Queueing Systems, 5 (1989) 381–392

LIMITS FOR QUEUES AS THE WAITING ROOM GROWS

Daniel P. HEYMAN
Bellcore, Red Bank, NJ 07701, U.S.A.

and

Ward WHITT
AT&T Bell Laboratories, Murray Hill, NJ 07974, U.S.A.

Received 21 July 1988; revised 6 June 1989

Abstract

We study the convergence of finite-capacity open queueing systems to their infinite-capacity counterparts as the capacity increases. Convergence of the transient behavior is easily established in great generality provided that the finite-capacity system can be identified with the infinite-capacity system up to the first time that the capacity is exceeded. Convergence of steady-state distributions is more difficult; it is established here for the GI/GI/c/n model with c servers, $n-c$ extra waiting spaces and the first-come first-served discipline, in which all arrivals finding the waiting room full are lost without affecting future arrivals, via stochastic dominance and regenerative structure.

Keywords: Queueing theory, limit theorems, approximation, truncation, finite waiting rooms, regenerative processes, stochastic comparisons.

1. Introduction

Consider an open queueing system with capacity n. When n is very large, we expect that the standard descriptive stochastic processes, such as the number of customers in the system at time t for $t \geq 0$, and their limiting steady-state distributions are very close to their counterparts in the same system with infinite capacity. Indeed, for simple models such as the M/M/c/n queue (c servers and $n-c$ extra waiting spaces) for which the steady-state distributions can be displayed explicitly, convergence of the steady-state distributions as $n \to \infty$ is easily verified (provided that the infinite-capacity model is stable). We establish convergence results here that do not depend on explicit expressions for the quantities of interest. We also give examples to show that some care is needed.

© J.C. Baltzer A.G. Scientific Publishing Company

In section 2 we establish very strong convergence (total variation) in great generality for the stochastic processes representing the transient behavior, provided that we can represent the finite-capacity system up to the first time that the capacity would be exceeded in terms of the infinite-capacity system. The real difficulty is obtaining convergence of the limiting steady-state distributions. In section 3 we show how the limits for the transient behavior in section 2 can be applied in the presence of regenerative structure to obtain convergence of the steady-state distributions. The results in section 3 also are very general, but it is necessary to control the behavior of the regeneration cycles in the n-capacity systems given that the capacity limit is exceeded. For the special case of single-facility loss systems, we provide a stochastic bound in section 4 that can be used to provide this control. Finally, in section 5 we combine the results of the previous sections to establish limits for the steady-state distributions of GI/GI/c/n loss systems as $n \to \infty$.

There is considerable related literature. The limits here express a form of model stability, continuity or robustness; see chapter 3 of Franken, König, Arndt and Schmidt [6], chapter 8 of Stoyan [15], chapter 4 of Borovkov [3], Brandt and Lisek [4], Kalashnikov and Rachev [9], Rachev [13] and Karr [10]. Perhaps more closely related is the literature about approximating countable-state Markov chains by finite-state Markov chains; see Wolf [19], Gibson and Seneta [7] and references cited there. In that context, the desired conclusion is that the steady-state distributions in the finite-state chains converge to the steady-state distribution of the infinite-state chain as the size of the state space grows. The classical paper by Ledermann and Reuter [12] established limits of this kind for birth-and-death processes.

2. Convergence of transient behavior

Let the stochastic process $[X_\infty, Y_\infty] \equiv \{[X_\infty(t), Y_\infty(t)]: t \geq 0\}$ describe an open queueing system with infinite capacity. (In this section the queueing system can be very general, e.g., a multi-class open queueing network. Indeed, we need not even have a queueing system.) The random variable $X_\infty(t)$ typically represents the number of customers in the system at time t, and the random variable $Y_\infty(t)$ typically represents other aspects of interest at time t, such as residual interarrival times and service times. The random variable $Y_\infty(t)$ might contain supplementary variables to makes $[X_\infty, Y_\infty]$ a Markov process, but need not. We assume that $X_\infty(t)$ is real-valued and $[X_\infty(t), Y_\infty(t)]$ takes values in a complete separable metric space S, endowed with the Borel σ-field, (which usually would be Euclidean space, but need not be) and that the sample paths of $[X_\infty, Y_\infty]$ are RCLL (right-continuous with left limits), so that $[X_\infty, Y_\infty]$ can be regarded as a random element of the function space $D[0, \infty)$; see Ch. 3 of Billingsley [2], Sec. 2 of Whitt [17] and Ch. 3 of Ethier and Kurtz [5].

