Asymptotic Methods in Probability and Statistics
B. Szyszkowicz (Editor)
© 1998 Elsevier Science B.V. All rights reserved.

Strong Approximations in Queueing Theory

Peter W. Glynn[a]

[a]Department of Engineering-Economic Systems and Operations Research,
Stanford University, Stanford, CA 94305-4023

Dedicated to Miklós Csörgő on the occasion of his 65th birthday

This paper discusses some of the merits of strong approximation ideas in developing diffusion approximations for queueing systems. Letting ρ be the utilization of the server, it is well known that as the queue is sent into heavy-traffic (i.e. $\rho \nearrow 1$), the system can be approximated by a diffusion process on spatial scales of order $(1-\rho)^{-1}$ and time scales of order $(1-\rho)^{-2}$. In this paper, we show how strong approximation methods permit one to validate the applicability of the diffusion approximation to the queue over other temporal and spatial scales. In addition, some pedagogical advantanges of the strong approximation approach are discussed, and two open problems for the strong approximation community are described.

1. Introduction

The use of strong approximation in the study of queues in heavy traffic originated with Rosenkrantz (1980). This mathematical tool is particularly convenient in this applications setting, because the dynamics of a queue can typically be expressed as a continuous functional of some family of additive processes that behave, roughly speaking, like random walk. Strong approximation principles effectively allow one to directly replace these additive processes by their corresponding Brownian approximations. Rigorous verification of a heavy-traffic diffusion limit then amounts to using elementary real variables arguments on a path-by-path basis.

In this paper, we discuss some of the advantages of using strong approximation machinery to study queueing systems. This paper's major contributions are:

1. a discussion in Section 3 of the relative pedagogical advantages of the strong approximation approach to the analysis of queueing systems in heavy traffic;

2. analysis in Section 4 of the temporal scales over which diffusion approximations to queues are typically valid;

3. a similar analysis in Section 5 of the spatial scales over which diffusions approximations to queues are generally valid;

4. a description in Section 6 of a couple of open problems for the strong approximation community that are relevant to certain theoretical issues that arise in the analysis of queues.

Some related discussion of the application of strong approximation methods in the analysis of queues appears in Glynn (1990) and Alex and Steinebach (1989), as well as in related references discussed elsewhere in this paper.

2. The CLT From a Strong Approximation Perspective

Let $S = (S(t) : t \geq 0)$ be a real-valued stochastic process. In the applications that we have in mind, S is typically a process that behaves like a random walk.

We say that S satisfies a central limit theorem (CLT) if there exist constants μ and σ such that

$$t^{-1/2}(S(t) - \mu t) \Rightarrow \sigma \mathrm{N}(0,1) \tag{2.1}$$

as $t \rightarrow \infty$, where $\Rightarrow$ denotes weak convergence (on $\mathbb{R}$) and $\mathrm{N}(0,1)$ is a mean-zero normal r.v. with unit variance. A stronger version of the CLT is the functional central limit theorem (FCLT). Let $B = (B(t) : t \geq 0)$ be a standard Brownian motion (so that $\mathbb{E}B(t) = 0$ and $\mathrm{Var}B(t) = t$), and set

$$\chi_\varepsilon(t) = \varepsilon(S(t/\varepsilon^2) - \mu t/\varepsilon^2)$$

for $\varepsilon > 0$. The process S is said to satisfy a FCLT if there exist constants μ and σ such that

$$\chi_\varepsilon \Rightarrow \sigma B \tag{2.2}$$

in $D[0,\infty)$ as $\varepsilon \downarrow 0$, where now the weak convergence is relative to the Skorohod topology on the space consisting of functions with domain $[0,\infty)$ that are right-continuous with left limits; see Ethier and Kurtz (1986) for details.

Finally, the process S is said to obey a strong approximation principle if there exists a probability space supporting a standard Brownian motion B and a process $S^* = (S^*(t) : t \geq 0)$ such that for some constants μ and σ

$$\begin{array}{ll} \text{i.)} & S^*(t) = \mu t + \sigma B(t) + o(t^{1/2}) \quad \text{a.s. as } t \rightarrow \infty; \\ \text{ii.)} & S^* \overset{\mathcal{D}}{=} S \quad (\overset{\mathcal{D}}{=} \text{ denotes "equality in distribution"}). \end{array} \tag{2.3}$$

Since $S^* \overset{\mathcal{D}}{=} S$, it is customary to take the view that the original probability space supporting S is itself rich enough to also support B, and to write the strong approximation (2.3) as

$$S(t) = \mu t + \sigma B(t) + o(t^{1/2}) \quad a.s. \tag{2.4}$$

as $t \rightarrow \infty$. It is well known that a strong approximation implies the FCLT, which in turn implies the CLT. Hence, in principle, a strong approximation requires stronger hypotheses than does either a FCLT or CLT. What then are the advantages to a strong approximation from an applications viewpoint?

Perhaps most fundamentally, it replaces the notion of weak convergence with a pathwise convergence statement. From a pedagogical standpoint, this means that if one is permitted to assume the strong approximation (2.4) as being given, any subsequent analysis building upon it can typically rely solely on elementary "real variables" path-by-path arguments. On the other hand, the FCLT (and even the CLT to some extent) requires introducing students to the notion of weak convergence and related topological issues. In particular, (2.2) cannot be rigorously discussed without reference to the Skorohod topology on $D[0, \infty)$.

A second important advantage to (2.4) is that the approximation for the r.v. $S(t)$ can be "read off" the strong approximation. Specifically, (2.4) clearly suggests using the r.v. $\mu t + \sigma B(t)$ as an approximation to $S(t)$. On the other hand, approximating $S(t)$ based on the CLT (2.1) is not as transparent to most students, since the statement of the result involves not $S(t)$ itself but the scaled/translated r.v. $t^{-1/2}(S(t) - \mu t)$.

Finally, from a mathematical viewpoint, (2.1) and (2.2) provide only information on the fluctuations of S over time scales of order $1/\varepsilon^2$ that are of order $1/\varepsilon$. On the other hand, (2.4) is a global statement that places some control on the behavior of S not only over finite time intervals, but over the entire infinite interval $[0, \infty)$.

