D.P. Heyman, M.J. Sobel, Eds., *Handbooks on OR & MS, Vol. 2*
© Elsevier Science Publishers B.V. (North-Holland) 1990

Chapter 4

Diffusion Approximations

Peter W. Glynn

Department of Operations Research, Stanford University, Stanford, CA 94305-4022, U.S.A.

1. Introduction

In this chapter, we shall give an overview of some of the basic applications of the theory of diffusion approximations to operations research. A diffusion approximation is a technique in which a complicated and analytically intractable stochastic process is replaced by an appropriate diffusion process. A diffusion process is a (strong) Markov process having continuous sample paths. Diffusion processes have a great deal of analytical structure and are therefore typically more mathematically tractable than the original process with which one starts. The approach underlying the application of diffusion approximations is therefore comparable to that underlying normal approximation for sums of random variables. In the latter setting, the central limit theorem permits one to replace the analytically intractable sum of random variables by an appropriately chosen normal random variable.

In this chapter, we shall describe some of the basic theory of weak convergence that underlies the method of diffusion approximation. We shall then survey various applications of this methodology to the approximation of complex queueing systems.

Because we are interested in developing approximations for the distribution of a process (considered as a random function of time), it is necessary for us to describe the basic elements of the theory of weak convergence in a function space. Sections 2 and 3 are therefore devoted to this topic. Section 4 discusses the most basic and easily understood of all diffusion approximations, namely the general principle that sums of random variables (when viewed as stochastic processes) can be approximated by Brownian motion: this result is known, in the literature, as Donsker's theorem. By using the close correspondence between random walk and the single-server queue, Section 5 develops the basic theory of weak convergence for the GI/G/1/∞ queue. This forms valuable background for the more complex diffusion approximations that appear in the network setting of Sections 7 and 8. Section 6 gives a brief overview of some of the basic analytical theory of diffusion processes. In particular, we show that a large number of interesting performance measures can be calculated as solutions to certain associated partial differential equations.

The next three sections describe various applications of the theory of weak convergence to a network of queues. In Sections 7 and 8, diffusion approximations for both open and closed networks of queues in heavy traffic are given. The complex behavior of the queueing model is replaced by a more tractable 'Brownian network'. By Brownian network, we refer to a diffusion process which is obtained by subjecting a Brownian motion to reflection at the boundaries of an appropriately defined region which occurs as the 'limiting state space' of the queueing network. Section 9 discusses further weak convergence theorems that describe the behavior of queueing networks in which the number of servers is large. The limit processes that arise in this setting are quite different from the Brownian networks of Sections 7 and 8.

Section 10 concludes the chapter with a brief description of the notion of conditional weak convergence theorems. Several conditional limit theorems for the single-server queue are described.

Although this chapter concentrates on describing the applications of diffusion approximations to queueing networks, there are additional operations research applications that have been enriched by the theory. We list several such areas here, to give the reader a flavor of the broad impact that these methods have had on operations research. One important example is the application of geometric Brownian motion to the optimization of financial gain from trading securities. The resulting Black–Scholes option pricing formula has had a major impact on the theory of finance; see Duffie and Protter (1988) for a description of the basic limit theorem. Another applications area that has been impacted by the theory of diffusion approximation is storage theory. Yamada (1984) obtains diffusion approximations for a class of storage systems characterized by a nonlinear release rule. The theory of diffusion approximation has also benefitted discrete-event simulation. For example, one important output analysis technique, known as standardized time series, has been developed on the basis of the approximation associated with Donsker's theorem. See Schruben (1983) and Glynn and Iglehart (1990) for details. The above is but a partial list of the various operations research applications to which the theory of diffusion approximation has contributed.

As stated above, this chapter focuses on providing an overview of some of the key features of diffusion approximation theory, as it applies to queues. It is not intended to serve as a historical perspective on the development of the field. The author apologizes, in advance, to the many contributors to the area that have not been adequately cited.

2. Weak convergence of stochastic processes

Suppose that $Y = \{Y(t): t \geq 0\}$ is a complex, analytically intractable stochastic process. The idea underlying a diffusion approximation is to find a diffusion process X such that the distribution of Y may be approximated by that of X. Specifically, we write this as

$$Y \overset{\mathrm{d}}{\approx} X \tag{2.1}$$

($\overset{\mathrm{d}}{\approx}$ denotes 'approximately equal in distribution to'). To make more precise sense of (2.1), the standard approach is to phrase the approximation in terms of a limit theorem. In other words, we suppose that there exists a sequence $\{X_n : n \geqslant 0\}$ of stochastic processes $X_n = \{X_n(t) : t \geqslant 0\}$ such that Y may be identified with X_n, where n is large. Then, the precise meaning of (2.1) is that the limit theorem

$$X_n \xrightarrow{\mathrm{w}} X \tag{2.2}$$

holds, where $\xrightarrow{\mathrm{w}}$ denotes 'weak convergence'. The remainder of this section is devoted to describing in more detail the notion of convergence (2.2); this type of convergence is also called 'convergence in distribution'.

We start by reviewing the notion of weak convergence of random variables on the real line. A sequence $\{X_n : n \geqslant 0\}$ of r.v.'s is said to converge weakly to X if

$$P\{X_n \leqslant x\} \to P\{X \leqslant x\} \tag{2.3}$$

as $n \to \infty$ at all continuity points x of the limit distribution function $P\{X \leqslant \cdot\}$. This convergence notion can be reformulated in many different ways. The reformulation which we shall find convenient here is given by the following proposition.

Proposition 1. *A sequence of r.v.'s $\{X_n : n \geqslant 0\}$ converges weakly to X if and only if there exists a probability space $(\Omega, \mathcal{F}, P)$ supporting a family of r.v.'s $\{X', X'_n : n \geqslant 0\}$ such that:*

(i) $X_n \overset{\mathrm{d}}{=} X'_n$ *(i.e. X_n and X'_n have the same distribution for each $n \geqslant 0$).*

(ii) $X \overset{\mathrm{d}}{=} X'$ *(i.e. X and X' have the same distribution).*

(iii) $X'_n \to X'$ *a.s. as $n \to \infty$ on $(\Omega, \mathcal{F}, P)$.*

To gain some insight into Proposition 1, observe that if one is given weak convergence of X_n to X, then one may construct the r.v.'s X'_n, X' in the following way. Let $(\Omega, \mathcal{F}, P)$ support uniform distribution on $[0, 1]$ and set

$$X'_n = F_n^{-1}(U)\,, \qquad X' = F^{-1}(U)\,,$$

where U is a uniform r.v., $F_n^{-1}(x) = \sup\{y : F_n(y) \leqslant x\}$, $F^{-1}(x) = \sup\{y : F(y) \leqslant x\}$. It is well known that $X'_n \overset{\mathrm{d}}{=} X_n$, $X' \overset{\mathrm{d}}{=} X$. Furthermore, it is easily verified that if $X_n \xrightarrow{\mathrm{w}} X$, then $F_n^{-1}(U) \to F^{-1}(U)$ a.s. as $n \to \infty$, yielding (iii). To start from (i)–(iii) and obtain weak convergence is even more trivial.

Proposition 1 indicates the approach that one needs to take in order to describe weak convergence of stochastic processes. Examining the proposition, everything generalizes easily from the r.v. context to the process setting (i.e. the setting in which the random elements X_n, X are random functions) except (iii). But (iii) shows that in order to describe weak convergence in a function space setting, it is first necessary to describe a notion of convergence in function space, so that $X_n' \to X'$ makes sense.

The most natural way to do this is to define a metric d on the function space, so that $X' \to X'$ means $d(X_n', X') \to 0$. We start by describing the function space. Most operations research applications involve the study of stochastic processes having sample paths that are right continuous with left limits. This leads us to consider here the function space

$$D_E[0,\infty) = \{\omega\colon [0,\infty) \to E \text{ such that } \omega(\cdot) \text{ is right continuous for every } t \geq 0 \text{ and has left limits at every } t > 0\}\,.$$

We shall further require the range E to be Euclidian space. This function space is well suited, for example, to the study of queueing processes, in which the process might correspond to the queue-length vector at time t in a queueing network. We turn now to a description of a metric d which is suitable for the space $D_E[0,\infty)$.