Our most important assumption concerns the way the finite-capacity systems are related to the infinite-capacity system. We assume that the system with capacity n can be constructed in terms of the infinite-capacity system up to the first time that the capacity exceeds n. Let T_n be the first passage time defined by

$$T_n = \inf\{t \geq 0: X_\infty(t) > n\}, \quad n \geq 1. \tag{2.1}$$

REPRESENTATION ASSUMPTION

We assume that $[X_n, Y_n]$ is defined on the same probability space as $[X_\infty, Y_\infty]$ and that $[X_n(t), Y_n(t)] = [X_\infty(t), Y_\infty(t)]$ for $0 \leq t < T_n$.

So far, we have said nothing about $[X_n(t), Y_n(t)]$ for $t \geq T_n$. However, the Representation Assumption already implies that the transient behavior of $[X_n, Y_n]$ converges to the transient behavior of $[X_\infty, Y_\infty]$ in a very strong sense. The RCLL property implies that for each positive t

$$\sup_{0 \leq s \leq t} \{X_\infty(s)\} < \infty \text{ w.p.1}, \tag{2.2}$$

p. 110 of Billingsley [2] and p. 70 of Whitt [17], which in turn guarantees that

$$\lim_{n \to \infty} T_n = \infty \text{ w.p.1}. \tag{2.3}$$

The Representation Assumption implies that

$$P([X_n(s), Y_n(s)] = [X_\infty(s), Y_\infty(s)], 0 \leq s \leq t) \geq P(T_n > t) \tag{2.4}$$

for all positive n and t. Properties (2.3) and (2.4) imply convergence in total variation for the probability distributions provided we restrict attention to bounded time intervals. Let π_t be the projection map from the function space $D[0, \infty)$ onto $D[0, t]$ defined by $\pi_t(x)(s) = x(s)$, $0 \leq s \leq t$. Then

$$\lim_{n \to \infty} \| \pi_t[X_n, Y_n] - \pi_t[X_\infty, Y_\infty] \|$$
$$\equiv \lim_{n \to \infty} \sup_A | P(\pi_t[X_n, Y_n] \in A) - P(\pi_t[X_\infty, Y_\infty] \in A) | = 0, \tag{2.5}$$

where A is a measurable subset of the function space $D[0, t]$. The total variation norm $\| \cdot \|$ in (2.5) applies to the *probability measures* induced by the random elements $\pi_t[X_n, Y_n]$ and $\pi_t[X_\infty, Y_\infty]$ on $D[0, t]$. An elementary consequence of (2.5) is the corresponding limit for the one-dimensional marginals, i.e.,

$$\lim_{n \to \infty} \| [X_n(t), Y_n(t)] - [X_\infty(t), Y_\infty(t)] \|$$
$$\equiv \lim_{n \to \infty} \sup_A | P([X_n(t), Y_n(t)] \in A) - P([X_\infty(t), Y_\infty(t)] \in A) | = 0, \tag{2.6}$$

where now A is a measurable subset of S. Another elementary consequence of (2.5) is that

$$[X_n, Y_n] \Rightarrow [X_\infty, Y_\infty] \text{ as } n \to \infty \text{ in } D[0, \infty), \tag{2.7}$$

where $\Rightarrow$ denotes convergence in distribution (weak convergence), and $D[0, \infty)$ is endowed with the Skorohod J_1 topology on $D[0, \infty)$; i.e.,

$$\lim_{n\to\infty} Ef([X_n, Y_n]) = Ef([X_\infty, Y_\infty]) \tag{2.8}$$

for all continuous bounded real-valued functions on $D[0, \infty)$.

