MATHEMATICS OF OPERATIONS RESEARCH
Vol. 20, No. 1, February 1995
Printed in U.S.A.

STRONG APPROXIMATIONS FOR TIME-DEPENDENT QUEUES

AVI MANDELBAUM AND WILLIAM A. MASSEY

A time-dependent $M_t/M_t/1$ queue alternates through periods of under-, over-, and critical loading. We derive period-dependent, pathwise asymptotic expansions for its queue length, within the framework of strong approximations. Our main results include time-dependent fluid approximations, supported by a functional strong law of large numbers, and diffusion approximations, supported by a functional central limit theorem. This complements and extends previous work on asymptotic expansions of the queue-length transition probabilities.

1. Introduction. The governing laws for the evolution of real-world queueing systems vary with time. Yet queueing research and practice, spanning a period of over nine decades, have been devoted almost exclusively to time-homogeneous models. Such models can indeed provide reasonable approximations for slowly varying systems. However, there are many time-dependent phenomena, such as rush hour or periodicity, that they fail to capture. Time-dependent models are difficult to analyze, even in a Markovian setting. Our goal therefore, is to develop a rigorous framework for their asymptotic approximations, starting in this paper with the $M_t/M_t/1$ queue.

A Markovian analysis of a time-homogeneous queueing system entails encoding its dynamics into Kolmogorov's forward (or backward) differential equations. Their solution yields the transition probabilities for the queueing model of the system. However, Kolmogorov's equations rarely have closed-form solutions, hence one resorts to steady state analysis. This reduces the problem from solving a set of differential equations to solving linear equations. The solution of the latter yields the steady state probabilities for the queueing model.

Time-dependent queueing systems also can be modelled by continuous-time Markov chains, but they must be time-inhomogeneous. Their transition probabilities solve Kolmogorov's equations as well, but one cannot expect explicit solutions in view of the complexity already encountered in the time-homogeneous case. Worse still, it is not immediately clear what constitutes a steady state analysis for time-inhomogeneous systems (at least when its evolution is not periodic; see for example Asmussen and Thorisson (1987), Bambos and Walrand (1989), Harrison and Lemoine (1977), Heyman and Whitt (1984), Lemoine (1989), Rolski (1981, 1990)). In particular, approximating the behavior of the system in the here and now by its behavior at time "infinity" is typically futile.

A time-inhomogeneous analogue to steady state analysis was proposed in the Ph.D. Thesis by Massey (1981) (see also Massey (1985) and Keller (1982)), where it was coined *uniform acceleration*. Here one scales all the average instantaneous transition rates of the Markovian model by a factor of $1/\epsilon$. As $\epsilon \downarrow 0$, each rate increases in

*Received November 9, 1992; revised July 14, 1993.
AMS 1991 subject classification. Primary: 60K25. Secondary: 68M20, 90B22.
OR/MS Index 1978 subject classification. Primary: 681 Queues. Secondary: 684 Queues/Approximations, 694 Queues/Limit theorems.
Key words. Time dependent queues, nonstationary queues, strong approximations, fluid approximations, diffusion approximations, asymptotic analysis.

0364-765X/95/2001/0033/$01.25

absolute terms, or is accelerated, but the ratio of any two rates relative to each other is held fixed.

Uniform acceleration enables a *dynamic* asymptotic analysis of time-inhomogeneous queueing models, which yields asymptotic expansions that vary over time. Moreover, when applied to time-inhomogeneous Markovian systems, it reduces to either steady state or heavy traffic analysis. (See the examples in §§4.1 and 4.4.) In Massey (1985), uniform acceleration gave rise to an asymptotic expansion of the *transition probabilities* for the queue length process of a time dependent $M/M/1$ queue, hereafter denoted by $M_t/M_t/1$. This provided a rigorous foundation to the earlier work of Newell (1968) and Keller (1982). It also led to the proper notion of a time-dependent traffic intensity parameter, namely $\rho^*(t)$ defined in (3.1) below and elaborated on in §7.

The purpose of this paper is to complement and refine Newell (1968), Massey (1981, 1985) and Keller (1982). We do this through an asymptotic analysis of the queue length *sample paths*, within the unifying framework of the strong approximation theorems, introduced by Komlós, Major and Tusnády (1976). Strassen was the first to prove a strong approximation result, then Skorohod introduced his embedding of random walks in a Brownian motion, and Keifer used it to establish answers to questions about best convergence rates (see Csörgő and Révész (1981) and Csörgő and Horvath (1993) for a survey on the historical evolution of the subject). However, the framework operate within is the one by Komlós, Major, and Tusnàdy (1976). Specifically, in §2, we apply uniform acceleration directly to the sample paths (2.1) of the queue $M_t/M_t/1$. The outcome is the asymptotic expansion (2.7). Its derivation relies on a *functional strong law of large numbers* (FSLLN, Theorem 2.1) and a *functional central limit theorem* (FCLT, Theorem 2.2). Both theorems are consequences of the strong approximation results in Theorem 2.3. The FSLLN limit (2.4) is deterministic and, as shown later, has the interpretation of a fluid flow system. Viewing the original model as a *microscopic* description, this deterministic fluid model provides a *macroscopic* fluid approximation of the queue which, furthermore, is the zeroth-order term in the asymptotic expansion (2.7) of its sample paths. The stochastic FCLT limit (2.2) then deserves to be referred to as a *mesoscopic* first-order refinement of the fluid model.

During its evolution, the $M_t/M_t/1$ queue can alternate between underloaded, critically loaded and overloaded phases. These phases are determined by its fluid approximation, and the phase transitions are summarized in Figure 3.1. Moreover, the asymptotic expansion (2.7) can be localized to each phase, and this is outlined in §3 and substantiated in §§8–10. In §4, we specialize our results to the time-homogeneous $M/M/1$ queue and to two periodic models. Finally, §§5 and 6 are devoted to proving Theorems 2.1–2.3 and their supporting assertions. Of special importance is Lemma 5.2, which provides the sample-path intuition behind our main asymptotic expansion (2.7).

The literature on time-inhomogeneous models, like the $M_t/M_t/1$ queue, is not vast. Insight and calculations have been commonly based on either approximations (Luchak (1956), Newell (1968), Keller (1982), Massey (1981, 1985), Rothkopf and Oren (1979), for example), or simulation (Green, Kolesar, and Svornos (1991) for example). Exact results are rarely available, with the notable exception of networks with Poisson arrivals and infinite server nodes (see Eick, Massey and Whitt (1993a, b) as well as Massey and Whitt (1993)). For a textbook treatment of some aspects of time-dependent queues, see Hall (1991), for example. Our paper focuses only on Poissonian single-stations, but we are also studying time-inhomogeneous Markovian networks (Mandelbaum and Massey, in preparation), for which the current paper is a prerequisite. Our framework also accommodates more general point processes, as in Chap-

ter 10 of Lipster and Shiryaev (1989). This latter work employs uniform acceleration of *state-dependent* queues, as also in Anulova (1989), Krichagina, Lipster, and Puhalski (1988), and Yamada (1984). Finally we note that the first application of the Komlós, Major and Tusnády theorem is due to Rosenkrantz (1980). The results of paper suggest that it might be possible to obtain estimates, in terms of ϵ, on the rates of convergence of the distributions of certain functionals to their limits.

Notations. Denote by $D = D[0,\infty)$ the space of all functions $x\colon [0,\infty) \to \mathbb{R}^1$ such that $x(0) = 0$, x is right-continuous at 0, and x is either right- or left-continuous at every $t > 0$. (This is a slight deviation from the common convention in which functions in D are taken right-continuous.) For $x \in D$, define $\bar{x}$ to be the *upper envelope* of the function x, that is

$$\bar{x}(t) \equiv \sup_{0 \leqslant s \leqslant t} x(s), \qquad t \geqslant 0.$$

The *completed graph* $\Gamma(x)$ of x is defined to be the subset of $[0,\infty) \times \mathbb{R}$ such that

$$\Gamma(x) = \{(t,\gamma) | x(t-) \leqslant \gamma \leqslant x(t+)\}, \tag{1.1}$$

with the convention $x(0-) = 0$. A *parametric representation* of $\Gamma(x)$ is a function $(\tau, g)\colon [0,\infty) \to \Gamma(x)$ which is onto, continuous, and τ is nondecreasing. A sequence x_n is M_1*-convergent* to x if there exist parametric representations (τ_n, g_n) of x_n which converge, uniformly on compact subsets of $[0,\infty)$, to some parametric representation (τ, g) of x. Formally, for all $t > 0$, $\|\tau_n - \tau\|_t \vee \|g_n - g\|_t \to 0$, as $n \uparrow \infty$, where

$$\|x\|_t = \sup_{0 \leqslant s \leqslant t} |x(s)|,$$

for any $x \in D$. M_1-convergence induces the M_1-topology on D, under which D is a Polish space (Pomerade (1976)). It is weaker than the more prevalent J_1-topology, which happens to be too strong for our purposes. (See the concluding paragraph of §2 for an elaboration.) M_1-convergence is metrizable, for example (as in Whitt (1980)) by defining

$$d(x_1, x_2) = \int_0^\infty e^{-t}[1 \wedge d_t(x_1, x_2)]\, dt,$$

where $d_t(x_1, x_2) = \inf(\|\tau_1 - \tau_2\|_t + \|g_1 - g_2\|_t)$, the infimum being taken over all possible parametrizations (τ_i, g_i) of x_i, for $i = 1, 2$.

Consider a family $\{x^\epsilon | \epsilon > 0\}$ and a function y, all elements in D. For some real-valued function $f(\epsilon)$, the little-o notation

$$x^\epsilon = f(\epsilon) y + o(f(\epsilon)),$$

stands for

$$\lim_{\epsilon \downarrow 0} d\left(\frac{1}{f(\epsilon)} x^\epsilon, y\right) = 0.$$

This is equivalent to $d_t(x^\epsilon/f(\epsilon), y) \to 0$, for almost all $t \geqslant 0$ or, in words, $\lim_{\epsilon \downarrow 0} x^\epsilon/f(\epsilon) = y$ in M_1. The big-O notation

$$x^\epsilon = f(\epsilon) y + O(f(\epsilon)),$$

means that for some $\epsilon_0 > 0$,

$$\sup_{0<\epsilon<\epsilon_0} \left\| \frac{1}{f(\epsilon)} x^\epsilon - y \right\| < \infty,$$

where

$$\|x\| = \int_0^\infty e^{-t}[1 \wedge \|x\|_t]\, dt.$$

When used with stochastic processes X^ϵ and Y, defined on a common probability space with sample paths in D, little-o and big-O indicate the above types of asymptotic behavior for almost all sample paths. Finally, M_1-convergence for stochastic processes, say $\lim_{\epsilon \downarrow 0} X^\epsilon/f(\epsilon) = Y$, holds if and only if there exists realizations of X^ϵ and Y, on a common probability space, for which M_1-convergence holds almost surely (Skorohod 1956, Pomerade 1976, Ethier and Kurtz 1986).

2. General asymptotic expansions. Our model for the queue-length process of an $M_t/M_t/1$ queue is taken to be

$$Q(t) \equiv X(t) - \inf_{0 \leqslant s \leqslant t} X(s), \qquad t \geqslant 0. \tag{2.1}$$

Here,

$$X(t) \equiv N^+\left(\int_0^t \lambda(r)\, dr\right) - N^-\left(\int_0^t \mu(r)\, dr\right), \tag{2.2}$$

N^+ and N^- are two independent Poisson processes with unit rate, λ is a continuous nonnegative function, μ is continuous and positive, and $Q(0) = 0$ is assumed for simplicity. The process X represents the difference between the cumulative number of actual arrivals and *potential* departures for Q. The jumps of Q coincide with those of X whenever Q is strictly positive. Otherwise Q equals zero and $X(t)$ equals its infimum over the time interval $[0, t]$. In this case, a unit increase in X causes a unit increase in Q. However, a unit decrease for X is matched by a unit decrease for the running infimum for X, which results in no change for Q. (Only in this last case, is a potential departure not realized.)

We derive asymptotic expansions of Q by uniformly accelerating its instantaneous transition rates. Formally, for each $\epsilon > 0$, introduce a stochastic process Q^ϵ by

$$Q^\epsilon(t) \equiv X^\epsilon(t) - \inf_{0 \leqslant s \leqslant t} X^\epsilon(s), \qquad t \geqslant 0, \tag{2.3}$$

in which

$$X^\epsilon(t) \equiv N^+\left(\frac{1}{\epsilon}\int_0^t \lambda(r)\, dr\right) - N^-\left(\frac{1}{\epsilon}\int_0^t \mu(r)\, dr\right).$$

THEOREM 2.1 (FSLLN). *The following functional strong law of large numbers holds for* Q^ϵ:

$$\lim_{\epsilon \downarrow 0} \epsilon Q^\epsilon(t, \omega) = Q^{(0)}(t) \quad a.s.$$

where

$$Q^{(0)}(t) \equiv \int_0^t [\lambda(r) - \mu(r)]\, dr - \min_{0 \leqslant s \leqslant t} \int_0^s [\lambda(r) - \mu(r)]\, dr, \tag{2.4}$$

and the convergence is uniform on compact subjects of $t \geqslant 0$.

