On the Heavy-Traffic Limit Theorem for $GI/G/\infty$ Queues

Ward Whitt

Advances in Applied Probability, Vol. 14, No. 1 (Mar., 1982), 171-190.

Stable URL:
http://links.jstor.org/sici?sici=0001-8678%28198203%2914%3A1%3C171%3AOTHLTF%3E2.0.CO%3B2-R

Advances in Applied Probability is currently published by Applied Probability Trust.

Adv. Appl. Prob. **14**, 171–190 (1982)
Printed in N. Ireland
0001-8678/82/01/0171-20$02.25

ON THE HEAVY-TRAFFIC LIMIT THEOREM FOR $GI/G/\infty$ QUEUES

WARD WHITT,* *Bell Laboratories*

Abstract

A revealing alternate proof is provided for the Iglehart (1965), (1973)–Borovkov (1967) heavy-traffic limit theorem for $GI/G/s$ queues. This kind of heavy traffic is obtained by considering a sequence of $GI/G/s$ systems with the numbers of servers and the arrival rates going to ∞ while the service-time distributions are held fixed. The theorem establishes convergence to a Gaussian process, which in general is not Markov, for an appropriate normalization of the sequence of stochastic processes representing the number of customers in service at arbitrary times. The key idea in the new proof is to consider service-time distributions that are randomly stopped sums of exponential phases, and then work with the discrete-time vector-valued Markov chain representing the number of customers in each phase of service at arrival epochs. It is then easy to show that this sequence of Markov chains converges to a multivariate O–U (Ornstein–Uhlenbeck) diffusion process by applying simple criteria in Stroock and Varadhan (1979). The Iglehart–Borovkov limit for these special service-time distributions is the sum of the components of this multivariate O–U process. Heavy-traffic convergence is also established for the steady-state distributions of $GI/M/s$ queues under the same conditions by exploiting stochastic-order properties.

$GI/G/s$ QUEUE; HEAVY TRAFFIC; LIMIT THEOREM; DIFFUSION PROCESS; ORNSTEIN–UHLENBECK PROCESS; APPROXIMATIONS; CONGESTION MODELS

1. Introduction and summary

The purpose of this paper is to provide some complements to the Iglehart (1965), (1973a)–Borovkov (1967) heavy-traffic limit theorem for $GI/G/\infty$ service systems. Heavy traffic is achieved by considering a sequence of $GI/G/\infty$ service systems with the arrival rates going to ∞ while the service-time distributions are held fixed. The Iglehart–Borovkov limit theorem establishes convergence to a Gaussian process for an appropriate normalization of the sequence of stochastic processes representing the number of customers in service; see (2.2) and (2.3).

The Iglehart–Borovkov theorem also applies to $GI/G/s$ queueing systems if the number of servers also goes to infinity sufficiently fast. Then the s-server

Received 16 April 1980; revision received 9 April 1981.

* Postal address: WB-1A350, Bell Laboratories, Holmdel, NJ 07733, U.S.A.

queueing systems are asymptotically indistinguishable from infinite-server systems; i.e., the probability that all servers are busy during any interval goes to 0. An interesting alternative limiting procedure in which the probability that all servers are busy does not go to 0 is introduced and investigated by Halfin and Whitt (1981).

Perhaps the most significant feature of the Iglehart–Borovkov limit theorem is the prominent role played by the service-time distribution. If the service-time distribution is exponential or is exponential except for a mass at 0, then the limiting Gaussian process is the Ornstein–Uhlenbeck diffusion process, i.e., the Gaussian process is Markov. However, if the service-time distribution is not one of these special cases, then the limiting Gaussian process is not a diffusion process, i.e., it is not Markov. Moreover, then the limit depends on the entire service-time distribution, not just the mean and variance. This suggests that in the search for approximations of many-server queueing models it is not always appropriate to use only the means and variances of the interarrival and service-time distributions. The limit theorem also indicates when the simpler diffusion approximation should be appropriate, namely, either when the number of servers is relatively small or the service-time distribution is nearly exponential. Hence the limit theorem is a useful complement to direct heuristic diffusion approximations for many-server queueing models; see Section IV of Newell (1973) and Halachmi and Franta (1978).

A Markovian limit is significant not only because it means a relatively more tractable approximation, but also because it means that the evolution of the system can be described approximately if we keep track only of the number of customers present. A non-Markovian limit suggests that we could better describe the evolution of the system by keeping track of more information (if it is available). The relevant extra information here is the proportion of those customers present at any time that have been in service for a time less than or equal to t, for all $t \geqq 0$. In the standard heavy-traffic limit theorem in which the traffic intensities converge to one from below while the number of servers is held fixed, this extra information has an asymptotically negligible contribution, but in the Iglehart–Borovkov heavy-traffic limit theorem with non-exponential service-time distributions this information remains relevant in the limit. The main idea here is to look at this extra information. In general, the extra information is a function of t, which means the state space would become uncountably infinite if we incorporated it all. It would be nice to obtain heavy-traffic theorems from this general Markov framework, but we do not yet know how to do this. What we do here is represent such additional information approximately via a discrete-time vector-valued Markov stochastic process. To accomplish this, we consider service-time distributions that consist of an appropriate random (finite) number of phases, with the successive phase lengths being i.i.d. exponential random variables. It is significant that this class of service-time distributions is dense in the family of all probability distributions on the positive real line. Hence, at least in principle, there is no significant loss of generality in this assumption. (In fact, it should be possible to exploit the denseness to verify convergence for general service-time distributions, but we have not yet been able to do this.) The discrete-time vector-valued Markov process is obtained by considering the number of customers in each phase at successive arrival epochs. Our main result (Theorem 3) is that this sequence of Markov processes, appropriately normalized, converges weakly to a multivariate O–U (Ornstein–Uhlenbeck) diffusion process. Since the arrival rate goes to ∞, we are able to show that a similar limit also holds for the associated sequence of continuous-time non-Markov processes representing the number of customers in each service phase at arbitrary times.

While this O–U diffusion limit is somewhat complicated, it has promise for approximations. With m service phases, the m-dimensional limit process is characterized by two $m \times m$ matrices of real numbers: an $m \times m$ infinitesimal mean matrix with $2m-1$ positive entries and an $m \times m$ infinitesimal covariance matrix with $3m-2$ positive entries, all of which depend on $m+2$ parameters (the mean and variance of the interarrival time, the mean of the exponential service phase and $m-1$ phase transition probabilities). Since the multivariate O–U process has been widely studied, it is relatively easy to work with. For example, it is the solution of a linear stochastic differential equation in the narrow sense; see Chapter 8 of Arnold (1974). See Section 2 of Iglehart and Lalchandani (1973) for a nice overview. Interesting properties of the one-dimensional O–U process obtained when the service times are exponentially distributed are discussed by Beekman (1974) and tables have been prepared by Keilson and Ross (1975).

Our limit theorem also provides an alternative proof and explanation for the Iglehart–Borovkov limit theorem. The number of customers in the system is clearly the sum of the numbers in each service phase. Hence, the general Gaussian limit can be interpreted as the sum of the components of a multivariate O–U process. The limit is Gaussian because the Gaussian property of the O–U process is preserved by linear mappings, but it is not Markov because the Markov property is not preserved.

While the proof here is less general and more cumbersome than Borovkov's, we believe it is conceptually elementary and revealing. It is easy to understand why the non-Markov Gaussian limit arises and why it has the form it does. Moreover, by expressing the convergence in terms of a sequence of Markov chains approaching a diffusion process, we are able to invoke a well-developed theory; see Chapter 11 of Stroock and Varadhan (1979) or p. 459 of Gikhman and Skorohod (1969). In order to establish weak convergence in the function space setting, it suffices to show that the infinitesimal means and covariances converge. The theory is sufficiently well developed in this setting that the usual checks to fit a heuristic approximation are essentially all that is needed to rigorously establish the general weak-convergence theorem. As soon as we introduce the exponential phases, we convert a single-station infinite-server non-Markovian service system into a Markovian network of infinite-server service stations. The multivariate O–U limit could be anticipated because of the extensive analysis of related systems; see Iglehart (1968), Schach (1971), Iglehart and Lalchandani (1973), McNeil and Schach (1973) and Lemoine (1978), especially Section 3.3. It should also be apparent that the multivariate O–U limit here extends easily to generalized Jackson networks of infinite-server stations. In order to keep within the Markov framework, we assume all but one external arrival process is Poisson, with the non-Poisson arrival process being a renewal process. (This assumption is not necessary for the theorem, but it is for our method of proof. By considering interarrival times consisting of a random number of exponential stages, we could let all the external arrival processes be renewal processes.) We let the arrival rate in the renewal arrival process and also possibly the Poisson arrival processes go to ∞ while holding all the service-time distributions fixed. We then represent each service-time distribution as appropriate random sums of exponential phases. Let the customers be routed through the original network independently according to fixed probabilities, but note that more complicated routings could be incorporated by introducing new classes of customers and making the state space still bigger; see Kelly (1976). The discrete-time vector-valued Markov process of interest depicts the number of customers in each phase of service at each station. By essentially the same reasoning, the sequence of these Markov processes converges to a multivariate O–U process under the same heavy-traffic conditions. The limits for such general networks obviously are very complicated, but even these limits may prove useful. For example, it might be a good strategy to simulate simple random-walk analogues of the multivariate O–U processes to gain insight into the behavior of complex networks of queues. The heavy-traffic limit theorem for networks of infinite-server stations just described extends the Iglehart–Borovkov limit theorem the same way that Reiman (1977) extended the more familiar ($\rho \uparrow 1$) heavy-traffic limit theorems; see also Harrison (1978). The limit theorems here are more elementary because there are no boundaries associated with the limit processes.

