EMBEDDED RENEWAL PROCESSES IN THE GI/G/s QUEUE

WARD WHITT, Yale University

Abstract

The stable GI/G/s queue (p < 1) is sometimes studied using the "fact" that epochs just prior to an arrival when all servers are idle constitute an embedded persistent renewal process. This is true for the GI/G/1 queue, but a simple GI/G/2 example is given here with all interarrival time and service time moments finite and p < 1 in which, not only does the system fail to be empty ever with some positive probability, but it is never empty. Sufficient conditions are then given to rule out such examples. Implications of embedded persistent renewal processes in the GI/G/1 and GI/G/s queues are discussed. For example, functional limit theorems for time-average or cumulative processes associated with a large class of GI/G/s queues in light traffic are implied.

1. Introduction: The standard single-server queue

In this paper we investigate embedded renewal processes in the GI/G/s queue. Our original aim was to show how functional limit theorems for the GI/G/1 queue in light traffic recently proved by Iglehart (1971a) could be extended to the GI/G/s queue. Iglehart's (1971a) GI/G/1 argument does indeed apply with only minor modification to a large class of GI/G/s queues (Theorem 2.3), but it does not apply to all GI/G/s queues (Example 2.2). No doubt, this difficulty with the GI/G/s queue has been discovered before, but we believe it is worth additional attention. We first review embedded renewal process approaches to the GI/G/1 queue and then turn to the GI/G/s queue with s > 1.

Consider the standard single-server queueing system in which customers are served in order of their arrival without defections, but at first make no distribution or independence assumptions. The basic data for this system is a sequence of ordered pairs of non-negative random variables \( \{(u_n, v_n), n \geq 0\} \) defined on some underlying probability space \((\Omega, \mathcal{F}, P)\), where for \( n \geq 1 \) the variable \( u_n \) represents the interarrival time between the \( n \)th and \((n + 1)\)th customers and the variable \( v_n \) represents the service time of the \( n \)th customer. The variable \( u_0 \) denotes the time of the first customer.

Received in revised form 6 October 1971. Research partially supported by National Science Foundation Grant GK-27866.
until the first customer arrives (after $t = 0$) and the variable $v_0$ denotes the initial workload facing the server, which we interpret as the result of a 0th customer arriving at time $t = 0$. It is customary to define the sequence $\{W_n, n \geq 0\}$ of successive waiting times recursively by

\begin{equation}
W_{n+1} = [W_n + X_n]^+, \quad n \geq 0,
\end{equation}

where for any real number $x$, $[x]^+ = \max\{x, 0\}$, $X_n = v_n - u_n$, and $W_0 = 0$, cf. Feller ((1966), p. 193) or Lindley (1952). From (1.1), it can then be shown (by induction, for example) that

\begin{equation}
W_n = \max_{0 \leq k \leq n} \{S_n - S_k\}, \quad n \geq 0,
\end{equation}

(1.2)

\begin{equation}
= S_n - \min_{0 \leq k \leq n} S_k, \quad n \geq 0,
\end{equation}

where $S_k = X_0 + \cdots + X_{k-1}$ and $S_0 = 0$. (Note that (1.1) is a definition and (1.2) is a theorem.)

The relationship in (1.2) is the vehicle for obtaining results about $\{W_n, n \geq 0\}$. The sequence of waiting times is related to something which has been studied extensively—a sequence of partial sums. If $X_n$ is i.i.d., which is implied by the system being GI/G/1, then $S_n - S_k \sim S_{n-k}$ (where $\sim$ means equality in distribution) so that

\begin{equation}
P(W_n \leq x) = P(M_n \leq x), \quad n \geq 0,
\end{equation}

(1.3)

where $M_n = \max \{S_0, \ldots, S_n\}$. Moreover, then $\{S_n\}$ is a random walk, so the theory of random walks as contained for example in Spitzer (1964), Chung (1968), Chapter 8 and Feller (1966), Chapter 12) can be applied with considerable force. In fact, Lindley's (1952) early investigation of the GI/G/1 queue proceeded somewhat along these lines. If $\rho = EV/Eu < 1$ (if $X_n$ has finite expectation with $EX_n < 0$), then $W_n$ equals 0 infinitely often and $W_n \Rightarrow W$, where $\Rightarrow$ means weak convergence or convergence in law and $W \sim M$ is a finite random variable where $M = \sup_{k \geq 0} M_k$ has been studied extensively, cf. Chung (1968), Section 8.5). A comprehensive account exposing this structure in the GI/G/1 queue has been given by Kingman (1966), where more of the history can be found.