3. Heavy Traffic Approximations From a Strong Approximation Perspective

The single-server queue offers an excellent arena within which to illustrate the full power of strong approximation machinery. We will focus on the workload process $W = (W(t) : t \geq 0)$, where $W(t)$ is the amount of unfinished work in the system at time t. To define $W(t)$, let $S(t)$ be the cumulative amount of work to have arrived to the system by time t. For a typical single-server queue, $S(t)$ takes the form

$$S(t) = \sum_{i=1}^{A(t)} V_i,$$

where V_i is the total processing time of customer i and $A(t)$ is the cumulative number of arrivals to the system by time t. If $c > 0$ is the processing rate of the server and if $W(0) = 0$, then

$$W(t) = S(t) - ct - \min_{0 \leq u \leq t} [S(u) - cu]. \tag{3.1}$$

To get a sense of why the representation (3.1) is valid, note that whenever W is positive, then $dW(t) = dS(t) - c$, so that the change in workload is just the difference between the incoming work process and the processing rate (as expected). The running minimum in (3.1) serves as a "reflecting barrier" for W, keeping it non-negative.

In order that an approximation of $W(t)$ by a functional of Brownian motion be reasonable, it seems clear that the queue must be in "heavy traffic", so that the rate at which work arrives must nearly balance the rate c at which work is completed. (Otherwise, if the system is to be "stable", W will spend most of its time near the origin, and the behavior of W is primarily explained by its boundary behavior, so that the larger scale random fluctuations of S that look approximately Brownian play a relatively minor role.) The standard way to send a queue into "heavy traffic" is to consider a sequence of queueing systems. Let $S_n = (S_n(t) : t \geq 0)$ be the incoming work process in system n, and let c_n be the processing rate in system n. Put

$$\chi_n(t) = n^{-1/2}(S_n(nt) - c_n nt) \tag{3.2}$$

for $t \geq 0$, and let $e(t) = t$. One then requires that there exist constants a and σ such that

$$\chi_n(\cdot) \Rightarrow \sigma B(\cdot) - ae(\cdot) \tag{3.3}$$

in $D[0, \infty)$ as $n \to \infty$. If $S_n(t)/t \Rightarrow \mu_n$ as $t \to \infty$, (3.2) and (3.3) together suggest that

$$n^{1/2}(\mu_n - c_n) \to -a \tag{3.4}$$

as $n \to \infty$, and consequently the arrival rate (of work) must balance the processing rate (of work) in the n'th system to a factor of order $n^{-1/2}$.

Given (3.3), the heavy-traffic analysis of the corresponding workload process $W_n = (W_n(t) : t \geq 0)$ for system n is then straightforward. For $x \in D[0, \infty)$, set

$$f_t(x) = x(t) - \min_{0 \leq u \leq t} x(u).$$

Since $f_t(\cdot)$ is continuous in the Skorohod topology at any continuous function x, it follows from the continuous mapping principle (see Billingsley (1968)) that

$$n^{-1/2}W_n(nt) = f_t(\chi_n) \Rightarrow f_t(\sigma B(\cdot) - ae(\cdot)) \tag{3.5}$$

as $n \to \infty$. The process $\sigma B(t) - at - \min_{0 \leq u \leq t}[\sigma B(u) - au]$ is a reflecting Brownian motion process, and consequently (3.5) provides a diffusion approximation to W_n that is valid for large n.

The argument developed above is the approach that has been followed in much of the heavy-traffic literature for queues and queueing networks; see, for example, Iglehart and Whitt (1970) and Reiman (1989). As in the CLT setting discussed in Section 2, this approach suffers from the pedagogical defect that it cannot be rigorously discussed without reference to the Skorohod topology on $D[0, \infty)$, and related notions of weak convergence of probability measures. But a new and more serious problem arises in this heavy traffic setting. Specifically, the limit theorem is stated in terms of a sequence of systems that involves a parameter n with no obvious physical meaning. Consequently, it is often not obvious to practitioners as to how to develop an approximation for a "real-world" queue, based on the heavy-traffic limit (3.5). On the other hand, application of strong approximation ideas immediately suggests the correct approximation, as we shall see in a moment.

Suppose that S satisfies the strong approximation (2.4). Then, substituting (2.4) and (3.1), we get

$$W(t) = \sigma B(t) - (c - \mu)t - \min_{0 \leq u \leq t}[\sigma B(u) - (c - \mu)u] + o(t^{-1/2}) \quad a.s. \tag{3.6}$$

Relation (3.6) immediately suggests the following approximation for $W(t)$:

$$W(t) \overset{\mathcal{D}}{\approx} \sigma B(t) - (c - \mu)t - \min_{0 \leq u \leq t}[\sigma B(u) - (c - \mu)u], \tag{3.7}$$

where $\overset{\mathcal{D}}{\approx}$ denotes "has approximately the same distribution as" (and has no rigorous mathematical meaning). Set

$$Z(t;\sigma,a) = \sigma B(t) - at - \min_{0\le u\le t}[\sigma B(u) - au];$$

$Z(\cdot;\sigma,a)$ is a reflected Brownian motion (RBM) process. Hence, (3.7) yields the desired approximation of $W(\cdot)$ by a diffusion process. In contrast to the weak convergence argument followed earlier in this section, the approximation (3.7) contains no artificially introduced parameters (like the system index n). Furthermore, as in the CLT setting of Section 2, strong approximation leads to a statement about $W(t)$ itself (as opposed to some "normalized" version of W). In addition, (3.6) is a global statement about $W(t)$, rather than one that purely provides information about W on finite time intervals (as is the case for (3.5). For these reasons, as well as the pedagogical reasons mentioned earlier, strong approximation methodology has become increasingly popular as a tool for approximating queueing systems.