We now indicate how the rest of this paper is organized. We give the background and state our results for single infinite-server queueing stations in Section 2. We conclude with the proofs in Section 3. Additional supporting material is contained in an appendix, which has been omitted to save space; it is available from the author. It contains a limit theorem for different kinds of service-time distributions: service-time distributions that are mixtures of Erlang distributions; this class is also dense in the set of all possible service-time distributions. In fact, this class contains the class in Theorem 3 as a proper subset.

2. Background and new results

Consider a sequence of $GI/G/s$ systems indexed by n with $s_n \leqq \infty$. Let u_n and v_n be the generic interarrival times and service times in the nth system. Assume that $n^{\frac{1}{2}}[E(nu_n)-\lambda^{-1}]\rightarrow 0$, $0<\lambda<\infty$; $\mathrm{Var}\,(nu_n)\rightarrow\sigma^2$, and $0<\sigma^2<\infty$. (For example, we could have $u_n = un^{-1}$ for all n, with $Eu=\lambda^{-1}$ and $\mathrm{Var}\,(u)=\sigma^2$.) Let $v_n = v$ for all n with $Ev=\mu^{-1}$, $0<\mu<\infty$; $\mathrm{Var}\,(v)=\alpha^2$, $0<\alpha^2<\infty$; and $P(v\leqq x)=H(x)$. Note that $Ev_n/nEu_n \rightarrow \rho\equiv\lambda/\mu$.

We have yet to specify the number of servers for each n. We assume that

$$(s_n - n\lambda\mu^{-1})n^{-\frac{1}{2}}\rightarrow\infty \quad\text{as}\quad n\rightarrow\infty, \tag{2.1}$$

with $s_n=\infty$ for all n a possibility. This turns out to guarantee that the probability that all servers are busy at any time in a given finite time interval is asymptotically negligible. The two cases of principal interest are: (i) $s_n=\infty$ for all n and (ii) $s_n=n$ and $\rho<1$.

Let $Q_n(t)$ be the number of customers in the nth system at time t. We are interested in limit theorems for normalizations of the stochastic processes $Q_n\equiv\{Q_n(t), t\geqq 0\}$ as $n\rightarrow\infty$. Such a theorem was first proved by Iglehart (1965) for the special case of $M/M/s$ systems. He considered the normalized process

$$X_n(t)=n^{-\frac{1}{2}}[Q_n(t)-n\rho], \qquad t\geqq 0. \tag{2.2}$$

To state the result, let $\Rightarrow$ denote weak convergence (convergence in distribution), both for real-valued random variables and random elements of function spaces (stochastic processes). Iglehart proved that if $X_n(0)\Rightarrow X(0)$, then $X_n\Rightarrow X$ where X is the O–U diffusion process with initial distribution that of $X(0)$, infinitesimal mean $m(x)=-\mu x$ and infinitesimal variance $\sigma^2(x)=2\lambda$. Hence, for large n and t, $X_n(t)\sim N(0,-\sigma^2(1)/2m(1))=N(0,\rho)$ (see p. 134 of Arnold (1974)) and $Q_n(t)\sim N(n\rho, n\rho)$, where we use $\sim$ for approximately equal in distribution and $N(a,b)$ for a normally distributed random variable with mean a and variance b.

Borovkov (1967) subsequently proved a weak convergence theorem for the general $GI/G/s$ case and even more general systems. He considered the normalized process

$$Y_n(t)=n^{-\frac{1}{2}}[Q_n(t)-nh(t)], \qquad t\geqq 0, \tag{2.3}$$

where

$$h(t)=\lambda\int_0^t [1-H(s)]\,ds. \tag{2.4}$$

(Note that $h(\infty)=\rho=\lambda\mu^{-1}$.) He proved that if $Q_n(0)=0$ for all n, then $Y_n\Rightarrow Y$, where Y is the sum of two independent centered Gaussian processes Y_1 and Y_2, where Y_1 has covariance function

$$\gamma_1(s, s+t)=\lambda\int_0^s H(u)[1-H(t+u)]\,du, \qquad s, t\geqq 0, \tag{2.5}$$

and Y_2 is the stochastic integral

$$Y_2(t)=\sigma\lambda^{\frac{3}{2}}\int_0^t [1-H(t-u)]\,dB(u), \qquad t\geqq 0, \tag{2.6}$$

with B being Brownian motion. (Note that the stochastic integral in (2.6), when viewed as a process, is not of standard type because the integrand is a function of t.) As a consequence, Y_2 has covariance function

$$\gamma_2(s, s+t)=\sigma^2\lambda^3\int_0^s [1-H(t+u)][1-H(u)]\,du, \qquad s, t\geqq 0. \tag{2.7}$$

For further discussion and results, see Iglehart (1973a) and Section 5 of Iglehart (1973b). Borovkov (Remark 1 on p. 748) observed that Y_1 cannot be represented as a stochastic integral with respect to Brownian motion. Moreover, it is significant that in general Y_1, Y_2, and Y_1+Y_2 need not be Markov. In fact, Glynn (1982) has shown that $Y=Y_1+Y_2$ is Markov if and only if $1-H(t)=pe^{-\alpha t}$ for $\alpha>0$ and $0<p\leqq 1$, using the necessary and sufficient condition

$$\gamma(s, s+t+u)\gamma(s+t, s+t)=\gamma(s, s+t)\gamma(s+t, s+t+u), \qquad s, t, u\geqq 0,$$

from Doob (1953), p. 233. With this condition, it is easy to see that Y_1, Y_2, and Y_1+Y_2 are all Markov if H is of this form. In fact, Y_2 is just the Ornstein–Uhlenbeck process normalized to be at 0 at $t=0$, with infinitesimal mean $m(x)=-\mu x$ and infinitesimal variance $\sigma^2(x)=\sigma^2\lambda^3$; p. 349 of Breiman (1968). Since

$$\begin{aligned}\gamma_1(s, s+t)&=\lambda e^{-\mu t}(\mu^{-1}[1-e^{-\mu s}]-(2\mu)^{-1}[1-e^{-2\mu s}])\\ &\to(\rho/2)e^{-\mu t} \quad \text{as} \quad s\to\infty,\end{aligned}$$

we see that $\{Y_1(s+t), t\geqq 0\}$ converges as $s\to\infty$ to a stationary O–U process with infinitesimal mean $m(x)=-\mu x$ and infinitesimal variance $\sigma^2(x)=\lambda$. (This is rigorously demonstrated by an easy application of Theorem 11.1.4 of Stroock and Varadhan (1979).) Thus, $\{Y(s+t), t\geqq 0\}$ converges as $s\to\infty$ to a stationary O–U process with infinitesimal mean $m(x)=-\mu x$ and infinitesimal variance $(\lambda^3\sigma^2+\lambda)$. Hence, in $GI/M/\infty$ systems, for large n and t, $Q_n(t)\sim N(n\rho, n\rho(1+\sigma^2\lambda^2)/2)$.

The limit process Y_1+Y_2 also simplifies when the interarrival time is not necessarily exponential but its coefficient of variation is 1. Then $\sigma^2\lambda^3=\lambda$ and Y_1+Y_2 has covariance function

$$\gamma(s,s+t)=\gamma_1(s,s+t)+\gamma_2(s,s+t)=\lambda\int_0^s[1-H(t+u)]\,du. \tag{2.8}$$

We shall see later that this case also leads to simplifications in the limits for vector-valued Markov chains.