Now we would like to apply these results to various continuous-time processes associated with the GI/G/1 queue such as the queue length process $\{Q(t), t \geq 0\}$. For example, it is intuitively obvious that the epochs in continuous time corresponding to the events $\{W_n = 0\}$ constitute embedded renewal processes in all the continuous-time processes. However, this is not always demonstrated properly; a careful treatment has just been provided by Iglehart (1971a), Section 2. The idea is to exploit the theory of optional random variables as discussed in Chung (1968). With the independence associated with the GI/G/1 queue, the basic data can be
represented as a sequence of independent pairs of independent non-negative random variables \(\{(u_n, v_n), n \geq 0\}\). If we understand successive busy cycles in discrete time to be the successive numbers of customers arriving between epochs just prior to an arrival when the server is idle, then these discrete-time busy cycles can be represented as optional random variables relative to the sequence \(\{u_n, v_n\}\). (These are the variables \(a^k\) in Iglehart (1971a).) This special kind of optional random variable is called a ladder variable by Blackwell (1953) and Feller ((1968), Chapter 12). If \(EX_\infty < 0 (\rho < 1)\), then these optional variables are almost surely finite with finite expectation. The remaining processes can all be defined in terms of the basic data. Theorem 8.2.3 of Chung (1968) applied to an i.i.d. sequence of random vectors in \(R^2\) then implies that the busy cycles in continuous time are independent and identically distributed. The \(k\)th busy cycle (in continuous time!) can be obtained as a function of \(V_k\) in Theorem 8.2.3 of Chung (1968). For more details, see Iglehart (1971a). Although the result is what we would expect, the construction is extremely important for cleaning up the theory of the GI/G/1 queue.

The upshot of the discussion above is that the successive busy cycles of the GI/G/1 queue in discrete and continuous time form persistent renewal processes when \(\rho < 1\). Consequently, we have a very powerful tool to analyze the GI/G/1 queue. The embedded renewal process is the key to finding limiting distributions of the continuous-time processes, cf. Feller ((1968), p. 319 and (1966), p. 365) and Takács (1963); it is the key to proving time-average limit theorems, cf. Iglehart (1971a); and it is the key to obtaining extreme value theorems, cf. Iglehart (1971b). The great power of embedded renewal processes was probably first realized by Smith ((1955, (1958)). Recent extensions appear in Brown and Ross (1970) and Miller (1971). Applications to the alternating-priority queue have been made by Stidham (1971). The time average limits of Iglehart (1971a) may be thought of as limits for cumulative processes, cf. Smith (1955). It is easy to see that functional strong laws are in fact equivalent to ordinary strong laws, so much of Section 3 of Iglehart (1971a) follows directly from the individual ergodic theorem, cf. Kiefer and Wolfowitz (1956). Various moments must be finite for these time-average limit theorems. For this purpose, Wald's equation, Theorem 5 of Kiefer and Wolfowitz (1956), Theorems 8.4.3 and 8.4.4 of Chung (1968), and Theorem 3 of Heyde (1964) can be used, cf. Iglehart (1971a).

We remark that the independence assumptions can be relaxed for much of the above. Instead of independence, the basic data can be assumed to be stationary satisfying various mixing conditions. The individual ergodic theorem then implies that \(\{W_n = 0\}\) happens infinitely often with probability one. The busy cycles remain optional random variables, but the basic sequence \(\{u_n, v_n\}\) is no longer i.i.d. so that Theorem 8.2.3 of Chung (1968) does not apply. For treatments of stability, limiting distributions, time-average limits, and extreme value theorems in this setting,
see Loynes (1962a, b) and Whitt (1971). Work by Serfozo (1971) on semi-stationary processes also appears to be relevant to this approach.