Any 'good' metric d on the space $D_E[0,\infty)$ should be able to deal with the following two examples:

(a) Suppose that for all $T > 0$, $\|x_n - x\|_T \to 0$ as $n \to \infty$, where $\|\cdot\|_T$ is the uniform norm on [0, T] defined by

$$\|z\|_T = \sup\{|z(t)|\colon 0 \leq t \leq T\}\,.$$

Then, it should follow that $d(x_n, x) \to 0$ as $n \to \infty$.

(b) Given that $x_n(t) = I(t < 1 - 1/n)$, $x(t) = I(t < 1)$, x_n should converge to x in the metric d, in the sense that $d(x_n, x) \to 0$ as $n \to \infty$.

Example (a) states that if x_n converges to x 'nicely' (i.e. in the topology of uniform convergence on compact sets), this should imply convergence in the metric d. Example (b) requires that the metric d should be flexible enough to deal with the difficulties that arise in describing closeness of discontinuous functions. (Note that in (b), $\|x_n - x\|_T = 1$ for $T \geq 2$.)

The Skorohod metric d on $D_E[0,\infty)$ has these properties and is therefore well adapted to the study of $D_E[0,\infty)$. In order to deal with problem (b) above, one observes that $x_n = x \circ \lambda_n$ where $\lambda_n(t) = t + 1/n$. So, x_n is merely x in which the time scale is changed from $\lambda(t) = t$ to λ_n. Thus, the Skorohod metric judges two functions x and y to be 'close' if there exists a 'small' time deformation of y such that the time-deformed y is 'close' to x. See Ethier and Kurtz (1986, pp. 116–122) for a rigorous description of the Skorohod metric.

Assume that we have a family of processes $\{X, X_n\colon n \geq 0\}$, for which the sample functions live in $D_E[0,\infty)$ (i.e. $X \in D_E[0,\infty)$, $X_n \in D_E[0,\infty)$). Following Proposition 1, we adopt the following definition of weak convergence of X_n to X.

Definition 1. If X_n, $X \in D_E[0,\infty)$, we say that X_n converges weakly to X (written $X_n \xrightarrow{\mathrm{w}} X$) if there exists a probability space $(\Omega, \mathcal{F}, P)$ supporting a family of r.v.'s $\{X', X_n' \colon n \geq 0\}$ such that:

(i) $X_n \overset{\mathrm{d}}{=} X_n'$, $X \overset{\mathrm{d}}{=} X'$.

(ii) $d(X_n', X') \to 0$ a.s. as $n \to \infty$, where d is the Skorohod metric on $D_E[0,\infty)$.

We remark that the family $\{X', X_n' \colon n \geq 0\}$ is known as the *Skorohod representation* of $\{X, X_n \colon n \geq 0\}$.

The definition of weak convergence or $D_E[0,\infty)$ can be recast in a somewhat different form. Consider a function $h : D_E[0,\infty) \to D_E[0,\infty)$. The function h is said to be *continuous at* x if $d(h(x_n), h(x)) \to 0$ as $n \to \infty$ whenever $d(x_n, x) \to 0$. Let $S_h = \{x \in D_E[0,\infty) \colon h$ is continuous at $x\}$. Observe that if $d(X_n', X') \to 0$ a.s. with $P\{X' \in S_h\} = 1$, then $d(h(X_n'), h(X')) \to 0$ a.s., which in turn implies that $h(X_n) \xrightarrow{\mathrm{w}} h(X)$. We therefore obtain the following proposition, known as the *continuous mapping principle.*

Proposition 2. *Let S_h be the set of continuity points of a map $h : D_E[0,\infty) \to D_E[0,\infty)$. If $X_n \xrightarrow{\mathrm{w}} X$ as $n \to \infty$ and $P\{X \in S_h\} = 1$, then $h(X_n) \xrightarrow{\mathrm{w}} h(X)$ as $n \to \infty$.*

As we shall see later, the continuous mapping principle has a wealth of applications. A second variant of the continuous mapping principle is also useful. A function $h : D_E[0,\infty) \to \mathbb{R}^d$ is said to be continuous at $x \in D_E[0,\infty)$ if $\|h(x_n) - h(x)\| \to 0$ ($\|\cdot\|$ is Euclidian norm on $\mathbb{R}^d$) whenever $d(x_n, x) \to 0$. Again, set $S_h = \{x \in D_E[0,\infty) \colon h$ is continuous at $x\}$. The following result is the analog of Proposition 2 for this new class of h's.

Proposition 3. *Let S_h be the set of continuity points of the map $h : D_E[0,\infty) \to \mathbb{R}^d$. If $X_n \xrightarrow{\mathrm{w}} X$ as $n \to \infty$ and $P\{X \in S_h\} = 1$, then $h(X_n) \xrightarrow{\mathrm{w}} h(X)$ as $n \to \infty$. (The weak convergence of $h(X_n)$ is standard weak convergence on $\mathbb{R}^d$.)*

Because of the importance of this result, it is convenient to have simpler criteria for verifying that $P\{X \in S_h\} = 1$ for a given map h. Recall that in the diffusion approximation setting, the limit process X generally is a diffusion process having continuous paths. Therefore, in order to show that $P(X \in S_h\} = 1$, it suffices to prove that h is continuous at each continuous function $x \in D_E[0,\infty)$. The following proposition is useful in verifying the appropriate condition.

Proposition 4. *Suppose $x \in C_E[0,\infty)$, the space of continuous functions $x : [0,\infty) \to E$. If $x_n \in D_E[0,\infty)$, then $d(x_n, x) \to 0$ as $n \to \infty$ is equivalent to requiring that for each $T > 0$, $\|x_n - x\|_T \to 0$ as $n \to \infty$.*

Thus, if it is known that the limit process $X \in C_E[0, \infty)$, the validity of $P\{X \in S_h\} = 1$ may be verified by showing that $\|h(x_n) - h(x)\| \to 0$ as $n \to \infty$ for each family $\{x, x_n : n \geq 0\}$ such that:

(i) $x \in C_E[0, \infty)$.

(ii) For each $T > 0$, $\|x_n - x\|_T \to 0$ as $n \to \infty$.

As a consequence of this observation, it is often unnecessary to deal explicitly with the Skorohod metric d; instead, one can work with the more easily manipulated topology of uniform convergence on compact sets.

Specializing Proposition 3 to $d = 1$, Proposition 3 shows that if $X_n \xrightarrow{w} X$ in $D_E[0, \infty)$, then $h(X_n) \xrightarrow{w} h(X)$ in $\mathbb{R}$ whenever $h : D_E[0, \infty) \to \mathbb{R}$ is continuous (i.e. $S_h = D_E[0, \infty)$). The bounded convergence theorem thus implies that if $X_n \xrightarrow{w} X$, then

$$Eh(X_n) \to Eh(X) \tag{2.4}$$

for all bounded continuous $h : D_E[0, \infty) \to \mathbb{R}$. The following theorem states that (2.4) is in fact equivalent to weak convergence in $D_E[0, \infty)$; see Ethier and Kurtz (1986) for further discussion.

Theorem 1. *Suppose X_n, $X \in D_E[0, \infty)$ for $n \geq 0$. Then $X_n \xrightarrow{w} X$ as $n \to \infty$ if and only if $Eh(X_n) \to Eh(X)$ as $n \to \infty$ for all bounded continuous functions $h : D_E[0, \infty) \to \mathbb{R}$.*

In most references on weak convergence of stochastic processes, the reformulation of weak convergence suggested by Theorem 1 is in fact taken as the definition of weak convergence.

3. Verification criteria for weak convergence of stochastic processes

In Section 2, the notion of weak convergence of stochastic processes was made precise. Our goal in this section is to give some idea of what is involved mathematically in proving that a limit theorem $X_n \xrightarrow{w} X$ holds.

