}$ , as  $t \rightarrow \infty$ .*

*Proof.* See the Appendix. ■

### 3 Numerical study

In this section, we provide numerical results which illustrate the performance of the proposed gamma approximation. We simulate  $M/G/1 + M$  single-server queues and run each queue for sufficient amount of time. In doing simulation, we fix the customer arrival rate at  $\lambda = 1$  and consider eight different distributions for service time (cf. DE KOK and TIJMS (1985), NOBEL and TIJMS (2006), IRAVANI and BALCIOĞLU (2008)):

- (a) *Exponential(0.4)*,
- (b) *Deterministic distribution with service time fixed at 0.8*,
- (c) *Shifted exponential distribution; Exponential(0.3) + 0.5*,
- (d) *Weibull(1.5, 1) with shape parameter 1.5 and scale parameter 1*,
- (e) *Exponential(1.5)*,
- (f) *Erlang(2, 1.5)*,
- (g) *Hyper-exponential  $0.75\text{Exponential}(2) + 0.25\text{Exponential}(4)$ , and*

(h) *Lognormal*(0.4, 0.3).

Notice that scenarios (e)–(h) have workload greater than 1 and the queue length approaches infinity if no abandonment occurs. The patience time of each customer in the system is independently generated from the specified exponential distribution. The simulation runs are taken to be  $T = 100000$  time units.

Recall the estimation procedure described in (9). The estimate of  $\mathbb{P}(W_o > x | W_o > 0)$  has stochastic variations, and is denoted as  $p(W_o > x | W_o > 0)$  to distinguish from the true value. The standard error associated with  $p(W_o > x | W_o > 0)$  is obtained by

$$\sqrt{p(W_o > x | W_o > 0)(1 - p(W_o > x | W_o > 0)) / \mathbb{P}(R > x)}.$$

We then apply the gamma approximation method explained in Section 2.1, and approximate the tail probability  $\mathbb{P}(W_o > x | W_o > 0)$  by  $\bar{F}_{\hat{\alpha}, \hat{\beta}}(x)$ , where  $\bar{F}_{\hat{\alpha}, \hat{\beta}}(x)$  denotes the tail probability of gamma distribution with parameters  $\hat{\alpha}$  and  $\hat{\beta}$ . The stochastic variability of  $F_{\hat{\alpha}, \hat{\beta}}(x)$  can be evaluated by Taylor linearization or resampling method like bootstrap, see EFRON and TIBSHIRANI (1993). The bootstrap method requires less hand calculations and is readily adapted to other more complicated estimators, thus is chosen for our analysis. To assess the variation of  $F_{\hat{\alpha}, \hat{\beta}}(x)$ , we propose the following bootstrap procedure:

- (S1) Specify the number of bootstrap resamples as  $B$  (we take  $B = 1000$  here). For  $b = 1, \dots, B$ :
- (a) Denote the number of waiting customers as  $n_a$ , and draw a simple random with replacement sample of size  $n_a$  from all waiting customers to obtain the sequence of  $W_a^{(b)}$ .
  - (b) Set the gamma parameters  $\alpha^{(b)}$  and  $\beta^{(b)}$  using (7) and (8) based on the resampled waiting times  $W_a^{(b)}$ . Then calculate  $F_{\hat{\alpha}^{(b)}, \hat{\beta}^{(b)}}(x)$  given each pair of  $\alpha^{(b)}$  and  $\beta^{(b)}$ .
- (S2) Approximate the standard error of  $F_{\hat{\alpha}, \hat{\beta}}(x)$  by the standard deviation among  $F_{\hat{\alpha}^{(b)}, \hat{\beta}^{(b)}}(x)$ 's, or alternatively,  $\sqrt{\sum_{b=1}^B (F_{\hat{\alpha}^{(b)}, \hat{\beta}^{(b)}}(x) - F_{\hat{\alpha}, \hat{\beta}}(x))^2 / (B - 1)}$ . The two choices usually lead to similar results, with the latter estimate slightly greater than the former due to the fact that  $F_{\hat{\alpha}, \hat{\beta}}(x)$  is nonlinear in the estimated parameters. We have adopted the former approach in simulation.