2. Multi-server queues

Multi-server queues such as the GI/G/s queue have shown a remarkable resistance to analysis, cf. Kingman ([1966], Section 12). (The recent thesis of de Smit (1971) is devoted to the GI/G/s queue and has many references.) One successful approach has been through a vector-valued waiting time process \( \{W_n, n \geq 0\} \) introduced by Kiefer and Wolfowitz ((1955), (1956)). Assuming that customers are served by the first available server, with a specified mechanism to break ties, the vector \( W_n = (W_{n1}, \ldots, W_{ns}) \) is obtained by assigning all customers in the system at the \( n \)th arrival epoch to the servers who will eventually serve them. Then look at the resulting workload in service time facing each server and rearrange them so that \( W_{n1} \) is the workload facing the server with the lightest load, \( W_{n2} \) is the workload facing the server with the second lightest load, and so forth. Consequently, \( W_{n1} \) represents the actual waiting time of the \( n \)th customer (until he reaches the server). Formally, we define \( \{W_n, n \geq 0\} \) recursively in terms of the same basic data \( \{(u_n, v_n), n \geq 0\} \) in (1.1) by setting \( W_0 = (0, \ldots, 0) \) and

\[
W_{n+1} = \left[F(W_n + V_n) - U_n\right]^+,
\]

where \( V_n = (v_n, 0, \ldots, 0), U_n = (u_n, \ldots, u_n), [X]^+ = (x_1^+, \ldots, x_r^+) \) with \( x^+ = \max\{0, x\} \) for arbitrary \( X = (x_1, \ldots, x_r) \in \mathbb{R}^r \), and \( F: \mathbb{R}^r \to \mathbb{R}^r \) rearranges the components of \( X = (x_1, \ldots, x_r) \) in ascending order. Note that the sequence \( \{W_n, n \geq 0\} \) of actual waiting times is not Markov in the GI/G/s queue, but the sequence \( \{W_n, n \geq 0\} \) defined in (2.1) is. In fact, in the GI/G/s queue the sequence \( \{W_n, n \geq 0\} \) is a random walk in \( \mathbb{R}^r \) which is restricted by impenetrable barriers to the set of \((x_1, \ldots, x_r) \in \mathbb{R}^r \) such that \( 0 \leq x_1 \leq \cdots \leq x_r \).

Loynes (1962a) has obtained stability results for the \( s \)-server queue as well as the single-server queue. If the basic sequence \( \{(u_n, v_n), n \geq 0\} \) in (1.1) is only assumed to be strictly stationary and ergodic, if \( \rho = Ev/sEu < 1 \), and if the queue is initially empty, then \( W_n \) converges in distribution monotonically to a finite limit \( W \), which when used as the initial distribution makes the process stationary. If the queue is not initially empty, then no limiting distribution need exist (under the condition of stationarity). This intriguing possibility is illustrated by the following example.

**Example 2.1.** (Loynes (1962a), p. 516) Let \( s = 2 \) and \( u_n = 1 \) for all \( n \). Let the probability space contain two points, each having probability \( \frac{1}{2} \), at one of which \( v_{2n} = \frac{1}{2} \) and \( v_{2n+1} = 2 \) and at the other \( v_{2n} = 2 \) and \( v_{2n+1} = \frac{1}{2} \). Note that \( \rho = Ev/sEu = \frac{7}{8} < 1 \). Then \( W_{2n} = (0, a) \) and \( W_{2n+1} = (a - 1, \frac{1}{2}) \) at the former point and those interchanged for the latter form a stationary sequence satisfying...
for any \( a, 1 \leq a \leq \frac{1}{2} \). However, if \( W_1 = (0, a) \) where \( 1 < a \leq \frac{1}{2} \), then the sequence oscillates (is property substable).

It is well known that the arrival epochs after the system has been completely empty are regeneration points for the \( GI/G/s \) queue as well as the \( GI/G/1 \) queue. At these time points all the processes probabilistically restart themselves. Consequently, if it is possible to show that the system will be empty infinitely often with probability one, then much of the theory for the \( GI/G/1 \) queue would carry over to the \( GI/G/s \) queue. Some authors (Finch (1959), p. 330; Jacobs (1971), p. 103, not Ross (1970), p. 777) apparently believe this to be true when \( \rho < 1 \) and attribute it to Kiefer and Wolfowitz ((1955),(1956)). However, the complexity of the argument in Section 6 of Kiefer and Wolfowitz (1955) indicates that they were well aware that the simpler approach was not available to them. Unfortunately, a stable multi-server queue does not necessarily have the propitious property of returning to the idle state infinitely often. Example 2.1 illustrates this, as does the following example.

**Example 2.2.** Consider a \( GI/G/2 \) queue in which

\[
v_n - 2u_n < 0 < v_n - u_n
\]

with probability one. If \( W_0 = (0, x) \), where \( 0 \leq x_0 < u_n \) with probability one then \( W_{n+1} = (0, v_n - u_n) \) for all \( n \geq 0 \).