We first observe that the projection map $\pi_{t_1, \ldots, t_m} : D_E[0, \infty) \to \mathbb{R}^{md}$ (recall that $E = \mathbb{R}^d$) is continuous at every $x \in C_E[0, \infty)$, where

$$\pi_{t_1, \ldots, t_m}(x) = (x(t_1), \ldots, x(t_m)) .$$

Applying the continuous mapping principle expressed by Proposition 3, we obtain the following result: If $X \in C_E[0, \infty)$ and $X_n \xrightarrow{w} X$, then for every collection $t_1, \ldots, t_n$ of time indices, it must be that

$$(X_n(t_1), \ldots, X_n(t_m)) \xrightarrow{w} (X(t_1), \ldots, X(t_n)) \tag{3.1}$$

as $n \to \infty$. Relation (3.1) asserts that if the limit process X has continuous paths (as is the case for a diffusion limit process), then the finite-dimensional distributions of X_n must converge weakly to those of X.

One might expect that (3.1) is also sufficient to guarantee weak convergence of X_n to X, but this is not so, as the following example illustrates.

Example 1. For U uniform on $[0, 1]$, set

$$X_n(t) = \exp(-n(t - U)^2)$$

for $t \geq 0$. Note that for $0 \leq t < \cdots < t_m$,

$$(X_n(t_1), \ldots, X_n(t_m)) \overset{\mathrm{w}}{\to} (X(t_1), \ldots, X(t_m))$$

where $X(t) = 0$ for $t \geq 0$. Hence, the finite-dimensional distributions of X_n converge to those of X, where X_n, $X \in C_E[0, \infty)$. On the other hand, X_n does not converge weakly to X. To see this, consider $h(x) = \max\{|x(t)|: 0 \leq t \leq 1\}$, and observe that h is continuous at any $x \in C_E[0, \infty)$. Since $X \in C_E[0, \infty)$, the continuous mapping principle would assert that $h(X_n) \overset{\mathrm{w}}{\to} h(X)$ if $X_n \overset{\mathrm{w}}{\to} X$. However, $h(X_n) = 1$, whereas $h(X) = 0$, contradicting the weak convergence of X_n to X.

Thus, the notion of weak convergence in $D_E[0, \infty)$ requires something more than weak convergence of the corresponding finite-dimensional distributions. To argue that $X_n \overset{\mathrm{w}}{\to} X$ involves the following circle of ideas.

Definition 2. Given a family $\{P, P_n: n \geq 0\}$ of probability measures on $\Omega = D_E[0, \infty)$, we say that P_n *converges weakly* to P (written $P_n \overset{\mathrm{w}}{\to} P$) if

$$\int_\Omega h(\omega) P_n(\mathrm{d}\omega) \to \int_\Omega h(\omega) P(\mathrm{d}\omega)$$

for all bounded continuous functions $h : \Omega \to \mathbb{R}$.

By Theorem 1, $X_n \overset{\mathrm{w}}{\to} X$ if and only if $P_n \overset{\mathrm{w}}{\to} P$, where $P_n(\cdot) = P\{X_n \in \cdot\}$ and $P(\cdot) = P\{X \in \cdot\}$. It turns out that weak convergence of the probability measures P_n to P can be formulated in terms of convergence in a certain metric ρ. To be precise, let $\mathcal{P} = \{Q : Q$ is a probability measure on $D_E[0, \infty)\}$. There exists a metric ρ on $\mathcal{P}$ such that $P_n \overset{\mathrm{w}}{\to} P$ as $n \to \infty$ is equivalent to requiring that $\rho(P_n, P) \to 0$ as $n \to \infty$; the metric ρ is called the *Prohorov metric*. Thus, the notion of weak convergence has been recast in terms of convergence in a certain metric ρ.

A standard approach to proving that $x_n \to x$ when the x_n's and x are elements of a metric space is to proceed via the following two steps:

(i) Show that the sequence $\{x_n: n \geq 0\}$ is relatively compact, in the sense that every subsequence $x_{n'}$ has a further convergent subsequence $x_{n''}$.

(ii) Show that every convergent subsequence $x_{n'}$ of x_n must converge to x. The first step shows that every subsequence has a limit point, whereas the second step proves that the only possible limit point of $\{x_n: n \geqslant 0\}$ is x. Thus, the second step involves identifying the set of all possible limit points. Returning to the space $\mathcal{P}$, we wish to obtain a condition for identifying the set of limit points of a sequence $\{P_n: n \geqslant 0\}$.

Theorem 2. *Consider $\{P, P_n: n \geqslant 0\}$ where $P_n(\cdot) = P\{X_n \in \cdot\}$, $P(\cdot) = P\{X \in \cdot\}$ and X_n, $X \in D_E[0, \infty)$. If the finite-dimensional distributions of X_n converge to those of X (i.e. if (3.1) holds), then the only possible limit point of $\{P_n: n \geqslant 0\}$ is P.*

Thus, Theorem 2 shows that convergence of the finite-dimensional distributions is important in identifying the set of possible limit points of P_n.

Returning to step one of the convergence proof outlined, we need to obtain criteria guaranteeing relative compactness of a sequence $\{P_n: n \geqslant 0\}$ of probability measures in $\mathcal{P}$. The following theorem, due to Prohorov, throws the question of compactness in $\mathcal{P}$ back into determining compactness in $D_E[0, \infty)$.

Theorem 3. *Consider $\{P_n: n \geqslant 0\}$ where $P_n \in P$. Then, $\{P_n: n \geqslant 0\}$ is relatively compact in ρ (i.e. for every subsequence n', there exists a further subsequence n'' and a probability measure $P'' \in \mathcal{P}$ such that $\rho(P_{n''}, P'') \to 0$ as $n'' \to \infty$) if and only if for every $\varepsilon > 0$, there exists a compact set $K_\varepsilon \subseteq D_E[0, \infty)$ (i.e. compact in the metric d on $D_E[0, \infty)$) such that*

$$\inf_{n \geqslant 0} P_n(K_\varepsilon) \geqslant 1 - \varepsilon . \tag{3.2}$$

Because of the obvious importance of the notion (3.2), it has received a name.

Definition 3. A family $\{P_n \in \mathcal{P}: n \geqslant 0\}$ of probability measures on $D_E[0, \infty)$ is said to be *tight* if for every $\varepsilon > 0$, there exists a compact set $K_\varepsilon \subseteq D_E[0, \infty)$ such that $\inf_{n \geqslant 0} P_n(K_\varepsilon) \geqslant 1 - \rho$.

We can now state the conventional 'first-principles' approach used to prove a limit theorem of the form $X_n \xrightarrow{w} X$ when $X_n \in D_E[0, \infty)$, $X \in C_E[0, \infty)$. One first shows that the family $\{P_n: n \geqslant 0\}$ is tight, where $P_n(\cdot) = P\{X_n \in \cdot\}$, followed by proving that the finite-dimensional distributions of X_n converge to those of X. Because of the important role of tightness in proving limit theorems in $D_E[0, \infty)$, the characterization of the compact sets in $C_E[0, \infty)$ and $D_E[0, \infty)$ occupies a central place in the corresponding limit theory.

However, in many operations research settings, one can avoid the technical complications associated with the above 'first-principles' argument by making use of the continuous mapping principle. Suppose that one wishes to show that $X_n \xrightarrow{w} X$ and that one can represent X_n, X as $X_n = h(Y_n)$, $X = h(Y)$ for some function h such that $P\{Y \in S_h\} = 1$. Then, the weak convergence of X_n to X follows immediately if it is known (either by assumption or by previously developed theory) that $Y_n \xrightarrow{w} Y$. We will see this 'continuous mapping' approach illustrated in the queueing example developed in Section 5.

4. Donsker's theorem

We shall now describe the most important functional limit theorem in the theory of stochastic processes. Let $\{Z_n: n \geqslant 0\}$ be an i.i.d. sequence of $\mathbb{R}^d$-valued r.v.'s and set $S_n = Z_1 + \cdots + Z_n$ with $S_0 = 0$. Donsker's theorem is a limit theorem that describes the behavior of the $\mathbb{R}^d$-valued random walk $\{S_n: n \geqslant 0\}$ over long time intervals; this theorem can be viewed as a 'building block' for many of the other limit theorems developed in probability theory.