While we have established convergence of the processes (transient behavior) as $n \to \infty$ in great generality, we have yet to treat stationary or limiting distributions. The following example illustrates some of the difficulties.

EXAMPLE 2.1

Let the infinite capacity system be a simple M/M/1 queue with traffic intensity $\rho < 1$. Let the associated n-capacity system be the modification in which the system closes down, i.e., empties immediately with all service times set equal to 0, the instant an arriving customer finds n customers in queue. Clearly the representation assumption above holds, so that (2.2)–(2.9) hold, but the limiting distributions do not converge. □

3. Regenerative framework

We obtain convergence of steady-state distributions from convergence of the transient behavior by exploiting regenerative structure. We now assume for $n \leqslant \infty$ that the stochastic processes $[X_n, Y_n]$ are *regenerative* with generic *cycle times* C_n having non-lattice distributions with $0 < E(C_n) < \infty$, so that

$$[X_n(t), Y_n(t)] \Rightarrow [X_n(\infty), Y_n(\infty)] \text{ as } t \to \infty \tag{3.1}$$

and

$$E(f[X_n(\infty), Y_n(\infty)]) = \frac{E\left[\int_0^{C_n} f[X_n(t), Y_n(t)]\, dt\right]}{E(C_n)} \tag{3.2}$$

for any bounded measurable real-valued function f on the state space S; see section V.1 of Asmussen [1]. As in [1], the cycle times C_n are i.i.d., but there may be some dependence between cycles of the process; see theorem 5(b) here. The expectations on the right of (3.2) refer to the *zero-delayed* case; i.e., we are assuming a regeneration point at $t = 0$.

To connect $[X_n, Y_n]$ to $[X_\infty, Y_\infty]$, we also assume that these cycle times C_n are consistent with our basic Representation Assumption in section 2, in a sense to be defined below. Let 1_A denote the indicator function of the set A, i.e., $1_A(x) = 1$ if $x \in A$ and 0 otherwise.

EXTENDED REPRESENTATION ASSUMPTION

In addition to the Representation Assumption above, we assume that $1_{\{C_n < T_n\}} = 1_{\{C_\infty < T_n\}}$ and $C_n 1_{\{C_n < T_n\}} = C_\infty 1_{\{C_\infty < T_n\}}$ w.p.1.

As a consequence of the extended representation assumption,

$$E\left[\int_0^{C_n} f[X_n(t), Y_n(t)]\, dt\right] = E\left[1_{\{C_\infty < T_n\}} \int_0^{C_\infty} f[X_\infty(t)], Y_\infty(t)]\, dt\right] + E\left[1_{\{C_n \geq T_n\}} \int_0^{C_n} f[X_n(t), Y_n(t)]\, dt\right] \quad (3.3)$$

for any bounded measurable real-valued function f on S. We now can apply (3.3) to obtain a condition for convergence of the steady-state distributions in total variation, in the sense of (2.6).

THEOREM 1

If $E[C_n 1_{\{C_n \geq T_n\}}] \to 0$ as $n \to \infty$ in this regenerative framework with the extended representation assumption, then

$$\| [X_n(\infty), Y_n(\infty)] - [X_\infty(\infty), Y_\infty(\infty)] \| \to 0 \text{ as } n \to \infty.$$

Proof

Since f is bounded, say by M,

$$E\left[1_{\{C_n \geq T_n\}} \int_0^{C_n} f[X_n(t), Y_n(t)]\, dt\right] \leq ME\left[C_n 1_{\{C_n \geq T_n\}}\right],$$

so that the second term in (3.3) is asymptotically negligible by the assumption. By (2.3) and the Lebesgue dominated convergence theorem, the first term in (3.3) converges to $E[\int_0^{C_\infty} f[X(t), Y(t)]\, dt]$. By this argument, both the numerators and denominators on the right of (3.2) converge. Moreover, the convergence is uniform in f for f of the form 1_A. □

Of course, the condition in theorem 1 will not always be satisfied. We now present a sufficient condition that we will apply to GI/GI/c/n queues. The idea is to bound the quantities associated with the n-capacity systems by a quantity associated with the infinite-capacity system.