The proof of Theorem 2.1, as well as those of Theorems 2.2 and 2.3, are deferred to §5. Theorem 2.1 gives rise to the asymptotic expansion

$$Q^{\epsilon}(t, \omega) = \frac{1}{\epsilon} Q^{(0)}(t) + o\left(\frac{1}{\epsilon}\right) \quad \text{a.s.}, \tag{2.5}$$

from which the deterministic process $Q^{(0)}$ emerges as a first-order, *macroscopic*, fluid approximation for Q. Indeed, $Q^{(0)}$ can be animated as the fluid level in a buffer that is governed by the following dynamics (Chen and Mandelbaum 1991a): The buffer is empty at time $t = 0$. At time $t > 0$, the exogenous inflow rate is $\lambda(t)$, and the potential outflow rate is $\mu(t)$. Finally, the actual outflow rate is strictly below its potential only when the buffer is empty, in which case it coincides with the inflow rate. With this interpretation, the quantity

$$Y^{(0)}(t) \equiv - \min_{0 \leqslant s \leqslant t} \int_0^s [\lambda(r) - \mu(r)]\, dr,$$

represents the *cumulative potential outflow* that is lost prior to time t.

Now define Φ_t to be the set of all times s up to t at which the fluid level is zero, but no potential outflow is lost during $[s, t]$. Thus

$$\Phi_t \equiv \{0 \leqslant s \leqslant t | Q^{(0)}(s) = 0 \text{ and } Y^{(0)}(s) = Y^{(0)}(t)\}.$$

THEOREM 2.2 (FCLT). *The following functional central limit theorem holds for* Q^{ϵ}:

$$\lim_{\epsilon \downarrow 0} \sqrt{\epsilon} \left(Q^{\epsilon}(t) - \frac{1}{\epsilon} Q^{(0)}(t) \right) \stackrel{\mathrm{d}}{=} Q^{(1)}(t) \tag{2.6}$$

where

$$Q^{(1)}(t) \equiv W\left(\int_0^t [\lambda(r) + \mu(r)]\, dr \right) - \min_{s \in \Phi_t} W\left(\int_0^s [\lambda(r) + \mu(r)]\, dr \right),$$

$W = \{W(t) | t \geqslant 0\}$ *is standard Brownian motion, and the convergence is weak with respect to Skorohod's* M_1*-topology on* $D[0, \infty)$. *Here it is assumed that* $Q^{(1)}$ *has a finite number of discontinuities on any compact subset of* $[0, \infty)$.

The nature of the discontinuities for $Q^{(1)}$ will be elaborated on in Theorem 3.1. The FCLT refines (2.5) in *distribution*, and gives rise to the asymptotic expansion

$$Q^{\epsilon}(t) \stackrel{\mathrm{d}}{=} \frac{1}{\epsilon} Q^{(0)}(t) + \frac{1}{\sqrt{\epsilon}} Q^{(1)}(t) + o\left(\frac{1}{\sqrt{\epsilon}}\right), \tag{2.7}$$

from which the stochastic process $Q^{(1)}$ emerges as a second-order, *mesoscopic*,

diffusion approximation for the deviation of Q from its fluid approximation $Q^{(0)}$. Our FSLLN and FCLT results are consequences of

THEOREM 2.3 (STRONG APPROXIMATION). *The parametrized family* $\{Q^\epsilon | \epsilon > 0\}$ *can be realized on a probability space* $(\Omega, \mathcal{F}, \mathsf{P})$, *supporting two independent, standard Brownian motions* W^+ *and* W^- *in a way that*

$$Q^\epsilon(t,\omega) = \tilde{X}^\epsilon(t,\omega) - \min_{0 \leqslant s \leqslant t} \tilde{X}^\epsilon(s,\omega) + O(\log \epsilon) \quad a.s.$$

where

$$\tilde{X}^\epsilon(t) = \frac{1}{\epsilon}\int_0^t [\lambda(r) - \mu(r)]\, dr + W^+\left(\frac{1}{\epsilon}\int_0^t \lambda(r)\, dr\right) - W^-\left(\frac{1}{\epsilon}\int_0^t \mu(r)\, dr\right). \tag{2.8}$$

It is now possible to motivate the presence of M_1 in Theorem 2.2. Indeed, the process $\tilde{X}^\epsilon$ has continuous sample paths, and so does $Q^{(0)}$. Thus, up to a negligible $\sqrt{\epsilon}\, O(\log \epsilon)$ term, the left-hand side of (2.6) is continuous. The limit $Q^{(1)}$, on the other hand, need *not* be continuous (see Theorem 3.1 for a precise characterization). Since continuous functions can not converge to a discontinuous one in the commonly used J_1-topology (the "largest jump" functional is J_1-continuous), one must use M_1.

3. Local asymptotic expansions. We now refine our asymptotic analysis of the $M_t/M_t/1$ queue. Let $\rho(t) \equiv \lambda(t)/\mu(t)$, and define

$$\rho^*(t) \equiv \sup_{0 \leqslant s < t} \frac{\int_s^t \lambda(r)\, dr}{\int_s^t \mu(r)\, dr}, \qquad t > 0, \tag{3.1}$$

with the convention $\rho^*(0) = \lambda(0)/\mu(0)$. The quantity ρ^* is the $M_t/M_t/1$ *traffic intensity function* introduced in Massey (1981). It will be seen that the functions ρ and ρ^* summarize the information embodied in the fluid model that is relevant to accelerating the stochastic model. In particular, ρ^* identifies three exhaustive asymptotic regions for $M_t/M_t/1$ as follows:

- Underloaded: $\rho^*(t) < 1$.
- Critically Loaded: $\rho^*(t) = 1$.
- Overloaded: $\rho^*(t) > 1$.

An equivalent characterization of these asymptotic regions, in terms of Φ_t and $\rho(t)$, will be given in §7. We also show there, that ρ^* is a continuous function of t. By Lindelöf's theorem, the underloaded and overloaded regions decompose into a countable disjoint union of open intervals. The set of critically loaded times is closed. For our asymptotic expansions, we must further divide it into the following four subregions:

- Onset of Critical Loading: $\rho^*(t) = 1$, and there exists a sequence $l_n \uparrow t$ such that $\rho^*(l_n) < 1$ for all n.
- Middle of Critical Loading: $\rho^*(t) = 1$, $\rho^* \geqslant 1$ on some open interval containing t, and there exists a sequence $l_n \uparrow t$ such that $\rho^*(l_n) = 1$ for all n.
- End of Critical Loading: $\rho^*(t) = 1$, $\rho^* \geqslant 1$ on some open interval where t is its right endpoint, there exists a sequence $l_n \uparrow t$ such that $\rho^*(l_n) = 1$ for all n, and there exists a sequence $r_n \downarrow t$ such that $\rho^*(r_n) < 1$ for all n.
- End of Overloading: $\rho^*(t) = 1$, and $\rho^* > 1$ on some open interval where t is its right endpoint.

Whereas the $M/M/1$ has three static types of asymptotic behavior (see §4.1), the $M_t/M_t/1$ has six types, and a single process may alternate among all of them over

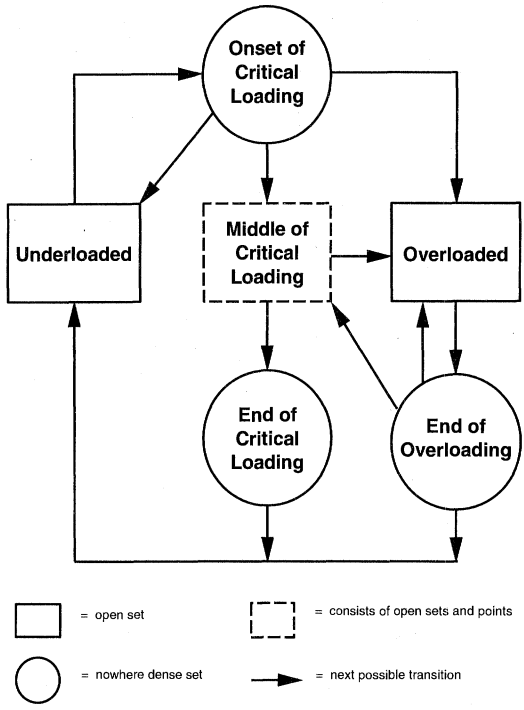

FIGURE 3.1. Phase transition diagram for the asymptotic regions.

time. Figure 3.1 consists of a diagram that shows possible phase transitions among asymptotic regions. A rectangle denotes a region that is an open set, in the case of underloaded and overloaded, or potentially has a nonempty interior with a dashed boundary, in the case of the middle of critical loading. (This assertion follows from the observation that any open subset of critically loaded times must always be in the middle of critical loading.) The circles denote closed sets that are nowhere dense.

The first theorem concerns sample-path properties of the asymptotic diffusion term.

THEOREM 3.1 (SAMPLE PATHS OF $Q^{(1)}$). *The process $Q^{(1)}$ is upper semicontinuous, almost surely. It is discontinuous at time t, with a nonzero probability, if and only if t is the end-point of overloading or critical loading. The set of such points is nowhere dense.*

The proof of Theorem 3.1 is deferred to the end of §7. We proceed with localizing our asymptotic expansions to the various regions.

THEOREM 3.2 (UNDERLOADED). *As the time points $t_1 < t_2 < \cdots$ vary through the underloaded region,*

$$\lim_{\epsilon \downarrow 0} Q^{\epsilon}(t_i) \overset{\mathrm{d}}{=} \tilde{Q}(t_i), \tag{3.2}$$

for all $i = 1, 2, \ldots$, where the $\tilde{Q}(t_i)$'s are mutually independent random variables, and the distribution of $\tilde{Q}(t_i)$ is geometric, with parameter $\rho(t_i)$.

The proof of Theorem 3.2 will be given in §8. The geometric limiting marginal distributions can be anticipated from Massey (1985), who proved

THEOREM 3.3 (TRANSITION PROBABILITIES). *Suppose that λ and μ are infinitely differentiable functions on the positive real line. As $\epsilon \downarrow 0$, $Q^\epsilon(t)$ converges in distribution to a probability measure if and only if $\rho^*(t) < 1$, at all $t > 0$. Moreover, the distribution of $Q^\epsilon(t)$ has then the following asymptotic series*:

$$\mathsf{P}(Q^\epsilon(t) = n) \cong \sum_{k=0}^{\infty} \epsilon^k \pi_n^{(k)}(t), \qquad \epsilon \downarrow 0,$$

where $\pi_n^{(0)}(t) = (1 - \rho(t))\rho(t)^n$, and the $\pi_n^{(k)}(t)$'s are, for each fixed integer $k \geqslant 1$, the unique solution to the equations

$$\mu(t)\pi_1^{(k+1)}(t) - \lambda(t)\pi_0^{(k+1)}(t) = \frac{d}{dt}\pi_0^{(k)}(t)$$

and

$$\lambda(t)\pi_{n-1}^{(k+1)}(t) + \mu(t)\pi_{n+1}^{(k+1)}(t) - (\lambda(t) + \mu(t))\pi_n^{(k+1)}(t) = \frac{d}{dt}\pi_n^{(k)}(t),$$

for all $n \geqslant 1$.

A simple consequence of Theorem 3.3 is

$$\lim_{\epsilon \downarrow 0} \mathsf{P}(Q^\epsilon(t) > 0) = \begin{cases} \rho(t) & \text{if } \rho^*(t) < 1, \\ 1 & \text{otherwise,} \end{cases}$$

which holds for all $t > 0$. Both Theorems 3.2 and 3.3 clarify the common practice of approximating the $M_t/M_t/1$ queue, when underloaded at time t, by the $M/M/1$ queue with traffic intensity $\rho = \rho(t)$. In particular, such an approximation is not justified when $\rho(t) < 1$ while $\rho^*(t) > 1$.

For the next theorem, define

$$Q^\epsilon(s,t) \equiv X^\epsilon(s,t) - \inf_{s \leqslant r \leqslant t} X^\epsilon(r,t), \qquad 0 \leqslant s \leqslant t, \tag{3.3}$$

where

$$X^\epsilon(s,t) \equiv \hat{N}_\epsilon^+(s,t) - \hat{N}_\epsilon^-(s,t),$$

$$\hat{N}_\epsilon^+(s,t) \equiv N^+\left(\frac{1}{\epsilon}\int_0^t \lambda(r)\,dr\right) - N^+\left(\frac{1}{\epsilon}\int_0^s \lambda(r)\,dr\right),$$

and $\hat{N}_\epsilon^-$ is defined similarly in terms of N^-. The process $\{Q^\epsilon(s,t) | t \geqslant s\}$ has the same distribution as Q^ϵ conditioned to be zero at time s.

THEOREM 3.4 (ONSET OF CRITICAL LOADING). *Consider a point t that is an onset of critical loading. If λ and μ are differentiable in a neighborhood about t, and*

$$\lambda(t + \tau) = \mu(t) + \frac{\lambda^{(k)}(t)}{k!}\tau^k + o(\tau^k),$$

for some $k \geqslant 1$, then for all pairs of real numbers $T_0 \leqslant T$, we have

$$Q^\epsilon\left(t + \epsilon^{1/(2k+1)}T_0, t + \epsilon^{1/(2k+1)}T\right) \stackrel{\mathrm{d}}{=} \frac{1}{\epsilon^{k/(2k+1)}}\left[\tilde{W}(T_0, T) - \inf_{T_0 \leqslant \sigma \leqslant T} \tilde{W}(T_0, \sigma)\right] + o\left(\frac{1}{\epsilon^{k/(2k+1)}}\right)$$

where

$$\tilde{W}(T_0, T) \equiv \frac{\lambda^{(k)}(t)}{(k+1)!}\left[T^{k+1} - T_0^{k+1}\right] + W(2\mu(t)(T - T_0)).$$

The proof of Theorem 3.4 will be given in §9. We have rescaled space and time so that the original queueing process, now evolving over $T \geqslant T_0$, converges to a reflected Brownian motion with polynomial drift. For example, with $k = 1$, we have $1/(2k+1) = k/(2k+1) = 1/3$ and the drift is quadratic. For $k = 2$, we have $1/(2k+1) = 1/5$, $k/(2k+1) = 2/5$, and the drift is cubic. These last two pairs of exponents for ϵ correspond to the time and queue length scalings given by Newell (1968) and (1982) for diffusion approximations at "transition through saturation" (page 270 of Newell 1982) and "a mild rush hour" (page 275) respectively.

