

A Monotonicity Result for a G/GI/c Queue with Balking or Reneging

Serhan Ziya*, Hayriye Ayhan**, Robert D. Foley**, Erol Peköz***

*Department of Statistics and Operations Research
University of North Carolina
CB# 3260, 213 Smith Building, Chapel Hill, NC 27599, U.S.A.
E-mail: ziya@unc.edu

**H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
765 Ferst Drive, Atlanta, GA 30332-0205, U.S.A.
E-mail: hayhan@isye.gatech.edu, rfoley@isye.gatech.edu

***School of Management
Boston University
595 Commonwealth Avenue, Boston, MA 02215, U.S.A.
E-mail: pekoz@bu.edu

October, 2006

Keywords: Monotonicity in queues, balking, reneging, loss systems, coupling.
AMS 2000 Subject Classification: Primary 60K25

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; this depends on how customers are deleted. For our model we show that randomly deleting arrivals independently with some fixed probability can never decrease the fraction of the remaining arrivals who get served.

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 where service takes a deterministic 2 units of time, and customers arrive at times 1, 2, 4, 6, 11, 12, 14, 16, 21, 22, 24, 26, \dots . In this system, only the customers arriving at times 2, 12, 22, \dots are lost, so 3/4 of customers get served. If we delete the customers arriving at time 1, 11, 21, \dots , then all of the remaining customers get served. If we instead delete the customers arriving at times 4, 14, 24, \dots , then only 2/3 of the remaining customers get served. Thus, the fraction of the remaining customers served can either rise or fall. For our model we show that 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 Bernoulli process with parameter p . 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 non-increasing function bounded below by 0 and above by 1 and $0 \leq p \leq 1$ is a system parameter. 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 where arriving customers who find less than m customers in the system are allowed to join with probability p .

In order to uniquely define the queue length process in 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 $E(p, t)$ denote the number of *eligible* customer arrivals by time t . Also define $J(p, t)$ as the number of customers who *join* the system by time t . Then, $J(p, t)/E(p, t)$ is the fraction of eligible customers who join the system and get served. We assume that for $0 \leq p \leq 1$, $\lim_{t \rightarrow \infty} J(p, t)/t$ and $\lim_{t \rightarrow \infty} B(t)/t$ exist and we define

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

as the long run fraction of eligible customers who join the system. Equivalently, we can write

$$L(p) = \frac{N(p)}{p} \tag{1}$$

where $N(p) = \lim_{t \rightarrow \infty} J(p, t)/t$ is the long-run average number of customers who join the system per unit time and the arrival rate $\lim_{t \rightarrow \infty} B(t)/t$ is set to 1 without loss of generality.

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

Theorem 2.1 $L(p)$ is non-increasing 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 $p_1 + p_2 \leq 1$), and is otherwise immediately lost. Let $\tilde{J}_i(p_1, p_2, t)$ be the number of type i customers who join the system by time t and define $\tilde{N}_i(p_1, p_2) = \lim_{t \rightarrow \infty} \tilde{J}_i(p_1, p_2, t)/t$ to be the long-run average number of type i customers who join the system per unit time.

Since service times are i.i.d. and every customer who joins the system is type 1 with probability $p_1 f(x)/(p_1 f(x) + p_2 f(x)) = p_1/(p_1 + p_2)$ which does not depend on x , we have

$$\frac{\tilde{N}_1(p, \varepsilon)}{\tilde{N}_1(p, \varepsilon) + \tilde{N}_2(p, \varepsilon)} = \frac{p}{p + \varepsilon}. \quad (2)$$

Also, by the rules of the construction we immediately have

$$\tilde{N}_1(p, \varepsilon) + \tilde{N}_2(p, \varepsilon) = \tilde{N}_1(p + \varepsilon, 0). \quad (3)$$

Combining (2) and (3), we get

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

and applying Lemma 2.1 below gives

$$\tilde{N}_1(p, 0)/p \geq \tilde{N}_1(p + \varepsilon, 0)/(p + \varepsilon)$$

or equivalently

$$N(p)/p \geq N(p + \varepsilon)/(p + \varepsilon).$$

Then, the result immediately follows from (1). \square

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

Proof The proof uses a coupling argument to establish $\tilde{J}_1(p, 0, t) \geq \tilde{J}_1(p, \varepsilon, t)$, which immediately implies the result.