It is worth noting that the "heavy traffic" assumption has no bearing on the validity of (3.6). Of course, from a rigorous mathematical viewpoint, the approximation (3.7) makes sense only so long as the approximating r.v. $Z(t;\sigma,c-\mu)$ is large relative to the error term $o(t^{1/2})$. It is here that the "heavy traffic" assumption is needed. As in the weak convergence argument given earlier, we consider a family of queueing systems in which the incoming work process S is fixed throughout (and satisfies (2.4)), and the processing rate c is permitted to decrease to μ (thereby sending the system into heavy-traffic). Set $\rho = \mu/c$ and note that $c \searrow \mu$ is equivalent to $\rho \nearrow 1$. Let $W(t;\rho)$ be the workload process associated with the ρ'th system. Hence, (3.6) can be re-written as

$$W(t;\rho) = \sigma B(t) - c(1-\rho)t - \min_{0\le u\le t}[\sigma B(u) - c(1-\rho)u] + o(t^{1/2}) \quad a.s. \tag{3.8}$$

where $o(t^{1/2})$ is uniform in ρ and c. But

$$Z(\cdot(1-\rho)^{-2};\sigma,c(1-\rho)) \overset{\mathcal{D}}{=} (1-\rho)^{-1}Z(\cdot;\sigma,c) \tag{3.9}$$

follows easily from standard scaling properties of Brownian motion. Consequently, we conclude that if we let $c \searrow \mu$, then

$$(1-\rho)W(\cdot(1-\rho)^{-2};\rho) \Rightarrow Z(\cdot;\sigma,\mu) \tag{3.10}$$

in $D[0,\infty)$. The limit theorem (3.10) is essentially equivalent to (3.5) (but contains no artificial system parameters). It asserts, as does (3.5), that when ρ is close to 1, then W can be approximated by an RBM on time scales of order $(1-\rho)^{-2}$, in which case the random fluctuations in W are of order $(1-\rho)^{-1}$.

Note that (3.10) suggests the approximation

$$\begin{aligned} W(t) &\overset{\mathcal{D}}{\approx} (1-\rho)^{-1}Z((1-\rho)^2t;\sigma,\mu) \\ &\overset{\mathcal{D}}{=} Z(t;\sigma,\mu(1-\rho)); \end{aligned} \tag{3.11}$$

this latter r.v. is not quite identical to the original approximation $Z(t;\sigma,c-\mu)$ suggested earlier in (3.7). (Of course, as $\rho \nearrow 1$, the approximations are asymptotically identical.)

In practice, (3.7) yields approximations that are somewhat better than those associated with (3.11), particularly for moderate values of the "traffic intensity" ρ. This tends to confirm the conclusion that strong approximations lead naturally to the "right" diffusion approximation.

The above argument guarantees that when ρ is close to 1, then RBM provides a good approximation to W over time scales of order $(1-\rho)^{-2}$ and spatial scales of order $(1-\rho)^{-1}$. In the next sections, we use strong approximation to study the question of whether RBM can be a good approximation over other temporal and spatial scales.

4. Heavy Traffic Time Scales

Given that the single-server queue is "balanced", in the sense that the station's service capacity c is close to its arrival rate μ, and "stable" (so that $c > \mu$), RBM provides a good approximation to W over time scales both shorter and longer than $(1-\rho)^{-2}$ (where $\rho = \mu/c$).

Proposition 1 *Suppose that S satisfies (2.4) and that $t \to \infty$ in such a way that $t(1-\rho)^2 \to 0$ as $\rho \nearrow 1$. Then,*

$$\frac{1}{\sqrt{t}} W(t;\rho) \Rightarrow Z(1;\sigma,0) \tag{4.1}$$

as $\rho \nearrow 1$.

Proof: Recall from (3.8) that

$$W(t;\rho) = Z(t;\sigma,c(1-\rho)) + o(t^{1/2}) \quad \text{a.s.}$$

Consequently,

$$\frac{1}{\sqrt{t}} W(t;\rho) - \frac{1}{\sqrt{t}} Z(t;\sigma,c(1-\rho)) \to 0 \quad \text{a.s.}$$

But, scaling properties of Brownian motion imply that

$$\frac{1}{\sqrt{t}} Z(t;\sigma,c(1-\rho)) \overset{\mathcal{D}}{=} Z(1;\sigma,c(1-\rho)\sqrt{t}).$$

Since $(1-\rho)\sqrt{t} \to 0$, the proposition is proved. □

The above proposition shows that $W(t;\rho)$ behaves like a driftless RBM over an initial time interval small relative to $(1-\rho)^{-2}$. It is also worth noting that $Z(1;\sigma,0) \overset{\mathcal{D}}{=} |\sigma B(1)|$. (See, for example, Karlin and Taylor (1975)).

To study time scales that are long relative to $(1-\rho)^{-2}$, we make the following assumption concerning S:

$$\begin{aligned} &S(t) = \textstyle\sum_{i=1}^{A(t)} V_i, \text{ where } A = (A(t) : t \geq 0) \text{ is a Poisson process with rate } \lambda > 0 \\ &\text{and } (V_n : n \geq 1) \text{ is an independent sequence of bounded i.i.d.r.v.'s.} \end{aligned} \tag{4.2}$$

This assumption serves to "streamline" the proofs below; qualitatively similar results undoubtedly hold under much weaker assumptions.

Since S has independent increments, we may view $S(n)$ as the sum of n i.i.d. compound Poisson r.v.'s, each having a moment generating function which converges in a neighborhood of the origin. Consequently, it is easily shown, by applying results of Komlós, Major, and Tusnády (1975, 1976), that S obeys the strong approximation principle

$$S(t) = \mu t + \sigma B(t) + O(\log t) \quad \text{a.s.} \tag{4.3}$$

as $t \to \infty$, when $\mu = \lambda \mathbb{E} V_1$ and $\sigma^2 = \lambda \mathbb{E} V_1^2$.

With the better behaved error term appearing in (4.3), we note that

$$W(t;\rho) = Z(t;\sigma, c(1-\rho)) + O(\log t) \quad \text{a.s.}$$

as $t \to \infty$, where $O(\log t)$ is uniform in c and ρ. Hence, provided that $(1-\rho)\log t \to 0$, we may conclude that

$$(1-\rho)W(t;\rho) - (1-\rho)Z(t;\sigma, c(1-\rho)) \to 0 \quad \text{a.s.}$$

as $\rho \nearrow 1$. But, as discussed in Section 3,

$$(1-\rho)Z(\cdot(1-\rho)^{-2};\sigma, c(1-\rho)) \overset{\mathcal{D}}{=} Z(\cdot;\sigma,c).$$

It follows that if $t \to \infty$ in such a way that $t(1-\rho)^2 \to \infty$ but $(1-\rho)\log t \to 0$, then

$$(1-\rho)W(t(1-\rho)^{-2};\rho) \Rightarrow Z(\infty;\sigma,\mu), \tag{4.4}$$

where $Z(\infty;\sigma,\mu)$ is a r.v. having the stationary distribution of $(Z(t;\sigma,\mu) : t \geq 0)$ (which necessarily exists if $\mu > 0$). In fact, it is well known that $Z(\infty;\sigma,\mu)$ is exponentially distributed with mean $\sigma^2/(2\mu)$; see, for example, Asmussen (1987).