PROPOSITION 2

If there exists a nonnegative random variable Z such that $E(Z) < \infty$ and

$$E\left[C_n 1_{\{C_n \geq T_n\}}\right] \leq E\left[[Z + C_\infty] 1_{\{C_\infty \geq T_n\}}\right] \quad (3.4)$$

for all n sufficiently large or, equivalently, if

$$E\left[E\left[C_n | 1_{\{C_n \geq T_n\}}\right]\right] \leq E\left[E\left[Z + C_\infty | 1_{\{C_\infty \geq T_n\}}\right]\right] \quad (3.5)$$

for all n sufficiently large, then

$$E\left[C_n 1_{\{C_n \geq T_n\}}\right] \to 0 \text{ as } n \to \infty.$$

Proof

Since $E(C_\infty) < \infty$ and $E(Z) < \infty$, we can apply (2.3) and the Lebesgue dominated convergence theorem to conclude that the right side of (3.4) converges to 0 as $n \to \infty$. □

One way to establish (3.4) is to construct all systems on a common probability space appropriately ordered. (Such a construction is possible whenever there is stochastic order.) In section 5 we apply the next proposition with $Z = 0$.

PROPOSITION 3

If there exists a nonnegative random variable Z with $E(Z) < \infty$ such that

$$C_n \leq Z + C_\infty \text{ w.p.1}, \tag{3.6}$$

then (3.4) holds.

Proof

The assumptions imply that

$$C_n 1_{\{C_n \geq T_n\}} \leq (Z + C_\infty) 1_{\{C_\infty \geq T_n\}} \text{ w.p.1},$$

from which (3.4) is immediate. □

4. General multi-server FCFS loss systems

Now assume that the infinite-capacity queueing system is a single facility with unlimited waiting space, c servers working in parallel and the FCFS (first-come first-served) queue discipline. Let the stochastic behavior be specified by a sequence $\{(u_k, v_k): k \geq 1\}$ of ordered pairs of nonnegative random variables, where u_k represents the interarrival time between the $(k-1)^{\text{st}}$ and k^{th} arrival, and v_k is the service time of the k^{th} arrival, but we make *no* independence or common-distribution assumptions. Hence, we have an A/A/c/∞ model (A for arbitrary, instead of G for general stationary or GI for renewal). Let the system start out empty at time 0. Other initial conditions can be introduced through the basic sequence $\{(u_k, v_k)\}$; i.e., if $u_k = 0$ for $1 \leq k \leq K$ with $u_{K+1} > 0$, then there are K customers in the system at time 0.

The n-capacity system can be defined in terms of the infinite capacity system, using the same sequence $\{(u_k, v_k)\}$, by letting $n - c$ be the size of the waiting room and stipulating that arrivals that find the waiting room full are lost (an

A/A/c/n system). If customer k is lost, he takes his service time v_k away with him. Obviously, we have represented the capacity-n system in terms of the infinite-capacity system as assumed in section 2, so that the limits described there apply here. With the additional structure in this section, we can also say what happens *after* T_n; we can conclude that the infinite-capacity system serves as a bound for the finite-capacity systems, as needed in proposition 3. We use the fact that all customers admitted to the capacity-n system have the same service times as their counterparts in the infinite-capacity system.

THEOREM 4

With the special construction, $X_n(t) \leqslant X_\infty(t)$ for all n and t.

