Queueing Models and
Service Management
Vol. 2, No. 2, page 154-172, 2019

QMSM

The Stationary Distribution of the M/M/1 Queue with Batch Reneging Triggered by Line Cutting

Nicholas O. Kirby* and Matthew O. Jones

Department of Mathematics and Statistics
Austin Peay State University
Clarksville, TN 37044, USA
(Received June 2018 ; accepted June 2019)

Abstract: We consider an extension of the $M/M/1$ queue where customers join the queue at uniformly random positions upon arrival. Immediately after, each customer who has been skipped over by a new arrival (or 'cut') leaves independently with the same probability. We provide necessary and sufficient criteria for stability, as well as closed-form stationary results for special cases, and an approximation method for the general case. We compare limiting distribution simulation output with the approximations for the general case.

Keywords: Foster's criterion, $M/M/1$ queue, priority queue, reneging, stationary distribution.

1. Introduction

Line-cutting occurs naturally in social systems for reasons of carelessness, selfish convenience, or even legitimate need. For example, arriving customers may be oblivious to the line structure and join at random positions. Or, an arrival may attempt to join the queue just behind the position of an acquaintance. Sometimes line cutting may be more acceptable or necessary, for instance, when travelers attempt to cut line in order to avoid missing an airplane flight. Or, patients waiting in line at a doctor's office may allow those in more urgent need of medical care to move toward the head of the line, or a triage protocol may even exist to facilitate this.

Our paper studies an extension of the $M/M/1$ queue where impatient customers jump or *cut* line upon arrival with the hope of receiving faster service. However, this cutting is observed by waiting customers who may become aggravated to the point of reneging. For example, consider three highway lanes. Drivers in the right lane jump into the middle traffic lane at random positions because the right lane is merging with the middle lane. The annoyed middle lane drivers then randomly depart the middle lane for the left lane.

One can view this model as a kind of priority queueing model where new arrivals are assigned uniform random priority ranks with respect to the current queue size. That is, each arrival takes position in the queue according to its rank, which is uniformly distributed on $\{0,1, \ldots,$ *queue size*$\}$, as opposed to being ranked according to previously assigned

*Corresponding author
Email : kirbyn@apsu.edu

independent and identically distributed (i.i.d.) marks. A new arrival that jumps to the head of the line preempts the service in progress. Customers behind the new arrival who have been cut become aggravated and leave, independently, with uniform probability a, whereas customers in line ahead of the arrival are unaffected. Aggravated customers who leave are assumed to do so instantaneously in batches[1]. Thus, our model can be viewed as an $M/M/1$ queue where customers cut line according to some relative priority assignment, and simultaneously trigger random queue-size-dependent reneging in batches.

Reneging models commonly assume customers join the queue with predetermined, i.i.d. maximum waiting times. If the time until service completion exceeds this maximum waiting time, the customer abandons the queue. Such models fail to account for specific reasons that trigger abandonment. Our line cutting model can be viewed as a combination of a kind of priority and reneging, and to our knowledge the study of reneging due to line cutting seems unaddressed in the literature.

The remainder of the paper is organized as follows. Section 2 provides a brief review of related literature. Section 3 gives the model framework and necessary and sufficient conditions for stability. Section 4 contains the stationarity results for the general and special cases. Section 5 includes a brief comparison of the analytical results with simulation output conclusions section, and section 6 concludes the paper.

2. Literature Review

Models of priority and reneging queues appeared in the 1950s, and have since been studied extensively in the literature. Analysis of priority queues seems to have originated with Cobham [9], and basic preemptive and nonpreemptive models appear in the popular texts by Gross *et al.* [13] and Baccelli and Bremaud [6]. For more recent priority models, we refer the reader to Wang *et al.* [29] and the references therein. Priority models typically assume a finite number of fixed priority classes. Our model differs from this in that the effective queue length dependent priorities (the cut positions) are simply the positions in the queue, and are reassigned upon each system change.

Reneging occurs for various reasons, often when customers in the queue become impatient and refuse to wait any longer. This phenomenon is prevalent in many situations including call centers, perishable goods inventory systems, and doctors' offices. Queueing models with reneging likely originated with Haight [14], and have since been studied extensively. The authors in Al-Seedy *et al.* [2] and Ammar [4, 5] give transient analysis results for queues with abandonment. The busy period for the $M/M/1$ with abandonments is studied in Ammar *et al.* [3], and the authors in Dimou *et al.* [10] study server vacations

[1] In reality, departures might occur in sequence over a time interval that is small with respect to the interarrival + service rate.

in abandonment queues. Abandonment in the $M/G/1$ queue is studied in Sherzer and Kerner [27], and Kapodistria [18] considers batch abandonments in the $M/M/1$. Diffusion approximations for reneging queues are studied in Liu [21] and Huang and Zhang [16]. Refer to Hasenbein and Perry [15] for a review of abandonment queue literature. As far as we know, none of the existing literature on abandonment queues accounts for reneging due to line cutting.

In Allon and Hanany [1], the authors describe scenarios where a large, finite population requires service from time to time (say, at a doctor's office). A customer waiting in the queue may allow future arrivals to cut line with the understanding that next time it could be him or her who requires more immediate attention. Their analysis indicates social policy deviation from strict FIFO by allowing cuts can mutually benefit customers as well as queue managers.

In order to analyze time-dependent properties for a FIFO queue with service and arrival rates that increase and decrease with queue size, respectively, Fralix [12] constructed the "knock-out queue". The knock-out queue is similar to our model in that a waiting customer may be removed from the queue upon an arrival. However, only one customer may be removed at a time from the knock-out queue, and the new arrival always jumps to the head of the line.