Hence, it is evident that if $\rho \approx 1$ and $(1-\rho)^{-2} \ll t \ll \exp((1-\rho)^{-1})$, then

$$W(t;\rho) \overset{\mathcal{D}}{\approx} \frac{\sigma^2}{2(1-\rho)\mu} \exp(1), \tag{4.5}$$

where $\exp(1)$ is an exponential r.v. with unit mean. Thus, the use of strong approximation leads, with little effort, to the conclusion that if t is large relative to $(1-\rho)^{-2}$ but small relative to $\exp((1-\rho)^{-1})$, then $W(t)$ may be approximated by the "steady-state" of RBM. Of course, given that steady-state approximations typically improve with larger t, one expects that approximating $W(t)$ by the steady-state distribution of RBM should require only that $t \gg (1-\rho)^{-2}$, without any additional restrictions whatsoever on the magnitude of t. Unfortunately, to establish this result requires tools that go beyond the theory of strong approximation. One way to do this is to use "change-of-measure" ideas that are widely applied in studying the r.v. $W(t;\rho)$.

Proposition 2 *Suppose that S satisfies (4.2) and that $t(1-\rho)^2 \to \infty$ as $\rho \nearrow 1$. Then,*

$$(1-\rho)W(t;\rho) \Rightarrow \frac{\sigma^2}{2\mu}\exp(1) \tag{4.6}$$

as $\rho \nearrow 1$.

Proof: We first note that for each fixed ρ, the independent increments structure of S implies that

$$W(t;\rho) = \max_{0\le u\le t}[S(u) - u\mu/\rho].$$

Hence, for $x > 0$,

$$\mathbb{P}(W(t;\rho) > (1-\rho)^{-1}x) = \mathbb{P}(T(\rho) \le t),$$

where $T(\rho) = \inf\{t \ge 0 : S(t) - \mu t/\rho > (1-\rho)^{-1}x\}$. Let $\mathbb{P}_\theta(\cdot)$ be the probability on the path-space of W under which S evolves according to a stationary independent increments process with

$$\mathbb{P}_\theta(S(1) \in dx) = \exp(\theta x - \psi(\theta))\mathbb{P}(S(1) \in dx)$$

for $x \ge 0$, where $\psi(\theta) = \log \mathbb{E}\exp(\theta S(1))$. Then,

$$\mathbb{P}(T(\rho) \le t) = \mathbb{E}_\theta[\exp(-\theta S(T(\rho)\wedge t) + (T(\rho)\wedge t)\psi(\theta)); T(\rho) \le t]. \tag{4.7}$$

Let $\theta^* = \theta^*(\rho) > 0$ solve the equation

$$\psi(\theta^*) = \theta^*\mu/\rho. \tag{4.8}$$

Note that as $\rho \nearrow 1$, (4.8) clearly has a unique solution θ^*. Furthermore, since $\psi(\theta) = \mu\theta + \sigma^2\frac{\theta^2}{2} + o(\theta^2)$ as $\theta \to 0$, it is clear that

$$\theta^* = 2\mu(1-\rho)/\sigma^2 + o(1-\rho) \tag{4.9}$$

as $\rho \nearrow 1$. Putting $\theta = \theta^*$ in (4.7), (4.8) yields

$$\mathbb{P}(T(\rho) \le t) = \mathbb{E}_{\theta^*}[\exp(-\theta^*(S(T(\rho)\wedge t) - \mu/\rho(T(\rho)\wedge t))); T(\rho) \le t].$$

But by definition of $T(\rho)$, $0 \le S(T(\rho)) - (\mu/\rho)T(\rho) - x(1-\rho)^{-1} \le b$, where b is the assumed upper bound on the V_i's (see (4.2)). So,

$$\begin{aligned} & \exp(-\theta^*(x(1-\rho)^{-1} + b))\mathbb{P}_{\theta^*}(T(\rho) \le t) \\ \le\ & \mathbb{P}(T(\rho) \le t) \\ \le\ & \exp(-\theta^* x(1-\rho)^{-1})\mathbb{P}_{\theta^*}(T(\rho) \le t). \end{aligned}$$

In view of the fact that $\theta^*(1-\rho)^{-1} \to 2\mu/\sigma^2$ as $\rho \nearrow 1$, the proposition's proof is therefore complete if we can argue that $\mathbb{P}_{\theta^*}(T(\rho) \le t) \to 1$ if $t(1-\rho)^2 \to \infty$. But

$$\begin{aligned} & \mathbb{P}_{\theta^*}(T(\rho) \le t) \\ \ge\ & \mathbb{P}_{\theta^*}(S(t) - \frac{\mu}{\rho}t > x(1-\rho)^{-1}) \\ =\ & 1 - \mathbb{P}_{\theta^*}(S(t) - \frac{\mu}{\rho}t \le x(1-\rho)^{-1}) \\ \ge\ & 1 - \mathbb{P}_{\theta^*}(|S(t) - \frac{\mu}{\rho}t - mt| > mt - x(1-\rho)^{-1}) \end{aligned}$$

where $m = m(\rho) = \mathbb{E}_{\theta^*}S(1) - \frac{\mu}{\rho} = \psi'(\theta^*) - \frac{\mu}{\rho} = \mu(1-\rho) + o(1-\rho)$. Note that if $t(1-\rho)^2 \to +\infty$, then $mt - x(1-\rho)^{-1} \sim mt$ as $\rho \nearrow 1$. Finally, Chebyshev's inequality yields

$$\begin{aligned} & \mathbb{P}_{\theta^*}(|S(t) - \frac{\mu}{\rho}t - mt|) \\ \leq \ & t\mathrm{Var}_{\theta^*}(S(1) - \mu/\rho)/(mt - x(1-\rho)^{-1})^2 \\ \sim \ & \psi''(\theta *)/m^2 t = (\sigma^2 + o(1))/m^2 t \to 0 \end{aligned}$$

as $\rho \nearrow 1$, proving the result. □

Proposition 2 is a clear counterpart to Proposition 1, establishing that the r.v. $W(t)$ can be well approximated by RBM over time scales that are long relation to $(1-\rho)^{-2}$.

