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[†]
National Institute of Statistical Sciences
Research Triangle Park
NC, USA

July 31, 2009

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. Simulation study is conducted to assess the accuracy of the approximation and it reveals the approximation performs satisfactorily under general condition on service time distributions.

Keywords and Phrases: impatient customers, waiting time distribution.

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 the arrival process. Let W_o denote the time that a customer needs to wait before receiving a

---

[*] chihoon@stat.colostate.edu
[†] jqwang@niss.org

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 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 TIJIMS (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 $I\!P(W_{\mathrm{o}} \leq x | W_{\mathrm{o}} > 0)$ for $x > 0$. The idea is based on using gamma distribution via matching the first and second moments of the actual waiting time. This approximation is discussed in Section 2.1 and numerical results in Section 3 indicate it performs satisfactorily on different service time distributions.

2 Approximation for waiting time probabilities

The waiting time probability distribution $I\!P(W_{\mathrm{o}} \leq x)$ 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

$$F(x) = P_0 + (1 - P_0) I\!P(W_{\mathrm{o}} \leq x | W_{\mathrm{o}} > 0), \ x \geq 0.$$

It suffices to calculate the system idle probability P_0 and approximate the conditional waiting time distribution $I\!P(W_{\mathrm{o}} \leq x | W_{\mathrm{o}} > 0), x > 0$. We note that in many practical cases the only available data are the *actual* waiting times that are right-censored: $W_a \equiv \min\{W_{\mathrm{o}}, R\}$. The main idea is to approximate $I\!P(W_{\mathrm{o}} \leq x | W_{\mathrm{o}} > 0)$ by the gamma distribution by calculating the first two moments of W_a and obtaining a set of parameter estimates for the gamma distribution. Details will be provided in Section 2.1.

NOBEL and TIJIMS (2006) deals with waiting time probabilities for $M/G/1$ retrial queue, inspiring us to think about waiting time approximation problem for queueing system with impatient customers. In practice, the gamma distribution is a flexible life distribution model that may offer a good fit to some 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 $\mathrm{E}[X] = \alpha\beta$, $\mathrm{Var}[X] = \alpha\beta^2$. For instance, the gamma distribution arises in situations where one is concerned about the waiting time for a finite number ($\alpha \in I\!N$) of independent events to occur, assuming that events occur at a constant rate $1/\beta > 0$. The distribution $\Gamma(k, \frac{1}{\lambda})$ 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 $\overline{B}(x)$ and $\overline{R}(x)$, 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 $\{W_{\text{o}}(t) : t \geq 0\}$ 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 $\lambda h P_0 \overline{B}(x)$ 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 $\lambda h \int_0^x \overline{B}(x-t)\overline{R}(t) f(t) dt$. The level-crossing method as described by COHEN (1977) and DOSHI (1992) implies that the conservation law holds in the long run:

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

Dividing both sides in (1) by h and letting $h \to 0$ we get

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

and the normalizing equation

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

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 (49) on page 175 of STANFORD (1979) provides an explicit formula for $f(x)$. See also BAE et al. (2001) and DE KOK and TIJIMS (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

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

where for any function $g(x)$, $\widetilde{g}(s) \equiv \int_0^\infty e^{-sx} dg(x)$. To obtain $\mathbb{P}(W_{\text{o}} \leq x | W_{\text{o}} > 0), x > 0$ one needs to evaluate the infinite series in (4) and invert $\widetilde{f}(s)$ at required x values, which in practice is quite difficult procedure to perform both analytically and numerically.

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 procedure we present below is similar to that of IRAVANI and BALCIOĞLU (2008). One can employ other discretization methods such as in NETRAVALI (1973), DEN ISEGER et al. (1997). We first specify a large enough integer N and small interval length Δ, such that the density of W_{o} beyond $N\Delta$ is negligible. Then we apply the trapezoidal rule to (2)-(3) at grid points $(0, \Delta, \ldots, N\Delta)$ and obtain the following linear equation system:

$$A\mathbf{f} = \mathbf{b}, \tag{5}$$

where $\mathbf{b}$ is $(N+1)\times 1$ vector with first element λ and the rest being 0, $\mathbf{f} = (f(0), f(\Delta), \ldots, f(N\Delta))^T$ represents the density function $f(x)$ evaluated at each grid point $0, \Delta, \ldots, N\Delta$. The matrix A is $(N+1)$-dimensional square matrix. We denote the k-th row of A as $\mathbf{a}_k$, then