Decisions of whether or not to abandon a queue often seem to coincide with customers receiving new information (such as another abandonment, service, or opening or closing of a service lane). Empirical studies on emergency room waiting rooms in Batt and Terwiesch [7], Bolandifar *et al.* [8] support this. Call center customers are shown to exhibit the same kind of behavior in Zolar *et al.* [30].

Optimal abandonment policies (if and when to renege at all) are studied in Mandelbaum and Shimkin [22], where it is shown that the $M/M/m$ queue in which customers are unaware of their queue position, that it is optimal for customers to either balk (renege immediately upon arrival), or wait until they receive service. However, in reality customers typically do not adopt such a policy, and abandonments typically occur after customers have waited for some time.

Customer abandonment is common in ticket queues. In a ticket queue, customers receive a number upon arrival, and may become aware of the current ticket number in service. However, due to prior abandonments, the difference between a customer's ticket number and current ticket in service may exceed the actual number of customers ahead. In Kuzu *et al.* [20], the authors analyze abandonment effects on ticket queue processes in which customers choose to renege or not based on constant updating of queue information (such as how many tickets are ahead of theirs, how long they have waited, how many customers have left the queue, etc.). They allow the patience levels of the customers in the queue to adapt as time passes, based on observations of the ticket counter. An empirical study suggests customers continuously monitor and react to the system state (e.g., choose to renege or not), and they construct models based around these assumptions. Note that in the ticket queue, a customer waits at most the difference between his/her ticket number and the current ticket number in service. However, in our model, a customer may have to wait for additional services to transpire, due to those who will cut line.

Our line-cutting model also differs from many traditional queueing models in the following way. In many traditional FIFO models and variations, if customer n arrives to the system at time $T_n < t$, then whether or not that customer remains in the system at time t is a function of the data (arrival times and service requirements) for customers who arrived strictly before time T_n. This is not the case in our model. A customer arriving at time T_n remains in the system at time $t > T_n$ if and only if she/he (i) has not yet been served and (ii) survives all arrivals and line-cuts that occur in $(T_n, t]$. The papers Durrett and Limic [11], Jones and Serfozo [17], and May and Nowak [23] consider models where particles (representing customers, or species in an ecosystem) arrive randomly in time to a region, and upon arrival are assigned i.i.d. marks. The particles are then removed from the system only at times of future arrivals with larger marks with a probability that depends on the mark value of particles considered for deletion. Our model differs from this setup mainly in that the marks correspond to relative positions in the queue, and will change each time there is a service or arrival. Our model also allows for traditional service as in the $M/M/1$ queue.

3. General Model

Customers arrive to a single server service station according to a time-homogeneous Poisson process with rate λ. Let $\tilde{X}(t)$ denote the number of customers in the queue at time t. A customer arriving at time t will see $\tilde{X}(t-) := \lim_{s \to t} X(s)$ customers waiting in the queue. However, instead of joining the back of the queue as in first-in-first-out disciplines, the new arrival joins the queue at one of the $1 + \tilde{X}(t-)$ positions between waiting customers (or at the ends) according to a discrete uniform distribution. Immediately proceeding the arrival, each customer who the new arrival jumps or 'cuts' becomes irritated by the added waiting time, and departs (reneges) independently of everything, with probability a. The customers are otherwise served one-at-a-time in order of arrival (FIFO) by a single server, and the services attempts occur according to a Poisson process with rate μ. Let $R(t)$ denote the number of customers that renege upon an arrival at time t. Then the continuous time Markov chain (CTMC) $\{\tilde{X}(t),\ t \geq 0\}$ updates according to the following dynamics:

$$\widetilde{X}(t) = \begin{cases} \widetilde{X}(t-) + 1 - R(t), & \text{if an arrival occurs at } t; \\ \widetilde{X}(t-) - \mathbb{1}\{\widetilde{X}(t-) \geq 1\}, & \text{if a service attempt occurs at } t. \end{cases}$$

Define the discrete time Markov chain (DTMC) $\{X_n,\ n \geq 0\}$ as follows:

$$\begin{aligned} X_0 &= 0 \qquad \text{with probability } 1, \\ X_{n+1} &= \begin{cases} X_n - \mathbb{1}\{X_n \geq 1\}, & \text{with probability } q, \\ X_n + 1 - R_{n+1}, & \text{with probability } 1-q, \end{cases} \end{aligned} \tag{1}$$

where $q = \mu/(\mu+\lambda)$ and $R_{n+1} \sim \text{binomial}(L_n, a)$, with $L_n \sim \text{discrete } U(0, X_n)$. That is, L_n denotes the number of customers the new arrival skips over, or cuts. Because the CTMC $\{\tilde{X}(t), t \geq 0\}$ evolves as the DTMC $\{X_n, n \geq 0\}$ subordinated to a Poisson process with rate $\lambda + \mu$, it suffices to study the embedded DTMC $\{X_n, n \geq 0\}$, which we do throughout. Note $\{X_n, n \geq 0\}$ is irreducible and aperiodic, and so the condition (1) results in no loss of generality.

Our first result shows that any reneging in our model (i.e., whenever $a > 0$) stabilizes the queue, regardless of the value of λ/μ (contrary to the $M/M/1$ queue).