Proposition 1 and 2 focus on the marginal distribution of W. Of course, strong approximation is ideally suited to studying more complex functionals of W. Note that the same argument as applied earlier in this section shows that if $(1-\rho)\log t \to 0$ as $\rho \nearrow 1$, then

$$\sup_{0 \leq u \leq t} |(1-\rho)W(u;\rho) - (1-\rho)Z(u;\sigma, c(1-\rho))| \to 0 \quad \text{a.s.} \tag{4.10}$$

as $t \to \infty$, permitting one to approximate suitably continuous functionals of W involving the path of W over time scales that are small relative to $\exp((1-\rho)^{-1})$ by the corresponding functional of RBM. For an example of such a computation, see Glynn and Whitt (1995).

The time scale $\exp((1-\rho)^{-1})$ is a critical time scale for such approximations. Specifically, one expects that (4.10) cannot be valid for longer time scales. This phenomenon is closely related to the solution of the "stochastic geyser problem"; see Bártfai (1966). Note that

$$\sup_{0 \leq u \leq t} (1-\rho)|S(t) - \mu t - \sigma B(t)| \to 0 \quad \text{a.s.} \tag{4.11}$$

provided that $(1-\rho)\log t \to 0$ as $\rho \nearrow 1$; this is the main theoretical ingredient that goes into proving (4.10). The problem is that if $(1-\rho)\log t \to \infty$, then

$$\sup_{0 \leq u \leq t} (1-\rho)|S(t) - \mu t - \sigma B(t)| \to +\infty \quad \text{a.s.}$$

for non-Brownian independent increments processes S, regardless of how the probability space used to support S and B is constructed. To see why, consider the functional (for $d > 0$)

$$(1-\rho) \max_{0 \leq k \leq [t - d \log t]} S(k + [d \log t]) - S(k) - \mu[d \log t]. \tag{4.12}$$

If (4.11) were to hold when $(1-\rho)\log t \to \infty$, the behavior of the above r.v. over such time scales must be determined solely by μ and σ^2. But the Erdős-Rényi law (see Csörgő and Révész (1981)) implies that (4.12) is almost surely asymptotic to $(1-\rho)d\alpha(d) \cdot \log t$, where

$$\alpha(d) = \sup\{x : \inf_{\theta} \exp(-\theta x)\mathbb{E}e^{\theta S(1)} \geq e^{-1/d}\}.$$

Hence, if $(1-\rho)\log t \to \infty$, (4.12) grows to infinity at a rate determined not just by μ and σ^2 but by the function $\alpha(\cdot)$. We conclude that the strong approximation breaks down for functionals that involve the path history over time scales $t \gg \exp((1-\rho)^{-1})$.

5. Heavy Traffic Spatial Scales

In this section, we are again interested in a single-server queue in heavy traffic. We shall be concerned here with the spatial scales over which such a queue can be approximated by a RBM when the time scale is of the conventional heavy-traffic order of magnitude, namely $(1-\rho)^{-2}$.

As seen earlier in Section 3, RBM clearly gives good approximations over spatial scales of order $(1-\rho)^{-1}$. The question is: How does RBM perform as an approximation to a single-server queue over both smaller and larger spatial scales than $(1-\rho)^{-1}$? We start with a discussion of the relevant behavior at smaller spatial scales.

Clearly, if $x = x(\rho) \ll (1-\rho)^{-1}$, then $\mathbb{P}(W(t(1-\rho)^{-2};\rho) \le x) \to 0$ as $\rho \nearrow 1$. Hence, by a good approximation, we are interested in finding conditions on x such that

$$\mathbb{P}(W(t(1-\rho)^{-2};\rho) \le x) \sim \mathbb{P}(Z(t(1-\rho)^{-2};\sigma, c(1-\rho)) \le x) \tag{5.1}$$

as $\rho \nearrow 1$. Our focus will be on the range of $p \in (0,1)$ for which (5.1) holds when $x = y(1-\rho)^{p-1}$ for $y > 0$ fixed. We can easily simplify the right-hand side of (5.1). Note that

$$\begin{aligned} & \mathbb{P}(Z(t(1-\rho)^{-2};\sigma, c(1-\rho)) \le x) \\ = \; & \mathbb{P}((1-\rho)^{-1}Z(t;\sigma,c) \le x) \\ = \; & \Phi\left(\frac{z+ct}{\sigma t^{1/2}}\right) - \exp(-2cz/\sigma^2)\Phi\left(\frac{-z+ct}{\sigma t^{1/2}}\right) \end{aligned}$$

where $\Phi(\cdot) = \mathbb{P}(\mathrm{N}(0,1) \le \cdot)$ and $z = y(1-\rho)^p$. (The transition distribution of RBM may be found, for example, in Harrison (1985), p.15.) Since $z \downarrow 0$ as $\rho \nearrow 1$, the latter probability is asymptotic to z multiplied by the value of the transition density at the origin. Consequently,

$$\mathbb{P}(Z(t(1-\rho)^{-2};\sigma, c(1-\rho)) \le x) \sim 2y(1-\rho)^p \left[\frac{1}{\sigma\sqrt{t}}\varphi\left(\frac{\mu}{\sigma}\sqrt{t}\right) + \frac{\mu}{\sigma^2}\Phi\left(\frac{\mu}{\sigma}\sqrt{t}\right)\right] \tag{5.2}$$

as $\rho \nearrow 1$, where $\varphi(\cdot)$ is the density of a $\mathrm{N}(0,1)$ r.v.