It is somewhat disconcerting, however, that in the $M/M/s$ case Borovkov's limit theorem does not reduce to Iglehart's. The reason of course is the different assumptions about the initial conditions. Iglehart assumes that $X_n(0)\equiv n^{-\frac{1}{2}}[Q_n(0)-n\rho]$ converges, whereas Borovkov assumes $Q_n(0)=0$. One purpose of this paper is to prove the following theorem.

Theorem 1. Consider the originally specified sequence of $GI/G/s$ systems with s_n and X_n satisfying (2.1) and (2.2). If the service-time distribution is exponential and $X_n(0)\Rightarrow X(0)$, then $X_n\Rightarrow X$, where X is the O–U process with initial distribution $X(0)$, infinitesimal mean $m(x)=-\mu x$ and infinitesimal variance $\sigma^2(x)=\lambda+\lambda^3\sigma^2$.

Theorem 1 also implies that $Q_n(t)\sim N(n\rho, n\rho(1+\lambda^2\sigma^2)/2)$ for large n and t. Moreover, if $X(0)$ is distributed as $N(0,\rho(1+\lambda^2\sigma^2)/2)$, then X is the stationary O–U process. (It is known that if a process is Markov, Gaussian, stationary, and continuous in probability, then it must be the O–U process; p. 350 of Breiman (1968).)

In addition, we establish convergence of the steady-state distributions. The proof here involves an interesting stochastic dominance argument which is of independent interest.

Theorem 2. For systems satisfying the assumptions of Theorem 1 and having $Q_n(t)\Rightarrow Q_n(\infty)$ as $t\to\infty$ for each n,

$$X_n(\infty)\equiv n^{-\frac{1}{2}}[Q_n(\infty)-n\rho]\Rightarrow N(0,\rho(1+\lambda^2\sigma^2)/2) \quad \text{as} \quad n\to\infty.$$

Remarks. (1) The theorems in this section are stated for the continuous-time stochastic process representing the number of customers in the system at an arbitrary time. Corresponding results exist for the discrete-time process representing the number of customers present at arrival epochs (not including the arrival). These results are obtained here in the proofs.

(2) With reference to the existence of a limiting distribution as $t\to\infty$ for each n, we mention that this is satisfied for $GI/G/s$ systems with $s\leqq\infty$ under

the moment conditions made at the outset if and only if u_n has a non-lattice distribution. (In the lattice case of course limits also exist if we go along the subsequence of multiples of the span.) For many systems, including $GI/M/s$ systems with $s \leqq \infty$, this can be easily demonstrated through regenerative process theory. In particular, if $P(u_n > v_n) > 0$, then there is an embedded renewal process with finite expected time between renewals. Moreover, it is possible to show that the time between renewals is non-lattice if and only if u_n is non-lattice. With the condition $P(u_n > v_n) > 0$, Theorem 2.2 of Whitt (1972) and Theorem 3.1 of Miller (1972) can be applied. The argument in Whitt (1972) easily extends to $GI/G/\infty$ systems. However, the general case is much more complicated. The general result for $GI/G/\infty$ queues has been established by Jagers (1968), Theorem 3; see also Kaplan (1975) and Pakes and Kaplan (1974). For $s < \infty$, the general result appears on p. 173 of Borovkov (1976). Miller and Sentilles (1975) have also established the general result for $s < \infty$ under the extra assumption that the interarrival-time and service-time distributions are atomless. A new proof also appears in Whitt (1981a). For an extensive treatment of the general case, including extensions covering non-renewal arrival processes, see Franken, König, Arndt, and Schmidt (1981).

(3) Various characterizations of these limiting distributions have also been established. In the $GI/G/\infty$ system, the limit $Q(\infty)$ has mean value ρ and probability generating function $\Phi(s) = Es^{Q(\infty)}$ given by

$$\Phi(s) = 1 - \lambda \int_0^\infty \Phi(s, t)[1 - H(t)](1 - s)\, dt, \tag{2.9}$$

where $\Phi(s, t) = Es^{Q(t)}$ with $Q(0) = 0$ and a regular interarrival time until the first arrival; see Jagers (1968), Kaplan (1975) and Pakes and Kaplan (1974). These references also serve as a reminder that much can be deduced about the $GI/G/\infty$ system from related stochastic models such as branching processes with immigration. For $s = \infty$, it is well known that $Q(\infty)$ has a Poisson distribution with parameter ρ if the interarrival times are exponential, see p. 18 of Ross (1970). As obviously must be the case with exponential interarrival times, Borovkov's (1967) limit theorem implies that $Q_n(t) \sim N(n\rho, n\rho)$ for large n and t. In other words, the approximating marginal distribution depends on the service-time distribution H only through its mean. However, it is only the one-dimensional marginal distributions which simplify. The limiting process for $M/G/\infty$ systems is not Markov and the covariance function of the limit process depends on the entire distribution H.

For $s < \infty$, characterizations of the time-dependent distribution as well as the limiting distribution have been obtained by transform methods by de Smit (1973a,b). Recently, Franken and his colleagues (1975), (1976), (1981) have obtained additional characterizations of the stationary distribution by applying the theory of point processes. Brillinger's (1974) study of identifiability for $s=\infty$ is in this spirit too.

Our primary interest is in many-server systems with non-exponential service times. What we do is let the service time consist of a random (finite) number of phases, with the length of each phase being exponentially distributed with mean β^{-1}. After completing phase k, each customer leaves the system with probability p_k and moves on to phase $k+1$ with probability $1-p_k$. It is significant that the class of service-time distributions of this form is dense in the family of all probability distributions on the positive real line; i.e., given any service-time distribution function H, there is a sequence of phase-type distributions $\{H_m, m\geqq 1\}$ with H_m converging weakly to H as $m\to\infty$ (denoted by $H_m\Rightarrow H$, which means $H_m(x)\to H(x)$ as $m\to\infty$ for each x which is a continuity point of H). Given H and the mean phase length β^{-1}, the obvious way to define the phase-transition probabilities is

$$(2.10)\qquad p_k=\frac{H((k+1)/\beta)-H(k/\beta)}{1-H(k/\beta)},\qquad k\geqq 1.$$

Let the number of phases be M. It is easy to see that the tail of the approximating c.d.f. H has the form

$$(2.11)\qquad 1-H(t)=\sum_{i=1}^{i=M}\prod_{k=1}^{k=i-1}(1-p_k)e^{-\beta t}\frac{(\beta t)^{i-1}}{(i-1)!},\qquad t\geqq 0.$$

We state the well-known convergence property in the following lemma (Schassberger (1973), p. 32).

Lemma 1. Consider a sequence of phase-type service distributions indexed by m. If $\beta_m=m$ and $m^{-1}M_m\to\infty$, then $H_m\Rightarrow H$.

We now fix the number of phases and the mean length of each phase and consider limits as $n\to\infty$ for the sequence of systems defined at the outset. However, now we are keeping track of the number of customers in each phase of service. In particular, let $Q_n^i(t)$ be the number of customers in the ith phase of the nth system at time t. We shall prove a limit theorem for the vector-valued process $\{\mathbf{Q}_n(t), t\geqq 0\}$, where $\mathbf{Q}_n(t)=[Q_n^1(t),\cdots,Q_n^M(t)]$. For $i=1,\cdots,M$, let $\mathbf{X}_n=[X_n^1,\cdots,X_n^M]$,

$$(2.12)\qquad X_n^i(t)=n^{-\frac{1}{2}}[Q_n^i(t)-n\alpha_i],\qquad t\geqq 0,$$

and

$$(2.13)\qquad \alpha_i=\lambda\beta^{-1}\prod_{k=1}^{i-1}(1-p_k).$$

Note that

$$\sum_{i=1}^{M} \alpha_i = \lambda \int_0^\infty [1-H(t)]\,dt = \lambda\mu^{-1} = \rho. \tag{2.14}$$

Our main result is the following generalization of Theorem 1.