We couple a (p, ε) 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, U_i be $U(0, 1)$ (uniform random variables taking values between 0 and 1) independent of everything else. First, if $V_i < p$, the arrival is labelled as type 1, and if $p < V_i < p + \varepsilon$ it is labelled as type 2. In the (p, ε) system, if the queue length is x at the time of the i th arrival, the customer joins the system as a type 1 customer if $U_i < f(x)$ and $V_i < p$, 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, 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 number of type 1 customers who join the (p, ε) and $(p, 0)$ systems by time t . For a given n , we will show that if we assume $A(t) \leq A'(t)$ holds for all $t < T_n$, then $A(T_n) \leq A'(T_n)$.

We consider two cases. First, suppose that $A(T_{n-1}) < A'(T_{n-1})$, then we must have $A(T_n) \leq A'(T_n)$ since $A(T_n) - A(T_{n-1}) \leq 1$.

Now, suppose that $A(T_{n-1}) = A'(T_{n-1})$. Let $Q(t)$ and $Q'(t)$ denote the queue lengths at time t in the (p, ε) system and the $(p, 0)$ system, respectively. Since type 1 service times are identical and in the same order in both systems and $A(T_{n-1}) = A'(T_{n-1})$ means that we have the same number of type 1 customers who join both systems prior to time T_n , the total of the service times of type 1 customers who join prior to time T_n will be the same in both systems. Furthermore, the assumption $A(t) \leq A'(t)$ for all $t < T_n$ implies that the type 1 customers who join have arrived no earlier in the (p, ε) 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 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 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 to instead serve this customer as soon as possible, and then idle the server for t time units after the customer finishes service. Repeating this interchange argument, it can be seen it is optimal to serve all customers as soon as possible. Then this implies that if customers arrive sooner, as is the case in the $(p, 0)$ system, it can not be worse – so we must have $Q(T_n) \geq Q'(T_n)$.

To conclude, since $f(\cdot)$ is non-increasing this means 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 $A(T_n) \leq A'(T_n)$. \square

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, and 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 the customer's service has not yet started within X_i time units of the customer's arrival where $\{X_j, j \geq 1\}$ are i.i.d. random variables.

As in Section 2, let $B(t)$ denote the number of arrivals to the system and $E(p, t)$ denote the number of eligible customers who arrive 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 enter service by time t . Then, $S(p, t)/E(p, t)$ is the fraction of admitted customers who enter the service.

We assume that for $0 \leq p \leq 1$, $\lim_{t \rightarrow \infty} S(p, t)/t$ and $\lim_{t \rightarrow \infty} B(t)/t$ exist and we define

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

to be the long run fraction of admitted customers who receive service. Equivalently, we can write

$$R(p) = \frac{M(p)}{p} \tag{4}$$

where $M(p) = \lim_{t \rightarrow \infty} S(p, t)/t$ is the long-run average number of customers who enters service per unit time and the arrival rate $\lim_{t \rightarrow \infty} B(t)/t$ is set to 1 without loss of generality. Then, the following result follows from Theorem 2.1 and Lemma 2.1.

Corollary 3.1 *$R(p)$ is non-increasing in p .*

Proof Consider a new $G/GI/c$ queueing system where arriving customers are accepted to the system with probability $pf(x)$, where x is the time this customer, if accepted would need to wait in the queue before starting service, and $f(x) = P(X_1 > x)$. Letting $\hat{M}(p)$ denote the long run number of customers who are accepted per unit time in this new system, it can be easily 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 amount of time a customer arriving at time t must wait in the queue (respectively in the (p, ε) system and the $(p, 0)$ system) before starting service. Then, from Theorem 2.1, it follows that $\hat{M}(p)/p$ is non-increasing in p , which immediately implies the result. \square

References

- SONDERMAN D. (1979). Comparing Multi-Server Queues with Finite Waiting Rooms, I: Same Number of Servers. *Advances in Applied Probability*. **11**, 439–447.