We now turn to the left-hand side of (5.1). Let $\varepsilon(\rho)$ be the "error term" r.v. in the strong approximation, namely,

$$\varepsilon(\rho) = \sup_{0 \le u \le t(1-\rho)^{-2}} |S(u) - \mu u - \sigma B(u)|.$$

We shall require that S satisfy (4.2), so that the results of Komlós, Major, and Tusnády (1975, 1976) imply that

$$\mathbb{P}(\varepsilon(\rho) > d\log((1-\rho)^{-1}) + r) \le Ke^{-\lambda r} \tag{5.3}$$

for some finite, positive constants d, K and λ. Then,

$$\mathbb{P}(\varepsilon(\rho) > (d + 2p/\lambda)\log((1-\rho)^{-1})) = O((1-\rho)^{2p})$$

as $\rho \nearrow 1$. Note that

$$\begin{aligned}
& \mathbb{P}(W(t(1-\rho)^{-2};\rho) \leq y(1-\rho)^{p-1}) \\
\leq\ & \mathbb{P}(Z(t(1-\rho)^{-2};\sigma, c(1-\rho)) \leq y(1-\rho)^{p-1} + (d+2p/\lambda)\log((1-\rho)^{-1})) \\
=\ & 2(y(1-\rho)^{p} + (d+2p/\lambda)\log((1-\rho)^{-1})(1-\rho)) \\
& \cdot\left[\frac{1}{\sigma\sqrt{t}}\varphi\left(\frac{\mu}{\sigma}\sqrt{t}\right) + \frac{\mu}{\sigma^2}\Phi\left(\frac{\mu}{\sigma}\sqrt{t}\right)\right] + o((1-\rho)^{p}) \\
=\ & 2y(1-\rho)^{p}\left[\frac{1}{\sigma\sqrt{t}}\varphi\left(\frac{\mu}{\sigma}\sqrt{t}\right) + \frac{\mu}{\sigma^2}\Phi\left(\frac{\mu}{\sigma}\sqrt{t}\right)\right] + o((1-\rho)^{p})
\end{aligned}$$

as $\rho \nearrow 1$, where the first equality follows from an argument similar to that used to obtain (5.2). A similar lower bound on $\mathbb{P}(W(t(1-\rho)^{-2};\rho) \leq y(1-\rho)^{p-1})$ can easily be computed. These two bounds, in conjunction with (5.2), prove the following result.

Proposition 3 *Suppose that S satisfies (4.2). Then, for each $p \in (0,1), y > o$,*

$$\begin{aligned}
& \mathbb{P}(W(t(1-\rho)^{-2};\rho) \leq y(1-\rho)^{p-1}) \\
\sim\ & \mathbb{P}(Z(t(1-\rho)^{-2};\sigma, c(1-\rho)) \leq y(1-\rho)^{p-1}) \qquad (5.4) \\
=\ & 2y(1-\rho)^{p}\left[\frac{1}{\sigma\sqrt{t}}\varphi\left(\frac{\mu}{\sigma}\sqrt{t}\right) + \frac{\mu}{\sigma^2}\Phi\left(\frac{\mu}{\sigma}\sqrt{t}\right)\right] + o((1-\rho)^{p})
\end{aligned}$$

as $\rho \nearrow 1$.

We conclude that the heavy traffic approximation is valid, in the sense described by Proposition 3, over spatial scales as small as $(1-\rho)^{p-1}$ for any $p > 0$.

We turn now to larger spatial scales. We wish to know the range of $p > 0$ for which

$$\begin{aligned}
& \mathbb{P}(W(t(1-\rho)^{-2};\rho) > y(1-\rho)^{-1-p}) \\
\sim\ & \mathbb{P}(Z(t(1-\rho)^{-2};\sigma, c(1-\rho)) > y(1-\rho)^{-1-p})
\end{aligned} \qquad (5.5)$$

is valid, as $\rho \nearrow 1$ (for $y > 0$ fixed). We follow the same approach as that used to obtain Proposition 3. Considering first the right-hand side of (5.5), observe that

$$\begin{aligned}
& \mathbb{P}(Z(t(1-\rho)^{-2};\sigma, c(1-\rho)) > y(1-\rho)^{-1-p}) \\
=\ & \mathbb{P}(Z(t;\sigma,c) > y(1-\rho)^{-p}) \\
=\ & \mathbb{P}(\mathrm{N}(0,1) > (z+ct)/\sigma t^{1/2}) + \exp(-2cz/\sigma^2)\mathbb{P}(\mathrm{N}(0,1) < (-z+ct)/\sigma t^{1/2}),
\end{aligned}$$

where $z = y(1-\rho)^{-p}$.

Recall that

$$\mathbb{P}(\mathrm{N}(0,1) > x) \sim \varphi(x)/x$$

as $x \to \infty$; see Feller (1968), p. 175. Then,

$$\begin{aligned}
&\mathbb{P}(Z(t(1-\rho)^{-2};\sigma,c(1-\rho)) > y(1-\rho)^{-1-p})\\
= \;& \mathbb{P}(\mathrm{N}(0,1) > (z+ct)/\sigma t^{1/2}) \cdot \left[1+\exp(-2cz/\sigma^2)\frac{\mathbb{P}(\mathrm{N}(0,1) > (z-ct)/\sigma t^{1/2})}{\mathbb{P}(\mathrm{N}(0,1) > (z+ct)/\sigma t^{1/2})}\right]\\
= \;& \mathbb{P}(\mathrm{N}(0,1) > (z+ct)/\sigma t^{1/2}) \cdot \left[1+\exp(-2cz/\sigma^2)\frac{\varphi((z-ct)/\sigma t^{1/2})}{\varphi((z+ct)/\sigma t^{1/2})} + o(1)\right]\\
= \;& \mathbb{P}(\mathrm{N}(0,1) > (z+ct)/\sigma t^{1/2}) \cdot \left[1+\exp\left(-2cz/\sigma^2 - \frac{(z-ct)^2}{2\sigma^2 t} + \frac{(z+ct)^2}{2\sigma^2 t}\right) + o(1)\right]\\
= \;& \mathbb{P}(\mathrm{N}(0,1) > (z+ct)/\sigma t^{1/2})(2+o(1)),
\end{aligned}$$

so that

$$\mathbb{P}(Z(t(1-\rho)^{-2};\sigma,c(1-\rho)) > y(1-\rho)^{-1-p}) \sim 2\mathbb{P}(\mathrm{N}(0,1) > (z+\mu t)/\sigma t^{1/2}) \tag{5.6}$$

as $\rho \nearrow 1$.