Theorem 3. Consider the originally specified sequence of $GI/G/s$ systems with s_n satisfying (2.1). Let the service-time distribution be phase-type as specified above and let $\boldsymbol{X}_n$ satisfy (2.12) and (2.13). If $\boldsymbol{X}_n(0) \Rightarrow \boldsymbol{X}(0)$, then $\boldsymbol{X}_n \Rightarrow \boldsymbol{X}$, where $\boldsymbol{X}$ is the M-dimensional Ornstein–Uhlenbeck process having initial distribution that of $\boldsymbol{X}(0)$, infinitesimal mean vector $m(\boldsymbol{x}) \equiv [m_1(x_1, \cdots, x_M), \cdots, m_M(x_1, \cdots, x_M)]' = A\boldsymbol{x} \equiv A(x_1, \cdots, x_M)'$ and infinitesimal covariances $\Sigma(\boldsymbol{x}) = \Sigma$, where A is the $M \times M$ matrix with elements

$$a_{ij} = \begin{cases} -\beta, & j = i \\ \beta(1-p_i), & j = i-1 \\ 0, & \text{otherwise} \end{cases} \tag{2.15}$$

and Σ is the symmetric $M \times M$ matrix with elements

$$\sigma_{ij} = \begin{cases} \lambda^3\sigma^2 + \lambda, & i = j = 1 \\ 2\alpha_i, & i = j > 1 \\ -\alpha_i, & i = j+1 \quad \text{or} \quad j = i-1 \\ 0, & \text{otherwise.} \end{cases} \tag{2.16}$$

The stationary covariance function of this limiting Gaussian process is

$$K(s, s+t) = \bar{K}(t) = (EX^i(s)X^j(s+t)) = e^{At}\bar{K}(0), \tag{2.17}$$

with i and j being superscripts rather than exponents, where

$$(e^{At})_{ij} = \begin{cases} \displaystyle\prod_{k=j}^{k=i-1} (1-p_k) \frac{(\beta t)^{i-j}}{(i-j)!} e^{-\beta t}, & j < i, \\ e^{-\beta t}, & j = i, \\ 0, & j > i, \end{cases} \tag{2.18}$$

and

$$\bar{K}(0)_{ij} = (\lambda^3\sigma^2 - \lambda)\left[\prod_{k=1}^{i-1} (1-p_k) \prod_{k=1}^{j-1} (1-p_k) \binom{i+j-2}{i-1} 2^{-(i+j-1)} \beta^{-1}\right] + \xi_{ij}$$

with

$$\xi_{ij} = \begin{cases} \lambda\beta^{-1} \displaystyle\prod_{k=1}^{i-1} (1-p_k), & i = j, \\ 0, & \text{otherwise.} \end{cases}$$

Remark. Note that in the setting of Theorem 3 the infinitesimal covariance matrix reduces to a simple diagonal matrix when the coefficient of variation of the interarrival time is 1. However, while the stationary variance matrix $\bar{K}(0)$ is diagonal, the stationary covariance matrix $\bar{K}(t)$ is not.

As an immediate consequence of Theorem 3 and the continuous mapping theorem, we obtain a variant of Borovkov's theorem for a dense family of service-time distributions. Let $Q_n(t)=Q_n^1(t)+\cdots+Q_n^M(t)$, $X_n(t)=X_n^1(t)+\cdots+X_n^M(t)$ and $X(t)=X^1(t)+\cdots+X^M(t)$, $t\geqq 0$.

Corollary (Borovkov). Let the service-time distribution function H be as in (2.11). If $\mathbf{X}_n(0)\Rightarrow\mathbf{X}(0)$ under the assumptions of Theorem 3, then $X_n\Rightarrow X$. If $\mathbf{X}(0)$ is distributed as $N(0,\bar{K}(0))$, then X is a stationary centered Gaussian process with covariance function

$$EX(s)X(s+t)=\lambda\int_0^\infty[1-H(t+u)]\,du+[\sigma^2\lambda^3-\lambda]\int_0^\infty[1-H(t+u)](1-H(u))\,du.$$

We illustrate Theorem 3 with an example.

Example. Suppose the service-time distribution satisfies

$$1-H(t)=e^{-t}+(1-p)te^{-t},\qquad t\geqq 0.$$

In other words, with probability p the service time consists of one exponential phase and with probability $(1-p)$ it consists of two exponential phases, where the exponential phases are independent with each having mean 1. The limiting Ornstein-Uhlenbeck diffusion process is thus two-dimensional with drift

$$A\mathbf{x}=\begin{pmatrix}-1 & 0\\ (1-p) & -1\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}m_1(x_1,x_2)\\ m_2(x_1,x_2)\end{pmatrix}$$

and covariance matrix

$$\Sigma=\begin{pmatrix}\lambda^3\sigma^2+\lambda & -\lambda(1-p)\\ -\lambda(1-p) & 2\lambda(1-p)\end{pmatrix}=\begin{pmatrix}\sigma_1^2(x_1,x_2) & \sigma_{12}^2(x_1,x_2)\\ \sigma_{21}^2(x_1,x_2) & \sigma_2^2(x_1,x_2)\end{pmatrix}.$$

From p. 133 of Arnold (1974), we see that the covariance function of the stationary process is

$$\begin{aligned}K(s,t)=\bar{K}(t-s)&=\begin{pmatrix}EX_1(s)X_1(t) & EX_1(s)X_2(t)\\ EX_2(s)X_1(t) & EX_2(s)X_2(t)\end{pmatrix},\\ &=e^{A(t-s)}\bar{K}(0),\qquad s\leqq t,\end{aligned}$$

where $\bar{K}(0)$ satisfies the matrix equation $A\bar{K}(0)+\bar{K}(0)A'=-\Sigma$, which in this case can be solved directly to obtain

$$\bar{K}(0)=\begin{bmatrix}\frac{1}{2}(\lambda^3\sigma^2+\lambda) & \frac{1}{4}(\lambda^3\sigma^2-\lambda)(1-p)\\ \frac{1}{4}(\lambda^3\sigma^2-\lambda)(1-p) & \frac{1}{4}(\lambda^3\sigma^2-\lambda)(1-p)^2+\lambda(1-p)\end{bmatrix}.$$

Alternatively, we can find $\bar{K}(0)$ from the integral

$$\bar{K}(0)=\int_0^\infty e^{At}\Sigma e^{A't}\,dt,$$

where A' is the transpose of A. Here

$$e^{At}=\sum_{n=0}^{\infty}\frac{A^n t^n}{n!}=\begin{bmatrix} e^{-t} & 0 \\ (1-p)te^{-t} & e^{-t}\end{bmatrix}$$

and, with $B(t)=e^{At}\sum e^{A't}$,

$$b_{11}(t)=(\lambda^3\sigma^2+\lambda)e^{-2t}$$

$$b_{12}(t)=b_{21}(t)=(\lambda^3\sigma^2+\lambda)(1-p)te^{-2t}-\lambda(1-p)e^{-2t}$$

$$b_{22}(t)=(\lambda^3\sigma^2+\lambda)(1-p)^2t^2e^{-2t}-2\lambda(1-p)^2te^{-2t}+2\lambda(1-p)e^{-2t}.$$

Here

$$\bar{K}(t)=\begin{bmatrix} \tfrac{1}{2}(\lambda^3\sigma^2+\lambda)e^{-t} & \tfrac{1}{4}(\lambda^3\sigma^2-\lambda)(1-p)e^{-t} \\ \lambda^3\sigma^2(1-p)e^{-t}[\tfrac{1}{2}t+\tfrac{1}{4}] & \tfrac{1}{4}(\lambda^3\sigma^2-\lambda)(1-p)^2e^{-t}(t+1) \\ +\lambda(1-p)e^{-t}[\tfrac{1}{2}t-\tfrac{1}{4}] & +\lambda(1-p)e^{-t} \end{bmatrix}.$$

The limiting stationary Gaussian (non-Markov) process for the normalization of the number of customers in the system is $Y=X_1+X_2$, which has covariance function

$$\begin{aligned}\gamma(0,t)&=EY(0)Y(t)=\bar{K}_{11}(t)+\bar{K}_{12}(t)+\bar{K}_{21}(t)+\bar{K}_{22}(t)\\ &=\lambda^3\sigma^2e^{-t}t[\tfrac{1}{2}(1-p)+\tfrac{1}{4}(1-p)^2]+\lambda e^{-t}t[\tfrac{1}{2}(1-p)-\tfrac{1}{4}(1-p)^2]\\ &\quad+\lambda^3\sigma^2e^{-t}(\tfrac{1}{2}(2-p)+\tfrac{1}{4}(1-p)^2)+\lambda e^{-t}(\tfrac{1}{2}(2-p)-\tfrac{1}{4}(1-p)^2)\\ &=\lambda\int_0^\infty[1-H(t+u)]\,du+(\sigma^2\lambda^3-\lambda)\int_0^\infty[1-H(t+u)](1-H(u))\,du,\end{aligned}$$

which is in agreement with the corollary above. The special case in which the service times have an Erlang distribution, i.e., $1-H(t)=e^{-t}+te^{-t}$, occurs when $p=0$. Then

$$\gamma(0,t)=e^{-t}[\lambda^3\sigma^2(3t+5)+\lambda(t+3)]/4,\qquad t\geqq 0.$$

On the other hand, when $p=1$, the service time is exponential and

$$\gamma(0,t)=e^{-t}(\lambda^3\sigma^2+\lambda)/2,\qquad t\geqq 0.$$

Theorem 3 is also of interest because of its applied value. It provides a means of incorporating the extra information about time in service in the approximation. For example, this can be done by dividing time into intervals of length m^{-1}. If the length of time a customer has been in service falls in the interval $[(k-1)m^{-1}, km^{-1}]$, then we can regard the customer as being in exponential phase k. As m gets large, the conditional distribution of the remaining approximate service time approaches the conditional distribution of the remaining actual service time. Moreover, if the completed service times are not monitored completely, then the exponential phases should be appropriate.