Our simulation results are presented in Tables 1, 2 and Figure 1. The two tables show service time distributions (a)–(d), with both exact and approximated tail probabilities along with the associated standard errors under *Exponential*(2) (Table 1) and *Exponential*(4) (Table 2) patience time distributions. By comparing the approximation error with the standard errors, one can tell whether the numerical difference is attributed to simulation error or inherent in the estimates. The relative approximation error under scenarios (e)–(h) are illustrated in Figure 1.

Notice that scenarios (a) and (c) satisfy the tail behavior requirement  $\overline{B}(x) \sim x^{-\theta_1} e^{-\theta_2 x}$ , and the gamma distribution approximates the conditional distribution of  $W_o$  very well. The differences between the approximated probabilities and true probabilities are mostly not significant compared with the random variations associated with each quantity. For scenario (b) with deterministic patience time, gamma approximation behaves reasonably well, even though the tail of service time distribution is zero. In scenario (d) where the tail of service time distribution is  $\overline{B}(x) \sim e^{-x^{1.5}}$  and lighter than the required tail behavior  $\overline{B}(x) \sim x^{-\theta_1} e^{-\theta_2 x}$ , the approximated probabilities are close to the exact values in the far tail. This observation holds under both patience time distributions.

Figure 1 presents our simulation results under  $M/G/1+M$  with service time distributions (e)–(h), where the workload is greater than one. The simulation results confirm the practical use of gamma approximation under various service time distributions. The “exact” tail probabilities can be approximated by those of gamma distributions reasonably well for scenarios (e)–(g), and the maximum relative error is between  $-10\%$  and  $10\%$ . The numerical difference between the exact and approximated probabilities are mostly insignificant when compared to the variations involved in the randomness of sampling. But the performance of gamma approximation deteriorates when the service time distribution is lognormal as in scenario (f), as the deviations between true and approximated probabilities are not only attributed to simulation errors. This is an indication that the performance of gamma approximation may vary according to the tail behavior of service time distribution.

Table 1: Conditional probabilities  $P(W_o > x | W_o > 0)$  under  $Exp(2)$  patience time distribution; “Exact” refers to the true tail probability and “Gamma” refers to the tail probability obtained by gamma approximation; the number in the bracket is the standard error associated with the quantity above and quantifies the Monte Carlo error.

$x$	(a) $Exp(0.4)$		(b) $Deterministic\ 0.8$		(c) $Exp(0.3) + 0.5$		(d) $Weibull(1.5, 1)$	
	Exact	Gamma	Exact	Gamma	Exact	Gamma	Exact	Gamma
0.3	0.593 (0.0029)	0.573 (0.0030)	0.812 (0.0018)	0.810 (0.0021)	0.808 (0.0018)	0.800 (0.0022)	0.839 (0.0017)	0.851 (0.0020)
0.6	0.340 (0.0033)	0.312 (0.0026)	0.567 (0.0026)	0.527 (0.0024)	0.564 (0.0026)	0.543 (0.0024)	0.673 (0.0025)	0.656 (0.0025)
0.9	0.190 (0.0032)	0.167 (0.0021)	0.321 (0.0029)	0.310 (0.0020)	0.370 (0.0030)	0.344 (0.0021)	0.522 (0.0030)	0.486 (0.0025)
1.2	0.104 (0.0029)	0.089 (0.0016)	0.188 (0.0028)	0.172 (0.0016)	0.228 (0.0030)	0.210 (0.0018)	0.388 (0.0034)	0.352 (0.0023)
1.5	0.057 (0.0025)	0.047 (0.0011)	0.095 (0.0024)	0.092 (0.0012)	0.135 (0.0029)	0.125 (0.0015)	0.278 (0.0037)	0.251 (0.0021)
1.8	0.029 (0.0021)	0.025 (0.0007)	0.047 (0.0020)	0.048 (0.0008)	0.077 (0.0026)	0.073 (0.0011)	0.193 (0.0038)	0.176 (0.0019)
2.1	0.014 (0.0018)	0.013 (0.0005)	0.021 (0.0016)	0.024 (0.0005)	0.042 (0.0023)	0.042 (0.0008)	0.131 (0.0037)	0.123 (0.0016)
2.4	0.007 (0.0014)	0.007 (0.0003)	0.009 (0.0012)	0.012 (0.0003)	0.022 (0.0019)	0.024 (0.0006)	0.088 (0.0036)	0.085 (0.0013)
2.7	0.004 (0.0012)	0.004 (0.0002)	0.003 (0.0008)	0.006 (0.0002)	0.011 (0.0016)	0.014 (0.0004)	0.056 (0.0035)	0.059 (0.0011)
3.0	0.001 (0.0009)	0.002 (0.0001)	0.002 (0.0007)	0.003 (0.0001)	0.0005 (0.0013)	0.0008 (0.0002)	0.038 (0.0033)	0.040 (0.0009)