Now introduce the process

$$\hat{W}(s, t) = W\left(\int_0^t [\lambda(r) + \mu(r)]\, dr\right) - W\left(\int_0^s [\lambda(r) + \mu(r)]\, dr\right), \qquad 0 \leqslant s \leqslant t.$$

The next two theorems are immediate consequences of Theorems 2.2 and 3.1.

THEOREM 3.5 (MIDDLE AND END OF CRITICAL LOADING). *As the point t varies through the middle or end of critical loading,*

$$Q^\epsilon(t) \stackrel{\mathrm{d}}{=} \frac{1}{\sqrt{\epsilon}} \sup_{s \in \Phi_t} \hat{W}(s, t) + o\left(\frac{1}{\sqrt{\epsilon}}\right),$$

where $t = \sup \Phi_t$.

Recall that $\sup_{s \in \Phi_t} \hat{W}(s, t) = Q^{(1)}(t)$ hence, by Theorem 3.1, it is continuous at t in the middle of critical loading, for almost all sample paths. Furthermore, if t belongs to an interval (l, r) of critical loading (which is by definition the middle of critical loading), and Φ_t is a closed interval, then over (l, t) the queueing process rescales to a driftless, reflected Brownian motion. On the other hand, for t at the end of critical loading, $Q^{(1)}$ is still left-continuous for almost every sample path, but it is not right-continuous at t with a positive probability.

THEOREM 3.6 (OVERLOADED). *As t varies through the overloaded region,*

$$Q^\epsilon(t) \stackrel{\mathrm{d}}{=} \frac{1}{\epsilon}\int_{t^*}^t [\lambda(r) - \mu(r)]\, dr + \frac{1}{\sqrt{\epsilon}}\left[\sup_{s \in \Phi_t} \hat{W}(s, t^*) + \hat{W}(t^*, t)\right] + o\left(\frac{1}{\sqrt{\epsilon}}\right).$$

where $t^* = \sup \Phi_t < t$.

With this theorem, one can demonstrate the phenomenon that each time of peak arrival rate lags behind a time of peak congestion, or peak queue length. Assume for simplicity, that μ is constant. Now suppose that $Q^{(0)}$ attains a positive local maximum

at t, a time of overloading. Note that (t^*, t) is the maximal time interval prior to t over which overloading occurs. (In particular, t^* is a time for the onset of overloading.) It follows that if $l^* = \sup \Phi_l$ for every intermediate time $l \in (t^*, t)$, then $l^* = t^*$ and so $\lambda(t) = \mu$ must hold. Since $\lambda(t^*) = \mu$ as well, λ must attain some maximal value between t^* and t.

THEOREM 3.7 (END OF OVERLOADING). *If overloading ends at time t, and λ and μ are continuous in a neighborhood of t, then for all T,*

$$Q^\epsilon(t + \sqrt{\epsilon}\, T) \stackrel{d}{=} \frac{1}{\sqrt{\epsilon}} \left[\sup_{s \in \Phi_t - \{t\}} \hat{W}(s, t) + (\lambda(t) - \mu(t))T \right]^+ + o\left(\frac{1}{\sqrt{\epsilon}}\right).$$

The proof of Theorem 3.7 will be given in §10. Theorems 3.5 and 3.7 clarify an issue not addressed thoroughly in either Massey (1981) or Newell (1968). Proposition 6.3 of Massey (1981) hints at this behavior by showing that the mean queue length at the end of overloading, can grow no faster than $1/\sqrt{\epsilon}$. Note that at the end of overloading $\lambda(t) - \mu(t) < 0$. Hence the leading asymptotic term for the end of overloading grows linearly in T, for sufficiently negative T, but converges to zero as $T \uparrow +\infty$.

4. Examples.

4.1. The $M/M/1$ queue. For the time-homogeneous $M/M/1$ queue, $\lambda(t) \equiv \lambda$ and $\mu(t) \equiv \mu$. Our formulas then reduce to

$$Q^{(0)}(t) = (\lambda - \mu)t - \min_{0 \leq s \leq t} (\lambda - \mu)s$$

and

$$Q^{(1)}(t) = W((\lambda + \mu)t) - \min_{s \in \Phi_t} W((\lambda + \mu)s),$$

where

$$\Phi_t = \{0 \leq s \leq t | (\lambda - \mu) \cdot (t - s) = Q^{(0)}(t)\}.$$

For all $t > 0$, our theorems now reduce to:

$$Q^\epsilon(t) = \frac{1}{\epsilon}(\lambda - \mu)^+ t + o\left(\frac{1}{\epsilon}\right) \quad \text{a.s.}$$

and

(4.1)

$$Q^\epsilon(t) \stackrel{d}{=} \begin{cases} \dfrac{1}{\epsilon}(\lambda - \mu)t + \dfrac{1}{\sqrt{\epsilon}} W((\lambda + \mu)t) + o\left(\dfrac{1}{\sqrt{\epsilon}}\right) & \text{if } \lambda > \mu, \\ \dfrac{1}{\sqrt{\epsilon}} \left[W((\lambda + \mu)t) - \min\limits_{0 \leq s \leq t} W((\lambda + \mu)s) \right] + o\left(\dfrac{1}{\sqrt{\epsilon}}\right) & \text{if } \lambda = \mu, \\ \tilde{Q} + o(1) & \text{if } \lambda < \mu, \end{cases}$$

where $\tilde{Q}$ is a geometrically distributed random variable with parameter $\rho = \lambda/\mu$. Notice that (4.1) summarizes the asymptotic results for Theorems 4.6.14 and 4.6.16 of

Prabhu (1980). When $\lambda = \mu$, the limit $Q^{(1)}$ of $\sqrt{\epsilon}Q^{\epsilon}$ is the well-studied, reflected Brownian motion (RBM). Time-homogeneous models can, in fact, be analyzed in a considerably wider scope. For example, the above results for $\lambda \geqslant \mu$, are a special case of the heavy traffic limit theorems for multiple channel queues in Iglehart and Whitt (1970, Theorems 2.1 and 2.2). Also, Chen and Mandelbaum (1991b), following Reiman (1984), generalize most of the above to nonparametric Jackson networks. There λ, the effective arrival rate to a node, must first be calculated by appropriate traffic equations. A node is called nonbottleneck if $\lambda < \mu$, balanced bottleneck if $\lambda = \mu$ and strict bottleneck if $\lambda > \mu$. The asymptotic expansions (4.1) are established only for bottlenecks.

4.2. A specific $M_t/M_t/1$ queue. The present example illustrates the relationship between the arrival rates λ, the service rates μ, the evolution of the set-valued function Φ_t, the fluid model $Q^{(0)}$, and the diffusion approximation $Q^{(1)}$. In Figure 4.1, the middle graph is that of λ and μ, where μ is assumed to be constant. The graph immediately above it is that of a and its upper envelope $\bar{x}$, with

$$a(t) = \int_0^t [\mu(s) - \lambda(s)]\, ds.$$

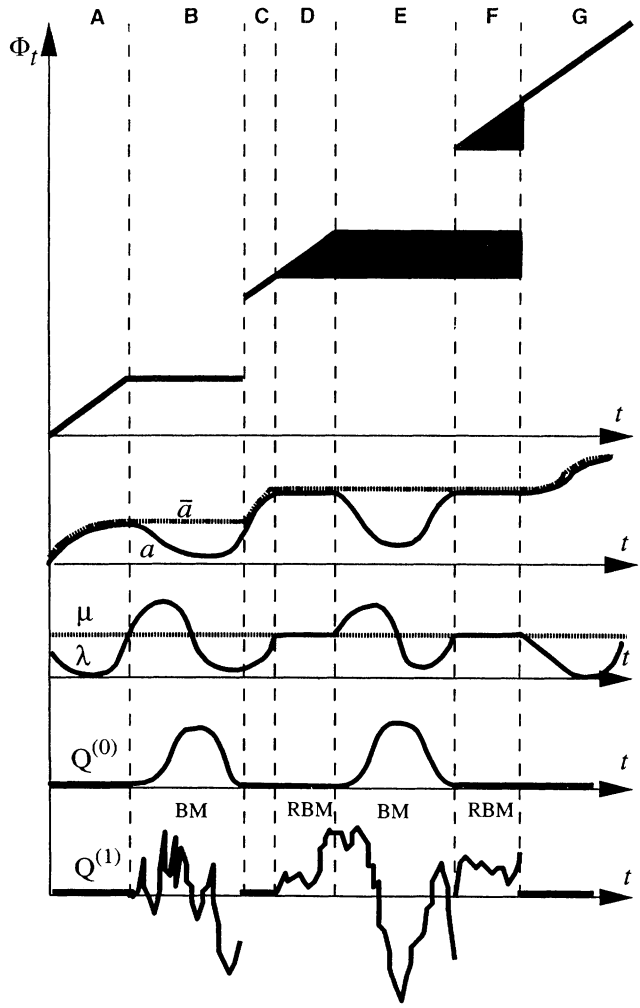

FIGURE 4.1. Graphs for Φ_t, a and $\bar{a}$, λ and μ, $Q^{(0)}$, and a realization of $Q^{(1)}$.

The fluid model $Q^{(0)}(t)$, plotted on the graph that is second from the bottom, is equal to $\bar{a}(t) - a(t)$, which is the vertical gap between the curves $\bar{a}$ and a. The topmost graph is a plot of t versus Φ_t, where Φ_t is represented as a subset of the vertical line that passes through t. Notice that the set of points in Φ_t is held fixed when t passes through time of overloading, and new time points are added to it during times of critical loading, but all points are lost except for the current time points during times of underloading. Finally, the bottom most graph is a realization of the diffusion term $Q^{(1)}$. Notice that it evolves, depending on the regime, as either zero, Brownian motion or reflected Brownian motion. If the time interval regions in Figure 4.1 are indicated by the letters A through G, and the boundary point between regions is denoted as AB for example, then we can characterize the boundary points and regions as follows:

- Underloaded: A, C
- Onset of Critical Loading: AB, CD
- Middle of Critical Loading: D, F, DE
- End of Critical Loading: FG
- Overloaded: B, E
- End of Overloading: BC, EF

When interpreting Figure 4.1, readers may find it useful to consult Proposition 7.2, which characterizes the regions in terms of either $\rho^*(t)$ or both Φ_t and $\rho(t)$.

4.3. A simple periodic queue, and Lindley's equations. Let $\lambda(t) = \lambda(0) + A\sin(2\pi t/T)$, $t \geqslant 0$, for some period $T > 0$ and amplitude $A > 0$. Take $\mu(t) \equiv \mu$ to be constant. Assume that $\lambda(0) \geqslant A$, to maintain λ nonnegative. Then the queue is permanently underloaded when $\mu > \lambda(0) + A$, and overloaded after a finite time when $\mu < \lambda(0)$. For the case $\lambda(0) \leqslant \mu \leqslant \lambda(0) + A$, the queue alternates periodically between over- and underloading. The case $\lambda(0) = \mu$ is particularly simple and pleasant to deal with. The fluid model $Q^{(0)}$ is then positive at all times, except for times $t_n = nT$, $n = 1, 2, \ldots$, which are end-of-periods for λ. Accordingly, the queue is always overloaded, except at the $\{t_n\}$'s, which are end-of-overloading epochs. Furthermore,

$$Q^{(1)}(t) = B(t) - \min_{n \leqslant t/T} B(t_n), \qquad t \geqslant 0,$$

where $B(t) = W[2\mu t + (AT/2\pi)(1 - \cos(2\pi t/T))]$, $t \geqslant 0$. In recursive form,

$$Q^{(1)}(t) = \left[B(t) - B(t_n)\right] + Q^{(1)}(t_n), \qquad t \in [t_n, t_{n+1}),$$

and

$$Q^{(1)}(t_{n+1}) = \left[Q^{(1)}(t_{n+1} -)\right]^+ = \left[B(t_{n+1}) - B(t_n) + Q^{(1)}(t_n)\right]^+,$$

for $n \geqslant 0$, starting with $Q^{(1)}(0) = 0$. We recognize this last recursion as Lindley's equations with the i.i.d. increments

$$\xi_n = B(t_{n+1}) - B(t_n) \stackrel{\mathrm{d}}{=} N(0, 2T).$$

It follows that

$$Q^{(1)}(t_n) \stackrel{\mathrm{d}}{=} \max\{0, \xi_1, \xi_1 + \xi_2, \ldots, \xi_1 + \cdots + \xi_n\}, \qquad n \geqslant 0,$$

for all $n \geqslant 0$; hence

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} Q^{(1)}(nT) \stackrel{\mathrm{d}}{=} \sqrt{2\mu T} \max_{0\leqslant t\leqslant 1} W(t).$$

4.4. The general periodic case and long-range planning. Consider λ and μ in (2.1) and (2.2) which are periodic. Suppose that λ has period-length T; denote by

$$\bar{\lambda} = \frac{1}{T}\int_0^T \lambda(s)\, ds,$$

the average inflow per period, and define $\bar{\mu}$ similarly. We accelerate the time evolution by a factor of $n = 1, 2, \ldots$. Formally, this entails looking at $Q^n(t) \equiv Q(nt)$, $t \geqslant 0$, constructed from $X^n(t) = X(nt)$. Letting $n \uparrow \infty$ yields the $M/M/1$-like asymptotic expansions

$$Q^n(t) = n(\bar{\lambda} - \bar{\mu})^+ t + o(n) \quad \text{a.s.}$$

and

$$Q^n(t) \stackrel{\mathrm{d}}{=} \begin{cases} n(\bar{\lambda} - \bar{\mu})t + \sqrt{n}\, W\big((\bar{\lambda} + \bar{\mu})t\big) + o(\sqrt{n}) & \text{if } \bar{\lambda} > \bar{\mu}, \\ \sqrt{n}\left[W\big((\bar{\lambda} + \bar{\mu})t\big) - \min\limits_{0\leqslant s\leqslant t} W\big((\bar{\lambda} + \bar{\mu})s\big)\right] + o(\sqrt{n}) & \text{if } \bar{\lambda} = \bar{\mu}. \end{cases}$$

In real-world terms, these latter asymptotic expansions can be thought of as being appropriate for strategic planning, at the corporate level of decision making, where long-range goals are formulated. The time-horizon is then typically measured in quarters or years; hence details must be suppressed. In contrast, our (2.7) expansion might be suitable for operational/regulatory control, at the shop-floor level of decision making. In this context of decision making, details count, since they must be responded to within a time-horizon of weeks or days, perhaps even hours or seconds.