Proof

For any sample path, we can represent $\{X_n(t): t \geqslant 0\}$ in terms of $\{X_\infty(t): t \geqslant 0\}$ in two steps: first, by replacing the service times of all customers that would be lost by 0 and, second, by not counting these customers. Since customers with 0 service times do not affect the time in system of other customers, the second step is clearly consistent with the claim. For the first step, we use known monotonicity properties for A/A/c/∞ systems, in particular, theorem 8 and the following remark in Whitt [18]: Making the service times smaller can only reduce $X_\infty(t)$. To see this directly, recall that

$$D_n = U_n + W_n + v_n, \tag{4.1}$$

where D_n is the departure epoch, $U_n = u_1 + \cdots + u_n$ is the arrival epoch, and W_n the waiting time before beginning service of the n^{th} arrival. Decreasing some of the service times v_n causes D_n to decrease or remain the same, because U_n is unchanged and W_n was shown to be a nondecreasing function of $(v_1, \ldots, v_{n-1})$ by Kiefer and Wolfowitz [11]. Since all arrival epochs are the same and all departure epochs are ordered, all queue lengths are ordered. □

Recall that a set of probability measures on a complete separable metric space is tight if for each $\epsilon > 0$ there exists a compact subset K such that $P(K) > 1 - \epsilon$ for all P in the set; pp. 9, 37 of Billingsley [2]. Tightness guarantees that every sequence of probability measures from the set has a weak convergent subsequence (with a proper limit).

COROLLARY 1

If $\{X_\infty(t): t \geqslant 0\}$ is tight, then $\{X_n(t): t \geqslant 0, n \geqslant 0\}$ is tight, so that every sequence $\{X_{n_k}(t_k): k \geqslant 0\}$ has a weak convergent subsequence.

We say that one real-valued random variable X_1 is stochastically less than or equal to another, and write $X_1 \leqslant_{st} X_2$, if $P(X_1 > t) \leqslant P(X_2 > t)$ for all t; see chapter 1 of Stoyan [15].

COROLLARY 2

(a) If $X_\infty(t) \Rightarrow X_\infty(\infty)$ and $X_n(t) \Rightarrow X_n(\infty)$ as $t \to \infty$, then $X_n(\infty) \leqslant_{st} X_\infty(\infty)$.

(b) If (a) holds for all n, then $\{X_n(\infty): n \geqslant 1\}$ is tight.

It is of course also of interest to compare the finite-capacity systems for different capacity sizes. We would like to conclude that $X_{n_1}(t) \leqslant X_{n_2}(t)$ when $n_1 < n_2 < \infty$ with the special construction, but this is not true in general, as we show below. However, it is possible to show that the epoch of the k^{th} admitted arrival and the k^{th} departure (not the departure epoch of the k^{th} arrival) occur sooner in the system with larger capacity. This is verified by a minor modification of theorem 1 of Sonderman [14].

EXAMPLE 4.1

To see that we need not have $X_1(t) \leqslant X_2(t)$ for all t or

$$\overline{X}_1 \equiv \lim_{t\to\infty} t^{-1}\int_0^t X_1(s)\,ds \leqslant \lim_{t\to\infty} t^{-1}\int_0^t X_2(s)\,ds \equiv \overline{X}_2, \tag{4.2}$$

consider a D/A/1/∞ model in which $u_k = 1$, $v_{2k} = 2$ and $v_{2k+1} = \epsilon$ for all k and a small positive ϵ $(0 < \epsilon < 1/2)$. For the D/A/1/1 model constructed from it, $X_1(t) = 1$ for all $t \geqslant 2$, so that $\overline{X}_1 = 1$ in (4.2). For the D/A/1/2 model, $X_2(t) = 2$ for $3 \leqslant t < 4$; $X_2(t) = 1$ for $1 \leqslant t < 1+\epsilon$, $2 \leqslant t < 3$ and $4 \leqslant t < 4+\epsilon$; $X_2(t) = 0$ for $0 \leqslant t < 1$, $1+\epsilon \leqslant t < 2$ and $4+\epsilon \leqslant t < 5$. Moreover, the form of $X_2(t)$ in the interval $[1+4k, 5+4k]$ is independent of k, so that $\overline{X}_2 = (3+2\epsilon)/4$ in (4.2). For $\epsilon < 1/2$, $\overline{X}_2 < \overline{X}_1$.