Proposition 1. *The $\{\tilde{X}(t), t \geq 0\}$ process is positive recurrent iff $a > 0$ or $\lambda < \mu$.*

The proof uses the following special case of Foster's Criterion called Pake's Lemma, which is Theorem 4.12 of Kulkarni [19]:

Theorem 2. *Pake's Lemma. Let $\{X_n, n \geq 0\}$ be an irreducible DTMC on $S = \{0,1, \ldots\}$. Define the drift function d to be*

$$d(i) := E[X_{n+1} - X_n \mid X_n = i],\ i \in S. \tag{2}$$

Then the DTMC is positive recurrent if both of the following hold:

(i) $d(i) < \infty$ *for all* $i \in S$*;*

(ii) $\limsup_{i\in S} d(i) < 0.$

Proof of Proposition 1. It suffices to prove stability for the DTMC $\{X_n,\ n \geq 0\}$. Fix $a > 0$ and set $\mu = 0$. The drift function $d(i)$ from (3) is

$$\begin{aligned} d(i) &= E[X_{n+1} - X_n \mid X_n = i] \\ &= 1 - E[\textit{number of departures at time } n+1 \mid X_n = i] \\ &= 1 - \sum_{k=1}^{i} \frac{a \cdot k}{i+1} = 1 - \frac{a \cdot i}{2}, \end{aligned} \tag{3}$$

and conditions (i) and (ii) of Pake's lemma are easily satisfied. Positive recurrence for the case when $\mu > 0$ follows by observing the drift function for the more general case is bounded above by $d(\cdot)$. In the case that $a = 0$, the model reduces to the $M/M/1$ queue, which requires $\lambda < \mu$ for positive recurrence.

We note in passing that when $\mu = 0$,

$$\begin{aligned} i > 2/a \quad &\Rightarrow \quad d(i) \le 0; \\ i = 2/a \quad &\Rightarrow \quad d(i) = 0; \\ i < 2/a \quad &\Rightarrow \quad d(i) \ge 0, \end{aligned}$$

and the process is symmetrically mean-reverting in the sense that $d(2/a-k) = -d(2/a+k)$ whenever $2/a \pm k \in \mathbb{Z}^+$.

We have the following.

Theorem 3. *The limiting distribution for $\{\tilde{X}(t), t \ge 0\}$ exists and equals the unique stationary distribution. Furthermore, if $\tilde{X}$ has this stationary distribution, then $\mu = 0$ implies $E\tilde{X}(t) \to E\tilde{X} = 2/a$.*

The proof uses the following dominated convergence theorem for convergence in distribution (see Serfozo [26]).

Theorem 4. *Suppose $X_n \xrightarrow{d} X$ in $\mathbb{R}$ and there exists a random variable Y with $|EY| < \infty$ so that Y stochastically dominates $|X_n|$ for each n:*

$$P(|X_n| \le x) \ge P(Y \le x), \quad x \ge 0. \tag{4}$$

Then $E|X|$ exists and $\lim_{n\to\infty} EX_n = EX$.

Proof of Theorem 3. As usual, it suffices to prove the discrete-time analogs of these statements. The first statement (convergence in distribution) follows from standard Markov chain theory (see [19], [26]) because $\{X_n, n \ge 0\}$ is irreducible, aperiodic and positive recurrent whenever $a > 0$. Next, let X be a random variable with this stationary distribution. By the Markov property and (3),

$$\begin{aligned} E[X_{n+k} \mid X_n = i] &= E[E[X_{n+k} \mid X_{n+k-1}] \mid X_n = i] \\ &= E[1 - \frac{a}{2} X_{n+k-1} + X_{n+k-1} \mid X_n = i] \\ &= 1 + (1 - a/2) E[X_{n+k-1} \mid X_n = i] \\ &= \sum_{j=0}^{k-1} (1-a/2)^j + i(1-a/2)^k \\ &\to 2/a \end{aligned}$$

as $k \to \infty$, independent of i.

It remains to show $EX = 2/a$, which we do using Theorem 4. Let $Y = k$ with probability $(1-\rho)\rho^k$, $\rho \in [0,1)$, $k \in \{0, 1, \ldots\}$. The variable Y has the stationary distribution of the $M/M/1$ queue with $\rho = \lambda/\mu$, and $EY = \rho/(1-\rho)$. Then

$$P(Y \ge k) = \rho^k. \tag{5}$$

Let

$$p_{i,i+1} := P(X_{n+1} = i+1 \mid X_n = i) = \frac{1-(1-a)^{i+1}}{a(i+1)}.$$

Note for $k \le n$ that $X_n \ge k$ implies at least one transition from i to $i+1$ for each $i \in \{0, \ldots, k-1\}$ in the first n transitions because $X_0 = 0$. Then

$$P(X_n \ge k) \le \prod_{i=1}^{k-1} p_{i,i+1}. \tag{6}$$

Then for fixed a, choosing $\rho \in [p_{1,2}, 1)$ for (5) satisfies (4) for $\{X_n, n \ge 0\}$ and Y because $p_{i,i+1}$ decreases in i. When $k > n$, $P(X_n \ge k) = 0$ since $X_0 = 0$ with probability 1.

4. The Stationary Distribution

4.1. Probability generating function

We now study the stationary distribution for $\{X_n, n \ge 0\}$. Let $C_n = c$ or s if the change at time n is triggered by a customer arrival or service attempt, respectively. Let X_n denote the state of the system after all events occur at time n. For instance, if $C_{n+1} = s$, then $X_{n+1} = (X_n - 1)\mathbb{1}\{X_n \ge 1\}$, If $C_{n+1} = c$, then $X_{n+1} = X_n + 1 - R_{n+1}$, where as before, $R_{n+1} \sim$ binomial (L_n, a) and $L_n \sim$ discrete $U(0, X_n)$. Define $q := \mu/(\lambda + \mu)$. For the remainder of this section we assume the process is stationary: $\{X_n, \ldots, X_{n+k}\} \overset{d}{=} \{X_m, \ldots, X_{m+k}\}$ for all m, n, k. Let G denote the probability generating function for X_0.