5. Proof of the main theorems.

Proof of Theorems 2.1–2.3. If N^+ is a unit rate Poisson process, a standard Brownian motion W^+ can be realized with it on a probability space $(\Omega, \mathcal{F}, P)$, such that

$$\kappa^+(\omega) \equiv \sup_{t\geqslant 0} \frac{|N^+(t,\omega) - t - W^+(t,\omega)|}{\log(2 \vee t)} < \infty \quad \text{a.s.}$$

THEOREM 5.1 (STRONG APPROXIMATIONS FOR LÉVY PROCESSES). *Let M be a Lévy process with $E[\exp M(1)] < \infty$. Then M can be realized on a probability space such that*

$$\sup_{t\geqslant 0} \frac{|M(t,\omega) - \mu t - W(\sigma^2 t, \omega)|}{\log(2 \vee t)} < \infty \quad a.s.,$$

where $\mu = E[M(1)]$, $\sigma^2 = \mathrm{Var}[M(1)]$, and W is a standard Brownian motion. Consequently,

$$M(t,\omega) = \mu t + W(\sigma^2 t, \omega) + O(\log t) \quad a.s.$$

for large t.

PROOF. See Corollary 5.5 of Chapter 7 in Ethier and Kurtz (1986). □

Now let N^- be another unit rate Poisson process independent of N^+ with corresponding standard Brownian motion W^-, which is also independent of W^+. Just like κ^+, we construct κ^- out of N^- and W^-. For all $\epsilon > 0$, we have

$$\left| N^+\left(\frac{1}{\epsilon}\int_0^s \lambda(r)\,dr\right) - \frac{1}{\epsilon}\int_0^s \lambda(r)\,dr - W^+\left(\frac{1}{\epsilon}\int_0^s \lambda(r)\,dr\right)\right|$$

$$\leqslant \kappa^+ \cdot \log\left(2 \vee \frac{1}{\epsilon}\int_0^t \lambda(r)\,dr\right) \quad \text{a.s.}$$

for all $0 \leqslant s \leqslant t$. We have a similar pathwise inequality for $N^-((1/\epsilon)\int_0^s \mu(r)\,dr)$, and combining these results gives us

$$\text{(5.1)} \quad \left| X^\epsilon(s) - \frac{1}{\epsilon}\int_0^s [\lambda(r) - \mu(r)]\,dr - W^+\left(\frac{1}{\epsilon}\int_0^s \lambda(r)\,dr\right) + W^-\left(\frac{1}{\epsilon}\int_0^s \mu(r)\,dr\right)\right|$$

$$\leqslant \kappa \cdot \left[\log\left(2 \vee \frac{1}{\epsilon}\int_0^t \lambda(r)\,dr\right) + \log\left(2 \vee \frac{1}{\epsilon}\int_0^t \mu(r)\,dr\right)\right] \quad \text{a.s.}$$

for all $0 \leqslant s \leqslant t$, where $\kappa = \max(\kappa^+, \kappa^-)$.

From this bound, the functional strong law of large numbers limit for Q^ϵ (Theorem 2.1) follows since we can rewrite Q^ϵ as

$$Q^\epsilon(t) = \sup_{0 \leqslant s \leqslant t} [X^\epsilon(t) - X^\epsilon(s)],$$

and combine this with the following results: First, if a and b are bounded functions on $[0, t]$ then

$$\text{(5.2)} \quad \left| \sup_{0 \leqslant s \leqslant t} a(s) - \sup_{0 \leqslant s \leqslant t} b(s)\right| \leqslant \sup_{0 \leqslant s \leqslant t} |a(s) - b(s)|.$$

Second, we have

$$\lim_{\epsilon \downarrow 0} \epsilon W\left(\frac{1}{\epsilon}t\right) = 0 \quad \text{a.s.}$$

Finally, we get from (5.1),

$$\lim_{\epsilon \downarrow 0} \epsilon X^\epsilon(s) = \int_0^s [\lambda(r) - \mu(r)]\,dr \quad \text{a.s.}$$

which establishes the FSLLN (Theorem 2.1).

Theorem 2.3 follows from (5.2) and (5.1), since

$$\sup_{0 \leqslant t \leqslant T} \left| Q^\epsilon(t) - \max_{0 \leqslant s \leqslant t} \left[\tilde{X}^\epsilon(t) - \tilde{X}^\epsilon(s)\right]\right| \leqslant 2\kappa \cdot (|\log \epsilon| + \sigma(T)) \quad \text{a.s.}$$

where $\tilde{X}^\epsilon$ is defined by (2.8), and

$$\sigma(T) = \log 4 + \left|\log\left(\int_0^T \lambda(r)\,dr\right)\right| + \left|\log\left(\int_0^T \mu(r)\,dr\right)\right|.$$

Finally, for the FCLT (Theorem 2.2), we observe that for all $\epsilon > 0$,

$$\tilde{X}^{\epsilon}(s) \stackrel{\mathrm{d}}{=} \frac{1}{\epsilon}\int_0^t [\lambda(r) - \mu(r)]\,dr + \frac{1}{\sqrt{\epsilon}} W\left(\int_0^t [\lambda(r) + \mu(r)]\,dr\right), \qquad t \geqslant 0.$$

Here W is a standard Brownian motion, and equality holds between distributions of stochastic processes (equivalently, between all their finite-dimensional distributions). It then follows that

(5.3)

$$\max_{0\leqslant s\leqslant t}\left[\tilde{X}^{\epsilon}(t) - \tilde{X}^{\epsilon}(s)\right] \stackrel{\mathrm{d}}{=} \max_{0\leqslant s\leqslant t}\left[\frac{1}{\epsilon}\int_s^t [\lambda(r) - \mu(r)]\,dr + \frac{1}{\sqrt{\epsilon}}\left(W\left(\int_0^t [\lambda(r) + \mu(r)]\,dr\right) - W\left(\int_0^s [\lambda(r) + \mu(r)]\,dr\right)\right)\right].$$

Our main asymptotic expansion is an immediate consequence of the fundamental lemma below, and the sample path continuity of W.

LEMMA 5.2 (FUNDAMENTAL LEMMA). *Let x and y be real-valued continuous functions on $[0,\infty)$. Then with respect to Skorohod's M_1-topology,*

(5.4)

$$\max_{0\leqslant s\leqslant t}\left(\frac{1}{\epsilon}x(s) + \frac{1}{\sqrt{\epsilon}}y(s)\right) = \frac{1}{\epsilon}\max_{0\leqslant s\leqslant t} x(s) + \frac{1}{\sqrt{\epsilon}}\max_{s\in\Phi_t} y(s) + o\left(\frac{1}{\sqrt{\epsilon}}\right), \quad \text{as } \epsilon\downarrow 0,$$

where

$$\Phi_t \equiv \left\{0 \leqslant s \leqslant t \mid x(s) = \max_{0\leqslant u\leqslant t} x(u)\right\}. \tag{5.5}$$

Here we assume that $\tilde{y}$, given by

$$\tilde{y}(t) = \max_{s\in\Phi_t} y(s), \tag{5.6}$$

has a finite number of discontinuities in every finite subinterval of $[0,\infty)$.

REMARKS. (1) The relation (5.4) amounts to

$$\lim_{\epsilon\downarrow 0}\sqrt{\epsilon}\left[\left(\overline{\frac{1}{\epsilon}x + \frac{1}{\sqrt{\epsilon}}y}\right) - \frac{1}{\epsilon}\bar{x}\right] = \tilde{y}, \quad \text{in } M_1.$$

It is equivalent to

$$\lim_{\epsilon\downarrow 0}\frac{1}{\epsilon}\left(\overline{x + \epsilon y} - \bar{x}\right) = \tilde{y}, \tag{5.7}$$

as well as

$$\lim_{n\uparrow\infty}\left(\overline{nx + y} - n\bar{x}\right) = \tilde{y}, \tag{5.8}$$

both in M_1, and it implies

$$\lim_{\epsilon \downarrow 0} \epsilon \left(\overline{\frac{1}{\epsilon} x + \frac{1}{\sqrt{\epsilon}} y} \right) = \bar{x}, \qquad \text{u.o.c.}$$

(2) The representation (5.7) is that of a directional derivative of the function $x \to \bar{x}$, at the point x, in the direction y. The representation (5.8) is the one we find most convenient to actually prove, and we do it in the next section.

(3) The restriction on the discontinuities of $\tilde{y}$ stems from our method of proof. We have no reason to suspect that Lemma 5.2 fails to hold for arbitrary *continuous* x and y. (Some form of continuity seems to be needed. For example, with

$$x(t) = \begin{cases} t, & 0 \leqslant t \leqslant 1, \\ 2 - t, & t \geqslant 1, \end{cases} \qquad y(t) = \begin{cases} 1, & 0 \leqslant t < 1, \\ 0, & t \geqslant 1, \end{cases}.$$

the left-hand side of (5.8) is identically 1 while $\tilde{y}(t) = 0$ for $t \geqslant 1$.)

To apply Lemma 5.2 in our situation, note first that each sample path on the right-hand side of (5.3) is of the form

$$\frac{1}{\sqrt{\epsilon}} \left[\varphi(x + \sqrt{\epsilon}\, y) - \varphi(x) \right], \tag{5.9}$$

with

$$\varphi(x) = x - \underline{x},$$

and

$$x(t) = \int_0^t [\lambda(r) - \mu(r)]\, dr, \qquad y(t) = W\left(\int_0^t [\lambda(r) + \mu(r)]\, dr \right).$$

Letting $\epsilon \downarrow 0$ in (5.9), and relying on the continuity of y, gives

$$\lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} [\varphi(x + \epsilon y) - \rho(x)] = \lim_{\epsilon \downarrow 0} \left[y - \frac{1}{\epsilon} \underline{(x + \epsilon y - \underline{x})} \right]$$

$$= y - \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \underline{(x + \epsilon y - \underline{x})} = y - y_*,$$

where

$$y_*(t) = \min_{s \in \Phi_t} y(s),$$

and

$$\Phi_t = \{0 \leqslant s \leqslant t | a(t) - a(s) = a(t) - \underline{a}(t)\}$$

$$= \left\{ 0 \leqslant s \leqslant t | \int_s^t [\lambda(r) - \mu(r)]\, dr = Q^{(0)}(t) \right\}$$

$$= \{0 \leqslant s \leqslant t | Q^{(0)}(s) = 0 \quad \text{and} \quad Y^{(0)}(s) = Y^{(0)}(t)\}.$$

This concludes the proofs of Theorems 2.1–2.3. □

6. Proof of the fundamental lemma. Our goal in the present section is to establish the monotone convergence

$$\tilde{y}_n \equiv \overline{nx + y} - n\bar{x} \downarrow \tilde{y}, \quad \text{in } M_1, \tag{6.1}$$

as $n \uparrow \infty$ and under the assumptions of Lemma 5.2. We start with monotone pointwise convergence, then proceed with a careful path-analysis of $\tilde{y}_n$ and $\tilde{y}$ that culminates in M_1-convergence.

LEMMA 6.1 (MONOTONICITY). $\overline{(n+1)x + y} - (n+1)\bar{x} \leqslant \overline{nx + y} - n\bar{x}$.