Table 2: Conditional probabilities  $P(W_o > x | W_o > 0)$  under  $Exp(4)$  patience time distribution; “Exact” refers to the true tail probability and “Gamma” refers to the tail probability obtained by gamma approximation; the number in the bracket is the standard error associated with the quantity above and quantifies the Monte Carlo error.

$x$	(a) $Exp(0.4)$		(b) $Deterministic\ 0.8$		(c) $Exp(0.3) + 0.5$		(d) $Weibull(1.5, 1)$	
	Exact	Gamma	Exact	Gamma	Exact	Gamma	Exact	Gamma
0.3	0.619 (0.0084)	0.608 (0.0097)	0.847 (0.0047)	0.851 (0.0049)	0.838 (0.0015)	0.839 (0.0018)	0.879 (0.0013)	0.897 (0.0016)
0.6	0.375 (0.0090)	0.360 (0.0075)	0.634 (0.0067)	0.617 (0.0062)	0.631 (0.0021)	0.620 (0.0022)	0.746 (0.0019)	0.745 (0.0022)
0.9	0.229 (0.0085)	0.212 (0.0065)	0.412 (0.0074)	0.414 (0.0059)	0.455 (0.0024)	0.434 (0.0020)	0.619 (0.0023)	0.598 (0.0023)
1.2	0.137 (0.0075)	0.124 (0.0055)	0.283 (0.0073)	0.266 (0.0051)	0.320 (0.0024)	0.294 (0.0018)	0.502 (0.0025)	0.470 (0.0022)
1.5	0.075 (0.0062)	0.073 (0.0044)	0.175 (0.0067)	0.166 (0.0042)	0.220 (0.0023)	0.195 (0.0015)	0.399 (0.0027)	0.364 (0.0020)
1.8	0.041 (0.0050)	0.043 (0.0034)	0.110 (0.0059)	0.101 (0.0033)	0.146 (0.0021)	0.128 (0.0012)	0.314 (0.0027)	0.279 (0.0019)
2.1	0.024 (0.0042)	0.025 (0.0025)	0.063 (0.0050)	0.061 (0.0024)	0.092 (0.0019)	0.083 (0.0010)	0.240 (0.0027)	0.212 (0.0017)
2.4	0.014 (0.0034)	0.014 (0.0017)	0.037 (0.0041)	0.036 (0.0018)	0.057 (0.0016)	0.053 (0.0007)	0.180 (0.0026)	0.160 (0.0015)
2.7	0.006 (0.0024)	0.008 (0.0012)	0.019 (0.0033)	0.021 (0.0012)	0.034 (0.0014)	0.034 (0.0006)	0.133 (0.0025)	0.121 (0.0013)
3.0	0.004 (0.0020)	0.005 (0.0008)	0.009 (0.0025)	0.012 (0.0008)	0.019 (0.0011)	0.021 (0.0004)	0.099 (0.0024)	0.090 (0.0011)