3. Proofs

We give no separate proof of Theorem 1 because it is a special case of Theorem 3.

3.1. *Proof of Theorem* 2. We first work with the discrete-time Markov chains $\{Q_n^A(k), k \geqq 0\}$ representing the number of customers in the nth system at arrival epochs. It is well known and easy to show that $Q_n^A(k) \Rightarrow Q_n^A(\infty)$ as $k \to \infty$ for each n. Let X_n^A be the associated normalized continuous-time process, defined by

$$X_n^A(t) = n^{-\frac{1}{2}}[Q_n^A([nt]) - n\rho], \qquad t \geqq 0. \tag{3.1}$$

We show that for these processes it suffices to prove that the sequence of normalized steady-state distributions $\{X_n^A(\infty)\}$ is tight; see Section 6 of Billingsley (1968). If $\{X_n^A(\infty)\}$ is tight, then by Prohorov's theorem, Theorem 6.1 of Billingsley (1968), the sequence $\{X_n^A(\infty)\}$ has a convergent subsequence $\{X_{n'}^A(\infty)\}$. If we let $Q_{n'}^A(0)$ be distributed as $Q_{n'}^A(\infty)$, then $\{Q_{n'}^A(k), k \geqq 0\}$ is a strictly stationary process for each n' and, by the proof of Theorem 1, $X_{n'}^A \Rightarrow X^A$ in the function space $D[0, \infty)$, where $X^A(0)$ has the distribution of the weak convergence limit of $\{X_{n'}^A(\infty)\}$. However, since $\{Q_{n'}^A(k), k \geqq 0\}$ is strictly stationary for each n' and $X_{n'}^A \Rightarrow X^A$, the limit process X^A must be strictly stationary too. Hence, the limit of $\{X_{n'}^A(\infty)\}$ is the unique stationary distribution of X^A. Since every subsequence of the sequence $\{X_n^A(\infty)\}$ which converges must converge to this same limit, the sequence $\{X_n^A(\infty)\}$ itself must converge to this limit.

Hence, to complete the proof for $\{X_n^A(\infty)\}$ it suffices to show that the sequence $\{X_n^A(\infty)\}$ is tight. We shall show this by bounding $X_n^A(\infty)$ above and below by random variables we already know converge weakly. In particular, we shall construct random variables $L_n(\infty)$ and $U_n(\infty)$ such that $L_n(\infty) \leqq_{\mathrm{st}} X_n^A(\infty) \leqq_{\mathrm{st}} U_n(\infty)$, where $\leqq_{\mathrm{st}}$ denotes stochastic order; i.e.,

$$P(L_n(\infty) \geqq x) \leqq P(X_n^A(\infty) \geqq x) \leqq P(U_n(\infty) \geqq x)$$

for all x and n, and

$$L_n(\infty) \Rightarrow L(\infty) \quad \text{and} \quad U_n(\infty) \Rightarrow U(\infty) \quad \text{as} \quad n \to \infty.$$

It is easy to see that this implies that $\{X_n^A(\infty)\}$ is tight.

We construct the stochastically bounding random variables $L_n(\infty)$ and $U_n(\infty)$ by constructing stochastically bounding stochastic processes $\{L_n(t), t \geqq 0\}$ and $\{U_n(t), t \geqq 0\}$ such that $L_n(t) \Rightarrow L_n(\infty)$ and $U_n(t) \Rightarrow U_n(\infty)$ as $t \to \infty$ for each n and $L_n(\infty) \Rightarrow L(\infty)$ and $U_n(\infty) \Rightarrow U(\infty)$. Since stochastic order is preserved under weak convergence, Proposition 3 of Kamae, Krengel, and O'Brien (1977), we obtain $L_n(\infty)\leqq_{\mathrm{st}} X_n^A(\infty)\leqq_{\mathrm{st}} U_n(\infty)$ for all n as desired. We let

$$L_n(t)=n^{-\frac{1}{2}}[L_n^A([nt])-n\rho],\qquad t\geqq 0,$$

and

$$U_n(t)=n^{-\frac{1}{2}}[U_n^A([nt])-n\rho],\qquad t\geqq 0,$$

paralleling the way X_n^A was defined in terms of Q_n^A in (3.1). We now construct discrete-time Markov chains $\{L_n^A(k), k\geqq 0\}$ and $\{U_n^A(k), k\geqq 0\}$ with the property that $L_n^A(k)\leqq_{\mathrm{st}} Q_n^A(k)\leqq_{\mathrm{st}} U_n^A(k)$ for all n and k.

We construct the upper bound chains $\{U_n^A(k)\}$ by modifying the $GI/M/\infty$ systems. For each n, we keep the same arrival process and put an impenetrable reflecting lower barrier at $n\rho+c\sqrt{n}$, $c>0$. Moreover, we let the departure rate be $(n\rho+c\sqrt{n})\mu$ for all states in the restricted state space of the nth system. For each n, the modified process behaves like a constant $n\rho+c\sqrt{n}$ plus a stable $GI/M/1$ queue with traffic intensity $(n\rho+c\sqrt{n})\mu(EU_n)^{-1}$. In particular, for each n, the sequence $\{U_n^A(k)\}$ satisfies the same functional equation as the embedded queue-length process in the $GI/M/1$ queue, namely, $\xi_{k+1}=\max\{0, \xi_k+\eta_k\}$, where η_k is independent of ξ_k and $\{\eta_k\}$ is i.i.d. In the $GI/M/1$ context, η_k is 1 minus the number of departures generated by a Poisson process in an interarrival interval. Any one of Theorems 5–7 in Whitt (1981) implies a sample path ordering between the chains $\{Q_n^A(k)\}$ and $\{U_n^A(k)\}$, which implies the desired stochastic ordering and extends immediately to the normalized continuous-time processes X_n^A and U_n. Alternatively, in this Markov-chain setting we could obtain the sample path ordering by checking the criteria for comparing Markov chains in O'Brien (1975). Because of the $GI/M/1$ structure, we know that $U_n^A(k)\Rightarrow U_n^A(\infty)$ as $k\to\infty$ for each n. Finally, $U_n(\infty)\Rightarrow U(\infty)$ by virtue of existing heavy-traffic limit theorems for the steady-state distributions of single-server queues. (Note that the associated sequence of $GI/M/1$ traffic intensities converges to 1 as $n\to\infty$.) In particular, Kingman's (1962) early result for waiting times applies because the embedded queue-length process in a $GI/M/1$ queue has the same structure as the waiting-time sequence in a $GI/G/1$ queue.

A similar construction yields appropriate lower bound chains. Again, for each n, keep the same arrival process but now put an impenetrable reflecting upper barrier at $n\rho-c\sqrt{n}$, $c>0$. Moreover, let the departure rate be $(n\rho-c\sqrt{n})\mu$ for all states in the restricted state space. For each n, the negative of the modified chain behaves like a constant plus the embedded chain at departure epochs of an $M/G/1$ queue. In particular, for each n, the sequence $\{-L_n(k)+n\rho-c\sqrt{n}\}$ satisfies the functional equation $\xi_{k+1}=\max\{\xi_k, 1\}+\eta_k$, where η_k is independent of ξ_k and $\{\eta_k\}$ is i.i.d. In the $M/G/1$ context, η_k represents the number of arrivals generated by a Poisson process in a service time minus 1.