PROOF. Equivalently

$$\overline{(n+1)x + y} - \overline{nx + y} \leqslant \bar{x},$$

which is obtained by substituting $f = (n+1)x + y$ and $g = nx + y$ into

$$\bar{f} - \bar{g} \leqslant \overline{f - g} \tag{6.2}$$

and we are done. □

LEMMA 6.2 (POINTWISE CONVERGENCE). *Let $s_n(t)$ be a point in $[0, t]$ where $nx + y$ attains its maximum over that interval. Then, as $n \uparrow \infty$,*

$$y[s_n(t)] \to \tilde{y}(t), \qquad n\{x[s_n(t)] - \bar{x}(t)\} \to 0, \quad \textit{for all } t \geqslant 0, \tag{6.3}$$

hence the convergence (6.1) *holds pointwise.*

PROOF. We have

$$\begin{aligned} y[s_n(t)] &\geqslant y[s_n(t)] + n\{x[s_n(t)] - \bar{x}(t)\}, \\ &\geqslant y(s) + n[x(s) - \bar{x}(t)], && \text{for } 0 \leqslant s \leqslant t, \\ &= y(s), && \text{for } s \in \Phi_t, \end{aligned} \tag{6.4}$$

implying that

$$\varliminf_{n \to \infty} y[s_n(t)] \geqslant \tilde{y}(t), \qquad t \geqslant 0. \tag{6.5}$$

We now verify

$$\varlimsup_{n \to \infty} y[s_n(t)] \leqslant \tilde{y}(t), \qquad t \geqslant 0, \tag{6.6}$$

which, combined with (6.5) and (6.4), will end the proof. First, observe that the limit $s_\infty(t)$ of any convergent subsequence $s_{n'}(t) \to s_\infty(t)$ must belong to Φ_t. (Otherwise $n\{x[s_{n'}(t)] - \bar{x}(t)\} \to -\infty$, contradicting (6.4).) Next, take a subsequence of the bounded sequence $\{y[s_n(t)]\}$ that converges to $\varlimsup_{n \to \infty} y[s_n(t)]$. Without loss of generality, $s_n(t) \to s_\infty(t) \in \Phi_t$; hence

$$\varlimsup_{n \to \infty} y[s_n(t)] = y[s_\infty(t)] \leqslant \tilde{y}(t), \tag{6.7}$$

and this completes the proof. □

LEMMA 6.3 (UPPER SEMICONTINUOUS LIMIT. *The function $\tilde{y}$ is upper semicontinuous and the convergence in* (6.1) *is uniform over compact subsets of continuity points for $\tilde{y}$.*

PROOF. Monotone decreasing limits of continuous functions must be upper semicontinuous. Monotone convergence of continuous functions to a continuous limit must be uniform on compacts, by Dini's theorem. □

For $x \in D$ let $I_{s,t}^{\alpha,\beta}(x)$ equal the number of intersections of x with the strip $[\alpha, \beta]$ during the time interval $[s, t]$. This number equals N if it is possible to find $N+1$ points $t_0 < t_1 < \cdots < t_N$ in $[s, t]$ with the property that either

$$x(t_0) \leqslant \alpha, \qquad x(t_1) \geqslant \beta, \qquad x(t_2) \leqslant \alpha, \ldots, \tag{6.8}$$

or

$$x(t_0) \geqslant \beta, \qquad x(t_1) \leqslant \alpha, \qquad x(t_2) \geqslant \beta, \ldots, \tag{6.9}$$

and it is impossible to find $N+2$ points with this property. The number of intersections is arbitrarily taken to be -1 if no such N exists (which occurs when x stays within (α, β) during $[s, t]$).

LEMMA 6.4 (SKOROHOD 1956, §2.2.11, p. 267). *Given x_n, $x \in D$, then $x_n \to x$ in M_1 if and only if*

$$\lim_{n \uparrow \infty} I_{s,t}^{\alpha,\beta}(x_n) = I_{s,t}^{\alpha,\beta}(x), \qquad n \uparrow \infty, \tag{6.10}$$

for all $0 \leqslant s \leqslant t$ which are points of continuity of the limit x, and for almost all $\alpha < \beta \in \mathbb{R}^1$.

We have already verified that the convergence in §6.1 is pointwise monotone convergence of continuous functions. But this need not imply M_1-convergence. Indeed, simple counterexamples can be constructed with $x(t) = 1_{[1,\infty)}(t)$, and $x_n \downarrow x$ in a way that $I_{1/2,3/2}^{1/4,1/2}(x_n) \geqslant 3$ while $I_{1/2,3/2}^{1/4,1/2}(x) = 1$. However, Lemma 6.3 guarantees (6.10) for $[s, t]$ over which $\tilde{y}$ is continuous. Furthermore, the assumed isolated discontinuities of $\tilde{y}$ allows one to verify (6.10) only for $[s, t]$ in which $\tilde{y}$ has a *single* point of discontinuity, and this will now be carried out.

LEMMA 6.5 (CONTINUITY CONDITIONS). *Given a continuous function x, the function $\tilde{y}$ is continuous at $t \geqslant 0$, for all continuous functions y, if and only if one of the following mutually exclusive statements holds*:

(i) $t \notin \Phi_t$.

(ii) $\Phi_t = \{t\}$.

(iii) $t \in \Phi_t \neq \{t\}$, *t is not isolated in Φ_t, and $\Phi_t \subseteq \Phi_r$ for some $r > t$.*

These conditions can be recast in terms of x as:

(i) $x(t) < \bar{x}(t)$.

(ii) $x(l) < x(t) = \bar{x}(t)$, *for all $l \in [0, t)$.*

(iii) $x(l) = x(t) = \bar{x}(t)$ *for some $l < t$, $x(l_n) = \bar{x}(t)$ for some sequence $l_n \uparrow t$, and $\bar{x}(t) = \bar{x}(r)$ for some $r > t$.*

LEMMA 6.6 (LEFT DISCONTINUITY CONDITIONS). *The following three statements are equivalent*:

(i) *The function $\tilde{y}$ is left discontinuous at $t > 0$, for some continuous function y.*

(ii) $t \in \Phi_t \neq \{t\}$, *and the point t is isolated in Φ_t.*

(iii) *For some $s \in [0, t)$ and all $l \in (s, t)$,*

$$x(s) = x(t) = \bar{x}(t) \quad \text{and} \quad x(l) < \bar{x}(t), \tag{6.11}$$

in which case the discontinuity is of the form

$$\tilde{y}(l) = \tilde{y}(t-) < \tilde{y}(t+) = y(t). \tag{6.12}$$

LEMMA 6.7 (RIGHT DISCONTINUITY CONDITIONS). *The following three statements are equivalent*:

(i) *The function $\tilde{y}$ is right discontinuous at $t > 0$, for some continuous function y.*

(ii) *$t \in \Phi_t \neq \{t\}$, and $\Phi_r \subseteq (t, r]$ for all $r > t$.*

(iii) *For some $0 \leqslant l < t$ and all $r > t$,*

$$x(l) = x(t) = \bar{x}(t) < \bar{x}(r) \tag{6.13}$$

in which case, the discontinuity is of the form

$$\tilde{y}(t-) = \tilde{y}(t) > \tilde{y}(t+) = y(t). \tag{6.14}$$

Proofs of Lemmas 6.5, 6.6 and 6.7. If $t \notin \Phi_t$, then $x(t) < \bar{x}(t)$. Since x is continuous, then x is strictly less than $\bar{x}$ on an open neighborhood of t. It follows that the set Φ_s equals Φ_t for all s in this open neighborhood. Therefore $\tilde{y}$ is constant on this open set, and so $\tilde{y}$ is continuous at t.

If $t \in \Phi_t$, then $x(t) = \bar{x}(t)$. Now suppose that $\Phi_t = \{t\}$. For all $s \geqslant 0$, Φ_s is a compact subset of $[0, s]$. For some continuous function y, there exists a $u_s \in \Phi_s$ such that $y(u_s) = \tilde{y}(s)$. To prove that $\tilde{y}$ is continuous at t, it is sufficient to show that $\lim_{s \to t} u_s = t$. Let $\{s_n | n \geqslant 1\}$ be a sequence such that $\lim_{n\to\infty} s_n = t$ and $\lim_{n\to\infty} u_{s_n} = v$, for some v. By compactness, we can always find such a sequence. Since $0 \leqslant u_{s_n} \leqslant s_n$ for all $n \geqslant 1$, we have $0 \leqslant v \leqslant t$. On the other hand,

$$x(v) = \lim_{n\to\infty} x(u_{s_n}) = \lim_{n\to\infty} \bar{x}(s_n) = \bar{x}(t).$$

Consequently, $v \in \Phi_t$ and so $v = t$. Hence in general, the limit exists and must equal t.

If $t \in \Phi_t \neq \{t\}$, and t is *not* an isolated point in Φ_t, then there exists a strictly increasing sequence $\{l_n | n \geqslant 1\}$ where $\lim_{n\to\infty} l_n = t$ such that $l_n \in \Phi_t$ for all $n \geqslant 1$. Moreover, $\Phi_{l_n} = \Phi_t \cap [0, l_n]$ since $\bar{x}$ is an increasing function, and $x(l_n) = \bar{x}(t)$ by definition. The sequence of sets Φ_{l_n} form an ascending chain, so $\tilde{y}(l_n)$ is a monotone increasing sequence with $\tilde{y}(l_n) \leqslant \tilde{y}(t)$. Hence $\lim_{n\to\infty} \tilde{y}(l_n)$ exists, and

$$\lim_{n\to\infty} \tilde{y}(l_n) \geqslant y(l_n)$$

for all n since $l_n \in \Phi_{l_n}$. But

$$\tilde{y}(t) = \tilde{y}(l_n) \vee \sup_{s \in \Phi_t \cap (l_n, t]} y(s) = \lim_{n\to\infty} \tilde{y}(l_n) \vee y(t) = \lim_{n\to\infty} \tilde{y}(l_n).$$

So for any l arbitrarily close to t from below, we can find some $l_n \leqslant l$ and consequently

$$\tilde{y}(l_n) \leqslant \tilde{y}(l) \leqslant \tilde{y}(t).$$

This shows that $\tilde{y}$ is left-continuous at t.

Conversely, if $t \in \Phi_t \neq \{t\}$, and t is isolated, let $\tau \equiv \sup(\Phi_t - \{t\}) < t$. This gives us $\Phi_r = \Phi_\tau$ for all $\tau \leqslant r < t$, and $\Phi_t = \Phi_\tau \cup \{t\}$. We can find a continuous function y

such that $y(t)$ is greater than any value of y on Φ_τ. So for all continuous functions y, $\tilde{y}$ will not necessarily be left-continuous at t, but

$$\tilde{y}(t-) = \lim_{l \uparrow t} \tilde{y}(l) = \tilde{y}(\tau) \leqslant \tilde{y}(t). \tag{6.15}$$

If $t \in \Phi_t \neq \{t\}$, and there exists some $r > t$ such that $\Phi_t \subseteq \Phi_r$, it follows that $\Phi_s \subseteq \Phi_t$ for all $t \leqslant s \leqslant r$. From this we get $\Phi_t = \Phi_r \cap [0, t]$ and

$$\tilde{y}(t) \leqslant \tilde{y}(r) \leqslant \tilde{y}(t) \vee \max_{t \leqslant s \leqslant r} y(s),$$

which gives us right-continuity for $\tilde{y}$ at t.

Conversely, if for all $r > t$, we have that Φ_t is not a subset of Φ_r, then $x(t) < \max_{t \leqslant s \leqslant r} x(s)$. This means that Φ_r is a subset of $(t, r]$ and so

$$\tilde{y}(t+) = \lim_{r \downarrow t} \tilde{y}(r) = y(t) \leqslant \tilde{y}(t). \tag{6.16}$$

Since $\Phi_t \neq \{t\}$, there exists some continuous function y such that $y(t) \neq \tilde{y}(t)$; hence $\tilde{y}$ is not necessarily right-continuous at t.

Finally, every statement made here in terms of Φ_t can be restated in terms of x. □

LEMMA 6.8. *The function $\tilde{y}$ has left limits at all $t > 0$, right limits at all $t \geqslant 0$, $\tilde{y}(0) = x(0)$ and*

$$\tilde{y}(t) = \tilde{y}(t-) \vee \tilde{y}(t+), \qquad t > 0. \tag{6.17}$$

Equivalently, $\tilde{y}$ is an upper semicontinuous function in D.

PROOF. At $t > 0$, $\tilde{y}$ is either left-continuous or $\tilde{y}$ is flat to the left of t. In any case, $\tilde{y}(t-)$ exists. At $t \geqslant 0$, either $\tilde{y}$ is right-continuous or (6.13) prevails. When the latter applies, let $r_n \downarrow t$. Then $\tilde{y}(r_n) = y(s_n)$ for some $s_n \in \Phi_{r_n} \cap (t, r_n]$. Consequently, $s_n \to t$, which implies $\tilde{y}(r_n) = y(s_n) \to y(t+)$. Again, $\tilde{y}(t+)$ exists in all cases. In fact, a review of the various alternatives analyzed in Lemmas 6.6 and 6.7 reveals that, for all $t > 0$, either $\tilde{y}(t) = \tilde{y}(t+)$ or $\tilde{y}(t) = \tilde{y}(t-)$. Upper semicontinuity now yields (6.17). □

LEMMA 6.9. *The function $\tilde{y}$ is flat to the left of its discontinuities.*

PROOF. Fix $t > 0$, a point of discontinuity for $\tilde{y}$. The assertion is clear if $t \in \Phi_t$ is isolated (left-discontinuity). Otherwise (right-discontinuity), let $l_n \in \Phi_t$ be such that $l_n \uparrow t$. Then

$$\tilde{y}(l_n) \uparrow \tilde{y}(t-) = \tilde{y}(t) > \tilde{y}(t+) = y(t).$$

Now $\tilde{y}(l_n) = y(u_n)$ for some $u_n \in \Phi_{l_n} \subseteq \Phi_t$. Let u be a cluster point of $\{u_n\}$. Then $u < t$ since $y(u)$, being a cluster point of $\{\tilde{x}(l_n)\}$, satisfies $y(u) = \tilde{y}(t) \geqslant y(t)$. Finally, $u \in \Phi_t$; hence $u \in \Phi_u$ and $\Phi_\cdot$ is monotone nondecreasing over $[u, t]$. One deduces that $\tilde{y}(u) \geqslant y(u) = \tilde{y}(t) \geqslant \tilde{y}(u)$, revealing a point $u < t$ for which $\tilde{y}(u) = \tilde{y}(t)$. Since $\tilde{y}$ is monotone over $[u, t]$, we are done. □

Intermediate summary. We have shown that
(i) $\tilde{y}(t) = \tilde{y}(t-) \vee \tilde{y}(t+)$ at all $t > 0$; $\tilde{y}(0) = y(0)$.
(ii) $\tilde{y}$ is flat to the left of its jumps.