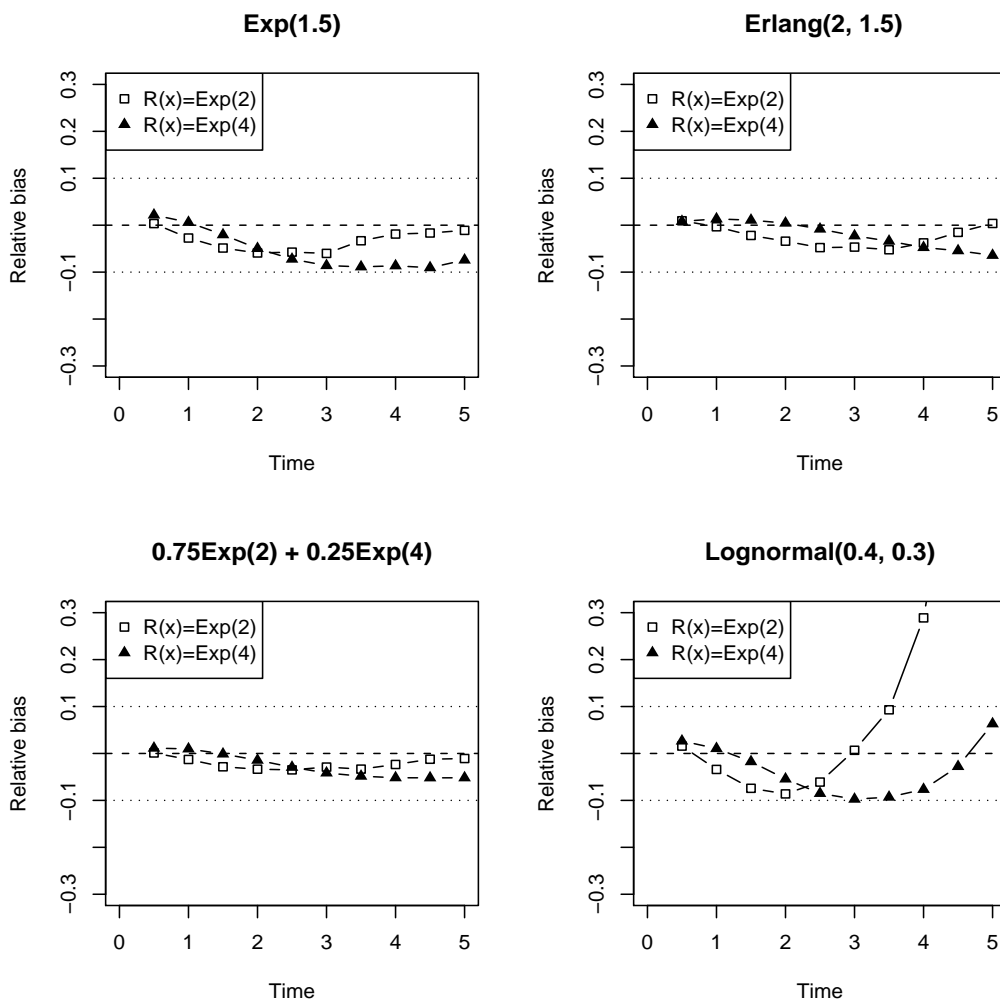


Figure 1: Relative bias of gamma approximated tail probabilities for *virtual* waiting distribution, under two different patience time distributions,  $Exp(2)$  and  $Exp(4)$ . The relative bias is computed as the ratio of the approximated tail probability to the exact probability less one. The dash and two dotted lines represent 0, and  $\pm 10\%$ , respectively.

## 4 Conclusion

We have described a novel methodology for approximating waiting time distributions for single-server queueing systems in the presence of customer abandonment. The key contribution of the paper is to propose a gamma approximation to the distribution of *virtual* waiting times, i.e., the amount of time a customer needs to wait before receiving the service. Due to customer abandonment, the *actual* waiting times are subject to censoring, and its distribution can not be directly approximated with precision. But the *virtual* waiting time distribution can be well approximated by gamma distribution, as shown in the preceding sections.