We first review some of the classical limit theory for the random walk $\{S_n: n \geqslant 0\}$. We start with the strong law of large numbers, which states that if $E\|Z_n\| < \infty$, then

$$n^{-1}S_n \to \mu \quad \text{a.s.} \tag{4.1}$$

as $n \to \infty$, where $\mu = EZ_n$. The law of large numbers (4.1) can be refined by appealing to the central limit theorem, which asserts that if $E\|Z_n\|^2 < \infty$, then

$$n^{1/2}(n^{-1}S_n - \mu) \xrightarrow{w} \Sigma^{1/2}\mathrm{N}(0, I) \tag{4.2}$$

where $\mathrm{N}(0, I)$ is an $\mathbb{R}^d$-valued multivariate normal r.v. with mean vector 0 and the identity as covariance matrix, and $\Sigma^{1/2}$ is the square root of the covariance matrix $\Sigma = EZ_n^{\mathrm{t}}Z_n - EZ_n^{\mathrm{t}} \cdot EZ_n$. (We assume here that Z_n is a row vector.)

The idea is now to look for process-valued versions of (4.1) and (4.2). First, consider $\bar{X}_n(t) = n^{-1}S_{[nt]}$, where $[x]$ denotes the greatest integer less than or equal to x. It can be shown that the limit theorem (4.1) implies that

$$d(\bar{X}_n, \bar{X}) \to 0 \quad \text{a.s.} \tag{4.3}$$

as $n \to \infty$, where $\bar{X}(t) = \mu t$; (4.3) is called the *functional strong law of large numbers*.

To obtain a functional form of the central limit theorem, we use the same scaling as in (4.2) and consider the stochastic process

$$X_n(t) = n^{1/2}(n^{-1}S_{[nt]} - \mu t) = n^{1/2}(\bar{X}_n(t) - \bar{X}(t)) .$$

Note that one unit of time in the process X_n corresponds to n time units of the random walk, and that one spatial unit of X_n is equivalent to $n^{1/2}$ spatial units of the random walk.

To identify the limit behavior of X_n, note that (4.2) proves that

$$X_n(t) \xrightarrow{\mathrm{w}} \Sigma^{1/2}\mathrm{N}(0, t \cdot I) \tag{4.4}$$

for each $t \geq 0$. Furthermore, the stationary independent increments of X_n allows one to extend (4.4) to the finite-dimensional distributions of X_n: For $0 \leq t_1 < t_2 < \cdots < t_n$,

$$(X_n(t_1), \ldots, X_n(t_n)) \xrightarrow{\mathrm{w}} (X(t_1), \ldots, X(t_m))$$

as $n \to \infty$, where $(X(t_1), \ldots, X(t_m))$ is a Gaussian (i.e. multivariate normal) random vector with mean and covariance described by

$$EX(t_i) = 0, \quad 1 \leq i \leq m, \qquad EX^{\mathrm{t}}(t_i)X(t_j) = \Sigma \min(t_i, t_j). \tag{4.5}$$

The only process supported on $D_E[0, \infty)$ with finite-dimensional distributions as described by (4.5) is the Brownian motion process. Thus, if $X \in D_E[0, \infty)$, it must be that $X \overset{\mathrm{d}}{=} \Sigma^{1/2}B$, where $B(\cdot)$ is a standard Brownian motion process on $\mathbb{R}^d$ having the following properties:

(i) $B(\cdot) \in C_E[0, \infty)$.

(ii) $B(\cdot)$ has stationary independent increments.

(iii) $B(t) \overset{\mathrm{d}}{=} \mathrm{N}(0, t \cdot I)$.

The following result, known both as Donsker's theorem and the *functional central limit theorem* (FCLT), is a precise statement of the limit behavior of the process $\{X_n : n \geq 0\}$.

Theorem 4. *If $E\|Z_n\|^2 < \infty$, then $X_n \xrightarrow{\mathrm{w}} \Sigma^{1/2}B$ as $n \to \infty$ in $D_E[0, \infty)$.*

We remark that Theorem 4 is the prototypical diffusion approximation, in that $\Sigma^{1/2}B$ is a diffusion process. (In fact, it is fair to say that B is the most fundamental diffusion process.) Theorem 4 illustrates several important features that are common to diffusion approximations. First, observe that the approximation suggested by the limit theorem takes the form

$$S_{[nt]} \overset{\mathrm{d}}{\approx} \mu nt + n^{1/2}\Sigma^{1/2}B(t).$$

Thus, the limit process that is used to approximate $S_{[nt]}$ depends on the random walk only through the mean vector $\mu = EZ_n$ and the covariance matrix Σ of Z_n.

Secondly, a great deal of information about X_n may be inferred from Theorem 4 by taking advantage of the continuous mapping principle. We illustrate the power of this idea by offering the following examples.

Example 2. Assume the random walk is real-valued (i.e. $d = 1$) with $\mu = 0$ and consider the following functional of $\{S_n : n \geq 0\}$: $\max\{S_k : 0 \leq k \leq n\}$. We observe that this functional may be expressed in terms of an appropriately chosen mapping h defined on X_n namely

$$\max\{S_k: 0 \leqslant k \leqslant n\} = n^{1/2} h(X_n)$$

where $h(x) = \max\{x(t): 0 \leqslant t \leqslant 1\}$. By applying Proposition 4, it is easily verified that h is continuous at any $x \in C_E[0, \infty)$. Thus, since Brownian motion has continuous paths, we conclude that $P\{X \in S_h\} = 1$, where $X = \sigma B$ and $\sigma^2 = \text{var } Z_n$. Hence, by Proposition 3 and Theorem 4, $h(X_n) \overset{\text{w}}{\rightarrow} h(\sigma B)$ and we obtain the approximation

$$\max\{S_k: 0 \leqslant k \leqslant n\} \overset{\text{d}}{\approx} \sigma n^{1/2} \max\{B(t): 0 \leqslant t \leqslant 1\}\,.$$

But by using the 'reflection principle' (see Karlin and Taylor, 1975), one can explicitly calculate the distribution of $h(B)$, namely

$$P\{\max\{B(t): 0 \leqslant t \leqslant 1\} < x\} = \sqrt{\frac{2}{\pi}} \int_x^\infty \exp(-\tfrac{1}{2}t^2)\,\mathrm{d}t\,. \tag{4.6}$$

Example 3. For the same random walk as in Example 2, consider $\#\{k: 0 \leqslant k \leqslant n,\ S_k \geqslant 0\}$. Observe that

$$\#\{k: 0 \leqslant k \leqslant n,\ S_k \geqslant 0\} = nh(X_n)$$

where

$$h(x) = \int_0^1 I(x(t) \geqslant 0)\,\mathrm{d}t\,.$$

Again, one can easily verify that $P\{\sigma B \in S_h\} = 1$, by which we may conclude that $h(X_n) \overset{\text{w}}{\rightarrow} h(\sigma B) = h(B)$. Again, it is possible, using properties of Brownian motion, to explicitly calculate the distribution of $h(B)$:

$$P\{h(B) \leqslant x\} = \frac{2}{\pi} \arcsin x^{1/2}\,, \quad 0 \leqslant x \leqslant 1\,.$$

This limit theorem comprises the well known 'arcsin law' for random walk.

Example 4. For the random walk $\{S_n: n \geqslant 0\}$ of Example 2, consider the hitting time

$$T(a) = \min\{k \geqslant 0: S_k \geqslant a\}\,, \quad a > 0\,.$$

To analyze the distribution of the hitting time, observe that

$$\{T(a) > n\} = \{\max\{S_k: 0 \leqslant k \leqslant n\} < a\}\,.$$

Hence,

$$\{T(an^{1/2}) > [xn]\} = \{\max\{S_k: 0 \leq k \leq [nx]\} < an^{1/2}\}$$

and one can appeal to Example 2 to show that

$$P\{T(an^{1/2}) > [xn]\} \rightarrow P\{\max\{B(t): 0 \leq t \leq x\} < a/\sigma\} ,$$

which can be calculated from (4.6).