Consider a complete *excursion* of x from $\bar{x}$ which starts at s and ends at t. By this we mean that s is a point of left-increase for $\bar{x}$ ($\bar{x}(l) < \bar{x}(s)$ for all $0 \leqslant l < s$, denoted by $\bar{x} \uparrow s -$), and that t is a point of right-increase for $\bar{x}$ ($\bar{x}(t) < \bar{x}(r)$ for all $r > t$, denoted by $\bar{x} \uparrow t +$). For such an excursion,

(i) $\tilde{y}(s) = y(s)$; s is a point of continuity for $\tilde{y}$;

(ii) $\tilde{y}(\cdot)$ is monotone nondecreasing over $[s, t]$, possibly with jumps.

(iii) Finally,

$$y(t) > \tilde{y}(t-) \Rightarrow \tilde{y}(t-) < \tilde{y}(t) = \tilde{y}(t+) = y(t), \quad \text{left-discontinuity at } t;$$

$$y(t) < \tilde{y}(t-) \Rightarrow \tilde{y}(t-) = \tilde{y}(t) > \tilde{y}(t+) = y(t), \quad \text{right-discontinuity at } t;$$

$$y(t) = \tilde{y}(t-) \Rightarrow \tilde{y}(t-) = \tilde{y}(t) = \tilde{y}(t+) = y(t), \quad \text{continuity at } t.$$

Before going on, readers may find it useful to draw the graph of

$$y_n(u) = n[x(u) - \bar{x}(t)] + y(u), \qquad s \leqslant u \leqslant t, \tag{6.18}$$

over an excursion $[s, t]$ as above, and then analyze the convergence $\bar{y}_n \downarrow \tilde{y}$ over $[s, t]$, as $n \uparrow \infty$. (A key observation is that $y_n(u) = y(u)$, for all $u \in \Phi_t$ and all $n \geqslant 1$. Thus, the component $n[x(u) - \bar{x}(t)]$, $s \leqslant u \leqslant t$, decomposes into negative excursions of x from $\bar{x}$, each of which is "hanging" on successive points in Φ_t.)

LEMMA 6.10 (M_1-CONVERGENCE AROUND LEFT-DISCONTINUITIES). *Let $t > 0$ be such that $\tilde{y}(t) = \tilde{y}(t+) > \tilde{y}(t-)$. Then the convergence* (6.10) *applies over a neighbourhood of t.*

PROOF. Choose $\epsilon > 0$ so that $\bar{x}$ and $\tilde{y}$ are flat over $[t - \epsilon, t]$, and $\tilde{y}$ has no discontinuities over $(t, t + \epsilon]$. For such ϵ,

(i) $\bar{y}_n$ is monotone nondecreasing over $[t - \epsilon, t]$, and

(ii) $\bar{y}_n \downarrow \tilde{y}$ uniformly over $[t, t + \epsilon]$.

This verifies (6.10) over $[t - \epsilon, t + \epsilon]$. □

LEMMA 6.11 (M_1-CONVERGENCE AROUND RIGHT-DISCONTINUITIES). *Let $t > 0$ be such that $\tilde{y}(t) = \tilde{y}(t-) > \tilde{y}(t+)$. Then the convergence* (6.10) *applies over a neighbourhood of t.*

PROOF. We are going momentarily to check that:

(i) There exists $\epsilon > 0$ for which the $\bar{y}_n$'s are all flat over $[t - \epsilon, t]$; and

(ii) For all n large enough, either $\bar{y}_n$ is monotone decreasing over $[t, \infty)$ or there exists a sequence $r_n \downarrow t$ for which $d\bar{y}_n = -n\, d\bar{x}$ over $[t, r_n)$, $\bar{y}_n(r_n) = y(r_n)$, and

$$\sup_{t < s \leqslant r_n + \delta} \bar{y}_n(s) \leqslant \sup_{t < s \leqslant r_n + \delta} y(s), \quad \text{for all } \delta > 0. \tag{6.19}$$

It follows that the only way to violate (6.10) is to have $\alpha < \beta \in \mathbb{R}^1$ and a sequence $u_n > r_n$, $u_n \downarrow t$, such that $\bar{y}_n(r_n) \leqslant \alpha$ and $\bar{y}_n(u_n) \geqslant \beta$. This, however, gives rise to $\underline{\lim}_{s \downarrow t} y(s) \leqslant \alpha$ and $\overline{\lim}_{s \downarrow t} y(s) \geqslant \beta$, which contradicts the existence of $y(t+)$.

To verify (i), recall that $\tilde{y}$ is flat to the left of t, choose $\epsilon, \delta > 0$ so that for all n large enough,

$$n[x(s) - \bar{x}(t)] + y(s) \leqslant \tilde{y}(s) - \delta \leqslant \bar{y}_n(s) - \delta, \qquad s \in [t - \epsilon, t],$$

deduce that

$$nx + y \leqslant \overline{nx + y} - \delta, \quad \text{on } [t - \epsilon, t], \tag{6.20}$$

and conclude that $\overline{nx + y}$ is flat over $[t - \epsilon, t]$. By Lemma 6.6, $\bar{x}$ is flat to the left t, which establishes (i).

The continuity of x and y guarantees (6.20), perhaps with a smaller $\delta > 0$, also on a neighbourhood to the right of t. It follows that $d\bar{y}_n = -n\, d\bar{x}$ on that neighbourhood. If $\tilde{y}_n(\cdot)$ has points of increase beyond time t, there must exist $u > t$ for which

$$\overline{(nx + y)}(u) = nx(u) + y(u) \leqslant n\bar{x}(u) + y(u),$$

thus $\tilde{y}_n(u) \leqslant y(u)$. With $\tilde{y}_n(t) \geqslant \tilde{y}(t) > \tilde{y}(t-) = y(t)$, continuity implies the existence of $r \in [t, u]$ where $\tilde{y}_n(r) = y(r)$. Along the same lines

(iii) If there exists N for which $\tilde{y}_N(r) = y(r)$ at some $r > t$, then for each $n \geqslant N$ there exists $r_n > t$ such that $\tilde{y}_n(r_n) = y(r_n)$.

Indeed, $\tilde{y}_n(r)$ decreases in n, consequently $y(r) \geqslant \tilde{y}_n(r)$ for all $n \geqslant N$, and since $\tilde{y}_n(t) > y(t)$, there exists $u \in (t, r]$ at which $\tilde{y}_n(u) = y(u)$. Finally, let

$$r = \inf\{s > t : \overline{(nx + y)}(s) = (nx + y)(s)\} < \infty.$$

Then for all $\delta > 0$, as in Lemma 6.1,

$$\begin{aligned}\overline{(nx + y)}(r + \delta) - n\bar{x}(r + \delta) &= \sup_{t < s \leqslant r + \delta} (nx + y)(s) - \sup_{t < s \leqslant r + \delta} nx(s) \\ &\leqslant \sup_{t < s \leqslant r + \delta} y(s),\end{aligned}$$

which concludes the proof of Lemma 6.11. □

EXAMPLE. We now give an example to show that the error term $o(1/\sqrt{\epsilon})$ in Lemma 5.2 is tight. Take,

$$x_n(s) \equiv -\alpha \cdot s^{n+1} \quad \text{and} \quad y(s) \equiv -\beta \cdot (t - s)$$

where α and β are positive constants. By taking the derivative with respect to s of $(1/\epsilon)x_n(s) + (1/\sqrt{\epsilon})y(s)$, we get for sufficiently small ϵ, that this sum of functions will realize its maximum on $[0, t]$ at

$$s = \left(\frac{\beta\sqrt{\epsilon}}{\alpha \cdot (n + 1)}\right)^{1/n}.$$

This in turn gives us

$$\max_{0 \leqslant s \leqslant t} \left(\frac{1}{\epsilon} x_n(s) + \frac{1}{\sqrt{\epsilon}} y(s)\right) = -\frac{\beta t}{\sqrt{\epsilon}} + \frac{1}{\sqrt{\epsilon^{1-1/n}}} \frac{\beta n}{n + 1} \left(\frac{\beta}{\alpha \cdot (n + 1)}\right)^{1/n}.$$

For arbitrary n, we see that the $o(1/\sqrt{\epsilon})$ description for the error is tight. Notice also that $\tilde{y}(t) = -\beta t$, which is continuous in t, and the error is independent of t. This illustrates the uniform convergence of the expansion in ϵ.

7. Properties of the traffic intensity parameter. This section is devoted to establishing useful properties of ρ^*. They will be used to complete the proof of Theorem 3.1, at the end of the section.

PROPOSITION 7.1. *The function ρ^* is continuous on $[0, \infty)$.*

REMARK. For the unstable example given in Heyman and Whitt (1984), λ is a step function, μ is constant, and ρ^* is a discontinuous function of time.

PROOF OF PROPOSITION 7.1. Define a bivariate function R on the set $\{(s,t)|0 \leqslant s \leqslant t\}$ such that

$$R(s,t) = \begin{cases} \int_s^t \lambda(r)\,dr / \int_s^t \mu(r)\,dr & \text{if } s < t, \\ \rho(t) & \text{if } s \geqslant t. \end{cases} \tag{7.1}$$

Since λ and μ are continuous, then $\int_s^t \lambda(r)\,dr$ and $\int_s^t \mu(r)\,dr$ are continuously differentiable as functions of s and

$$\lim_{s \uparrow t} \frac{\int_s^t \lambda(r)\,dr}{\int_s^t \mu(r)\,dr} = \rho(t). \tag{7.2}$$

It follows that R is continuous on the closed set $\{(s,t)|0 \leqslant s \leqslant t\}$ and *uniformly* continuous on the compact set $\{(s,t)|0 \leqslant s \leqslant t \leqslant T\}$ for some fixed T.

Now $\rho^*(t) = \sup_{0 \leqslant s \leqslant t} R(s,t)$, so for all t and t' we have

$$\begin{aligned} |\rho^*(t) - \rho^*(t')| &\leqslant \left| \sup_{0 \leqslant s \leqslant t} R(s,t) - \sup_{0 \leqslant s \leqslant t'} R(s,t') \right| \\ &\leqslant \left| \sup_{0 \leqslant s \leqslant t} R(s,t) - \sup_{0 \leqslant s \leqslant t} R(s,t') \right| \\ &\quad + \left| \sup_{0 \leqslant s \leqslant t} R(s,t') - \sup_{0 \leqslant s \leqslant t} R(s,t') \right| \\ &\leqslant \sup_{0 \leqslant s \leqslant t} |R(s,t) - R(s,t')| + \left| \sup_{0 \leqslant s \leqslant t} R(s,t') - \sup_{0 \leqslant s \leqslant t} R(s,t') \right|. \end{aligned}$$

By the uniform continuity of R, the above bounds give us $\lim_{t \to t'} |\rho^*(t) - \rho^*(t')| = 0$, and this completes the proof. □

The sets Φ_t for $t \geqslant 0$, are determined by the fluid limit. In the proposition below, we use this relation to describe the three main asymptotic regions in terms of the fluid approximation:

PROPOSITION 7.2. *The following statements hold for all $t > 0$:*
(i) *We have $\rho^*(t) < 1$ if and only if $\Phi_t = \{t\}$ and $\rho(t) < 1$.*
(ii) *We have $\rho^*(t) = 1$ if and only if $t \in \Phi_t$ and either $\Phi_t \neq \{t\}$ or $\rho(t) = 1$ occurs.*
(iii) *We have $\rho^*(t) > 1$ if and only if $t \notin \Phi_t$.*

PROOF. It is sufficient to prove only the first and third statements since the second one will follow from them. Recall that

$$Q^{(0)}(t) = \sup_{0 \leqslant s \leqslant t} \int_s^t [\lambda(r) - \mu(r)]\,dr$$

and Φ_t consists of the times s such that $\int_s^t [\lambda(r) - \mu(r)]\,dr = Q^{(0)}(t)$ for all $0 \leqslant s \leqslant t$.

It follows that $\Phi_t = \{t\}$ if and only if $\int_s^t [\lambda(r) - \mu(r)]\,dr < 0$ for all $0 \leqslant s \leqslant t$. Combining this with

$$\rho(t) = \lim_{s \uparrow t} \frac{\int_s^t \lambda(r)\,dr}{\int_s^t \mu(r)\,dr},$$

gives us the first statement.