$$\begin{aligned}
\mathbf{a}_1 &= \left(1 + \frac{\lambda\Delta}{2}, \lambda\Delta, \ldots, \lambda\Delta, \frac{\lambda\Delta}{2}\right), \\
\mathbf{a}_k &= \lambda\Delta\left(\frac{1}{2}\overline{R}(0)\overline{B}(k\Delta), \overline{R}(\Delta)\overline{B}((k-1)\Delta), \ldots, \overline{R}((k-1)\Delta)\overline{B}(\Delta), \frac{1}{2}\overline{R}(k\Delta) - 1, 0, \ldots, 0\right) \\
&\quad +\mathbf{v},
\end{aligned}$$

for $k = 2, \ldots, N+1$, where $\mathbf{v}$ is a vector of 0's except the first element being $\overline{B}(k\Delta)$ and the $(k+1)$-st element being -1. Solve the equation system (5) to obtain $\mathbf{f} = A^{-1}\mathbf{b}$, which gives the conditional density of W_{o} evaluated at each grid point. Then we can calculate the system idle probability as $P_0 = \frac{f(0)}{\lambda}$. We point out that both Laplace transform (4) and numerical approximation (5) approaches are sufficiently complex and computationally expensive, and must be adapted to specific service time distributions. As a practical alternative, we propose a simple approximation scheme of 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 $x \to \infty$, $\overline{B}(x) \sim x^{-\theta_1} e^{-\theta_2 x}$ for some $\theta_1, \theta_2 \in (0, \infty)$, where $f(x) \sim g(x)$ means $\lim_{x\to\infty} \frac{f(x)}{g(x)} = c$ for some finite constant c. Then choosing an appropriate set of parameters θ_1, θ_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. Hence for large $x > 0$, one can anticipate that the solution $f(x)$ of (2) can be well approximated by the density of gamma random variable. Recall the *actual* waiting time $W_a \equiv \min\{W_{\mathrm{o}}, R\}$. Let $\gamma = \mathrm{E}[R]$ so that R is $\Gamma(1, \gamma)$ random variable. Then one can check by assuming W_{o} follows $\Gamma(\alpha, \beta)$ that

$$\mathrm{E}[W_a] = \gamma\left(1 - \left(\frac{\gamma}{\gamma+\beta}\right)^{\alpha}\right), \tag{6}$$

$$\mathrm{E}[W_a^2] = 2\gamma^2\left(1 - \left(\frac{\gamma}{\gamma+\beta}\right)^{\alpha}\right) - 2\gamma\alpha\beta\left(\frac{\gamma}{\gamma+\beta}\right)^{\alpha+1}. \tag{7}$$

We use (6) and (7) to set the parameters α and β of the gamma distribution. More specifically, we replace $\mathrm{E}[W_a]$ and $\mathrm{E}[W_a^2]$ by sample moments $\overline{W}_a$ and $\frac{1}{N}\sum_{n=1}^{N} W_{a,n}^2$, respectively, and seek to solve equations (6) and (7). It is often difficult to get the solution to the estimating equations to converge in the form of (6) and (7). The reason for this is that the shape parameter α is in the exponent and a small change in α may result in magnified change in the exponentiation. A reparameterization fixes the convergence problem. We define $\eta \equiv \log\left(\frac{\gamma}{\gamma+\beta}\right)$, then one can show that the aforementioned estimating equations are equivalent to the following:

$$\begin{aligned} \log\left(1 - \frac{\overline{W}_a}{\gamma}\right) &= \alpha\eta && (8) \\ C + \alpha\gamma(1 - \exp(\eta)) &= 0, && (9) \end{aligned}$$

where $C = \frac{\frac{1}{N}\sum_{n=1}^{N} W_{a,n}^2 - 2\gamma\overline{W}_a}{2(\gamma - \overline{W}_a)}$. The two equations (8) and (9) in α and η are always solvable and have a unique solution, and then we can transform back to estimated β after solving for α and η. We denote the estimated set of 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 idea of moment matching gamma approximation is not restricted to $M/G/1+M$ queuing systems, but can be readily generalized to arbitrary known patience time distribution. In the general case, one can analytically derive the first and second moment of W_a as functions of gamma parameters α, β 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 (6) and (7) 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 γ^{-1} given data up to time $t \in (0, \infty)$ by

