*J. Appl. Prob.* **43,** 1201–1205 (2006)
*Printed in Israel*


# A MONOTONICITY RESULT FOR A G/GI/*c* QUEUE WITH BALKING OR RENEGING

SERHAN ZIYA,* *University of North Carolina*
HAYRIYE AYHAN ** *** AND
ROBERT D. FOLEY,** **** *Georgia Institute of Technology*
EROL PEKÖZ,***** *Boston University*

**Abstract**

In a G/GI/*c* loss system with balking, reneging, or limited waiting space, deleting some of the arriving customers can either increase or decrease the fraction of the remaining arrivals who get served, depending on how customers are deleted. We present a model in which the random deletion of arrivals independently and with some fixed probability can never decrease the fraction of the remaining arrivals who get served.

*Keywords:* Monotonicity in queues; balking; reneging; loss system; coupling

2000 Mathematics Subject Classification: Primary 60K25

## 1. Introduction

In a G/GI/*c* loss system with balking, reneging, or limited waiting space, deleting some of the arriving customers can either increase or decrease the fraction of the remaining arrivals who get served. For example, consider a G/D/1/1 system in which service takes a deterministic two units of time, and customers arrive at times $1, 2, 4, 6, 11, 12, 14, 16, 21, 22, 24, 26, \ldots$. In this system, only the customers arriving at times $2, 12, 22, \ldots$ are lost, so three-quarters of the customers get served. If we delete the customers arriving at times $1, 11, 21, \ldots$, then all of the remaining customers get served. If we instead delete the customers arriving at times $4, 14, 24, \ldots$, then only two-thirds of the remaining customers get served. Thus, the fraction of the remaining customers served can either rise or fall, depending on which customers we delete. We present a model in which, if customers are deleted *randomly* with some fixed probability, the fraction of the remaining customers who get served can never decrease.

## 2. G/GI/*c* queue with balking or limited waiting space

Consider a standard G/GI/*c* queue working in a first-come–first-served fashion with the following additional conditions. Arriving customers are classified as eligible according to a

Received 18 February 2004; revision received 3 October 2006.
* Postal address: Department of Statistics and Operations Research, University of North Carolina, CB# 3260, 213 Smith Building, Chapel Hill, NC 27599, USA. Email address: ziya@unc.edu
** Postal address: H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, 765 Ferst Drive, Atlanta, GA 30332-0205, USA.
*** Email address: hayhan@isye.gatech.edu
**** Email address: rfoley@isye.gatech.edu
***** Postal address: School of Management, 595 Commonwealth Avenue, Boston, MA 02215, USA.
Email address: pekoz@bu.edu

Bernoulli process with parameter $p$, $0 \leq p \leq 1$. This is a system parameter. Ineligible customers are lost, while eligible customers finding the queue length equal to $x$ join the system with probability $f(x)$, where $f(\cdot)$ is a nonincreasing function bounded below by 0 and above by 1. In other words, eligible customers who find $x$ customers in the queue balk with probability $1 - f(x)$. Note that setting $f(x) = 0$ for $x \geq m$ limits the total number of customers in the system to $m$, and further setting $f(x) = 1$ for $0 \leq x \leq m - 1$ reduces the system to a standard G/GI/$c$/$m$ queue which arriving customers who find less than $m$ customers in the system are allowed to join with probability $p$.

In order to define the queue length process uniquely in the case of simultaneous occurrence of service completion(s) and customer arrival(s), we assume that Sonderman's (1979) event-order assumption holds. That is, we assume that, at any time $t$, first all scheduled departures are allowed to occur and then new arrivals are allowed to enter the system. If any of these new arrivals have zero service time and are next to be serviced, they are allowed to depart, allowing for more arrivals. Continuing in this way, all events that can occur at time $t$ are allowed to occur.

Let $B(t)$ denote the number of customer arrivals and $E(p, t)$ the number of *eligible* customer arrivals by time $t$. Also, define $J(p, t)$ as the number of customers who have *joined* the system by time $t$. Then $J(p, t)/E(p, t)$ is the fraction of eligible customers who have joined the system by time $t$. We assume that, for $0 \leq p \leq 1$, $\lim_{t\to\infty} J(p, t)/t$ and $\lim_{t\to\infty} B(t)/t$ exist, and we define

$$L(p) = \lim_{t\to\infty} \frac{J(p, t)}{E(p, t)} = \lim_{t\to\infty} \frac{J(p, t)}{pB(t)}$$

as the long-run fraction of eligible customers who join the system. Equivalently, we can set the arrival rate $\lim_{t\to\infty} B(t)/t$ to 1 without loss of generality and write

$$L(p) = \frac{N(p)}{p}, \tag{2.1}$$

where $N(p) = \lim_{t\to\infty} J(p, t)/t$ is the long-run average number of customers who join the system per unit time.