For the third statement, we observe that $t \notin \Phi_t$ if and only if $Q^{(0)}(t) > 0$. This is equivalent to having $\int_s^t [\lambda(r) - \mu(r)]\,dr > 0$ for some $0 \leqslant s \leqslant t$, which is equivalent to $\rho^*(t) > 1$. □

LEMMA 7.3. *If $t_n \uparrow t$ with $\rho^*(t_n) = 1$, then $\rho(t) = 1$.*

PROOF. Since ρ^* is continuous, we have $\rho^*(t) = 1$. By the definition of ρ^*, we either have $\rho(t) = 1$, or there exists some $0 \leqslant s^* < t$ such that

$$\int_{s^*}^t \lambda(r)\,dr = \int_{s^*}^t \mu(r)\,dr. \tag{7.3}$$

On the other hand, $\rho^*(t) = 1$ means that for all n,

$$\int_{t_n}^t \lambda(r)\,dr \leqslant \int_{t_n}^t \mu(r)\,dr. \tag{7.4}$$

For all sufficiently large n, we have $s^* \leqslant t_n \leqslant t$, and $\rho^*(t_n) = 1$ gives us

$$\int_{s^*}^{t_n} \lambda(r)\,dr \leqslant \int_{s^*}^{t_n} \mu(r)\,dr. \tag{7.5}$$

Combining (7.4) and (7.5) with (7.3), we get for all sufficiently large n,

$$\int_{t_n}^t \lambda(r)\,dr = \int_{t_n}^t \mu(r)\,dr.$$

Therefore, we have

$$\rho(t) = \lim_{n \to \infty} \frac{\int_{t_n}^t \lambda(r)\,dr}{\int_{t_n}^t \mu(r)\,dr} = 1,$$

which completes the proof. □

LEMMA 7.4. *If $\rho^* > 1$ on (t_1, t_2) with $\rho^*(t_1) = \rho^*(t_2) = 1$, then*

$$\int_{t_1}^{t_2} \lambda(r)\,dr = \int_{t_1}^{t_2} \mu(r)\,dr.$$

PROOF. By the definition of $Q^{(0)}(t)$ and Φ_t, we have

$$Q^{(0)}(t) = \int_{t^*}^t [\lambda(r) - \mu(r)]\,dr$$

where $t^* = \sup \Phi_t$. If $t_1 < t < t_2$, then $t \notin \Phi_t$, $t^* < t$, and $\Phi_{t^*} = \Phi_t$. Since $t^* \in \Phi_{t^*}$,

we must have $\rho^*(t^*) \leqslant 1$ and $t^* \leqslant t_1$. By the continuity of $Q^{(0)}$, we have

$$\int_{t_2^*}^{t_2} \lambda(r)\, dr = \int_{t_2^*}^{t_2} \mu(r)\, dr.$$

where $t_2^* \leqslant t_1$. However $\rho^*(t_1) = \rho^*(t_2) = 1$, so we must have

$$\int_{t_1}^{t_2} \lambda(r)\, dr \leqslant \int_{t_1}^{t_2} \mu(r)\, dr \quad \text{and} \quad \int_{t_2^*}^{t_1} \lambda(r)\, dr \leqslant \int_{t_2^*}^{t_1} \mu(r)\, dr.$$

Combining these two inequalities with the previous equality gives us the lemma. □

LEMMA 7.5. *If* $\rho^* \geqslant 1$ *on* $[t_1, t_2]$ *with* $\rho^*(t_1) = \rho^*(t_2) = 1$, *then*

$$\int_{t_1}^{t_2} \lambda(r)\, dr = \int_{t_1}^{t_2} \mu(r)\, dr.$$

PROOF. The set $A \equiv \{t | \rho^*(t) > 1\} \cap [t_1, t_2]$ is open. By Lindelöf's theorem, it must equal a countable union of disjoint intervals where ρ^* equals one on the endpoints. By our previous lemma, we have

$$\int_A [\lambda(r) - \mu(r)]\, dr = 0.$$

So it is sufficient for us to prove that

$$\int_{[t_1, t_2]-A} [\lambda(r) - \mu(r)]\, dr = 0. \tag{7.6}$$

This complement of A relative to $[t_1, t_2]$, is a closed subset of the times when ρ^* equals one. With the exception of a countable number of points (the endpoints of the open intervals in A), every point t in this set has a sequence within the set where $t_n \uparrow t$. By Lemma 7.3, this means that $\rho(t) = 1$ for all such t or $\lambda(t) = \mu(t)$, and the lemma follows. □

PROOF OF THEOREM 3.1. By Proposition 7.2 and Lemma 6.5, we have that $Q^{(1)}(t)$ is continuous at t whenever $\rho^*(t) \neq 1$. It now remains to be shown that each subcase for $\rho^*(t) = 1$ implies continuity or discontinuity.

If the onset of critical loading occurs, then there exists a sequence $s_n \uparrow t$ with $\rho^*(s_n) < 1$ for all n. Now $\rho^*(t) = 1$ implies $t \in \Phi_t$. If $s \in \Phi_t$ and $s < t$, then

$$\Phi_{s_n} \subseteq \Phi_{s_m} \quad \text{for all } s < s_n < s_m < t.$$

But $\rho^*(s_n) < 1$ implies $\Phi_{s_n} = \{s_n\}$, which means that $s_n = s_m$. By contradiction, we have proved that $\Phi_t = \{t\}$. By Lemma 6.5, this makes $Q^{(1)}$ continuous at t.

If $t_n \uparrow t$ where $\rho^*(t_n) = 1$ and $\rho^* \geqslant 1$ on $[t_1, t]$, then by Lemma 7.5, we have

$$\int_{t_n}^{t} [\lambda(r) - \mu(r)]\, dr = 0.$$

It follows that $t_n \in \Phi_{t_n} \subseteq \Phi_t$, and so $t_n \in \Phi_t$ for all n. As a consequence, $t \in \Phi_t \neq \{t\}$, and t is not an isolated point. It follows from Lemma 6.5, that $Q^{(1)}$ is left-continuous whenever t is in the middle or end of critical loading.

If in addition, $\rho^* \geqslant 1$ on an open interval with t as the left endpoint, then there exists some $s > t$ such that $\Phi_t \subseteq \Phi_s$ which makes $Q^{(1)}$ continuous at t if it is a time for the middle of critical loading.

If t is at the end of critical loading, then there exists as sequence such that $s_n \downarrow t$ with $\Phi_{s_n} = \{s_n\}$. Since $t \in \Phi_t$ but $\Phi_t \neq \{t\}$, then Φ_t cannot be a subset of Φ_s for all $t < s$. Therefore, $Q^{(1)}$ cannot be right-continuous at t for almost all sample paths.

Finally, if t is the end of overloading, then $\Phi_t \neq \{t\}$, but t is an isolated point of Φ_t, so $Q^{(1)}$ cannot be left-continuous at t on almost all sample paths. □

8. Underloaded or local equilibrium.

PROOF OF THEOREM 3.2. We will show by induction on N, that for all $t_1 < \cdots < t_N$, with $\rho^*(t_i) < 1$, $i = 1, \ldots, N$, we get

$$\lim_{\epsilon \downarrow 0} \mathsf{P}(Q^\epsilon(t_1) \geqslant n_1, \ldots, Q^\epsilon(t_N) \geqslant n_N) = \prod_{i=1}^{N} \rho(t_i)^{n_i}.$$

Case [$N = 1$]. There exists some $\delta > 0$ such that ρ^* is strictly less than one on the closed interval $[t_1 - \delta, t_1]$. By (2.3) and (3.3), we can rewrite $Q^\epsilon(t_1)$ as

$$Q^\epsilon(t_1) = [X^\epsilon(t_1) - X^\epsilon(t_1 - \delta) + Q^\epsilon(t_1 - \delta)] \vee Q^\epsilon(t_1 - \delta, t_1).$$

By (2.2), we have

$$\lim_{\epsilon \downarrow 0} \epsilon[X^\epsilon(t_1) - X^\epsilon(t_1 - \delta)] = \int_{t_1 - \delta}^{t_1} [\lambda(r) - \mu(r)]\, dr < 0 \quad \text{a.s.}$$

and by Theorem 2.1 we have

$$\lim_{\epsilon \downarrow 0} \epsilon Q^\epsilon(t_1 - \delta) = 0.$$

Combining these two results gives us

$$\lim_{\epsilon \downarrow 0} [X^\epsilon(t_1) - X^\epsilon(t_1 - \delta) + Q^\epsilon(t_1 - \delta)] = -\infty \quad \text{a.s.}$$

which finally leads to

$$\lim_{\epsilon \downarrow 0} [Q^\epsilon(t_1) - Q^\epsilon(t_1 - \delta, t_1)] = 0 \quad \text{a.s.} \tag{8.1}$$

So it is sufficient to show that $Q^\epsilon(t_1 - \delta, t_1)$ converges to a geometric distribution with parameter $\rho(t_1)$.

Let Q^+ be an $M/M/1$ queue length process with arrival and service rates

$$\lambda^+(\delta) = \sup_{t_1 - \delta \leqslant s < t_1} \lambda(s) \quad \text{and} \quad \mu^+(\delta) = \inf_{t_1 - \delta \leqslant s < t_1} \mu(s)$$

respectively. Similarly, define Q^- to be an $M/M/1$ queue length process with arrival and service rates

$$\lambda^-(\delta) = \inf_{t_1 - \delta \leqslant s < t_1} \lambda(s) \quad \text{and} \quad \mu^-(\delta) = \sup_{t_1 - \delta \leqslant s < t_1} \mu(s)$$

respectively. By stochastic dominance, we have

$$Q^-(\delta/\epsilon) \leqslant_{\text{st}} Q^\epsilon(t_1 - \delta, t_1) \leqslant_{\text{st}} Q^+(\delta/\epsilon).$$

So for all $n \geqslant 0$, we have

$$\mathsf{P}(Q^-(\delta/\epsilon) \geqslant n) \leqslant \mathsf{P}(Q^\epsilon(t_1 - \delta, t_1) \geqslant n) \leqslant \mathsf{P}(Q^+(\delta/\epsilon) \geqslant n).$$

Taking the limit as $\epsilon \downarrow 0$, we get

$$\left(\frac{\lambda^-(\delta)}{\mu^-(\delta)}\right)^n \leqslant \underline{\lim_{\epsilon \downarrow 0}} \mathsf{P}(Q^\epsilon(t - \delta, t) \geqslant n) \leqslant \overline{\lim_{\epsilon \downarrow 0}} \mathsf{P}(Q^\epsilon(t - \delta, t) \geqslant n) \leqslant \left(\frac{\lambda^+(\delta)}{\mu^+(\delta)}\right)^n.$$

By (8.1), this means that for all $n \geqslant 0$,

$$\left(\frac{\lambda^-(\delta)}{\mu^-(\delta)}\right)^n \leqslant \underline{\lim_{\epsilon \downarrow 0}} \mathsf{P}(Q^\epsilon(t_1) \geqslant n) \leqslant \overline{\lim_{\epsilon \downarrow 0}} \mathsf{P}(Q^\epsilon(t_1) \geqslant n) \leqslant \left(\frac{\lambda^+(\delta)}{\mu^+(\delta)}\right)^n.$$

But δ can be made arbitrarily small. Since

$$\lim_{\delta \downarrow 0} \lambda^+(\delta) = \lim_{\delta \downarrow 0} \lambda^-(\delta) = \lambda(t),$$

and a similar relationship holds between $\mu^+(\delta)$, $\mu^-(\delta)$, and $\mu(t)$, we then get

$$\lim_{\epsilon \downarrow 0} \mathsf{P}(Q^\epsilon(t_1) \geqslant n) = \rho(t_1)^n.$$

Case $[N \to N + 1]$. Observe that if $t_1 < \cdots < t_N < t_{N+1}$, where $\rho^*(t_i) < 1$ for all $i = 1, \ldots, N + 1$, then there exists some $\delta > 0$ such that ρ^* is strictly less than one on $[t_{N+1} - \delta, t_{N+1}]$, and $t_N < t_{N+1} - \delta$. From this it follows that $Q^\epsilon(t_{N+1} - \delta, t_{N+1})$ is independent of the random vector $(Q^\epsilon(t_1), \ldots, Q^\epsilon(t_N))$. By induction hypothesis, we then have for all $n_i \geqslant 0$,

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \mathsf{P}(Q^\epsilon(t_1) \geqslant n_1, \ldots, Q^\epsilon(t_N) \geqslant n_N, Q^\epsilon(t_{N+1}) \geqslant n_{N+1}) \\
&\quad = \lim_{\epsilon \downarrow 0} \mathsf{P}(Q^\epsilon(t_1) \geqslant n_1, \ldots, Q^\epsilon(t_N) \geqslant n_N, Q^\epsilon(t_{N+1} - \delta, t_{N+1}) \geqslant n_{N+1}) \\
&\quad = \lim_{\epsilon \downarrow 0} \mathsf{P}(Q^\epsilon(t_1) \geqslant n_1, \ldots, Q^\epsilon(t_N) \geqslant n_N) \cdot \mathsf{P}(Q^\epsilon(t_{N+1} - \delta, t_{N+1}) \geqslant n_{N+1}) \\
&\quad = \prod_{i=1}^{N} \rho(t_i)^{n_i} \cdot \rho(t_{N+1})^{n_{N+1}} \\
&\quad = \prod_{i=1}^{N+1} \rho(t_i)^{n_i},
\end{aligned}$$

and this completes the proof. □

9. Onset of critical loading.

LEMMA 9.1. *Let λ and μ satisfy the hypotheses of Theorem* 3.4. *We then have for our new time scale T_0 and T (not necessarily positive) with $T_0 \leqslant T$,*

$$\lim_{\epsilon \downarrow 0} e^{k/(2k+1)} X^{\epsilon}\left(t_1 + e^{1/(2k+1)} T_0, t_1 + e^{1/(2k+1)} T\right)$$

$$= \frac{\lambda^{(k)}(t_1)}{(k+1)!}\left[T^{k+1} - T_0^{k+1}\right] + W\left(2\mu(t_1)(T - T_0)\right) \quad a.s.$$

PROOF. We have in general

$$\epsilon^b X^{\epsilon}(t_1 + \epsilon^a T_0, t_1 + \epsilon^a T)$$

$$\stackrel{\mathrm{d}}{=} \epsilon^{b-1} \int_{t_1+\epsilon^a T_0}^{t_1+\epsilon^a T} [\lambda(r) - \mu(r)]\, dr + W\left(\epsilon^{2b-1} \int_{t_1+\epsilon^a T_0}^{t_1+\epsilon^a T} [\lambda(r) + \mu(r)]\, dr\right)$$

$$+ O(\epsilon^b \log \epsilon),$$

where a and b are positive constants that are to be determined. By our hypothesis,

$$\epsilon^{b-1} \int_{t_1+\epsilon^a T_0}^{t_1+\epsilon^a T} [\lambda(r) - \mu(r)]\, dr = \frac{\lambda^{(k)}(t_1)}{(k+1)!}\left[T^{k+1} - T_0^{k+1}\right]\epsilon^{a(k+1)+b-1}$$

$$+ o(\epsilon^{a(k+1)+b-1})$$

and

$$\epsilon^{2b-1} \int_{t_1+\epsilon^a T_0}^{t_1+e^a T} [\lambda(r) + \mu(r)]\, dr = 2\mu(t_1)(T - T_0)\epsilon^{a+2b-1} + o(\epsilon^{a+2b-1}).$$

Our goal is now to obtain Brownian motion with drift in the limit as ϵ approaches zero. The leading terms for the drift and the Brownian motion expressions will be nonzero only when we employ the conditions

$$a(k+1) + b = 1 \quad \text{and} \quad a + 2b = 1.$$

These conditions yield a unique solution for a and b, namely

$$a = \frac{1}{2k+1} \quad \text{and} \quad b = \frac{k}{2k+1}$$

which completes the proof. □

10. End of overloading.

PROOF OF THEOREM 3.7. From Theorem 2.2, it follows that

$$\lim_{\epsilon \downarrow 0} \sqrt{\epsilon}\, Q^{\epsilon}(t) \stackrel{\mathrm{d}}{=} \max_{s \in \Phi_t} \hat{W}(s, t).$$

Now let τ be a nonnegative number. We will use τ for positive T, and $-\tau$ for negative T.