Example 5. This example will serve to illustrate the point made at the end of Section 3, namely that many functional limit theorems can be derived by judicious use of the continuous mapping principle, thereby avoiding a 'first principles' argument. For the random walk of Example 2, set

$$\tilde{S}_n(t) = \begin{cases} S_{[nt]} - tS_n , & 0 \leq t \leq 1 , \\ 0 , & t > 1 , \end{cases} \qquad \tilde{X}_n(t) = n^{-1/2}\tilde{S}_n(t) .$$

Note that $\tilde{X}_n = h(X_n)$, where $h : D_E[0,\infty) \rightarrow D_E[0,\infty)$ is given by $h(x)(t) = x(t) - tx(1)$ for $0 \leq t \leq 1$ and $h(x)(t) = 0$ for $t > 1$. It is easily verified, using Proposition 4, that h is continuous at any $x \in C_E[0,\infty)$, so that Proposition 2 allows us to conclude that $h(X_n) \xrightarrow{w} h(\sigma B)$. We remark that the process $B(t) - tB(1)(0 \leq t \leq 1)$ is the so-called *Brownian bridge* process (also known as 'tied-down' Brownian motion).

An additional remark, related to Theorem 4, is in order. For an appropriately continuous h, the continuous mapping principle proves that $h(X_n) \xrightarrow{w} h(\sigma B)$ as n$\rightarrow \infty$. One way to calculate the distribution of $h(\sigma B)$ is to use this limit theorem to observe that the limit random element $h(\sigma B)$ is invariant with respect to the particular choice of the random walk $\{S_n: n \geq 0\}$ (modulo $\sigma^2 = \text{var } Z_n$). Thus, one may calculate the distribution of $h(\sigma B)$ by choosing a particularly simple random walk (symmetric nearest neighbor random walk on the integers is a typical choice), and taking the limit of $h(X_n)$ for the chosen random walk. This approach to calculating the distribution of $h(\sigma B)$ explains why Theorem 4 is often called the *invariance principle*.

A number of important extensions to the basic FCLT described by Theorem 4 have been made over the years. The flavor of most of these extensions has been to prove limit theorems for the partial sum process in which the summand sequence $\{Z_n: n \geq 0\}$ is no longer i.i.d., but instead allows for some kind of dependence and (possibly) some mild form of non-stationarity. Since partial sum processes often serve as 'inputs' to various stochastic processes arising in operations research (e.g. if the Z_i's correspond to inter-arrival times in a queue, the S_n's are the arrival times), we will now describe some of the extensions that are available. We specialize to the setting in which the Z_i's are real-valued, to avoid complications in stating results, but note that similar limit theorems hold in the $\mathbb{R}^d$-valued vector context.

The generic form of the FCLT involves a statement that, for a given (real-valued) sequence $\{Z_n: n \geq 0\}$, there exist finite constants μ and σ such that

$$X_n \xrightarrow{w} \sigma B \tag{4.7}$$

in $D_E[0, \infty)$, where

$$X_n(t) = n^{1/2}(n^{-1}S_{[nt]} - \mu t) .$$

The extensions of the FCLT that are available include:

(a) $\{Z_n : n \geq 0\}$ is a *stationary mixing* process. Here, $\{Z_n: n \geq 0\}$ is a strictly stationary sequence which satisfies a mixing condition. Roughly speaking, a mixing hypothesis states that events which occur at widely separated time points are asymptotically independent of one another. This is generally stated mathematically as a condition of the form

$$P\{Z_0^m \in A, Z_{m+n}^{m+n+r} \in B\} \to P\{Z_0^m \in A\}P\{Z_0^r \in B\}$$

as $n \to \infty$, where $Z_i^j = (Z_i, \ldots, Z_j)$ for $i \leq j$. Assuming that $EZ_n^2 < \infty$ and that $\Sigma_{k=0}^{\infty} |\text{cov}(Z_0, Z_k)| < \infty$, the relevant FCLT's yield limit theorems of the form (4.7), where μ, σ are given by

$$\mu = EX_n , \qquad \sigma^2 = \text{var}\, Z_0 + 2 \sum_{k=1}^{\infty} \text{cov}(Z_0, Z_k) .$$

For a more precise formulation of these results, see Ethier and Kurtz (1986, pp. 350–353).

(b) $\{S_n: n \geq 0\}$ is a *martingale* sequence with stationary martingale differences. (See Chapter 3.) If $\{S_n: n \geq 0\}$ is a martingale sequence for which the differences $\{D_n: n \geq 1\}$ $(D_n = S_n - S_{n-1})$ are strictly stationary, then (4.7) holds if $ED_n^2 < \infty$ and μ, σ are given by $\mu = 0$, $\sigma^2 = ED_n^2$.

(c) $\{Z_n: n \geq 0\}$ is a *regenerative* sequence with regeneration times $0 \leq T_0 < T_1 < \cdots$. If

$$E(T_1 - T_0)^2 < \infty , \qquad E\left(\sum_{n=T_0}^{T_1-1} |Z_n|\right)^2 < \infty ,$$

then (4.7) is valid with

$$\mu = E \sum_{n=T_0}^{T_1-1} Z_n \Big/ E(T_1 - T_0)$$

and

$$\sigma^2 = E\left(\sum_{n=T_0}^{T_1-1} (Z_n - \mu)\right)^2 \Big/ E(T_1 - T_0) ;$$

see Glynn and Whitt (1987).

(d) $\{Z_n: n \geqslant 0\}$ is a real-valued functional of a time-homogeneous positive recurrent *Markov chain*. Thus, $Z_n = f(W_n)$ where $\{W_n: n \geqslant 0\}$ is a time-homogeneous Markov chain satisfying a suitable recurrence condition; let $\pi(\cdot)$ be the associated stationary distribution of $\{W_n: n \geqslant 0\}$. We remark that there is no need here for W_n to be discrete-valued. Under certain regularity hypotheses (see Nummelin, 1984 for a precise description of the results), (4.7) holds with

$$\mu = E_\pi f(W_n) ,$$

$$\sigma^2 = E_\pi(f(W_0) - \mu)^2 + 2 \sum_{k=1}^{\infty} E_\pi(f(W_0) - \mu)(f(W_k) - \mu)$$

($E_\pi(\cdot)$ is the expectation on the path space of $\{W_n: n \geqslant 0\}$ under which W_0 has distribution π.)

The above FCLT's are an indication of the basic robustness of the Brownian motion approximation for the partial sum process. They indicate that, regardless of the fine structure of the summands, that partial sum processes over long time intervals behave (when suitably normalized) like Brownian motions.

We conclude this section with the discussion of a very important and powerful theorem which states, in a very strong sense, that random walk is well approximated by Brownian motion. The type of theorem that we shall discuss here is called a *strong approximation theorem*, since it is phrased not in a distributional sense (as in (4.7)) but in an 'almost sure' sense.

The following result, which is due to Komlós, Major and Tusnady (1975), is a particularly sharp form of the strong approximation theorem for random walk.

Theorem 5. *Let $\{Z_n: n \geqslant 1\}$ be a sequence of i.i.d. real-valued r.v.'s satisfying $E\exp(\alpha Z_n) < \infty$ for α in an open neighborhood of zero. Set $S_0 = 0$ and $S_n = Z_1 + \cdots + Z_n$ for $n \geqslant 1$. Then there exists a probability space $(\Omega, \mathscr{F}, P)$ supporting a standard Brownian motion $\{B(t): t \geqslant 0\}$ and a sequence $\{S'_n: n \geqslant 0\}$ such that:*

(i) *$\{S'_n: n \geqslant 0\} \overset{\mathrm{d}}{=} \{S_n: n \geqslant 0\}$ (i.e. the sequence $\{S'_n\}$ shares the same distribution as $\{S_n\}$).*

(ii) *For every x and n,*

$$P\left\{\max_{0 \leqslant k \leqslant n} |S'_k - k\mu - \sigma B(k)| > C \log n + x\right\} < K \mathrm{e}^{-\lambda x}$$

where $\mu = EZ_n$, $\sigma^2 = \operatorname{var} Z_n$, and K, C and λ are positive constants depending only on the distribution of Z_n.

It is easily verified (use the Borel–Cantelli lemma) that (ii) implies that

$$S'_n = n\mu + \sigma B(n) + \mathrm{O}(\log n) \quad \text{a.s.} \tag{4.8}$$

(The sequence $\mathrm{O}(\log n)$ represents a sequence $\{R_n: n \geq 0\}$ of r.v.'s such that $|R_n| \leq A \log n + B$ where A, B are finite-valued r.v.'s.) This result is sharp, in the sense that if $S'_n = n\mu + \sigma B(n) + \mathrm{o}(\log n)$ a.s., then Z_n has a $\mathrm{N}(\mu, \sigma^2)$ distribution so that the $\mathrm{o}(\log n)$ term can be taken to be zero.