In Example 2.2 note that any \( x_0 > 0 \) will not do. If \( x_0 = 2v_0 \), then \( W_2 = (0, 0) \). Note that bounded interarrival times and service times for which \( \rho = \frac{Ev}{sEu} < 1 \) can be used to get (2.2) so that the counterexample could prevail even if moments of all orders exist. However, it is easy to get sufficient conditions to rule out such pathologies.

**Theorem 2.1.** If \( \rho < s^{-1} \) in the \( GI/G/s \) queue, then

\[
P\{W_n = (0, \ldots, 0) \text{ infinitely often} \} = 1
\]

**Proof.** First observe in (2.1) that if \( V'_n \geq V_n \) for all \( n \), then \( W'_n \geq W_n \) for all \( n \). Let \( V'_n = (v_n, \ldots, v_n) \). Then \( \{W'_n, n \geq 0\} \) is just the \( s \)-fold copy of a single-server queue with interarrival times \( u_n \) and \( v_n \). Then apply the result in Section 1. Such order-preserving properties are discussed at length in Jacobs (1971). They are also used by Kiefer and Wolfowitz (1955).

**Theorem 2.2.** If \( \rho = \frac{Ev}{sEu} < 1 \) and \( P\{u_n - v_n > 0\} > 0 \) in the \( GI/G/s \) queue, then

\[
P\{W_n = (0, \ldots, 0) \text{ infinitely often} \} = 1
\]

**Proof.** We apply the ingenious device of Kiefer and Wolfowitz (1955). We domi-
nate the sequence \( \{ W_n, n \geq 0 \} \) by a lattice process and then apply available theorems for denumerable-state discrete-time Markov chains. Following Section 6 of [12], we let \( u'_n = c[u_n/c] \) and \( v'_n = c[v_n/c] + c \), where \( c \) is chosen sufficiently small that \( p' = Ev'/sEv' < 1 \) and \( P\{u'_n - v'_n > 0\} > 0 \). Kiefer and Wolfowitz (1955) have shown that such discrete-time Markov chains have only one ergodic set which is positive recurrent (persistent non-null) and aperiodic. Since \( P\{u'_n - v'_n > 0\} > 0 \), it is possible to move to the left in each coordinate in any step. Consequently, for any \((x_1, \ldots, x_d)\), there exists a \( k > 0 \) such that

\[
P\{W_{n+k} = (0, \ldots, 0) \mid W_n = (x_1, \ldots, x_d)\} > 0.
\]

This means that \((0, \ldots, 0)\) will be visited infinitely often with probability one, so \((0, \ldots, 0)\) is not transient. Since the ergodic set is positive recurrent, the expected time between visits to \((0, \ldots, 0)\) is finite for \( \{W_n\} \). Finally \( W_n \leq W'_n \).

Under the assumptions of Theorem 2.2, optional random variables which are finite with probability one can be defined relative to \( \{u_n, v_n\} \). In this case

\[
\alpha^1 = \min\{k \geq 1 : W_k = (0, \ldots, 0)\},
\]

with \( \alpha^1 = +\infty \) if the minimum is never attained. It is necessary to observe that \( \{\alpha^1 = m\} \) is contained in the augmented \( \sigma \)-field generated by the first \( m \) pairs in the sequence \( \{u_n, v_n\} \) for any \( m \geq 1 \), cf. Iglehart ((1971a), Section 2). Then Theorem 8.2.3 of Chung (1968) can be applied and the busy cycles in continuous time can be defined. Just as in Iglehart (1971a), these are

\[
(2.4) \quad \xi_k = u_{\beta_{k-1} + \cdots + \beta_k},
\]

where \( \beta_k = \alpha^1 + \cdots + \alpha^k \), \( \alpha^k = \alpha^{k-1} \circ \tau \), and \( \tau \) is the shift.

Once this set up has been established, associated results such as functional limit theorems can be obtained, although there appears to be no easy way to represent the constants explicitly. For example, consider the queue length process \( \{Q(t), t \geq 0\} \). We give a somewhat imprecise statement of the available results, providing additional details in the proof. We state the ordinary limit theorems in \( R^1 \) which can be obtained from the functional versions by projections, cf. Iglehart (1971a).