As given, the D/A/1/∞ model above is not stable, but it can easily be made stable by inserting periodic blocks of 0 service times. This would reduce both $\overline{X}_1$ and $\overline{X}_2$, but leave $\overline{X}_2 < \overline{X}_1$. Moreover, as constructed, the processes $X_n(t)$ do not converge in distribution as $t \to \infty$. To obtain such convergence, we can perturb the model above slightly. In particular, consider the GI/A/1/∞ model in which u_k is uniformly distributed in $[1-\delta, 1+\delta]$ for very small δ and

$$\begin{aligned} P(v_{2k} = 2 \text{ and } v_{2k+1} = \epsilon \text{ for all } k) &= P(v_{2k} = \epsilon \text{ and } v_{2k+1} = 2 \text{ for all } k) \\ &= 1/2. \end{aligned}$$

Then $X_1(t) \Rightarrow X_1(\infty)$ and $X_2(t) \Rightarrow X_2(\infty)$ with $E[X_1(\infty)] \geqslant E[X_2(\infty)]$. □

5. The GI/GI/c/n loss model

Finally, consider the classical GI/GI/c/n loss model, i.e., the A/A/c/n model of section 4 with the additional assumption that $\{u_k\}$ and $\{v_k\}$ are independent sequences of i.i.d random variables with $E(u_1) < \infty$ and $E(v_1) < \infty$;

see chapter XI of Asmussen [1]. As usual, let $\rho = E(v_1)/cE(u_1)$. For $n \leqslant \infty$, let $Y_n(t)$ be a vector representing the number of busy servers, the residual service times of each and the residual interarrival time at time t, so that $[X_n, Y_n]$ is a Markov process. We say that a distribution function F is *spread out* if, for some n, the n-fold convolution F^{*n} satisfies $F^{*n}(x) \geqslant G(x) \geqslant 0$ for all x, where G is not identically zero and is absolutely continuous (has a density) with respect to Lebesgue measure; see p. 140 of [1].

THEOREM 5

Suppose that $\rho < 1$. If (a) $P(u_1 > v_1) > 0$ and u_1 has a non-lattice distribution or (b) u_1 has a spread-out distribution, then

$$[X_n(t), Y_n(t)] \Rightarrow [X_n(\infty), Y_n(\infty)] \text{ as } t \to \infty \text{ for } \textit{each } n \leqslant \infty \tag{5.1}$$

and

$$\| [X_n(\infty), Y_n(\infty)] - [X_\infty(\infty), Y_\infty(\infty)] \| \to 0 \text{ as } n \to \infty. \tag{5.2}$$

Proof

(a) For each $n \leqslant \infty$, the epochs just prior to an arrival in the empty state constitute regeneration points for the processes $[X_n, Y_n]$ and the cycle times C_n have nonlattice distributions with $E(C_n) < \infty$, so that (5.1) holds; for the case $n = \infty$, see Whitt [16] or proposition 3.2 on p. 187 and corollaries 2.5 and 2.8 on pp. 251–252, of Asmussen [1]. In this case, the process cycles as well as the cycle time C_∞ are i.i.d. Using the special construction in section 4, $X_n(t) \leqslant X_\infty(t)$ for all n and t, so that the result above holds for $n < \infty$ too, with $C_n \leqslant C_\infty$ w.p.1. Indeed, with the special construction in Section 4, we could use the regeneration points associated with $n = \infty$ for all n. (Then the regeneration points for $n < \infty$ would be a subset of the epochs just prior to an arrival in the empty state.) Hence, we can apply proposition 3, to obtain the condition of theorem 1, which implies (5.2).