It is easily verified that (4.8) implies Theorem 4. Set

$$X'_n(t) = n^{1/2}(n^{-1}S'_{[nt]} - \mu)$$

and note that by (i), $X'_n \overset{\mathrm{d}}{=} X_n$. By (4.8) and basic properties of Brownian motion,

$$\|X'_n(\cdot) - n^{-1/2}\sigma B(n\,\cdot)\|_T = \mathrm{O}(\log n/n^{1/2}) \quad \text{a.s.} \tag{4.9}$$

for any $T > 0$. But $n^{-1/2}B(n\,\cdot) \overset{\mathrm{d}}{=} B(\cdot)$. Hence, Theorem 4 follows from (4.9) by letting $n \to \infty$.

One important application of Theorem 5 is to (easily) obtain rates of convergence for various limit theorems related to Donsker's theorem. Consider Example 2, in which it is shown that $h(X_n) \xrightarrow{\mathrm{w}} h(\sigma B)$ where $h(x) = \max\{x(t): 0 \leq t \leq 1\}$. By (i), $h(X_n) \overset{\mathrm{d}}{=} h(X'_n)$. Straightforward analysis, using (ii), then proves that

$$P\{h(X'_n) \leq x\} = P\{h(\sigma B) \leq x\} + \mathrm{O}(\log n/n^{1/2})\,.$$

As in the case of Theorem 4, there exist extensions of strong approximation results to dependent sequences; a good reference for such theorems is Philipp and Stout (1975).

5. Weak convergence theorems for the single-server queue

In this section, we indicate how the weak convergence theory of Sections 2 through 4 can be applied to obtain limit theorems for the single-server queue. To be precise, we consider a single GI/G/1/∞ queueing system in which customer 0 arrives at time $T_0 = 0$, finds a free server, and experiences a service time V_0. The nth customer arrives at time T_n and experiences a service time V_n. Setting $U_n = T_n - T_{n-1}$ $(n \geq 1)$, we assume that each of the sequences $\{U_n: n \geq 1\}$ and $\{V_n: n \geq 0\}$ are i.i.d. and independent of one another. The three processes that we shall consider in the section are:

W_n = waiting time (i.e. excluding service) of the nth customer;

$Q(t)$ = number of customers in the system (i.e. including the server) at time t;

$D(t)$ = the cumulative number of customers departing the system in $[0, t]$.

Finally, let $EU_n \equiv \lambda^{-1}$, $EV_n \equiv \mu^{-1}$, where $0 < \lambda$, $\mu < \infty$. We shall first discuss the relevant theory when the traffic intensity $\rho = \lambda/\mu < 1$. In this 'light traffic' setting, the queue is stable in the sense that $\{W_n: n \geq 0\}$ and $\{Q(t): t \geq 0\}$ are stochastically bounded. In fact, it can be shown that $W_n \xrightarrow{w} W$ as $n \to \infty$, where the characteristic function of W is given by

$$E\exp(\mathrm{i}tW) = \exp\left\{\sum_{n=1}^{\infty} n^{-1}[E(\exp(\mathrm{i}tS_n^+) - 1)\right\},$$

and

$$S_n = \sum_{i=0}^{n-1} V_i - \sum_{i=1}^{n} U_i \quad (n \geq 1)$$

with $S_0 = 0$ ($S_n^+ = \max(S_n, 0)$). Furthermore, if we additionally assume that U_n has a non-lattice distribution, then there exists a proper r.v. Q such that $Q(t) \xrightarrow{w} Q$ as $t \to \infty$. Given the well-behaved nature of W_n (and $Q(t)$), it is easy to show that there do not exist sequences $\{a_n\}$ and $\{b_n\}$ for which the random functions $\{(W_{[nt]} - a_n t)/b_n\}$ (and $\{(Q(nt) - a_n t)/b_n\}$) converge weakly to a non-degenerate process in $D_{\mathbb{R}}[0, \infty)$. This suggests that we should change our point of view, to consider the *cumulative* processes

$$\sum_{j=1}^{n} W_j \quad \text{and} \quad \int_0^t Q(s)\,\mathrm{d}s\,.$$

The analysis of these cumulative processes depends on the fact that both $\{W_n: n \geq 0\}$ and $\{Q(t): t \geq 0\}$ are regenerative; this allows us to apply the FCLT for regenerative processes (discussed in Section 4) to obtain limit behavior for the cumulative processes. To precisely state the relevant theorems, let $\eta = \inf\{n \geq 1: W_n = 0\}$ and let $T = \inf\{t > v_0: Q(t-) = 0, Q(t) = 1\}$; η and T are regeneration times for $\{W_n: n \geq 0\}$ and $\{Q(t): t \geq 0\}$ respectively.

Theorem 6. *Suppose $\rho < 1$. If*

$$EU_1^2 < \infty\,, \quad EV_0^2 < \infty\,, \quad E\eta^2 < \infty\,, \quad ET^2 < \infty\,,$$

$$E\left(\sum_{k=0}^{\eta-1} W_k\right)^2 < \infty \quad \textit{and} \quad E\left(\int_0^T Q(s)\,\mathrm{d}s\right)^2 < \infty\,,$$

then

$$n^{1/2}\left(\sum_{k=0}^{[nt]} n^{-1} W_k - tEW\right) \xrightarrow{w} \sigma_1 B(t)\,,$$

$$n^{1/2}\left(n^{-1}\int_0^{nt} Q(s)\,\mathrm{d}s - tEQ\right) \xrightarrow{w} \sigma_2 B(t)\,,$$

in $D_{\mathbb{R}}[0,\infty)$, *where*

$$\sigma_1^2 = E\Big(\sum_{k=0}^{\eta-1} (W_k - EW)\Big)^2 \Big/ E\eta$$

and

$$\sigma_2^2 = E\Big(\int_0^T (Q(s) - EQ)\,\mathrm{d}s\Big)^2 \Big/ ET\,.$$

Furthermore,

$$EW = E\sum_{k=0}^{\eta-1} W_k \Big/ E\eta \quad \textit{and} \quad EQ = E\int_0^T Q(s)\,\mathrm{d}s \Big/ ET\,.$$

Recall that when $\rho < 1$, the translation constants EW and EQ appearing in Theorem 6 are related via Little's formula: $EQ = \lambda EW$. A similar 'Little-type' result holds for the variance constants σ_1^2 and σ_2^2; for details, see Glynn and Whitt (1986).

Turning now to the departure process, we observe that $D(t) = N(t) - Q(t)$ where $N(t)$ = number of arrivals by time t. Note that $N(t)$ is essentially the renewal process corresponding to the sequence $\{U_n: n \geqslant 1\}$. To handle $N(t)$, we use a basic idea from the theory of weak convergence of stochastic processes, namely the method of 'random time change'. Evidently, $N(t)$, when viewed on the time scale of the T_n's is deterministic, by which we mean that $N(T_n) = n + 1$. Similarly, the sequence $\{T_n: n \geqslant 0\}$, when viewed through $N(\cdot)$'s time scale, is basically non-random, in that $T_{N(s)} \approx s$. Thus, $\{T_n\}$ and $\{N(t)\}$ are 'inverse' processes to one another. This important idea is central to the study of $N(\cdot)$.

To exploit these 'time change' ideas in the weak convergence setting, observe that a 'time change' corresponds to composition of processes. Thinking of Φ_n as a change in time scale (i.e. from t to $\Phi_n(t)$), we need to consider when $X_n \circ \Phi_n$ converges weakly to $X \circ \Phi$.

In the following propositions, we endow the product space with the product metric induced from the component spaces (i.e., a sequence converges in the product space if and only if the components converge in their respective spaces).

Proposition 5. *Suppose X_n is E-valued and $\Phi_n : [0,\infty) \to [0,\infty)$ is non-decreasing. If $(X_n, \Phi_n) \xrightarrow{w} (X, \Phi)$ in $D_E[0,\infty) \times D_{\mathbb{R}}[0,\infty)$ where $(X, \Phi) \in C_E[0,\infty) \times C_{\mathbb{R}}$, then $X_n \circ \Phi_n \xrightarrow{w} X \circ \Phi$ in $D_E[0,\infty)$.*

See Billingsley (1968, Section 17) for additional details. To establish the joint convergence in Proposition 5, the following tool is often used.