The validity of the gamma approximation has been confirmed by numerical studies. But the practitioners should be cautioned of applying gamma approximation blindly, since the tail behavior of service time distribution as well as the relative magnitude of system input parameters affect the performance of gamma approximation. Although we have conducted extensive simulations to examine the goodness-of-fit, rigorous theoretical work is certainly warranted as guidelines for the proposed methodology. Alternative distributions like the Weibull, lognormal, or mixture of gamma distributions can be potentially used for approximating the *virtual* waiting time distribution and it would be of particular interest to study the relative performance of various distribution families in our present context.

The present paper focuses on exponentially distributed patience time distributions, but our methodology is readily applicable to the  $M/G/1 + G$  queue with general patience time distribution, where both theoretical and empirical work are needed to better explore the main idea.

## Appendix

**Proof of Theorem 1:** Notice that

$$r(t) = \frac{\sum_{n=1}^{A(t)} 1_{\{R_n \leq W_{o,n}\}}}{\sum_{n=1}^{A(t)} \min(R_n, W_{o,n})},$$

where  $A(t)$  denotes the total number of arriving customers by time  $t$ . In view of renewal reward theorem (cf. Theorem 3.6.1 in ROSS (1996)), we have with probability 1 that

$$r(t) = \frac{\frac{1}{t} \sum_{n=1}^{A(t)} 1_{\{R_n \leq W_{o,n}\}}}{\frac{1}{t} \sum_{n=1}^{A(t)} \min(R_n, W_{o,n})} \rightarrow \frac{\mathbb{P}(R \leq W_o)}{\mathbb{E} \min(R, W_o)} \quad \text{as } t \rightarrow \infty,$$

and also

$$\mathbb{E}[r(t)] \rightarrow \frac{\mathbb{P}(R \leq W_o)}{\mathbb{E} \min(R, W_o)} \quad \text{as } t \rightarrow \infty.$$

It remains to show

$$\frac{\mathbb{P}(R \leq W_o)}{\mathbb{E} \min(R, W_o)} = \frac{1}{\mathbb{E}[R]}.$$

Assuming  $W_o$  follows  $\Gamma(\alpha_W, \beta_W)$  we have

$$\begin{aligned}
E \min(R, W_o) &= \int_0^\infty (1 - F_{\min(R, W_o)}(t)) dt \\
&= \int_0^\infty \mathbb{P}(R \geq t) \mathbb{P}(W_o \geq t) dt \\
&= \int_0^\infty e^{-\frac{t}{\gamma}} \left[ \int_t^\infty f_{\Gamma(\alpha_W, \beta_W)}(y) dy \right] dt \\
&= \int_0^\infty f_{\Gamma(\alpha_W, \beta_W)}(y) \left[ \int_0^y e^{-\frac{t}{\gamma}} dt \right] dy \\
&= \frac{\gamma}{\Gamma(\alpha_W) \beta_W^{\alpha_W}} \left[ \int_0^\infty e^{-\frac{y}{\beta_W}} y^{\alpha_W-1} (1 - e^{-\frac{y}{\gamma}}) dy \right] \\
&= \frac{\gamma}{\Gamma(\alpha_W) \beta_W^{\alpha_W}} \left[ \Gamma(\alpha_W) \beta_W^{\alpha_W} - \Gamma(\alpha_W) \left( \frac{\gamma \beta_W}{\gamma + \beta_W} \right)^{\alpha_W} \right] \\
&= \gamma \left( 1 - \left( \frac{\gamma}{\gamma + \beta_W} \right)^{\alpha_W} \right).
\end{aligned}$$