Proposition 5. *The probability generating function* G *satisfies*

$$\begin{aligned} 2aqG(0)(t-1) = {} & at(t-1)(t-q)G'(t) \\ & + [t(a-t^2) + q(t^3 + at - 2a)]G(t) \\ & + (1-a)(1-q)t^3 G(a + (1-a)t). \end{aligned} \tag{7}$$

Proof. We begin by obtaining an expression for the probability generating function $G_{X_1|X_0}$ of the process at time 1 given the process has the stationary distribution at time 0, by conditioning on the type of change triggered C_1.

$$G_{X_1|X_0}(t) = qE\left[G_{X_1|X_0,C_1=s}(t)\right] + (1-q)E\left[G_{X_1|X_0,C_1=c}(t)\right]. \tag{8}$$

The first expectation on the right-hand-side of (8) is

$$\begin{aligned} E\left[G_{X_1|X_0,C_1=s}(t)\right] &= E\left[t^{(X_0-1)}\mathbb{1}\{X_0 \ge 1\} + \mathbb{1}\{X_0 = 0\}\right] \\ &= t^{-1}E\left[t^{X_0}\mathbb{1}\{X_0 \ge 1\}\right] + P(X_0 = 0) \\ &= t^{-1}\left(G_{X_0}(t) - P(X_0 = 0)\right) + P(X_0 = 0) \\ &= t^{-1}G_{X_0}(t) + (1 - t^{-1})P(X_0 = 0). \end{aligned} \tag{9}$$

Next let $U \sim U(0,1)$ and $V_i = \mathbb{1}\{\text{customer } i \text{ survives the next arrival}\}$, where customer i is the customer who currently sees $i-1$ customers ahead in line. The expectation in the second term on the right-hand-side of (8) is

$$
\begin{aligned}
G_{X_1|X_0,C_1=c}(t) &= E[t^{X_1} \mid X_0, C_1 = c] \\
&= E\left[t^{(1+\sum_{i=1}^{X_0} V_i)} \mid X_0, C_1 = c \right] \\
&= tE\left[E\left[t^{\sum_{i=1}^{\lfloor U(X_0+1)\rfloor} V_i} \times t^{\sum_{\lceil U(X_0+1)\rceil}^{X_0} V_i} \mid U, X_0, C_1 = c \right] \mid X_0, C_1 = c \right] \\
&= tE\left[\left(\prod_{i=1}^{\lfloor U(X_0+1)\rfloor} E[t^{V_i} \mid U, X_0, C_1 = c] \right) \right. \\
&\qquad \left. \times \left(\prod_{\lceil U(X_0+1)\rceil}^{X_0} E[t^{V_i} \mid U, X_0, C_1 = c] \right) \mid X_0, C_1 = c \right],
\end{aligned}
$$

where a product of the form $\prod_{k=n+1}^{n} K_n$ is taken to be 1. Simplifying and letting Y be a discrete uniform random variable on $\{0,1, \ldots, X_0\}$,

$$
\begin{aligned}
G_{X_1|X_0,C_1=c}(t) &= tE\left[\left(\prod_{i=1}^{\lfloor U(X_0+1)\rfloor} (a + t(1-a)) \right) \left(\prod_{i=\lceil U(X_0+1)\rceil}^{X_0} t \right) \mid X_0, C_1 = c \right] \\
&= tE\left[\left(a + t(1-a) \right)^{\lfloor U(X_0+1)\rfloor} t^{(X_0 - \lfloor U(X_0+1)\rfloor)} \mid X_0, C_1 = c \right] \\
&= tE\left[\left(a + t(1-a) \right)^{Y} t^{(X_0 - Y)} \mid X_0, C_1 = c \right] \\
&= tE\left[\left(at^{-1} + (1-a) \right)^{Y} t^{X_0} \mid X_0, C_1 = c \right].
\end{aligned}
$$

Then

$$EG_{X_1|X_0,C_1=c}(t) = E\left[t^{(X_0+1)}E\left[\left(at^{-1}+(1-a)\right)^Y \mid X_0\right]\right]$$
$$= E\left[\frac{t^{(X_0+1)}}{X_0+1}\sum_{k=0}^{X_0}\left(at^{-1}+(1-a)\right)^k\right]$$
$$= E\left[\frac{t^{(X_0+1)}}{X_0+1}\cdot\frac{1-\left(at^{-1}+(1-a)\right)^{X_0+1}}{a(1-t^{-1})}\right]. \tag{10}$$

Using (9) and (10) in (8), setting $G_{X_1} = G_{X_0} = G$, and noting $G(0) = P(X_0 = 0)$ gives

$$G(t) = qt^{-1}G(t) + q(1-t^{-1})G(0) \tag{11}$$
$$+(1-q)E\left[\frac{t^{(X_0+1)}}{X_0+1}\cdot\frac{1-\left(at^{-1}+(1-a)\right)^{X_0+1}}{a(1-t^{-1})}\right]. \tag{12}$$

Multiplying through (12) by $a(1-t^{-1})$, differentiating with respect to t, and then multiplying through by t^3 gives (7).

We solve some special cases before considering the solution for the general model.

4.2. Model with no service

The case when $\mu = 0$ approximates the situation when $\lambda \gg \mu$, for instance, when a line for ticket sales grows extremely long before the service window opens.