Proposition 6. *Suppose that $X_n' \in D_{E'}[0,\infty)$ and that there exists (deterministic) $x' \in D_{E'}[0,\infty)$ such that $d(X_n', x') \xrightarrow{w} 0$ in $\mathbb{R}$. (This is equivalent to $X_n' \xrightarrow{w} x'$ in*

$D_{E'}[0,\infty)$.) *If* $X_n \xrightarrow{w} X$ *in* $D_E[0,\infty)$, *then* $(X_n, X_n') \xrightarrow{w} (X, x')$ *in* $D_E[0,\infty) \times D_{E'}[0,\infty)$.

To apply these results to $N(\cdot)$, one first generalizes the strong law for renewal processes to a functional strong law:

$$\|n^{-1}N(nt) - \lambda t\|_T \xrightarrow{w} 0 \quad \text{a.s.}$$

as $n \to \infty$, for each $T > 0$. Since λt is deterministic, Proposition 6 and Donsker's theorem for the inter-arrival times yields the weak convergence relation

$$(n^{1/2}(n^{-1}T_{[ns]} - \lambda^{-1}s), n^{-1}N(ns)) \xrightarrow{w} (\sigma_A B(s), \lambda s)$$

where $\sigma_A^2 = \operatorname{var} U_n$; the random time change result provides

$$n^{1/2}(n^{-1}t_{N(ns)} - \lambda^{-1}n^{-1}N(ns)) \xrightarrow{w} \sigma_A B(\lambda s)\,.$$

Since $T_{N(ns)} \approx ns$, one obtains the following FCLT for $N(\cdot)$:

$$n^{1/2}(n^{-1}N(ns) - \lambda s) \xrightarrow{w} \sigma_A \lambda^{3/2} B(s) \tag{5.1}$$

(we have used the fact that $B(\alpha\,\cdot) \overset{d}{=} \alpha^{1/2}B(\cdot)$). To deal with the approximation $T_{N(ns)} \approx ns$, one uses the following 'converging-together' result.

Proposition 7. *If* $X_n \xrightarrow{w} X$ *in* $D_E[0,\infty)$ *and* $d(X_n, X_n') \xrightarrow{w} 0$ *in* $\mathbb{R}$, *then* $X_n' \xrightarrow{w} X$ *in* $D_E[0,\infty)$.

(For Propositions 6 and 7, we have implicitly assumed that X_n and X_n' are defined on the same probability space.) Propositions 5 through 7 follow directly from the Skorohod representation and the fact that the Skorohod topology reduces to the topology of uniform convergence when the limit elements are continuous. The ideas exploited above, namely those of 'continuous mapping', 'random time change', and 'convergence together', are the three main tools used in proving most of the diffusion approximations that arise in operations research. The limit theorem (5.1) is a particularly simple application of these ideas, but is nevertheless somewhat typical of how diffusion approximation limit theorems are proved in operations research.

We return now to the departure process $D(t)$. Since the queue-length process $Q(\cdot)$ is stochastically bounded, the next result follows from (5.1) and the 'converging together' proposition.

Theorem 7. *Suppose* $\rho < 1$ *and* $EU_n^2 < \infty$. *Then,*

$$n^{1/2}(n^{-1}D(nt) - \lambda t) \xrightarrow{w} \sigma_A \lambda^{3/2} B(t)$$

in $D_{\mathbb{R}}[0,\infty)$.

This concludes our discussion of the stable case in which $\rho < 1$; we turn now to analysis of the unstable GI/G/1/∞ queue, in which the traffic intensity $\rho > 1$. The crucial point here is that when a queue is oversaturated, there exists some (random) time β after which the server is never idle. One now observes that if $S(t)$ is the renewal counting process corresponding to the service times $V_0, V_1, \ldots$, then $S(t)$ can be interpreted as the total number of customers that would exit by t if the server were in constant operation. Thus, $D(t) \approx S(t)$ and $Q(t) \approx N(t) - S(t)$, from which the next result follows (use (5.1) and the independence of $N(\cdot)$, $S(\cdot)$).

Theorem 8. *Suppose* $\rho < 1$ *and* $EU_n^2 < \infty$, $EV_n^2 < \infty$. *Then*

$$n^{1/2}(n^{-1}D(nt) - \mu t) \xrightarrow{\mathrm{w}} \sigma_S \mu^{3/2} B(t)\,,$$

$$n^{1/2}(n^{-1}Q(nt) - (\lambda - \mu)t) \xrightarrow{\mathrm{w}} (\sigma_A^2 \lambda^3 + \sigma_S^2 \mu^3)^{1/2} B(t)\,,$$

in $D_{\mathbb{R}}[0, \infty)$, *where* $\sigma_S^2 = \operatorname{var} V_n$.

An immediate consequence of Theorem 8 is the following application of the continuous mapping principle: if $\sigma_A^2, \sigma_S^2 > 0$, then

$$P\{Q(n) \leq n(\lambda - \mu) + n^{1/2}(\sigma_A^2 \lambda^3 + \sigma_S^2 \mu^3)^{1/2} x\}$$
$$\rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\tfrac{1}{2}t^2)\, \mathrm{d}t$$

as $n \rightarrow \infty$. To handle the waiting times, we use the fact that the waiting time sequence $\{W_n : n \geq 0\}$ satisfies the recursion $W_{n+1} = [W_n + X_{n+1}]^+$ (with $W_0 = 0$), where $X_n = V_{n-1} - U_n$. Since the system is oversaturated, $W_n \rightarrow \infty$ a.s. and is hence positive with high probability for large n. Thus, $W_{n+1} \approx W_n + X_{n+1}$ (i.e. the positive part operator can be dropped). This implies that $W_{n+1} \approx S_{n+1}$, where $S_{n+1} = X_1 + \cdots + X_{n+1}$. A limit theorem for W_n can then be obtained from the FCLT for S_n.

Theorem 9. *Suppose* $\rho > 1$ *and* $EU_n^2 < \infty$, $EV_n^2 < \infty$. *Then*

$$n^{1/2}(n^{-1}W_{[nt]} - (\mu^{-1} - \lambda^{-1})t) \xrightarrow{\mathrm{w}} (\sigma_A^2 + \sigma_S^2)^{1/2} B(t)$$

in $D_{\mathbb{R}}[0, \infty)$.

We now describe the approximation theorems that are available in the setting of 'heavy traffic', where $\rho = 1$. Returning to the waiting time sequence $\{W_n : n \geq 0\}$, the recursion for the W_n's can be solved to yield $W_n = S_n - \min\{S_k; 0 \leq k \leq n\}$, from which it follows that

$$n^{-1/2}W_{[nt]} = n^{-1/2}S_{[nt]} - \min\{n^{-1/2}S_{[ns]} : 0 \leq s \leq t\} = f(X_n)(t)$$

where $X_n(t) = n^{-1/2}S_{[nt]}$ and $f : D_{\mathbb{R}}[0,\infty) \to D_{\mathbb{R}}[0,\infty)$ is given by

$$f(x)(t) = x(t) - \min\{x(s) : 0 \leq s \leq t\} . \tag{5.2}$$

By Donsker's theorem, $S_n \xrightarrow{w} (\sigma_A^2 + \sigma_S^2)^{1/2}B$. On the other hand, it is easily verified that f is continuous in the Skorohod topology at any continuous function $x \in C_{\mathbb{R}}[0,\infty)$, so the continuous mapping principle yields the following theorem.

Theorem 10. *Suppose $\rho = 1$ and $EU_n^2 < \infty$, $EV_n^2 < \infty$. Then*

$$n^{-1/2}W_{[nt]} \xrightarrow{w} (\sigma_A^2 + \sigma_S^2)^{1/2} f(B)(t)$$

in $D_{\mathbb{R}}[0,\infty)$.

The mapping f appearing in (5.2) is called a 'reflection mapping'; such mappings arise naturally in the analysis of queues in heavy traffic. Thus, the following result for $Q(t)$, $D(t)$ should not come as a surprise.