For the case of $T > 0$, we have

$$Q^\epsilon(t + \sqrt{\epsilon}\tau) = \sup_{0 \leqslant s \leqslant t} \left[X^\epsilon(t + \sqrt{\epsilon}\tau) - X^\epsilon(s)\right]$$

$$\vee \sup_{t \leqslant s \leqslant t + \sqrt{\epsilon}\tau} \left[X^\epsilon(t + \sqrt{\epsilon}\tau) - X^\epsilon(s)\right]$$

$$= \left[X^\epsilon(t + \sqrt{\epsilon}\tau) - X^\epsilon(t) + Q^\epsilon(t)\right]$$

$$\vee \sup_{0 \leqslant \sigma \leqslant \tau} \left[X^\epsilon(t + \sqrt{\epsilon}\tau) - X^\epsilon(t + \sqrt{\epsilon}\sigma)\right].$$

The theorem follows for this case from observing that

$$\lim_{\epsilon \downarrow 0} \sqrt{\epsilon}\left[X^\epsilon(t + \sqrt{\epsilon}\tau) - X^\epsilon(t)\right] = (\lambda(t) - \mu(t))\tau + o(1) \quad \text{a.s.},$$

and

$$\sqrt{\epsilon} \sup_{0 \leqslant \sigma \leqslant \tau} X^\epsilon(t + \sqrt{\epsilon}\tau) - X^\epsilon(t + \sqrt{\epsilon}\sigma)$$

$$= \max_{0 \leqslant \sigma \leqslant \tau} (\lambda(t) - \mu(t))(\tau - \sigma) + o(1) = o(1).$$

The last step follows since $\lambda(t) \leqslant \mu(t)$ by hypothesis.

For $T < 0$, let $T = -\tau$. Since

$$Q^\epsilon(t - \sqrt{\epsilon}\tau) = X^\epsilon(t - \sqrt{\epsilon}\tau) - X^\epsilon(t) + \sup_{0 \leqslant s \leqslant t - \sqrt{\epsilon}\tau} \left[X^\epsilon(t) - X^\epsilon(s)\right]$$

and

$$\lim_{\epsilon \downarrow 0} \sqrt{\epsilon}\left[X^\epsilon(t - \sqrt{\epsilon}\tau) - X^\epsilon(t)\right] = -(\lambda(t) - \mu(t))\tau \quad \text{a.s.},$$

we need only show that

$$\lim_{\epsilon \downarrow 0} \sqrt{\epsilon} \cdot \sup_{0 \leqslant s \leqslant t - \sqrt{\epsilon}\tau} \left[X^\epsilon(t) - X^\epsilon(s)\right] \stackrel{\mathrm{d}}{=} \max_{s \in \Phi_t} \hat{W}(s, t) \vee (\lambda(t) - \mu(t))\tau.$$

To show this, let $Y^\epsilon(s) = \sqrt{\epsilon}[X^\epsilon(t) - X^\epsilon(s)]$. We then have

$$\sqrt{\epsilon} \cdot \sup_{0 \leqslant s \leqslant t - \sqrt{\epsilon}\tau} \left[X^\epsilon(t) - X^\epsilon(s)\right]$$

$$= \sup_{0 \leqslant s < t_1 + \delta} Y^\epsilon(s) \vee \sup_{t_1 + \delta \leqslant s \leqslant t - \delta} Y^\epsilon(s) \vee \sup_{t - \delta < s \leqslant t - \sqrt{\epsilon}\tau} Y^\epsilon(s),$$

where $t_1 \equiv \sup(\Phi_t - \{t\})$, and the constant $\delta > 0$ is sufficiently small. By Theorem 5.1, we have

$$Y^\epsilon(s) \stackrel{\mathrm{d}}{=} \frac{1}{\sqrt{\epsilon}} \int_s^t [\lambda(r) - \mu(r)]\, dr + \hat{W}(s,t) + o(1).$$

Combining this with Lemma 5.2, we get

$$\lim_{\epsilon \downarrow 0} \sup_{0 \leqslant s < t_1 + \delta} Y^\epsilon(s) \stackrel{\mathrm{d}}{=} \sup_{s \in \Phi_t - \{t\}} \hat{W}(s,t)$$

and

$$\lim_{\epsilon \downarrow 0} \sup_{t_1 + \delta \leqslant s \leqslant t - \delta} Y^\epsilon(s) \stackrel{\mathrm{d}}{=} -\infty,$$

so it remains to show that

$$\lim_{\delta \downarrow 0} \lim_{\epsilon \downarrow 0} \sup_{t - \delta < s \leqslant t - \sqrt{\epsilon}\tau} Y^\epsilon(s) \stackrel{\mathrm{d}}{=} (\lambda(t) - \mu(t))\tau. \tag{10.1}$$

Now, we can construct a process $\tilde{Y}^\epsilon(s)$, having the same distribution as $Y^\epsilon(s)$ such that

$$\begin{aligned}
&\overline{\lim_{\epsilon \downarrow 0}} \sup_{t - \delta < s \leqslant t - \sqrt{\epsilon}\tau} \tilde{Y}^\epsilon(s) \\
&\quad = \overline{\lim_{\epsilon \downarrow 0}} \sup_{t - \delta < s \leqslant t - \sqrt{\epsilon}\tau} \left[\frac{1}{\sqrt{\epsilon}} \int_s^t [\lambda(r) - \mu(r)]\, dr + \hat{W}(s,t) \right] \\
&\quad \leqslant \max_{t - \delta \leqslant s \leqslant t} \hat{W}(s,t) + \overline{\lim_{\epsilon \downarrow 0}} \sup_{t - \delta < s \leqslant t - \sqrt{\epsilon}\tau} \frac{1}{\sqrt{\epsilon}} \int_s^t [\lambda(r) - \mu(r)]\, dr \\
&\quad \leqslant \max_{t - \delta \leqslant s \leqslant t} \hat{W}(s,t) + \lim_{\epsilon \downarrow 0} \frac{1}{\sqrt{\epsilon}} \int_{t - \sqrt{\epsilon}\tau}^t [\lambda(r) - \mu(r)]\, dr \\
&\quad \leqslant \max_{t - \delta \leqslant s \leqslant t} \hat{W}(s,t) + (\lambda(t) - \mu(t))\tau \quad \text{a.s.}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
(\lambda(t) - \mu(t))\tau &= \lim_{\epsilon \downarrow 0} \frac{1}{\sqrt{\epsilon}} \int_{t - \sqrt{\epsilon}\tau}^t [\lambda(r) - \mu(r)]\, dr \\
&= \lim_{\epsilon \downarrow 0} \tilde{Y}^\epsilon(t - \sqrt{\epsilon}\tau) \\
&\leqslant \overline{\lim_{\epsilon \downarrow 0}} \sup_{t - \delta < s \leqslant t - \sqrt{\epsilon}\tau} \tilde{Y}^\epsilon(s) \quad \text{a.s.}
\end{aligned}$$

This establishes (10.1), since δ can be made arbitrarily small. □

Acknowledgement. The first author was partially supported by the Fund for the Promotion of Research at the Technion, and by Technion V.P.R. Fund—E. and J. Bishop Fund.

References

Anulova, S. V. (1989). Functional Limit Theorems for Networks of Queues. Abstract, IFAC Congress Tallinn-90.

Asmussen, S. and Thorisson, H. (1987). A Markov Chain Approach to Periodic Queues. *J. Appl. Probab.* **24** 215–225.

Bambos, N. and Walrand, J. (1989). On Queues with Periodic Inputs. *J. Appl. Probab.* **26** 381–389.

Chen, H. and Mandelbaum, A. (1991a). Discrete Flow Networks: Bottleneck Analysis and Fluid Approximations. *Math. Oper. Res.* **16** 408–446.

_____ and _____ (1991b). Discrete Flow Networks: Diffusion Approximations and Bottlenecks. *Ann. Probab.* **19** 1463–1519.

Csörgő, M. and Horvath, L. (1993). *Weighted Approximations in Probability and Statistics.* John Wiley and Sons, NY.

_____ and Révész, (1981). *Strong Approximations in Probability and Statistics.* Academic Press, NY.

Eick, S., Masey, W. A. and Whitt, W. (1993a). $M_t/G/\infty$ Queues with Sinusoidal Rates. *Man. Sci.* **39** 241–252.

_____ and _____ (1993b). The Physics of the $M_t/G/\infty$ Queue. *Oper. Res.* **41** 731–742.

Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes, Characterization and Convergence.* John Wiley and Sons, NY.

Green, L., Kolesar, P. and Svornos, A. (1991). Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems. *Oper. Res.* **39** 502–511.

Hall, Randolph W. (1991). *Queueing Methods for Services and Manufacturing.* Prentice Hall, NY.

Harrison, J. M. and Lemoine, A. J. (1977). Limit Theorems for Periodic Queues, *J. Appl. Probab.* **14** 566–576.

Heyman, D. P. and Whitt, W. (1984). The Asymptotic Behaviour of Queues with Time-Varying Arrival Rates. *J. Appl. Probab.* **21** 143–156.

Iglehart, D. L. and Whitt, W. (1970). Multiple Channel Queues in Heavy Traffic. I. *Adv. in Appl. Probab.* **2** 150–177.

Keller, J. B. (1982). Time-Dependent Queues. *SIAM Rev.* **24** 401–412.

Komlós, Major and Tusnàdy, (1976). An Approximation of Partial Sums of Independent RV's and the Sample DF. I, II. *ZW* **32** 111–131, **34** 33–58.

Krichagina, E. V., Lipster, R. Sh. and Puhalski, A. A. (1988). Diffusion Approximation for the System with Arrival Process Depending on Queue and Arbitrary Service Distribution. *Theor. Probab. Appl.* **33** 124–135.

Lemoine, A. J. (1989). Waiting Time and Workload in Queues with Periodic Poisson Input. *J. Appl. Probab.* **26** 390–397.

Lipster, R. Sh. and Shiryaev, A. N. (1989). Theory of Martingales. *Dordrecht, Holland: Kluwer.*

Luchak, G. (1956). The Solution of the Single Channel Queueing Equations Characterized by a Time-Dependent Arrival Rate and a General Class of Holding Times. *Oper. Res.* **4** 711–732.

Mandelbaum, A. and Massey, W. A. Strong Approximations for Time-Dependent Queueing Networks. (in preparation).

Massey, W. A. (1981). Nonstationary Queues. Thesis, Stanford University.

_____ (1985). Asymptotic Analysis of the Time Dependent M/M/1 Queue. *Math. Oper. Res.* **10** 305–327.

_____ and Whitt, W. (1993). Networks of Infinite-Server Queues with Nonstationary Poisson Input. *Queueing Systems Theory Appl.* **13** 183–250.

Newell, G. F. (1968). Queues with Time-Dependent Arrival Rates. I, II, III. *J. Appl. Probab.* **5** 436–451 (I); 570–590 (II); 591–606 (III).

_____ (1982). *Applications of Queueing Theory.* Chapman and Hall (Second Edition).

Pomarede, J. L. (1976). A Unified Approach via Graphs to Skorohod's Topologies on the Function Space *D.* Ph.D. Thesis, Dept. of Statistics, Yale University.

Prabhu, N. U. (1980). *Stochastic Storage Processes: Queues, Insurance Risk, and Dams.* Springer Verlag, Berlin and NY.

Reiman, M. (1984). Open Queueing Networks in Heavy Traffic. *Math. Oper. Res.* **9** 441–458.

Rolski, T. (1981). Queues with Non-Stationary Input Stream: Ross's Conjecture. *Adv. Appl. Probab.* **13** 603–618.

_____ (1990). Queues with Nonstationary Inputs. *Queueing Systems* **5** 113–130.

Rothkopf, M. H. and Oren, S. S. (1979). A Closure Approximation for the Nonstationary M / M / s Queue. *Man. Sci.* **25** 522–534.

Rosenkrantz, W. (1980). On the Accuracy of Kingman's Heavy Traffic Approximation in the Theory of Queues. *Zeit. Wahr. verw. Geb.* **51** 115–121.

Skorohod, A. V. (1956). Limit Theorems for Stochastic Processes. *Theory Probab. Appl.* **1** 261–290.

Yamada, K. (1984). Diffusion Approximations for Storage Processes with General Release Rules. *Math. Oper. Res.* **9** 459–470.

_____ (1986). Multi-Dimensional Bessel Processes as Heavy Traffic Limits of Certain Tandem Queues. *Stochastic Process. Appl.* **23** 35–56.

Whitt, W. (1980). Some Useful Functions for Functional Limit Theorems. *Math. Oper. Res.* **5** 67–85.

A. Mandelbaum: Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa, 32000 Israel; e-mail: avim@techunix.technion.ac.il

W. A. Massey: AT & T Bell Laboratories, Murray Hill, New Jersey 07974