(b) Now epochs just prior to an arrival in the empty state might not occur infinitely often, but the processes $[X_n, Y_n]$ become Harris recurrent Markov processes, and thus regenerative processes for which (3.2) holds; see pp. 126, 150, 252 of [1]. For the case $n = \infty$, the cycle times can be arrival epochs associated with the arrival indices that serve as the regeneration points for the discrete-time vector waiting time process constructed in lemma 2.4 on p. 250 of [1]. We use the fact that the final c arriving customers in the cycle enter service immediately upon arrival and all previous customers are gone by the end of the cycle; see corollary 2.8 on p. 252 of [1], where total-variation convergence as in (5.2) is established for (5.1) with $n = \infty$. In this case, the process cycles need not be independent. By using the construction in the proof of theorem 4, i.e., by assigning lost customers zero service times, we can construct the regeneration points in such a way that the extended representation assumption in section 3 is

satisfied; note that the vector waiting time sequences are also ordered under the construction in section 4. Indeed, after constructing all processes on the sample space, we can use the regeneration points associated with $n = \infty$ for $n < \infty$. The stationary distribution at these points is independent of n for $n \geq c$. Finally, the regenerative cycle distributions are spread out by proposition 3.2 on p. 187 of [1]. The rest of the proof is as in (a). □

Let the workload at time t be the sum of all remaining service times of customers in the system at time t. We can obtain convergence of the limiting workload distributions as $n \to \infty$ directly from theorem 5 by including in $Y_n(t)$ the remaining service time of each of the $X_n(t)$ customers in the system at time t.

By essentially the same argument, we can obtain convergence as $n \to \infty$ of the limiting distributions for associated embedded sequences, e.g., for the number in system just prior to the n^{th} arrival and the waiting time of the n^{th} arrival. If we focus on external arrivals, then we can apply theorem 4, because the arrival epochs will be the same for all n. (A lost customer departs the same time it arrives.) Under the condition of theorem 5(a), the cycle times C_n now consist of the number of arrivals between successive epochs just prior to an arrival in the empty state, and the integral in (3.2) is replaced by a sum. If we want to consider only arrivals that actually enter the system, then C_n becomes smaller, so that we still have $C_n \leq C_\infty$ w.p.1.

6. Embedded Markov chains in M/G/1 and GI/M/1

The queueing systems M/G/1 and GI/M/1 can also be analyzed by focusing on embedded discrete-time Markov chains, but even when we have discrete-time Markov chains the established theory ([7], [19]) does not always apply. For the M/G/1/n queue, the transition matrix of the usual embedded chain (looking just after departures) is a truncation of the transition matrix for $n = \infty$. Moreover, the transition matrix for $n = \infty$ has upper-Hessenberg form, so that previous theory establishes that the limiting distributions converge as $n \to \infty$; see [7]. On the other hand, for GI/M/1/n the transition matrix of the usual embedded chain (looking just before arrivals) is not a truncation of the transition matrix for $n = \infty$, so that the previous theory does not apply. We remark that sections 2 and 3 can also be applied directly to approximate general infinite-state Markov chains by associated finite-state truncations; then we do not need Y_∞ and Y_n.

7. Extensions

The results in sections 2 and 3 remain valid if T_n in (2.1) is a more general hitting time, provided that (2.3) holds. Moreover, $X_\infty(t)$ need not be real-valued.

For example, in a queueing network $X_\infty(t)$ might be the vector representing the number of customers at each queue and T_n might be the first time any queue length exceeds n. Of course, to establish convergence of steady-state distributions in other models, we must establish the condition in theorem 1. It is also not essential that the sample paths of $[X_\infty, Y_\infty]$ be RCLL; this is a convenient regularity condition to obtain (2.2) and appropriate measurability.

Finally, it is not essential to have the independence associated with the regenerative framework in section 3; it suffices to have (3.2) and the extended representation assumption. In a general stationary framework, (3.2) can often be obtained from the stochastic mean value theorem; see (1.5.3) on p. 45 of Franken et al. [6]. (Then we might not have (3.1); i.e. $[X_n(\infty), Y_n(\infty)]$ might have the stationary distribution without there being a limiting distribution.)