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

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 γ^{-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, $r(t) \to \gamma^{-1}$, as $t \to \infty$,*
- *(ii) $\mathrm{E}[r(t)] \to \gamma^{-1}$, as $t \to \infty$.*

Proof. See Appendix. ■

3 Numerical study and discussion

In this section, we provide empirical 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 use six different distributions for service time: (a) $Exp(0.4)$, (b) deterministic distribution with service time fixed at 0.8, (c) $Exp(1.5)$, (d) $Erlang(2, 1.5)$, (e) hyper-exponential $0.75Exp(2) + 0.25Exp(4)$ and (f) $Lognormal(0.4, 0.3)$. Notice that scenarios (c)-(f) 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 $Exp(2)$. The simulation runs are taken to be $T = 100000$ time units.

In simulation, the true tail probability of offered waiting time $I\!P(W_o > x|W_o > 0)$ is computed by taking the proportion of customers with actual waiting time greater than x, and then use the formula $I\!P(W_o > x|W_o > 0) = I\!P(W_a > x|W_a > 0)/I\!P(R > x)$. The estimate of $I\!P(W_o > x|W_o > 0)$ 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 $\sqrt{p(W_o > x|W_o > 0)(1 - p(W_o > x|W_o > 0))}/I\!P(R > x)$.

We then apply the gamma approximation method explained in Section 2.1, and approximate the tail probability $I\!P(W_o > x|W_o > 0)$ by $\overline{F}_{\hat{\alpha},\hat{\beta}}(x)$. 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 Taylor linearization requires to estimate the variance of $\hat{\alpha}, \hat{\beta}$ as solution to (6) and (7), and then linearize $F_{\hat{\alpha},\hat{\beta}}(x)$ using the first derivative of $F_{\alpha,\beta}(x)$ with respect to the parameters. The bootstrap method requires less hand calculations and is readily adapted to other more complicated estimators. To assess the variation of $F_{\hat{\alpha},\hat{\beta}}(x)$, we adopt the bootstrap procedure and here are the detailed steps:

(S1) Consider the actual waiting time W_a of all n_a waiting customers, and draw a simple random sample of size n_a with replacement to obtain the sequence of W_a^*.

(S2) Set the gamma parameters α^* and β^* using (6) and (7) based on the resampled waiting times W_a^*. Then calculate $F_{\hat{\alpha}^*,\hat{\beta}^*}(x)$ given each pair of α^* and β^*.

(S3) Repeat (S1) and (S2) B times (we take $B = 1000$ here), and approximate the standard error of $F_{\hat{\alpha},\hat{\beta}}(x)$ by the standard deviation among $F_{\hat{\alpha}^*,\hat{\beta}^*}(x)$'s.

Tables 1 and 2 present our simulation results under $M/G/1+M$ with six service time distributions under $Exp(2)$ patience time distribution. Scenarios (a) and (c)-(e) 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 nonsignificant compared with the stochastic 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. 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 solely explained by standard errors. This is an indication that the performance of gamma approximation may vary according to the tail behavior of service time distribution.

Tables 3 and 4 present our simulation results under $M/G/1 + M$ with $Exp(4)$ patience time distribution. 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 all scenarios (a)-(e), especially scenarios (a) and (b). A close examination of the farther tail probabilities indicates satisfactory performance of gamma approximation, e.g., the "exact" tail probabilities at $x = 5$ are very well approximated by those of gamma distributions for scenarios (c)-(e).

Table 1: Conditional probabilities $I\!P(W_o > x|W_o > 0)$ 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.

	$Exp(0.4)$		*Deterministic* 0.8	
x	Exact	Gamma	Exact	Gamma
0.3	0.593	0.573	0.812	0.810
	(0.0029)	(0.0030)	(0.0018)	(0.0021)
0.6	0.340	0.312	0.567	0.527
	(0.0033)	(0.0026)	(0.0026)	(0.0024)
0.9	0.190	0.167	0.321	0.310
	(0.0032)	(0.0021)	(0.0029)	(0.0020)
1.2	0.104	0.089	0.188	0.172
	(0.0029)	(0.0016)	(0.0028)	(0.0016)
1.5	0.057	0.047	0.095	0.092
	(0.0025)	(0.0011)	(0.0024)	(0.0012)
1.8	0.029	0.025	0.047	0.048
	(0.0021)	(0.0007)	(0.0020)	(0.0008)
2.1	0.014	0.013	0.021	0.024
	(0.0018)	(0.0005)	(0.0016)	(0.0005)
2.4	0.007	0.007	0.009	0.012
	(0.0014)	(0.0003)	(0.0012)	(0.0003)
2.7	0.004	0.004	0.003	0.006
	(0.0012)	(0.0002)	(0.0008)	(0.0002)
3.0	0.001	0.002	0.002	0.003
	(0.0009)	(0.0001)	(0.0007)	(0.0001)