**Theorem 2.3.** Let \( p = Ev'/sEv' < 1 \) and \( P\{u_n - v_n > 0\} > 0 \) in the GI/G/s queue. If appropriate moments are finite, then:

(a) there exists a finite random variable \( Q \) such that

\[
\lim_{t \to \infty} P\{Q(t) \leq x\} = P\{Q \leq x\}
\]
for all continuity points of the limit;

(b) \[ \lim_{t \to \infty} EQ(t) = EQ; \]

(c) \[ \lim_{t \to \infty} t^{-1} \int_0^t Q(s) ds = EQ \quad \text{a.e.}; \]

(d) \[ \lim_{t \to \infty} \mathbb{P} \left\{ \int_0^t Q(s) ds - tEQ \leq x \right\} = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy; \]

(e) \[ \limsup_{t \to \infty} \frac{\int_0^t Q(s) ds - tEQ}{(2t \log \log t)^{\frac{1}{2}}} = c \quad \text{a.e.} \]

Proof. (a) We need to exclude periodicity, which we can do by assuming that the interarrival times are non-lattice. No moments being finite are necessary other than the time to return to \((0, \ldots, 0)\), which we obtain from the proof of Theorem 2.2. We apply p. 365 of Feller (1966), cf. Example (a) of p. 366. Actually, the assumption that the interarrival times be non-lattice is not quite sufficient, but it is if we assume \(\{Q(t), t \geq 0\}\) has sample paths in \(D = D[0, \infty)\) with probability one, where \(D\) is the space of right-continuous real-valued functions on \([0, \infty)\) with left limits everywhere, cf. Miller ((1971), p. 21). Paths being in \(D\) is of course always necessary for the functional limit theorems.

(b) Theorem 2 of Kiefer and Wolfowitz (1956) implies that

\[ EW_n < EW < \infty \quad \text{if} \quad EV_n^2 < \infty, \]

where \(W_n\) is the waiting time of the \(n\)th customer \((W_n = W_{n+1})\) and \(W\) is the stationary waiting time. However,

\[ EW_n = \sum_{k=0}^{\infty} E\{W_n \mid Q_n = k\} P\{Q_n = k\} \]

\[ \geq \sum_{k=0}^{\infty} (kEV/s)P\{Q_n = k\} \]

\[ = (EV/s)EQ_n \]

\[ > (EV/s)(EQ(t) - 1), \]

where \(Q_n\) is the number of customers in the system at the epoch of the \(n\)th arrival and \(t\) is some point between the \(n\)th and \((n + 1)\)th arrival. As a consequence, if \(EV_n^2 < \infty\), then

\[ \sup_{t \geq 0} EQ(t) \leq (s/EV)EW + 1 < \infty, \]

so that \(\{Q(t), t \geq 0\}\) is uniformly integrable, which with (a) implies (b).
Embedded renewal processes in the GI/G/s queue

The existence of the limit is an easy consequence of adding up a random number of independent chunks, cf. Iglehart (1971a). This argument shows that the limit equals the expected value of the integral of the queue length process over one cycle divided by the expected length of one cycle. That this coincides with $E_\lambda$ can be deduced from (b). For an argument not using (b), see Section 2 of Brown and Ross (1970).

Again Iglehart's (1971a) argument can be used. The translation term can be deduced from (c). The constant $c$ is not known even for the GI/G/1 queue, cf. Iglehart (1971a), Theorem 4.1. However, it can be estimated easily from independent observations of successive busy cycles. In Iglehart's (1971a) notation, $c$ is the second moment of

$$(Y^{(2)}_k - \xi_k E_\lambda) = (Y^{(2)}_k - \xi_k E_\lambda Y^{(2)}_k / E_\lambda^2),$$

where $Y^{(2)}_k$ is the integral of the queue length process over the $k$th cycle.

Follow Section 5 of Iglehart (1971a). Similarly for $\lim\inf$. From the point of view of applications, it is not so bad that the constants $E_\lambda$ and $c$ in Theorem 2.3 cannot be expressed simply in terms of the interarrival time and service time random variables and their moments. The constants $E_\lambda$ and $c$ can be estimated from successive observations from the sequence of i.i.d. busy cycles, that is, as functions $V_k$ in Theorem 8.2.3 of Chung (1968) or Lemma 2.1 of Iglehart (1971). This seems to be a good point to apply computational methods and data analysis. If the object is estimation, then Theorem 2.3 establishes consistency and asymptotic normality.

References