As a consequence, to obtain (5.2), but not necessarily (5.1), instead of the independence conditions in section 5 we can assume that the sequence of ordered pairs $\{(u_k, v_k)\}$ is stationary and ergodic (the G/G/c/n model), provided that there exist appropriate renewing events or construction points, as in chapter 4 of Borovkov [3] or chapter 2 of Franken et al. [6], which are consistent with the extended representation assumption. The natural renewing events are arrivals to an empty system. As in the GI/GI/c/n model, there are always infinitely many of these simple renewing events for $c = 1$ but not necessarily for $c > 1$; see section 2.4 of [6] and section 4.7 of [3].However, the main point is that theorem 5 extends readily to a very large class of G/G/c/n models.

An interesting goal for further research is to obtain quantitative estimates of the distance between the distributions of $[X_n(\infty), Y_n(\infty)]$ and $[X_\infty(\infty), Y_\infty(\infty)]$. An approach that should prove useful for this purpose pointed out by a referee is the uniform-in-time continuity method for regenerative processes in section 4 of Kalashnikov [8].

References

[1] S. Asmussen, *Applied Probability and Queues* (Wiley, New York, 1987).

[2] P. Billingsley, *Convergence of Probability Measures* (Wiley, New York, 1968).

[3] A.A. Borovkov, *Asymptotic Methods in Queueing Theory* (Wiley, Chichester, 1984).

[4] A. Brandt and B. Lisek, On the continuity of G/GI/m queues, Math. Operationsforsch. Statist., Ser. Statist. 12 (1983) 577–587.

[5] S.N. Ethier and T.G. Kurtz, *Markov Processes, Characterization and Convergence* (Wiley, New York, 1986).

[6] P. Franken, D. König, U. Arndt and V. Schmidt, *Queues and Point Processes* (Akademie-Verlag, Berlin, 1981).

[7] D. Gibson and E. Seneta, Augmented truncations of infinite stochastic matrices, J. Appl. Prob. 24 (1987) 600–608.

[8] V.V. Kalashnikov, The analysis of continuity of queueing systems, In: *Probability Theory and Mathematical Statistics*, eds. K. Itô and J.V. Prokhorov, Lecture Notes in Mathematics 1021 (Springer-Verlag, New York, 1983) 268–278.

[9] V.V. Kalashnikov and S.T. Rachev, *Mathematical Methods for Construction of Queueing Models* (Nauka, Moscow, 1988) (in Russian, to be translated into English be Wadsworth and Brooks/Cole).
[10] A.F. Karr, Weak convergence of a sequence of Markov chains, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 33 (1975) 41–48.
[11] J. Kiefer and J. Wolfowitz, On the theory of queues with many servers, Trans. Amer. Math. Soc. 78 (1955) 1–18.
[12] W. Ledermann and G.E.H. Reuter, Spectral theory for the differential equations of simple birth and death processes, Phil. Trans. Roy. Soc., London A246 (1954) 321–369.
[13] S.T. Rachev, The problem of stability in queueing theory, Queueing Systems 4 (1989) 287–318.
[14] D. Sonderman, Comparing multi-server queues with finite waiting rooms, I: same number of servers, Adv. Appl. Prob. 11 (1979) 439–447.
[15] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models* (Wiley, Chichester, 1983).
[16] W. Whitt, Embedded renewal processes in the GI/G/s queue, J. Appl. Prob. 9 (1972) 650–658.
[17] W. Whitt, Some useful functions for functional limit theorems, Math. Opns. Res. 5 (1980) 67–85.
[18] W. Whitt, Comparing counting processes and queues, Adv. Appl. Prob. 13 (1981) 207–220.
[19] D. Wolf, Approximation of the invariant probability measure of an infinite stochastic matrix, Adv. Appl. Prob. 12 (1980) 710–726.