Proposition 6. *When* $\mu = 0$, *the stationary distribution is given by* $P(X_0 = n) = \dfrac{G^{(n)}(0)}{n!}$, *where* $G(t) = \sum_{n=0}^{\infty} b_n (t-1)^n$ *and*

$$b_n = (n+1)\prod_{k=1}^{n}\frac{1-(1-a)^k}{k-(1-a)(k+1)+(1-a)^{k+1}} \tag{13}$$
$$= \frac{1}{n!}\prod_{k=1}^{n}\frac{d}{da}\left\{\ln[k-(1-a)(k+1)+(1-a)^{k+1}]\right\}. \tag{14}$$

Proof. In this case (7) becomes

$$0 = at(t-1)G'(t) + (a-t^2)G(t) + (1-a)t^2G(a+(1-a)t), \tag{15}$$

with $G(1) = 1$ and $a \in (0,1]$.

First, note when $a = 1$, this reduces to

$$G'(t) = \frac{1+t}{t}G(t), \qquad G(1) = 1,$$

which is satisfied by setting $G(t) = te^{t-1}$.

To find the solution to (15) for arbitrary $0 < a \le 1$, we suppose that $G(t)$ can be represented as a power series centered at $t = 1$, so that

$$G(t) = \sum_{n=0}^{\infty} b_n (t-1)^n. \tag{16}$$

To simplify the calculation, we change variables by letting $s = t - 1$ and put $r = 1 - a$ (notice that $0 \le r < 1$). Then (15) and (16) imply

$$\begin{aligned}0 &= (1-r)s(s+1)\sum_{n=1}^{\infty} n b_n s^{n-1} + (1 - r - (s+1)^2)\sum_{n=0}^{\infty} b_n s^n + r(s+1)^2 \sum_{n=0}^{\infty} b_n (rs)^n \\ &= (1-r)[(1-r)b_1 - 2b_0]s + \sum_{n=2}^{\infty}\Big\{ b_{n-2}[r^{n-1} - 1] \\ &\quad + b_{n-1}[n - 3 - (n-1)r + 2r^n] + b_n[n - (n+1)r + r^{n+1}]\Big\} s^n.\end{aligned}$$

Thus

$$b_1 = \frac{2b_0}{1-r}$$

and for $n \ge 2$:

$$b_n = -\frac{n - 3 - (n-1)r + 2r^n}{n - (n+1)r + r^{n+1}} b_{n-1} + \frac{1 - r^{n-1}}{n - (n+1)r + r^{n+1}} b_{n-2}.$$

Because $G(1) = 1$, it follows that $b_0 = 1$ and

$$b_1 = \frac{2}{1-r},$$

so that $EX_0 = G'(1) = b_1 = 2/a$, which verifies the mean of the stationary distribution in Theorem 3.

Enumerating the ratios b_n / b_{n-1} leads to the following observation:

$$\frac{b_n}{b_{n-1}} = \frac{(n+1)(1-r^n)}{n[n - (n+1)r + r^{n+1}]}$$

for all $n \ge 1$, which can be proven easily by induction. It follows from the ratio test that the radius of convergence of the power series (16) is found to be

$$R = \lim_{n\to\infty}\left|\frac{b_n}{b_{n+1}}\right| = \lim_{n\to\infty}\frac{(n+1)[r^{n+2} - r(n+2) + (n+1)]}{(n+2)(1-r^{n+1})} = +\infty,$$

since $0 \le r < 1$. This serves as a post hoc justification of our use of the power series method.

Since

$$\begin{cases} b_0 = 1, \\ b_n = \dfrac{(n+1)(1-r^n)}{n[n-(n+1)r+r^{n+1}]} b_{n-1}, \end{cases} \tag{17}$$

it follows that

$$\begin{aligned} b_n &= \prod_{k=1}^{n} \frac{(k+1)(1-r^k)}{k[k-(k+1)r+r^{k+1}]} \\ &= (n+1)\prod_{k=1}^{n} \frac{1-r^k}{k-(k+1)r+r^{k+1}}. \end{aligned}$$

Following the convention that $\prod_{k=1}^{0} c_k = 1$, this closed form applies when $n = 0$ as well as for all $n > 0$, and this is (13). Also, the observation that

$$\frac{d}{da}[k-(1-a)(k+1)+(1-a)^{k+1}] = (k+1)[1-(1-a)^k]$$

suggests the alternative expression (14).

4.3. Model with no arrivals

When $q = 1$, there are no arrivals and (7) gives

$$t(t-1)G'(t) + 2G(t) - 2G(0) = 0.$$

Inspecting the point $t = 1$ and recalling $G(1) = 1$, we find that $G(0) = 1$, and $G(t)$ solves

$$t(t-1)G'(t) + 2G(t) = 2$$

subject to the initial condition $G(1) = 1$. Since the general solution to this linear differential equation is

$$G(t) = 1 + \frac{Ct^2}{(t-1)^2},$$

it follows that $C = 0$, and $G(t) = 1$. This is not surprising in that the equilibrium distribution for the model with only service is obviously $P(X_0 = 0) = 1$ and $P(X_0 = n) = 0$ for all $n \geq 1$.

4.4. Model with 100% reneging

When $a = 1$, each line cutter causes a mass exodus from the queue by those who were cut. In this case (7) reduces to the linear differential equation

$$t(t-q)G'(t) + [(q-1)t^2 + (q-1)t + 2q]G(t) - 2G(0)q = 0, \tag{18}$$

and we have the following.