Table 2: Conditional probabilities $\mathbb{P}(W_{\mathrm{o}} > x | W_{\mathrm{o}} > 0)$ 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.

	$Exp(1.5)$		$Erlang(2, 1.5)$		$0.75Exp(2) + 0.25Exp(4)$		$Lognormal(0.4, 0.3)$	
x	Exact	Gamma	Exact	Gamma	Exact	Gamma	Exact	Gamma
0.5	0.868	0.871	0.957	0.966	0.936	0.937	0.891	0.905
	(0.0015)	(0.0021)	(0.0008)	(0.0014)	(0.0010)	(0.0017)	(0.0014)	(0.0018)
1.0	0.732	0.712	0.897	0.894	0.858	0.847	0.734	0.709
	(0.0026)	(0.0026)	(0.0016)	(0.0026)	(0.0019)	(0.0025)	(0.0025)	(0.0027)
1.5	0.596	0.567	0.823	0.805	0.775	0.753	0.552	0.511
	(0.0037)	(0.0028)	(0.0026)	(0.0032)	(0.0029)	(0.0029)	(0.0036)	(0.0027)
2.0	0.473	0.445	0.735	0.710	0.686	0.663	0.383	0.350
	(0.0048)	(0.0029)	(0.0039)	(0.0036)	(0.0042)	(0.0032)	(0.0045)	(0.0025)
2.5	0.366	0.345	0.649	0.618	0.600	0.579	0.246	0.231
	(0.0059)	(0.0030)	(0.0054)	(0.0039)	(0.0057)	(0.0036)	(0.0051)	(0.0022)
3.0	0.283	0.266	0.557	0.531	0.517	0.502	0.148	0.149
	(0.0071)	(0.0030)	(0.0073)	(0.0040)	(0.0075)	(0.0040)	(0.0054)	(0.0018)
3.5	0.211	0.204	0.478	0.453	0.449	0.434	0.086	0.094
	(0.0083)	(0.0028)	(0.0094)	(0.0042)	(0.0095)	(0.0043)	(0.0055)	(0.0015)
4.0	0.159	0.156	0.398	0.383	0.383	0.374	0.045	0.058
	(0.0095)	(0.0026)	(0.0118)	(0.0043)	(0.0120)	(0.0045)	(0.0052)	(0.0011)
4.5	0.120	0.118	0.327	0.322	0.325	0.321	0.021	0.036
	(0.0109)	(0.0024)	(0.0145)	(0.0043)	(0.0148)	(0.0046)	(0.0046)	(0.0008)
5.0	0.091	0.090	0.268	0.269	0.278	0.275	0.009	0.022
	(0.0123)	(0.0021)	(0.0176)	(0.0043)	(0.0182)	(0.0047)	(0.0040)	(0.0006)

Table 3: Conditional probabilities $\mathbb{P}(W_{\mathrm{o}} > x | W_{\mathrm{o}} > 0)$ 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.

	$Exp(0.4)$		$Deterministic$ 0.8	
x	Exact	Gamma	Exact	Gamma
0.3	0.619	0.608	0.847	0.851
	(0.0084)	(0.0097)	(0.0047)	(0.0049)
0.6	0.375	0.360	0.634	0.617
	(0.0090)	(0.0075)	(0.0067)	(0.0062)
0.9	0.229	0.212	0.412	0.414
	(0.0085)	(0.0065)	(0.0074)	(0.0059)
1.2	0.137	0.124	0.283	0.266
	(0.0075)	(0.0055)	(0.0073)	(0.0051)
1.5	0.075	0.073	0.175	0.166
	(0.0062)	(0.0044)	(0.0067)	(0.0042)
1.8	0.041	0.043	0.110	0.101
	(0.0050)	(0.0034)	(0.0059)	(0.0033)
2.1	0.024	0.025	0.063	0.061
	(0.0042)	(0.0025)	(0.0050)	(0.0024)
2.4	0.014	0.014	0.037	0.036
	(0.0034)	(0.0017)	(0.0041)	(0.0018)
2.7	0.006	0.008	0.019	0.021
	(0.0024)	(0.0012)	(0.0033)	(0.0012)
3.0	0.004	0.005	0.009	0.012
	(0.0020)	(0.0008)	(0.0025)	(0.0008)

Table 4: Conditional probabilities $\mathbb{P}(W_o > x | W_o > 0)$ 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.