On the other hand, we have that

$$\begin{aligned}
\mathbb{P}(R \leq W_o) &= \int_0^\infty \int_0^y f_R(t) f_{\Gamma(\alpha_W, \beta_W)}(y) dt dy \\
&= \int_0^\infty [1 - e^{-\frac{y}{\gamma}}] f_{\Gamma(\alpha_W, \beta_W)}(y) dy \\
&= 1 - \frac{1}{\Gamma(\alpha_W) \beta_W^{\alpha_W}} \int_0^\infty y^{\alpha_W-1} e^{-\left(\frac{1}{\gamma} + \frac{1}{\beta_W}\right)y} dy \\
&= 1 - \frac{1}{\Gamma(\alpha_W) \beta_W^{\alpha_W}} \Gamma(\alpha_W) \left( \frac{\gamma \beta_W}{\gamma + \beta_W} \right)^{\alpha_W} \\
&= 1 - \left( \frac{\gamma}{\gamma + \beta_W} \right)^{\alpha_W}.
\end{aligned}$$

This completes the proof. ■

## Acknowledgments

The authors are grateful to the associate editor and the anonymous referees for carefully examining the paper and providing a number of important comments that led to several improvements. The second author would like to thank the research support from the Cross-Sector Research in Residence Program between the National Institute of Statistical Sciences (NISS) and National Agricultural Statistics Service (NASS).

## References

- ANCKER, C. J. and A. V. GAFARIAN (1963). Queuing with reneging and multiple heterogeneous servers. *Naval Research Logistics Quarterly*, 10, 125–149.
- BACELLI, F., P. BOYER, and G. HÉBUTERNE (1984). Single-server queues with impatient customers. *Advances in Applied Probability*, 16(4), 887–905.
- BACELLI, F. and G. HÉBUTERNE (1981). On queues with impatient customers. In *Performance '81 (Amsterdam, 1981)*, pp. 159–179. Amsterdam: North-Holland.
- BAE, J., S. KIM, and E. Y. LEE (2001). The virtual waiting time of the  $M/G/1$  queue with impatient customers. *Queueing Systems: Theory and Applications*, 38(4), 485–494.
- COHEN, J. W. (1977). On up- and downcrossings. *Journal of Applied Probability*, 14(2), 405–410.
- DALEY, D. J. (1965). General customer impatience in the queue  $GI/G/1$ . *Journal of Applied Probability*, 2, 186–205.
- DE KOK, A. G. and H. C. TIJMS (1985). A queueing system with impatient customers. *Journal of Applied Probability*, 22(3), 688–696.
- DEN ISEGER, P., M. SMITH, and R. DEKKER (1997). Computing compound distributions faster! *Insurance: Mathematics and Economics*, 20, 23–34.
- DOSHI, B. (1992). Level-crossing analysis of queues. In *Queueing and related models*, Volume 9 of *Oxford Statistical Science Series*, pp. 3–33. Oxford University Press, New York.
- EFRON, B. and R. TIBSHIRANI (1993). *An Introduction to the Bootstrap*. Chapman and Hall Ltd., New York.
- IRAVANI, F. and B. BALCIOĞLU (2008). Approximations for the  $M/GI/N + GI$  type call center. *Queueing Systems: Theory and Applications*, 58(2), 137–153.
- NETRAVALI, A. N. (1973). Spline approximation to the solution of the Volterra integral equation of the second kind. *Mathematics of Computation*, 27, 99–106.
- NOBEL, R. D. and H. C. TIJMS (2006). Waiting-time probabilities in the  $M/G/1$  retrial queue. *Statistica Neerlandica*, 60(1), 73–78.
- OLVER, F., D. LOZIER, R. BOISVERT, and C. W. CLARK (2010). *NIST Handbook of Mathematical Functions: Companion to the Digital Library of Mathematical Functions*. Cambridge University Press, New York.
- ROSS, S. M. (1996). *Stochastic processes* (Second edition). Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley and Sons, New York.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons, New York.

STANFORD, R. E. (1979). Reneging phenomena in single channel queues. *Mathematics of Operations Research*, 4(2), 162–178.