The stochastic ordering and the heavy-traffic limits for the steady-state distributions follow by essentially the same arguments. A specific heavy-traffic limit theorem to apply in this case is the one on p. 168 of Gnedenko and Kovalenko (1968). (Recall that the steady-state distributions at arrival epochs, departure epochs and arbitrary time points coincide for an $M/G/1$ queue.)

We have finished the proof for $\{X_n^A(\infty)\}$. We now show that $X_n^A(\infty)$ and $X_n(\infty)$ have the same limits by applying another stochastic dominance argument. Note that

$$(3.2)\quad Q_n^A(A_n(t)) - M(\mu[Q_n^A(A_n(t))+1], u^-_{nA_n(t)}) \leqq_{\mathrm{st}} Q_n(t) \leqq Q_n^A(A_n(t))+1$$

for each n and t, where $A_n(t)$ is the number of arrivals in $[0, t]$ in the nth system, $u^-_{nA_n(t)}$ is the elapsed portion of the interarrival time in progress at time t in the nth system, and $M(\mu, t)$ is a Poisson process with rate μ. The upper and lower bounds express the queue length at any time t in terms of the queue length at the most recent arrival epoch. The bounds represent the extreme cases of no departures (upper bound) and departures throughout the interarrival time at the highest possible (initial) rate. Since $Q_n^A(A_n(t))$ converges weakly as $t \to \infty$ to the same limit $Q_n^A(\infty)$ as $Q_n^A(k)$ when $k \to \infty$, we obtain the useful relationship $Q_n(\infty) \leqq_{\mathrm{st}} Q_n^A(\infty)+1$ for the upper bound. For the lower bound, note that $u^-_{nA_n(t)}$ converges to the stationary excess, say e_n, associated with u_n as $t \to \infty$; see Section 3.6 of Ross (1970). Hence,

$$Q_n(\infty) \geqq_{\mathrm{st}} Q_n^A(\infty) - M(\mu[Q_n^A(\infty)+1], e_n).$$

By the Markov inequality,

$$P(M(\mu[Q_n^A(\infty)+1], e_n) \geqq K) \leqq EM(\mu[Q_n^A(\infty)+1], e_n)/K.$$

However,

$$EM(\mu[Q_n^A(\infty)+1], e_n) = \mu[EQ_n^A(\infty)+1]Ee_n,$$

where

$$Ee_n \leqq C_1(\sigma^2+\lambda^{-2})/2n\lambda^{-1}$$

and

$$EQ_n^A(\infty) \leqq n\rho + n^{\frac{1}{2}}C_2$$

for all n for appropriate constants C_1 and C_2. (Use the upper bound $U_n^A(\infty)$ plus the moment conditions at the outset of Section 2 to generate C_2.) Hence, there is a constant C_3 such that $EM(\mu[Q_n^A(\infty)+1], e_n) \leqq C_3$ for all n, so that $M(\mu[Q_n^A(\infty)+1], e_n)$ is tight or stochastically bounded. Finally, we have established that

$$X_n^A(\infty) - Y_n^1 \leqq_{\mathrm{st}} X_n(\infty) \leqq_{\mathrm{st}} X_n^A(\infty) + Y_n^2,$$

where $X_n^A(\infty) \Rightarrow X^A(\infty)$, $Y_n^1 \Rightarrow 0$ and $Y_n^2 \Rightarrow 0$. By the convergence-together theorem, Theorem 4.1 of Billingsley (1968), $X_n^A(\infty) - Y_n^1 \Rightarrow X^A(\infty)$ and $X_n^A(\infty) + Y_n^2 \Rightarrow X^A(\infty)$. Because of the stochastic dominance, $X_n(\infty) \Rightarrow X^A(\infty)$ too.

3.2. *Proof of Theorem* 3. We start by working with the discrete-time Markov chains obtained by looking at the system only at arrival epochs. Let $Q_n^{Ai}(k)$ be the number of customers in the ith phase of service of the nth system at the epoch of the kth arrival (but not including the kth arrival). Let $\boldsymbol{Q}_n^A(k) = [Q_n^{A1}(k), \cdots, Q_n^{AM}(k)]$ and $\boldsymbol{X}_n^A = [X_n^{A1}, \cdots, X_n^{AM}]$, where

$$X_n^{Ai}(t) = n^{-\frac{1}{2}}[Q_n^{Ai}([nt]) - n\alpha_i\lambda^{-1}], \qquad t \geqq 0. \tag{3.3}$$

It is easy to see that $\{\boldsymbol{Q}_n^A(k), k \geqq 0\}$ is an irreducible aperiodic Markov chain for each n. Hence, in order to demonstrate the weak convergence of $\{\boldsymbol{X}_n^A\}$ to an appropriate diffusion process, we can apply Theorems 10.2.2 and 11.2.3 of Stroock and Varadhan (1979), i.e., it suffices to check the infinitesimal conditions given in (2.4)–(2.6) on p. 268 there. This involves relatively straightforward (but tedious) calculations, some of which are displayed below. The messiest technical point is showing that it suffices to assume that the total service rate is constant throughout each interarrival interval. Of course, the number of busy servers (and thus also the total service rate) may often change in an interarrival interval, but it is possible to show that the adjustment is asymptotically insignificant.

Having established the weak convergence $\boldsymbol{X}_n^A \Rightarrow \boldsymbol{X}^A$, we can get the weak convergence of $\boldsymbol{X}_n$ in (2.12) by performing a random time change; see Section 17 of Billingsley (1968) or Section 3 of Whitt (1980). Let $\{A_n(t), t \geqq 0\}$ be the arrival process in the nth system and let $B_n(t) = A_n(t)/n$, $t \geqq 0$. Then, by the initial assumptions on u_n, $B_n \xrightarrow{P} B$ in $D[0, \infty)$, where $B(t) = \lambda t$, $t \geqq 0$. Hence, $\boldsymbol{X}_n^A \circ B_n \Rightarrow \boldsymbol{X}^A \circ B = \boldsymbol{X}$, where $\circ$ is the composition map. The difference between $\boldsymbol{X}_n^A \circ B_n$ and $\boldsymbol{X}_n$ in each coordinate is dominated by the jumps of $\boldsymbol{X}_n^A$ in that coordinate, but the maximum jump in any coordinate in any bounded time interval converges to 0 because $\boldsymbol{X}_n^A$ has a limit with continuous paths. Hence, $d(\boldsymbol{X}_n^A \circ B_n, \boldsymbol{X}_n) \Rightarrow 0$ in $D([0, \infty), R^M)$, using the metric in Whitt (1980), say, so that $\boldsymbol{X}_n \Rightarrow \boldsymbol{X}$ by Theorem 4.1 of Billingsley (1968).

We conclude by displaying some of the calculations showing that the infinitesimal conditions in (2.4)–(2.6) on p. 268 of Stroock and Varadhan (1979) are satisfied for Theorem 3.

Let $D_n^i(x_i \mid u_n)$ represent the number of completed services from the ith phase in the nth system in some interarrival-time interval conditioned on the number of customers in phase i at the beginning of the interval being $n\alpha_i + n^{\frac{1}{2}}x_i$ and the length of the interval being u_n. Let η_{nj}^i be a random variable with value

1 if the jth customer to complete service in the ith phase of the nth system in some interarrival interval goes to the next phase, and value 0 if that customer leaves the system. Let $P(\eta^i_{nj}=1)=1-p_i$ and let the collection $\{\eta^i_{nj}$: $i=1,\cdots,M-1;\ j=1,\cdots\}$ be mutually independent.

Let $m_{ni}(x), \sigma^2_{ni}(x)$, and $\sigma^2_{nij}(x)$ be the infinitesimal means, variances, and covariances, respectively, for the Markov chain $\boldsymbol{X}^A_n$ in (3.3). In calculating these infinitesimal parameters, we act as if the total service rate is unchanged during the interarrival-time interval, i.e., we assume that $D^i_n(x_i \mid u_n)$ has a Poisson distribution with mean $\beta(n\alpha_i+n^{\frac{1}{2}}x_i)u_n$ given the interval u_n. It can be shown that the error resulting from this assumption is asymptotically negligible.