To analyze the left-hand side of (5.5), we again assume (4.2). Then, (5.3) yields the inequality

$$\mathbb{P}(\varepsilon(\rho) > d\log((1-\rho)^{-1}) + v(1-\rho)^{-q}) \le K\exp(-\lambda v(1-\rho)^{-q})$$

for $v, q > 0$. Hence, as in (5.6),

$$\begin{aligned}
&\mathbb{P}(W(t(1-\rho)^{-2};\rho) > y(1-\rho)^{-1-p})\\
\le \;& \mathbb{P}(Z(t(1-\rho)^{-2};\sigma,c(1-\rho)) > y(1-\rho)^{-1-p} - v(1-\rho)^{-q})\\
&+\mathbb{P}(\varepsilon(\rho) > v(1-\rho)^{-q}) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (5.7)\\
\le \;& \mathbb{P}(\mathrm{N}(0,1) > (y(1-\rho)^{-p} - v(1-\rho)^{1-q} + \mu t)/\sigma t^{1/2})(2+o(1))\\
&+K\exp(-\lambda v(1-\rho)^{-q}).
\end{aligned}$$

In order that the error term $\exp(-\lambda v(1-\rho)^{-q})$ be small relative to (5.6), we must choose $q > 2p$. On the other hand, in order that our "upper bound" $\mathbb{P}(\mathrm{N}(0,1) > (y(1-\rho)^{-p} - v(1-\rho)^{1-q} + \mu t)/\sigma t^{1/2})$ be asymptotic to (5.6), we are required to choose $q < 1-p$. Hence, we find that

$$\mathbb{P}(W(t(1-\rho)^{-2};\rho) > y(1-\rho)^{-1-p}) \sim \mathbb{P}(Z(t(1-\rho)^{-2};\sigma,c(1-\rho)) > y(1-\rho)^{-1-p})$$

as $\rho \nearrow 1$, provided that we can find a $q > 0$ with the property that $2p < q < 1-p$. So, p must lie on the interval $(0, 1/3)$, in order that the above argument be valid. We summarize our discussion with the following proposition.

Proposition 4 *Suppose that S satisfies (4.2). Then, for $0 < p < 1/3$, $y > 0$,*

$$\begin{aligned}
&\mathbb{P}(W(t(1-\rho)^{-2};\rho) > y(1-\rho)^{-1-p})\\
\sim \;& \mathbb{P}(Z(t(1-\rho)^{-2};\sigma,c(1-\rho)) > y(1-\rho)^{-1-p})\\
= \;& \mathbb{P}(\mathrm{N}(0,1) > (y(1-\rho)^{-p} + \mu t)/\sigma t^{1/2})(2+o(1))
\end{aligned}$$

as $\rho \nearrow 1$.

So, the heavy-traffic RBM approximation is valid over spatial scales as large as $(1-\rho)^{-4/3}$. This is a heavy-traffic analog to the CLT tail asymptotic given by, for example, Theorem 1, p. 549, of Feller (1971).

6. Concluding Remarks and Open Problems

In the previous three sections, we have focussed on the heavy-traffic behavior of the single-server queue, using strong approximation as our principal mathematical tool. Heavy-traffic theorems for networks of queues also exist. Specifically, we refer to the work of Horváth (1992) and Chen and Mandelbaum (1994), in which strong approximations for a family of single customer class queueing networks in heavy-traffic were developed.

Roughly speaking, the theory for a d station network goes as follows. Suppose that μ_i is the rate at which work flows into the i'th station of the network, and let c_i be the processing rate of the server at station i $(1 \leq i \leq d)$. Set $\rho_i = \mu_i/c_i$, and suppose that $\rho_i < 1$ for $1 \leq i \leq d$, so that there is adequate processing capacity at each station for its associated incoming workload. If

$$\rho = \min_{1\leq i\leq d} \rho_i \approx 1,$$

then the network is in "heavy traffic", and the network's temporal dynamics can be approximated by a d-dimensional reflecting Brownian motion on the orthant. Furthermore, the heavy-traffic time and spatial scaling is as in the single-station case. In particular, on time scales of order $(1-\rho)^{-2}$ and spatial scales of order $(1-\rho)^{-1}$, the network (vector) workload process behaves like an RBM (having no dependence on ρ).

Our analysis of Sections 4 and 5 suggest that:

i). the RBM network approximation to the marginal distribution of the vector workload process is valid provided t is large (but may be of smaller order than $(1-\rho)^{-1}$ when $t \ll (1-\rho)^{-2}$);

ii). the RBM network approximation to the vector workload process is valid for functionals that depend on a segment of the path history that is small relative to $\exp((1-\rho)^{-1})$;

iii). on time scales of order $(1-\rho)^{-2}$, the RBM provides accurate approximations to the marginal distribution of the network workload process over spatial scales of as small as $(1-\rho)^{-q}$ $(0 < q < 1)$ to as large as $(1-\rho)^{-4/3}$.

The current state of knowledge for multi-class network is more muddled. Such networks are poorly understood, relative to their single-class counterparts, in part because even the question of "stability" is not yet settled. For some insight into the "heavy-traffic" behavior of such multi-class networks, see Harrison (1995).

We conclude this section by providing a couple of open problems that may be of theoretical interest to the "strong approximation" community. Each of these problems is motivated by certain issues that arise in the application of strong approximation ideas to queueing systems.

In certain arguments, it would be convenient to invoke a "triangular array" version of the strong approximation principle in which logarithmic error terms appear (as in (4.3)). Such a result is needed, for example, in studying the behavior of a sequence of single-station systems going into heavy-traffic, in which the arriving work process is itself permitted to depend upon the traffic intensity ρ. (In Sections 3, 4, and 5 of this paper,

$W(\cdot;\rho)$ was constructed from a single arrival process defined independently of ρ.) A partial result in this direction would be the following.

Open Problem 1: *Investigate how the constants d, K, and λ appearing in (5.3) depend on the distribution of S. (This would potentially permit one to obtain some control on the strong approximation that is uniform as one "moves down" the triangular array.)*

The solution of this problem would also be useful in generalizing the results of Glynn and Whitt (1991) to the setting in which the consecutive customers passing through the network are permitted to have different processing time distributions.

A second theoretical problem arises as follows. In many applications, it is highly unrealistic to assume that the arriving work process has stationary independent increments (as we did earlier in (4.2)). In fact, a more widely accepted model assumes that

$$S(t) = \int_0^t f(X(s))\,ds,$$

where $X = (X(t) : t \geq 0)$ is a Markov process, and f is a real-valued functional defined on the states of X. In order that S look Brownian, one typically requires that X be positive recurrent in some sense. One widely used notion of recurrence is that of Harris recurrence (in continuous time). For example, d-dimensional (nice) diffusions are typically recurrent in this sense. In order to develop a strong approximation principle like (4.3), the most obvious theoretical approach is to attempt to reduce the analysis to that of a sum of i.i.d.r.v.'s via use of "regenerative" ideas; one can then directly apply the results of Komlós, Major, and Tusnády (1975, 1976) to obtain a strong approximation with logarithmic error bounds.