Proposition 7. *When* $a=1$ *the stationary distribution* $(p_0, p_1, \ldots)$ *of* $\{\tilde{X}(t), t \geq 0\}$ *is given by*

$$p_0 = \frac{q}{\sum_{k=0}^{\infty} \frac{(1-q)^{2k}}{(q^2+1)_k}}, \qquad p_1 = \frac{1-q}{\sum_{k=0}^{\infty} \frac{(1-q)^{2k}}{(q^2+1)_k}}, \qquad p_2 = \frac{(1-q)^2 \sum_{k=0}^{\infty} \frac{[(q-1)q]^k}{(q^2+1)_{k+1}}}{\sum_{k=0}^{\infty} \frac{(1-q)^{2k}}{(q^2+1)_k}},$$

and for $n \geq 3$

$$p_n = \frac{(q-2+n)p_{n-1} + (q-1)p_{n-2}}{q(n-2)}.$$

Proof. Supposing $G(t)$ is analytic at $t=0$, and applying the power series method reveals that for

$$G(t) = \sum_{n=0}^{\infty} p_n t^n,$$

since in this case $G(0) = p_0$,

$$\sum_{n=0}^{\infty} 2qp_n t^n + \sum_{n=1}^{\infty} [(q-1)p_{n-1} - nqp_n]t^n + \sum_{n=2}^{\infty} [p_{n-1}(n-1) + (q-1)p_{n-2}]t^n - 2qp_0 = 0.$$

Collecting like terms t^n and setting the corresponding coefficients to 0 gives the following recurrence relations:

$$\begin{aligned} n=0:&\quad 0 &&= 0, \\ n=1:&\quad 0 &&= qp_1 + (q-1)p_0, \\ n=2:&\quad 0 &&= qp_1 + (q-1)p_0, \\ n\geq 3:&\quad 0 &&= q(2-n)p_n + (q-2+n)p_{n-1} + (q-1)p_{n-2.} \end{aligned} \tag{19}$$

Note the equations corresponding to $n=1$ and $n=2$ are identical, which means there are solutions for nonzero choices of p_0. Furthermore, since p_2 is not determined by the $n=2$ equation, there are two parameters which must be chosen in order to close this system, namely, p_0 and p_2. Once p_0 and p_2 are determined, the power series can be iteratively constructed.

If instead we expand the power series about $t=q$, we find that for

$$G(t) = \sum_{n=0}^{\infty} c_n (t-q)^n,$$

and letting $s = t-q$,

$$0=[q(q^2+1)c_0-2G(0)q]+[q(q^2+2)c_1+(q-1)(2q+1)c_0]s$$
$$+\sum_{n=2}^{\infty}[q(q^2+1+n)c_n+(2q^2-q-2+n)c_{n-1}+(q-1)c_{n-2}]s^n.$$

Therefore

$$n=0:\quad c_0=\frac{2G(0)}{q^2+1},$$
$$n=1:\quad c_1=\frac{(1-q)(2q+1)}{q(q^2+2)}c_0=\frac{2(1-q)(2q+1)G(0)}{q(q^2+2)(q^2+1)},$$
$$n\geq 2:\quad c_n=\frac{1-q}{q(q^2+1+n)}c_{n-2}-\frac{2q^2-q-2+n}{q(q^2+1+n)}c_{n-1}.$$

By induction it can be proven that the ratios c_n/c_{n-1} satisfy

$$\frac{c_n}{c_{n-1}}=\frac{(n+1)(1-q)(2q+n)}{n(2q+n-1)(q^2+n+1)},$$

and therefore

$$\begin{cases} c_0=c_0 \\ c_n=\dfrac{(n+1)(1-q)(2q+n)}{n(2q+n-1)(q^2+n+1)}c_{n-1} \quad \text{for } n\geq 1. \end{cases}$$

This recurrence relation has solution for $n\geq 1$

$$\begin{aligned} c_n&=c_0\prod_{k=1}^{n}\frac{(k+1)(1-q)(2q+k)}{k(2q+k-1)(q^2+k+1)} \\ &=\frac{c_0(n+1)(1-q)^n(2q+n)}{2q}\prod_{k=1}^{n}\frac{1}{q^2+k+1} \\ &=\frac{c_0(n+1)(1-q)^n(2q+n)}{2q(q^2+2)_n}, \end{aligned}$$

where $(x)_n$ denotes the Pochhammer symbol defined by

$$(x)_0=1 \text{ and } (x)_n=x(x+1)\cdots(x+n-1) \quad \text{for } n\geq 1.$$

We note that this formula for c_n also holds when $n=0$.

The radius of convergence of a power series solution centered at $t=q$ with $q\neq 1$ is therefore found to be

$$R=\lim_{n\to\infty}\left|\frac{c_n}{c_{n+1}}\right|=\lim_{n\to\infty}\frac{(n+1)(n+2q)(n+q^2+2)}{(n+2)(1-q)(n+2q+1)}=+\infty.$$

To determine c_0, we must enforce $G(1)=1$. Hence

$$c_0\sum_{n=0}^{\infty}\frac{(n+1)(2q+n)}{2q(q^2+2)_n}(1-q)^{2n}=1,$$

i.e.,

$$\begin{aligned} c_0 &= \left(\sum_{n=0}^{\infty}\frac{(n+1)(2q+n)(1-q)^{2n}}{2q(q+2)_n}\right)^{-1} \\ &= \frac{2q}{(q^2+1)\sum_{k=0}^{\infty}\frac{(1-q)^{2k}}{(q^2+1)_k}}. \end{aligned} \tag{20}$$

It is noteworthy that the series

$$G(t)=c_0\cdot\sum_{n=0}^{\infty}\frac{(n+1)(1-q)^n(2q+n)}{2q(q^2+2)_n}(t-q)^n$$

can be put into a closed form by Mathematica, which allows us to quickly find exact coefficients of the Maclaurin series.