First, we obtain the limits for the infinitesimal means. For the first phase,

$$\begin{aligned} m_{n1}(x) &= E\{1-ED^1_n(x_1 \mid u_n)\}n^{\frac{1}{2}} \\ &= \{1-(n\lambda+\beta x_1 n^{\frac{1}{2}})Eu_n\}n^{\frac{1}{2}} \to x_1\beta\lambda^{-1}, \end{aligned}$$

using the condition $n^{\frac{1}{2}}[E(nu_n)-\lambda^{-1}]\to 0$ as $n\to\infty$. For any $i>1$,

$$\begin{aligned} m_{ni}(x) &= E\left\{E\left[\sum_{j=1}^{D^{i-1}_n(x_{i-1}\mid u_n)} \eta^{i-1}_{nj} - D^i_n(x_i \mid u_n)\right]\right\}n^{\frac{1}{2}} \\ &= [(n\alpha_{i-1}+x_{i-1}n^{\frac{1}{2}})\beta(1-p_{i-1})Eu_n-(n\alpha_i+x_i n^{\frac{1}{2}})\beta Eu_n]n^{\frac{1}{2}} \\ &\to x_{i-1}(1-p_{i-1})\beta\lambda^{-1}-x_i\beta\lambda^{-1}, \end{aligned}$$

using $\alpha_i=\alpha_{i-1}(1-p_{i-1})$.

Now we turn to the infinitesimal variances. For $i=1$, we have

$$\begin{aligned} \sigma^2_{n1}(x) &= E\{E[1-D^1_n(x_1 \mid u_n)]^2\} \\ &= E\{1-2ED^1_n(x_1 \mid u_n)+ED^1_n(x_1 \mid u_n)^2\} \\ &= 1-2(n\lambda\beta^{-1}+x_1 n^{\frac{1}{2}})\beta Eu_n+(n\lambda\beta^{-1}+x_1 n^{\frac{1}{2}})\beta Eu_n+(n\lambda\beta^{-1}+x_1 n^{\frac{1}{2}})^2\beta^2 Eu_n^2 \\ &\to \lambda^2\sigma^2+1. \end{aligned}$$

For any $i>1$,

$$\begin{aligned} \sigma^2_{ni}(x) &= E\left\{E\left[\sum_{j=1}^{D^{i-1}_n(x_{i-1}\mid u_n)} \eta^{i-1}_{nj} - D^i_n(x_i \mid u_n)\right]^2\right\} \\ &= E\left\{E\left[\sum_{j=1}^{D^{i-1}_n(x_{i-1}\mid u_n)} \eta^{i-1}_{nj}\right]^2 - 2E\left(\sum_{j=1}^{D^{i-1}_n(x_{i-1}\mid u_n)} \eta^{i-1}_{nj}\right)ED^i_n(x_i \mid u_n)\right. \\ &\qquad \left. +ED^i_n(x_i \mid u_n)^2\right\} \\ &= E\{ED^{i-1}_n(x_{i-1} \mid u_n)\operatorname{Var}(\eta^{i-1}_{n1}) \\ &\qquad +\operatorname{Var} D^{i-1}_n(x_{i-1} \mid u_n)(E\eta^{i-1}_{n1})^2+(ED^{i-1}_n(x_{i-1} \mid u_n)E\eta^{i-1}_{n1})^2 \\ &\qquad -2ED^{i-1}_n(x_{i-1} \mid u_n)E\eta^{i-1}_{n1}ED^i_n(x_i \mid u_n)+ED^i_n(x_i \mid u_n)^2\} \end{aligned}$$

$$
\begin{aligned}
&= (n\alpha_{i-1} + x_{i-1}n^{\frac{1}{2}})\beta E u_n p_{i-1}(1-p_{i-1}) \\
&\quad + (n\alpha_{i-1} + x_{i-1}n^{\frac{1}{2}})\beta E u_n (1-p_{i-1})^2 + \beta^2 (n\alpha_{i-1} + x_{i-1}n^{\frac{1}{2}})^2 E u_n^2 (1-p_{i-1})^2 \\
&\quad - 2(n\alpha_{i-1} + x_{i-1}n^{\frac{1}{2}})(1-p_{i-1})(n\alpha_i + x_i n^{\frac{1}{2}})\beta^2 E u_n^2 \\
&\quad + (n\alpha_i + x_i n^{\frac{1}{2}})\beta E u_n + (n\alpha_i + x_i n^{\frac{1}{2}})^2 \beta^2 E u_n^2 \\
&\rightarrow \alpha_i p_{i-1}\lambda^{-1} + (1-p_{i-1})\alpha_i\lambda^{-1} + \alpha_i^2(\sigma^2 + \lambda^{-2}) \\
&\quad - 2\alpha_i^2(\sigma^2 + \lambda^{-2}) + \alpha_i\lambda^{-1} + \alpha_i^2(\sigma^2 + \lambda^{-2}) \\
&= 2\alpha_i\lambda^{-1}.
\end{aligned}
$$

Finally, we consider the infinitesimal covariances:

$$
\begin{aligned}
\sigma_{n12}^2(x) &= E\left\{E\left[(1 - D_n^1(x_1 \mid u_n))\left(\sum_{j=1}^{D_n^1(x_1|u_n)} \eta_{nj}^1 - D_n^2(x_2 \mid u_n)\right)\right]\right\} \\
&= E \sum_{j=1}^{D_n^1(x_1|u_n)} \eta_{nj}^1 - E\left(D_n^1(x_1 \mid u_n) \sum_{j=1}^{D_n^1(x_1|u_n)} \eta_{nj}^1\right) \\
&\quad - ED_n^2(x_1 \mid u_n) + ED_n^1(x_1 \mid u_n) ED_n^2(x_2 \mid u_n) \\
&= (n\alpha_1 + x_1 n^{\frac{1}{2}})\beta E u_n (1-p_1) \\
&\quad - [(n\alpha_1 + x_1 n^{\frac{1}{2}})\beta E u_n + (n\alpha_1 + x_1 n^{\frac{1}{2}})^2 \beta^2 E u_n^2](1-p_1) \\
&\quad - (n\alpha_2 + x_2 n^{\frac{1}{2}})\beta E u_n + (n\alpha_1 + x_1 n^{\frac{1}{2}})(n\alpha_2 + x_2 n^{\frac{1}{2}})\beta^2 E u_n^2 \\
&\rightarrow \alpha_2\lambda^{-1} - \alpha_2\lambda^{-1} - \alpha_1\alpha_2(\sigma^2 + \lambda^{-2}) - \alpha_2\lambda^{-1} + \alpha_1\alpha_2(\sigma^2 + \lambda^{-2}) = -\alpha_2\lambda^{-1}. \\
&= -\alpha_2\lambda^{-1}.
\end{aligned}
$$

For $i > 1$, similarly, we have

$$
\begin{aligned}
\sigma_{ni,(i+1)}^2(x) &= E\left\{E\left[\left(\sum_{j=1}^{D_n^{i-1}(x_{i-1}|u_n)} \eta_{nj}^{i-1} - D_n^i(x_i \mid u_n)\right)\left(\sum_{j=1}^{D_n^i(x_i|u_n)} \eta_{nj}^i - D_n^{i+1}(x_{i+1} \mid u_n)\right)\right]\right\} \\
&= E\left\{E\left(\sum_{j=1}^{D_n^{i-1}(x_{i-1}|u_n)} \eta_{nj}^{i-1}\right) E\left(\sum_{j=1}^{D_n^i(x_i|u_n)} \eta_{nj}^i\right) - E\left(D_n^i(x_i \mid u_n) \sum_{j=1}^{D_n^i(x_i|u_n)} \eta_{nj}^i\right)\right. \\
&\quad \left. - E\left(\sum_{j=1}^{D_n^{i-1}(x_{i-1}|u_n)} \eta_{nj}^{i-1}\right) ED_n^{i+1}(x_{i+1} \mid u_n) + ED_n^i(x_i \mid u_n) ED_n^{i+1}(x_{i+1} \mid u_n)\right\} \\
&= (n\alpha_{i-1} + x_{i-1}n^{\frac{1}{2}})(1-p_{i-1})(n\alpha_i + x_i n^{\frac{1}{2}})(1-p_i)\beta^2 E u_n^2 \\
&\quad - [(n\alpha_i + x_i n^{\frac{1}{2}})\beta E u_n + (n\alpha_i + x_i n^{\frac{1}{2}})^2 \beta^2 E u_n^2](1-p_i) \\
&\quad - (n\alpha_{i-1} + x_{i-1}n^{\frac{1}{2}})(1-p_{i-1})(n\alpha_{i+1} + x_{i+1}n^{\frac{1}{2}})\beta^2 E u_n^2 \\
&\quad + (n\alpha_i + x_i n^{\frac{1}{2}})(n\alpha_{i+1} + x_{i+1}n^{\frac{1}{2}})\beta^2 E u_n^2 \\
&\rightarrow \alpha_i\alpha_{i+1}(\sigma^2 + \lambda^{-2}) - \alpha_{i+1}\lambda^{-1} - \alpha_i\alpha_{i+1}(\sigma^2 + \lambda^{-2}) \\
&\quad - \alpha_i\alpha_{i+1}(\sigma^2 + \lambda^{-2}) + \alpha_i\alpha_{i+1}(\sigma^2 + \lambda^{-2}) \\
&= -\alpha_{i+1}\lambda^{-1}.
\end{aligned}
$$