The difficulty is that processes X that are Harris recurrent in continuous time are not typically regenerative. To see this, let X be a d-dimensional diffusion, and suppose there exists a sequence of random times $T_0 < T_1 < T_2 < \cdots$ such that $((X(T_{i-1}+u) : 0 \leq u < \tau_i) : i \geq 1)$ $(\tau_i = T_i - T_{i-1})$ is i.i.d. Consequently, for each $i \geq 1$, $X(T_i-)$ is independent of $X(T_i+)$. But, by path continuity, $X(T_i-) = X(T_i+) = X(T_i)$. Hence, $X(T_i)$ is deterministic for $i \geq 1$. It follows that X must visit some fixed deterministic point infinitely often. In general, this is false for diffusions in dimensions two or higher. So, we conclude that such diffusions do not typically exhibit this type of regenerative structure.

Instead, it turns out that one can establish existence, for any Markov process that is Harris recurrent (in continuous time), of a sequence of random times $T_0 < T_1 < \cdots$ such that $\{(X(T_{i-1}+u) : 0 \leq u < \tau_i) : i \geq 1\}$ is identically distributed and 1-dependent; see Sigman (1990). Let

$$Y_i = \int_{T_{i-1}}^{T_i} f(X(s))\,ds$$

for $i \geq 0$ (setting $T_{-1} = 0$). Then, under mild conditions on (Y_1, τ_1),

$$\frac{1}{t}S(t) \to \mu \overset{\Delta}{=} \frac{\mathbb{E}Y_1}{\mathbb{E}\tau_1} \quad \text{a.s.}$$

as $t \to \infty$. Furthermore,

$$S(t) - \mu t \approx \sum_{i=1}^{N(t)} (Y_i - \mu\tau_i),$$

where $N(t) = \max\{n \geq -1 : T_n \leq t\}$. Evidently, then, a major step towards obtaining good strong approximations for a process S constructed from such a Harris recurrent X would be the solution of the following open problem:

Open Problem 2: *Let $(W_i : i \geq 1)$ be a sequence of identically distributed one-dependent r.v.'s. Find conditions on W_1 such that there exist constants μ and σ for which*

$$\sum_{i=1}^{n} W_i - n\mu = \sigma B(n) + O(\log n) \quad a.s.$$

for some standard Brownian motion B (on a suitably defined probability space).

With a positive solution to this second problem, strong approximation principles for a very wide class of recurrent Markov processes (having logarithmic error term) would potentially be made available.

Acknowledgments: The author would like to thank Jacky Lee for his comments and assistance in typing the paper. The author also gratefully acknowledges the support of the Army Research Office under Contract No. 1DLA486 and the National Science Foundation under Grant DMS-9704732.

REFERENCES

1. Rosenkrantz, W. (1980). On the accuracy of Kingman's heavy traffic approximation in the theory of queues. *Z. Wahrscheinlichkeitsth.* **51**, 115-121.
2. Glynn, P.W. (1990). Diffusion approximations. In *Handbook in Operations Research and Management Science II: Stochastic Models* (D.P. Heyman and M.J. Sobel, ed.) North-Holland, Amsterdam.
3. Alex, M. and J. Steinebach (1989). Invariance principles in queueing theory. *J. Appl. Prob.* **27**, 845-857.
4. Ethier, S.N. and T.G. Kurtz (1986). *Markov Processes: Characterization and Convergence.* John Wiley, New York.
5. Billingsley, P. (1968). *Weak Convergence of Probability Measures.* John Wiley, New York.
6. Iglehart, D.L. and W. Whitt (1970). Multiple channel queues in heavy traffic. I. *Adv. Appl. Prob.* **2**, 355-369.
7. Reiman, M.I. (1989). Open queueing networks in heavy traffic. *Math. Oper. Res.* **9**, 441-458.
8. Karlin, S. and H.M. Taylor (1975). *A First Course in Stochastic Processes*. Academic Press, New York.
9. Komlós, J., Major, P. and G. Tusnády (1975). An approximation of partial sums of independent R.V.'s and the sample DF. I. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **32**, 111-131.

10. Komlós, J., Major, P. and G. Tusnády (1976). An approximation of partial sums of independent R.V.'s and the sample DF. II. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **34**, 33-58.
11. Asmussen (1987). *Applied Probability and Queues*. John Wiley, New York.
12. Glynn, P.W. and W. Whitt (1995). Heavy-traffic extreme-value limits for queues. *Operations Research Letters* **18**, 107-111.
13. Bártfai, P. (1966). Die Bestimmung der zu einem wiederkehrenden Prozess gehörenden Verteilungsfunktion aus den mit Fenlern behafteten Daten einer Einzingen Relation. *Studia Sci. Math. Hung.* 1, 161-168.
14. Csörgő, M. and P. Révész (1981). *Strong Approximations in Probability and Statistics*. Academic Press, New York.
15. Harrison, J.M. (1985). *Brownian Motion and Stochastic Flow Systems*. John Wiley, New York.
16. Feller, W. (1968). *An Introduction to Probability Theory and its Applications, Volume* **I**. John Wiley, New York.
17. Feller, W. (1971). *An Introduction to Probability Theory and its Applications, Volume* **II**. John Wiley, New York.
18. Horváth, L. (1992). Strong approximations of open queueing networks. *Math. O.R.* **17**. 487-508.
19. Chen, H. and A. Mandelbaum (1994). Hierarchical modeling of stochastic networks, Part II: Strong approximations. In *Stochastic Modeling and Analysis of Manufacturing Systems* (D.D. Yao, ed.), 102-130.
20. Harrison, J.M. (1995). Balanced fluid models of multiclass queueing networks: a heavy traffic conjecture. In *Stochastic Networks* (F.P. Kelly and R.J. Williams, ed.), 1-20. Springer-Verlag, New York.
21. Glynn, P.W. and W. Whitt (1991). Departures from many queues in series. *Annals of Applied Probability* **1**, 546-572.
22. Sigman, K. (1990). One-dependent regeneration processes and queues in continuous time. *Math. O.R.* **15**, 175-189.