To close the loop in the discussion of the Maclaurin series for $G(s)$ studied above, the closed-form solution allows us to specify after some effort that

$$p_0=\frac{q}{\sum_{k=0}^{\infty}\frac{(1-q)^{2k}}{(q^2+1)_k}} \quad \text{and} \quad p_2=\frac{(1-q)^2\sum_{k=0}^{\infty}\frac{[(q-1)q]^k}{(q^2+1)_{k+1}}}{\sum_{k=0}^{\infty}\frac{(1-q)^{2k}}{(q^2+1)_k}}.$$

It follows that

$$p_1=\frac{1-q}{\sum_{k=0}^{\infty}\frac{(1-q)^{2k}}{(q^2+1)_k}},$$

and the recurrence relation $(19)_4$ allows us to reconstruct the probabilities p_n for $n\geq 3$.

4.5. The general setting

Having studied the cases in which $q=1$, $q=0$, and $a=1$, we now turn to the case when $a,q\in(0,1)$.

Proposition 8. *For $a,q \in (0,1)$, the stationary distribution $\{p_0, p_1, \ldots\}$ of $\{X(t),\ t \geq 0\}$ is*

$$p_n = \sum_{k=n}^{\infty} \binom{k}{n} b_k (-1)^{k-n},$$

where

$$b_0 = 1,$$

$$b_1 = \frac{2(1-2q+p_0 q)}{a(1-q)},$$

$$b_2 = \frac{3\{2p_0 q[1-a-(2-a)q]+2-a-4(2-a)q+(8-3a)q^2\}}{a^2(3-a)(1-q)^2},$$

and for $n \geq 3$

$$b_n = \frac{1}{(1-a)^{n+1}+(n+1)a-1}([1-(1-a)^{n-2}]b_{n-3} + \{3[1-(1-a)^{n-1}] - \frac{(n-2)a}{1-q}\}b_{n-2} + \{3[1-(1-a)^n] - \frac{a[2q-1+n(2-q)]}{1-q}\}b_{n-1}),$$

and p_0 *is chosen such that*

$$p_0 = \sum_{k=0}^{\infty} b_k (-1)^k.$$

Moreover, the stationary distribution satisfies:

$$p_1 = \frac{1-q}{q} p_0$$

and for $k \geq 0$,

$$(1-q)[1-(1-a)^{k+1}]p_k = a(k+1)p_{k+1} - a(k+1)(1+q)p_{k+2} + a(k+1)qp_{k+3} + (1-q)(1-a)^{k+1}\left[\sum_{n=k+1}^{\infty}\binom{n}{k} p_n a^{n-k}\right],$$

with

$$\sum_{k=0}^{\infty} p_k = 1.$$

Proof. Note that $p_0 = G(0)$. Suppose that

$$G(t) = \sum_{n=0}^{\infty} b_n (t-1)^n.$$

Then setting $s = t-1$, the general equation (7) requires that

$$
\begin{aligned}
2ap_0qs &= a[a(1-q)b_1 - 2(1-2q)b_0]s \\
&\quad + a\{a(3-a)(1-q)b_2 + 3[1-2q+a(1-q)]b_1 - 3(1-q)b_0\}s^2 \\
&\quad + \sum_{n=3}^{\infty}((1-q)[(1-a)^{n-2}-1]b_{n-3} + \{(n-2)a - 3(1-q)[1-(1-a)^{n-1}]\}b_{n-2} \\
&\quad + \{3(1-q)[(1-a)^n - 1] + a[2q-1+n(2-q)]\}b_{n-1} \\
&\quad + (1-q)[(1-a)^{n+1} + (n+1)a - 1]b_n)s^n,
\end{aligned}
$$

which gives values of the b_n above. Equating $G(t)=\sum_{n=0}^{\infty} p_n t^n = \sum_{n=0}^{\infty} b_n (t-1)^n$ gives the stationary probabilities p_n to be

$$
p_n = \sum_{k=n}^{\infty} \binom{k}{n} b_k (-1)^{k-n}.
$$

The relations for p_k can be found similarly upon substitution of $G(t)=\sum_{n=0}^{\infty} p_n t^n$ directly into (7).

5. Simulations and Approximations

We compared simulation output statistics with analytical approximations derived in the previous sections for various values of μ and a with λ fixed at 1, since the distribution of queue lengths depends only on the values of a and $q=\mu/(\lambda+\mu)$, and we have considered separately the case $q=1$, we are free to choose $\lambda=1$. We simulated 100,000 iterations of the DTMC in R software for $\mu \in \{2,1.5,1,0.1,0.01\}$, $a \in \{0.1,0.2, ...,0.9\}$, and $a=\mu \in \{0.1,0.2, \ldots,0.9\}$. Each time the chain was initialized near the mean. We found confidence intervals are of the Wilson-Agresti-Coull (WAC) type, which helps mitigate when true proportions are close to 0 or 1 (see Ott and Longnecker[24]). Also, each confidence interval is constructed at the 99.90909% individual confidence level so that we can apply the Bonferroni adjustment and make simultaneous intervals at the 99% family-wise level: for fixed μ and a, we can be about 99% confident that *all* eleven confidence intervals contain the true corresponding stationary probabilities. Following the traditional Wald confidence interval construction and using the individual confidence level of 99.90909% results in a minimum required sample size of 99,811 in order to achieve at most a 0.00525 margin of error (which is why we ran each simulation 100,000 steps):

$$
\textit{minimum required sample size} = \left\lceil 0.25 \times \left(\frac{z_{(1-.9990909)/2}}{0.00525} \right)^2 \right\rceil = 99{,}811.
$$

All but seventeen of the 594 intervals contain the analytical approximation. The tables of the approximations along with the confidence intervals can be found at the following website: ***https://sites.google.com/site/mattjones1204/publications***.