	$Exp(1.5)$		$Erlang(2, 1.5)$		$0.75Exp(2) + 0.25Exp(4)$		$Lognormal(0.4, 0.3)$	
x	Exact	Gamma	Exact	Gamma	Exact	Gamma	Exact	Gamma
0.5	0.920	0.940	0.988	0.996	0.970	0.981	0.942	0.967
	(0.0010)	(0.0013)	(0.0004)	(0.0003)	(0.0006)	(0.0008)	(0.0009)	(0.0008)
1.0	0.831	0.836	0.966	0.979	0.930	0.939	0.857	0.866
	(0.0016)	(0.0020)	(0.0007)	(0.0010)	(0.0011)	(0.0017)	(0.0015)	(0.0019)
1.5	0.738	0.723	0.936	0.946	0.885	0.884	0.744	0.731
	(0.0022)	(0.0024)	(0.0011)	(0.0018)	(0.0015)	(0.0023)	(0.0021)	(0.0024)
2.0	0.645	0.613	0.895	0.899	0.835	0.823	0.625	0.591
	(0.0027)	(0.0025)	(0.0016)	(0.0026)	(0.0020)	(0.0027)	(0.0026)	(0.0025)
2.5	0.553	0.513	0.849	0.842	0.781	0.758	0.505	0.462
	(0.0032)	(0.0024)	(0.0021)	(0.0031)	(0.0025)	(0.0030)	(0.0031)	(0.0023)
3.0	0.465	0.425	0.795	0.777	0.722	0.692	0.390	0.352
	(0.0036)	(0.0024)	(0.0027)	(0.0034)	(0.0031)	(0.0031)	(0.0034)	(0.0021)
3.5	0.383	0.349	0.735	0.710	0.661	0.629	0.290	0.263
	(0.0040)	(0.0024)	(0.0034)	(0.0035)	(0.0037)	(0.0032)	(0.0036)	(0.0019)
4.0	0.312	0.285	0.673	0.641	0.599	0.568	0.209	0.193
	(0.0043)	(0.0022)	(0.0041)	(0.0035)	(0.0043)	(0.0032)	(0.0036)	(0.0016)
4.5	0.254	0.231	0.607	0.574	0.539	0.511	0.144	0.140
	(0.0045)	(0.0021)	(0.0048)	(0.0035)	(0.0050)	(0.0032)	(0.0036)	(0.0014)
5.0	0.202	0.187	0.545	0.510	0.482	0.457	0.095	0.101
	(0.0048)	(0.0020)	(0.0056)	(0.0034)	(0.0057)	(0.0032)	(0.0034)	(0.0012)

Appendix

Proof of Theorem 1: Notice that

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

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

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

and also

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

It remains to show $\frac{\mathbb{P}(R \le W_o)}{\mathrm{E}\min(R,W_o)} = \frac{1}{\mathrm{E}[R]}$. Assuming W_o follows $\Gamma(\alpha_W, \beta_W)$ we have

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

On the other hand,

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

This completes the proof. ■

References

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 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. TIJIMS (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 Statist. Sci. Ser.*, pp. 3–33. New York: Oxford Univ. Press.

EFRON, B. and R. TIBSHIRANI (1993). *An Introduction to the Bootstrap.* Chapman & Hall Ltd.

IRAVANI, F. and B. BALCIOĞLU (2008). Approximations for the $M/GI/N+GI$ type call center. *Queueing Systems 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. TIJIMS (2006). Waiting-time probabilities in the $M/G/1$ retrial queue. *Statistica Neerlandica 60*(1), 73–78.

ROSS, S. M. (1996). *Stochastic processes* (Second ed.). Wiley Series in Probability and Statistics: Probability and Statistics. New York: John Wiley & Sons Inc.

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