In the remainder of this section, we prove the following theorem.

**Theorem 2.1.** *$L(p)$ is nonincreasing in $p$.*

*Proof.* Consider the system described above with the following additions. When a customer arrives and the queue length is equal to $x$, the customer joins the system and is labelled as a type-1 customer with probability $p_1 f(x)$, joins the system and is labelled as a type-2 customer with probability $p_2 f(x)$ (with $p_1, p_2 \geq 0$ and $p_1 + p_2 \leq 1$), or is otherwise immediately lost. Let $\tilde{J}_i(p_1, p_2, t)$ be the number of type-$i$ customers who have joined the system by time $t$, and define

$$\tilde{N}_i(p_1, p_2) = \lim_{t\to\infty} \frac{\tilde{J}_i(p_1, p_2, t)}{t}$$

to be the long-run average number of type-$i$ customers who join the system per unit time.

Since service times are independent and identically distributed and every customer who joins the system is type 1 with probability

$$\frac{p_1 f(x)}{p_1 f(x) + p_2 f(x)} = \frac{p_1}{p_1 + p_2}$$

(independent of $x$), we have

$$\frac{\tilde{N}_1(p,\varepsilon)}{\tilde{N}_1(p,\varepsilon)+\tilde{N}_2(p,\varepsilon)} = \frac{p}{p+\varepsilon}, \tag{2.2}$$

where $0 \le p \le 1$ and $0 \le \varepsilon \le 1-p$. Also, by the rules of the construction we immediately have

$$\tilde{N}_1(p,\varepsilon)+\tilde{N}_2(p,\varepsilon) = \tilde{N}_1(p+\varepsilon,0). \tag{2.3}$$

Combining (2.2) and (2.3), we get

$$\frac{\tilde{N}_1(p,\varepsilon)}{p} = \frac{\tilde{N}_1(p+\varepsilon,0)}{p+\varepsilon},$$

and using the fact that $\tilde{N}_1(p,0) \ge \tilde{N}_1(p,\varepsilon)$ (to be proved in Lemma 2.1, below) gives

$$\frac{\tilde{N}_1(p,0)}{p} \ge \frac{\tilde{N}_1(p+\varepsilon,0)}{p+\varepsilon}$$

or, equivalently,

$$\frac{N(p)}{p} \ge \frac{N(p+\varepsilon)}{p+\varepsilon}.$$

The result now follows immediately from (2.1).

**Lemma 2.1.** *We have* $\tilde{N}_1(p,0) \ge \tilde{N}_1(p,\varepsilon)$.

*Proof.* The proof uses a coupling argument to establish that $\tilde{J}_1(p,0,t) \ge \tilde{J}_1(p,\varepsilon,t)$, which immediately implies the result. We couple a $(p,\varepsilon)$ system together with a $(p,0)$ system, let both systems have the same arrival process, and let the service time of the $i$th type-1 customer to enter service be the same in both systems, for $i \in \{1,2,\dots\}$.

For the $i$th arrival, let $V_i$ and $U_i$ be uniform random variables taking values between 0 and 1, independent of everything else. First, if $V_i < p$ then the arrival is labelled as type 1, and if $p < V_i < p+\varepsilon$ then it is labelled as type 2. In the $(p,\varepsilon)$ system, if the queue length is $x$ at the time of the $i$th arrival, then the customer joins the system as a type-1 customer if $U_i < f(x)$ and $V_i < p$ and as a type-2 customer if $U_i < f(x)$ and $p < V_i < p+\varepsilon$. In the $(p,0)$ system, if the queue length is $x$ at the time of the $i$th arrival, then the customer joins the system as a type-1 customer if $U_i < f(x)$ and $V_i < p$.

Let $T_n$ be the time of the $n$th arrival which is labelled as type 1, whether or not the customer joins either system. For notational convenience, let $A(t) = \tilde{J}_1(p,\varepsilon,t)$ and $A'(t) = \tilde{J}_1(p,0,t)$ respectively denote the total numbers of type-1 customers who have joined the $(p,\varepsilon)$ and $(p,0)$ systems by time $t$. For a given $n$, we will show that if we assume $A(t) \le A'(t)$ to hold for all $t < T_n$, then $A(T_n) \le A'(T_n)$.