6. Conclusions

Batch reneging of impatient customers due to line cutting provides a stabilizing mechanism for the standard $M/M/1$ model—even in the case in which the arrival rate is greater than the service rate. We have found closed-form solutions for the distribution of steady state queue lengths for certain values of the model parameters by brute force. For the most general scenario, we found an approximation scheme for the distribution of queue lengths, which was in good agreement with estimates found through simulations for 35 combinations of the two essential parameters.

Acknowledgment

The authors wish to thank the anonymous reviewers for their helpful comments, which led to an improved paper.

References

[1] Allon, G. & Hanany, E. (2012). Cutting in line: social norms in queues. *Management Science*, 58, 493-506.

[2] Al-Seedy, R. O., El-Sherbiny, A. A., El-Sherbiny, S. A., & Ammar, S. I. (2009). Transient solution of the $M/M/c$ queue with balking and reneging. *Computers & Mathematics with Applications*, 57, 1280-1285.

[3] Ammar, S. I., Helan, M. M., & Amri, F. T. (2013). The busy period of an $M/M/1$ queue with balking and reneging. *Applied Mathematical Modelling*, 37, 9223-9229.

[4] Ammar, S. I. (2014). Transient analysis of a two-hetergeneous servers queue with impatient behavior. *Journal of the Egyptian Mathematical Society*, 22, 90-95.

[5] Ammar, S. I. (2015). Transient analysis of an $M/M/1$ queue with impatient behavior and multiple vacations. *Applied Mathematics and Computation*, 260, 97-105.

[6] Baccelli, F., & Bremaud, P. (2003). Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences. second edition, Springer.

[7] Batt, R., & Terwiesch, C. (2015). Waiting patiently: An empirical studey of queue abandonment in an emergency department. *Management Science*, 61, 39-59.

[8] Bolandifar, E., DeHoratius, N., Olsen, T., & Wiler, J. (2016). Modeling the behavior of patients who leave the ED without being seen. *Chicago Booth Research Paper*.

[9] Cobham, A. (1954). Priority assignment in waiting line problems. *Operations Research*, 2, 470-476.

[10] Dimou, S., Economou, A., & Fakinos, D. (2011). The single server vacation queueing model with geometric abandonments. *Journal of Statistical Planning and Inference*, 141, 2863-2877.

[11] Durrett, R., & Limic, V. (2002). A surprising Poisson process arising from a species competition model. *Stochastic Processes and their Applications*, 102, 301-309.

[12] Fralix, B. (2013). A time-dependent study of the knockout queue. *Probability in the Engineering and Informational Sciences*, 27, 309-317.

[13] Gross, D., Shortle, J. F., Thompson, J. M., &. Harris, C. M. (2008). Fundamentals of Queueing Theory. fourth edition, Wiley.

[14] Haight, F. A. (1959). Queueing with reneging. *Metrika*, 2, 186-197.

[15] Hasenbein, J., & Perry, D. (2013). Introduction: queueing systems special issue on queueing systems with abandonments. *Queueing Systems*, 75, 111-113.

[16] Huang, J., & Zhang, H. (2013). Diffusion approximations for open Jackson Networks with reneging. *Queueing Systems*, 74, 445-476.

[17] Jones, M., & Serfozo, R. (2007). Poisson limits of sums of point processes and a particle survivor model. *Annals of Applied Probability*, 17, 265-283.

[18] Kapodistria, S. (2011). The $M/M/1$ queue with synchronized abandonments. *Queueing Systems*, 68, 79-109.

[19] Kulkarni, V. G. (2017). Modeling and Analysis of Stochastic Systems. third edition, CRC Press.

[20] Kuzu, K., Gao, L., & Xu, S. H. (2017). To Wait or Not to Wait: The theory and practice of ticket queues. SSRN.

[21] Liu, X. (2019). Diffusion approximations for double-ended queues with reneging in heavy traffic. *Queueing Systems*, 91, 49-87.

[22] Mandelbaum, A., & Shimkin, N. (2000). A model for rational abandonments from invisible queues. *Queueing Systems*, 36, 141-173.

[23] May, R. M., & Nowak, M. A. (1994). Superinfection, metapopulation dynamics, and the evolution of diversity. *Journal of Theoretical Biology*, 170, 95-114.

[24] Ott, R. L., & Longnecker, M. (2015). An Introduction to Statistical Methods and Data Analysis. seventh edition, Cengage Learning.

[25] Ross, S. (1996). Stochastic Processes. second edition, Wiley.

[26] Serfozo, R. F. (2009). Basics of Applied Stochastic Processes. Springer.

[27] Sherzer, E., & Kerner, Y. (2018). Customers' abandonment strategy in an $M/G/1$ queue. *Queueing Systems*, 90, 65-87.

[28] Tilman, D. (1994). Competition and biodiversity in spatially structured habitats. *Ecology*, 75, 2-16.

[29] Wang, J. , Baron, O., & Scheller-Wolf, A. (2015). $M/M/c$ queue with two priority classes. *Operations Research*, 63, 733-749.

[30] Zolar, E., Mandelbaum, A., & Shimkin, N. (2002). Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Science*, 48, 566-583.