Finally, for $k > i+1$,

$$\begin{aligned}\sigma^2_{nik}(x) &= E\left\{\left(\sum_{j=1}^{D_n^{i-1}(x_{i-1}\mid u_n)} \eta_{nj}^{i-1} - D_n^i(x_i \mid u_n)\right)\left(\sum_{j=1}^{D_n^{k-1}(x_{k-1}\mid u_n)} \eta_{nj}^{k-1} - D_n^k(x_k \mid u_n)\right)\right\}\\ &= ([(n\alpha_{i-1} + x_{i-1}n^{\frac{1}{2}})(1-p_{i-1}) - (n\alpha_i + x_i n^{\frac{1}{2}})]\\ &\quad \times [(n\alpha_{k-1} + x_{k-1}n^{\frac{1}{2}})(1-p_{k-1}) - (n\alpha_k + x_k n^{\frac{1}{2}})])\beta^2 Eu_n^2\\ &= (x_{i-1}(1-p_{i-1}) - x_i)(x_{k-1}(1-p_{k-1}) - x_k)n\beta^2 Eu_n^2\\ &\to 0.\end{aligned}$$

We do not display calculations for the extra regularity condition (2.6) on p. 268 of Stroock and Varadhan (1979), but it is easy to check too.

References

Arnold, L. (1974) *Stochastic Differential Equations: Theory and Applications.* Wiley, New York.

Beekman, J. A. (1974) *Two Stochastic Processes.* Almqvist and Wiksell, Stockholm.

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

Borovkov, A. A. (1967) On limit laws for service processes in multi-channel systems. *Siberian Math. J.* **8,** 746–763.

Borovkov, A. A. (1976) *Stochastic Processes in Queueing Theory.* Springer-Verlag, New York.

Breiman, L. (1968) *Probability.* Addison-Wesley, Reading, Ma.

Brillinger, D. R. (1974) Cross-spectral analysis of processes with stationary increments including the stationary *GI/G/∞* queue. *Ann. Prob.* **2,** 815–827.

De Smit, J. H. A. (1973a) Some general results for many server queues. *Adv. Appl. Prob.* **5,** 153–169.

De Smit, J. H. A. (1973b) On the many server queue with exponential service times. *Adv. Appl. Prob.* **5,** 170–182.

Doob, J. L. (1953) *Stochastic Processes.* Wiley, New York.

Finkbeiner, D. T. (1966) *Introduction to Matrices and Linear Transformations,* 2nd edn. W. H. Freeman, San Francisco.

Franken, P. (1975) Stationary probabilities of states of queueing systems at different times. *Engineering Cybernetics* **1,** 84–89.

Franken, P. (1976a) Some applications of point processes in queueing theory, I (in German). *Math. Nachr.* **70,** 303–319.

Franken, P. (1976b) On the investigation of queueing and reliability models with the help of point processes. Department of Mathematics, Humboldt University, Berlin, DDR.

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

Gikhman, I. I. and Skorohod, A. V. (1969) *Introduction to the Theory of Random Processes.* W. B. Saunders, Philadelphia.

Glynn, P. W. (1982) On the Markov property of the *GI/G/∞* Gaussian limit. *Adv. Appl. Prob.* **14,** 191–194.

Gnedenko, B. V. and Kovalenko, I. N. (1968) *Introduction to Queueing Theory.* Israel Program for Scientific Translations, Jerusalem.

Halachmi, B. and Franta, W. R. (1978) A diffusion approximation to the multi-server queue. *Management Sci.* **24,** 522–529.

Halfin, S. and Whitt, W. (1981) Heavy-traffic limits for queues with many exponential servers. *Operat. Res.* **29,** 567–588.

HARRISON, J. M. (1978) The diffusion approximation for tandem queues in heavy traffic. *Adv. Appl. Prob.* **10,** 886–905.

IGLEHART, D. L. (1965) Limit diffusion approximations for the many server queue and the repairman problem. *J. Appl. Prob.* **2,** 429–441.

IGLEHART, D. L. (1968) Limit theorems for the multi-urn Ehrenfest model. *Ann. Math. Statist.* **39,** 864–876.

IGLEHART, D. L. (1973a) Weak convergence of compound stochastic processes. *Stoch. Proc. Appl.* **1,** 11–31.

IGLEHART, D. L. (1973b) Weak convergence in queueing theory. *Adv. Appl. Prob.* **5,** 570–594.

IGLEHART, D. L. AND LALCHANDANI, A. P. (1973) Diffusion approximations for complex repair systems. Technical Report No. 266-12, Control Analysis Corporation, 800 Welch Road, Palo Alto, California.

JAGERS, P. (1968) Age-dependent branching processes allowing immigration. *Theory Prob. Appl.* **13,** 225–236.

KAMAE, T., KRENGEL, U. AND O'BRIEN, G. L. (1977) Stochastic inequalities on partially ordered spaces. *Ann. Prob.* **5,** 899–912.

KAPLAN, N. (1975) Limit theorems for a $GI/G/\infty$ queue. *Ann. Prob.* **3,** 780–789.

KEILSON, J. AND ROSS, H. F. (1975) Passage time distributions for Gaussian Markov (Ornstein–Uhlenbeck) statistical processes. *Selected Tables in Math. Statist.* **3,** 233–327.

KELLY, F. P. (1976) Networks of queues. *Adv. Appl. Prob.* **8,** 416–432.

KINGMAN, J. F. C. (1962) On queues in heavy traffic. *J. R. Statist. Soc.* B **24,** 383–392.

KÖNIG, D., SCHMIDT, V. AND STOYAN, D. (1976) On some relations between stationary distributions of queue lengths and imbedded queue length in $G/G/s$ systems. *Math. Operationsforsch. Statist.* **7,** 577–586.

LEMOINE, A. J. (1978) Networks of queues—a survey of weak convergence results. *Management. Sci.* **24,** 1175–1193.

MCNEIL, D. R. (1973) Diffusion limits for congestion models. *J. Appl. Prob.* **10,** 368–376.

MCNEIL, D. R. AND SCHACH, S. (1973) Central limit analogues for Markov population processes. *J. R. Statist. Soc.* B **35,** 1–23.

MILLER, D. R. (1972) Existence of limits in regenerative processes. *Ann. Math. Statist.* **43,** 1275–1282.

MILLER, D. R. AND SENTILLES, F. D. (1975) Translated renewal processes and the existence of a limiting distribution for the queue length of the $GI/G/s$ queue. *Ann. Prob.* **3,** 424–439.

NEWELL, G. F. (1973) *Approximate Stochastic Behavior of n-Server Service Systems with Large n.* Lecture Notes in Economics and Mathematical Systems **87,** Springer-Verlag, Berlin.

O'BRIEN, G. L. (1975) The comparison method for stochastic processes. *Ann. Prob.* **3,** 80–88.

PAKES, A. G. AND KAPLAN, N. (1974) On the subcritical Bellman–Harris process with immigration. *J. Appl. Prob.* **11,** 652–668.

REIMAN, M. I. (1977) Queueing networks in heavy traffic. Technical Report No. 76, Department of Operations Research, Stanford University.

ROSS, S. M. (1970) *Applied Probability Models with Optimization Applications.* Holden-Day, San Francisco.

SCHACH, S. (1971) Weak convergence results for a class of multivariate Markov processes. *Ann. Math. Statist.* **42,** 451–465.

SCHASSBERGER, R. (1973) *Queueing Theory* (in German). Springer-Verlag, Berlin.

STROOCK, D. W. AND VARADHAN, S. R. S. (1979) *Multidimensional Diffusion Processes.* Springer-Verlag, New York.

WHITT, W. (1972) Embedded renewal processes in the $GI/G/s$ queue. *J. Appl. Prob.* **9,** 650–658.

WHITT, W. (1980) Some useful functions for functional limit theorems. *Math. Operat. Res.* **5,** 67–85.

WHITT, W. (1981) Comparing counting processes and queues. *Adv. Appl. Prob.* **13,** 207–220.

WHITT, W. (1981a) Existence of limiting distributions in the $GI/G/s$ queue. *Math. Operat. Res.* **6.**