We consider two cases. First suppose that $A(T_{n-1}) < A'(T_{n-1})$. Then we must have $A(T_n) \le A'(T_n)$, since $A(T_n) - A(T_{n-1}) \le 1$. Now suppose that $A(T_{n-1}) = A'(T_{n-1})$. Let $Q(t)$ and $Q'(t)$ respectively denote the queue lengths at time $t$ in the $(p,\varepsilon)$ system and in the $(p,0)$ system. Since type-1 service times are identical and in the same order in both systems and (since $A(T_{n-1}) = A'(T_{n-1})$) the same number of type-1 customers have joined both systems prior to time $T_n$, the combined service time of type-1 customers who join prior to time $T_n$ will be the same in both systems. Furthermore, the assumption that $A(t) \le A'(t)$ for all $t < T_n$

implies that the type-1 customers who join have arrived no earlier in the $(p, \varepsilon)$ system, and, since they are served in the same order, we must have $Q(T_n) \geq Q'(T_n)$.

To see why $Q(T_n) \geq Q'(T_n)$, imagine a decision-maker who decides when customers enter service and who may keep servers idle. The goal of the decision-maker is to minimize $Q(T_n)$ with the restriction that customers must be served in the order in which they arrive. Suppose that the decision-maker decides to idle a server for $t$ time units while a customer waits. It is easily seen that it is at least as good instead to serve this customer as soon as possible, and idle the server for $t$ time units after the customer finishes service. Repeating this interchange argument, it can be seen that it is optimal to serve all customers as soon as possible. This implies that it cannot be worse if customers arrive sooner, as is the case in the $(p, 0)$ system, so we must have $Q(T_n) \geq Q'(T_n)$.

To conclude, the fact that $f(\cdot)$ is nonincreasing implies that if the customer arriving at time $T_n$ is the $i$th customer to arrive and $U_i < f(Q(T_n))$, then we also have $U_i < f(Q'(T_n))$ and, so, $A(T_n) \leq A'(T_n)$.

## 3. G/GI/*c* queue with reneging

Consider a standard G/GI/$c$ queue working in a first-come–first-served fashion and according to Sonderman's (1979) event-order assumption, with the following additional conditions. An arriving customer is allowed to join with probability $p$ and is otherwise lost. The $i$th customer to join the system waits to be served but abandons the system if his service has not started within $X_i$ time units of his arrival, where $\{X_j,\ j \geq 1\}$ are independent, identically distributed random variables.

As in Section 2, let $B(t)$ denote the number of arrivals to the system and $E(p, t)$ the number of eligible customers who have arrived by time $t$. Note that in this model all the eligible customers enter the system. Also, define $S(p, t)$ as the number of customers who have entered service by time $t$. Then $S(p, t)/E(p, t)$ is the fraction of admitted customers who enter service. We assume that, for $0 \leq p \leq 1$, $\lim_{t\to\infty} S(p, t)/t$ and $\lim_{t\to\infty} B(t)/t$ exist, and we define

$$R(p) = \lim_{t\to\infty} \frac{S(p, t)}{E(p, t)} = \lim_{t\to\infty} \frac{S(p, t)}{pB(t)}$$

to be the long-run fraction of admitted customers who receive service. Equivalently, we can set the arrival rate $\lim_{t\to\infty} B(t)/t$ to 1 without loss of generality and write

$$R(p) = \frac{M(p)}{p}, \tag{3.1}$$

where $M(p) = \lim_{t\to\infty} S(p, t)/t$ is the long-run average number of customers who enter service per unit time. The following result then follows from Theorem 2.1 and Lemma 2.1.

**Corollary 3.1.** *$R(p)$ is nonincreasing in $p$.*

*Proof.* Consider a new G/GI/$c$ queueing system to which arriving customers are accepted with probability $pf(x) = p\,\mathrm{P}(X_1 > x)$, where $x$ is the time the arriving customer, if accepted, would need to wait in the queue before starting service. Letting $\hat{M}(p)$ denote the long-run number of customers who are accepted per unit time in this new system, it can easily be seen that $M(p) = \hat{M}(p)$. For this new system, the argument in the proof of Lemma 2.1 also goes through if $Q(t)$ and $Q'(t)$ are defined to be the respective amounts of time a customer arriving at time $t$ must wait in the queue in the $(p, \varepsilon)$ system and in the $(p, 0)$ system before

starting service. Then, from Theorem 2.1, it follows that $\hat{M}(p)/p$ is nonincreasing in $p$, which immediately implies the result.

## References

SONDERMAN, D. (1979). Comparing multi-server queues with finite waiting rooms. I: Same number of servers. *Adv. Appl. Prob.* **11,** 439–447.