Theorem 11. *Suppose $\rho = 1$ and $EU_n^2 < \infty$, $EV_n^2 < \infty$. Then*

$$n^{-1/2}Q(nt) \xrightarrow{w} (\lambda^3\sigma_A^2 + \mu^3\sigma_S^2)^{1/2} f(B)(t) ,$$
$$n^{1/2}(n^{-1}D(nt) - \lambda t) \xrightarrow{w} g(\lambda^3\sigma_A^2 B_1, \mu^3\sigma_S^2 B_2)(t) ,$$

in $D_{\mathbb{R}}[0,\infty)$, where B_1, B_2 are independent real-valued standard Brownian motions and $g : D_{\mathbb{R}^2}[0,\infty) \to D_{\mathbb{R}}[0,\infty)$ is defined by

$$g(x, y)(t) = y(t) + \inf\{x(s) - y(s) : 0 \leq s \leq t\} .$$

To use the above approximations, one needs to study the distribution of the 'reflected' process $f(B)$. (Such processes are also known as 'regulated Brownian motion'.) It can be shown that $f(B) \overset{d}{=} |B|$ so that the continuous mapping principle yields results of the following kind: if σ_A^2, $\sigma_S^2 > 0$ and $\rho = 1$, then

$$P\{W_n > x(\sigma_A^2 + \sigma_S^2)^{1/2} n^{1/2}\}$$
$$\to P\{|B(1)| > x\} = \sqrt{\frac{2}{\pi}} \int_x^\infty \exp(-\tfrac{1}{2}t^2)\,dt ,$$

$$P\left\{\max_{0 \leq t \leq n} Q(t) \leq (\lambda^3\sigma_A^2 + \mu^3\sigma_S^2)^{1/2} x n^{1/2}\right\}$$
$$\to P\left\{\max_{0 \leq t \leq 1} |B(t)| \leq x\right\}$$
$$= 1 - \frac{4}{\pi}\sum_{k=1}^{\infty} \frac{(-1)^k}{2k+1} \exp\{-[\tfrac{1}{8}\pi^2(2k+1)^2/x^2]\} ,$$

as $n \to \infty$.

The question also arises as to whether Theorems 10 and 11 are true diffusion approximations, in the sense that the limit processes are diffusions. In particular, is $f(B)$ a diffusion process? The path continuity of $f(B)$ is obvious. To see that $f(B)$ has the Markov property, one notes that if $x(0)=0$, then $f(x)=\hat{f}(x)$ where

$$\hat{f}(x)(t)=x(t)-\min\{x(s)\wedge 0 : 0\leq s\leq t\}\,.$$

Since $B(0)=0$, $f(B)=\hat{f}(B)$. Furthermore, if $t>u$, then $\hat{f}(x)(t)=\hat{f}(\theta_u x)(t-u)$ where $(\theta_u x)(s)=x(s+u)-x(u)+\hat{f}(x)(u)$. Thus, the distribution of $\hat{f}(B)(t)$, conditional on $\hat{f}(B)(s)$ for $s\leq u$, is identical to the distribution of $\hat{f}(x+B)(t-u)$ where $x=\hat{f}(B)(u)$ (the independent increments property of B was used here). Thus, the reflection functional f preserves the Markov property in this setting.

We turn now to the statement of certain variants of the above heavy traffic results, which hold for stable and unstable queues in which $\rho\approx 1$. The idea here is to consider a sequence of GI/G/1/∞ queueing systems in which the arrival and service distributions vary with the queue. Suppose that in the nth system, the inter-arrival and service time r.v.'s are given by the sequences $\{U_{nj}\colon j\geq 1\}$ and $\{V_{nj}\colon j\geq 0\}$. Set $\lambda_n^{-1}=EU_{nj}$, $\mu_n^{-1}=EV_{nj}$, $\sigma_{A_n}^2=\operatorname{var}U_{nj}$, $\sigma_{S_n}^2=\operatorname{var}V_{nj}$, and assume that $0<\lambda_n^{-1}$, $\mu_n^{-1}<\infty$. We further require that:

$$\begin{array}{ll}
\text{(i)} & \lambda_n\to\lambda\,,\quad 0<\lambda<\infty\,,\\
\text{(ii)} & \sigma_{A_n}^2\to\sigma_A^2\,,\quad \sigma_A^2>0\,,\\
\text{(iii)} & \mu_n\to\mu\,,\quad 0<\mu<\infty\,,\\
\text{(iv)} & \sigma_{S_n}^2\to\sigma_S^2\,,\quad \sigma_S^2>0\,,\\
\text{(v)} & \text{there exists } \varepsilon>0 \text{ such that } \sup\limits_{n\geq 0} E(|U_{nj}|^{2+\varepsilon}+|V_{nj}|^{2+\varepsilon})<\infty\,,\\
\text{(vi)} & n^{1/2}(\lambda_n-\mu_n)\to c\,,\quad -\infty\leq c\leq\infty\,.
\end{array}\qquad(5.3)$$

Just as in the case of the ordinary central limit theorem for r.v.'s, there exists a version of Donsker's theorem that governs doubly indexed families of r.v.'s (see Prohorov, 1956). To be precise, let $e:[0,\infty)\to\mathbb{R}$ be given by $e(t)=t$ and consider

$$\chi_n(t)=n^{1/2}\Big(\sum_{j=1}^{[nt]} n^{-1}U_{nj}-\lambda_n^{-1}e(t)\Big)\,.$$

Then, $\chi_n\xrightarrow{w}\sigma_A B$ in $D_{\mathbb{R}}[0,\infty)$; a similar FCLT holds for the service times. By applying continuous mapping ideas similar to those used in the $\rho=1$ setting, we obtain the following limit theorem. ($W^n_\cdot$, $Q^n(\cdot)$, $D^n(\cdot)$ are the waiting time, queue length, and departure processes associated with the nth queue).

Theorem 12. *Assume* (5.3)(i)–(vi).

(a) *If* $-\infty < c < \infty$, *then*

$$n^{-1/2}Q^n(nt) \xrightarrow{w} f((\lambda^3\sigma_A^2 + \mu^3\sigma_S^2)^{1/2}B + ce)(t)\,,$$

$$n^{-1/2}(n^{-1}D^n(nt) - \mu_n t) \xrightarrow{w} g(\lambda^3\sigma_A^2 B_1 + ce,\ \mu^3\sigma_S^2 B_2)(t)\,,$$

in $D_{\mathbb{R}}[0,\infty)$, *where* B_1, B_2 *are independent standard Brownian motions.*

(b) *If* $c = \infty$, *then*

$$n^{1/2}(n^{-1}Q^n(nt) - (\lambda_n - \mu_n)t) \xrightarrow{w} (\lambda^3\sigma_A^2 + \mu^3\sigma_S^2)^{1/2}B(t)\,,$$

$$n^{1/2}(n^{-1}D^n(nt) - \mu_n t) \xrightarrow{w} \mu^{3/2}\sigma_S B(t)\,,$$

in $D_{\mathbb{R}}[0,\infty)$.

(c) *If* $c = -\infty$, *then*

$$n^{-1/2}Q^n(nt) \xrightarrow{w} 0\,,$$

$$n^{1/2}(n^{-1}D^n(nt) - \lambda_n t) \xrightarrow{w} \lambda^{3/2}\sigma_A B(t)\,,$$

in $D_{\mathbb{R}}[0,\infty)$.

A similar result holds for the waiting times.

Theorem 13. *Assume* (5.3)(i)–(v) *and replace* (vi) *by* $n^{1/2}(\mu_n^{-1} - \lambda_n^{-1}) \to c$, $-\infty \leqslant c \leqslant \infty$. *Then*

$$n^{-1/2}W^n_{[nt]} \xrightarrow{w} f((\sigma_A^2 + \sigma_S^2)^{1/2}B + ce)(t) \quad \text{if } -\infty < c < \infty\,,$$

$$n^{-1/2}(n^{-1}W^n_{[nt]} - (\mu_n^{-1} - \lambda_n^{-1})t) \xrightarrow{w} (\sigma_A^2 + \sigma_S^2)^{1/2}B(t) \quad \text{if } c = \infty\,,$$

$$n^{-1/2}W^n_{[nt]} \xrightarrow{w} 0 \quad \text{if } c = -\infty\,.$$

For further details on the above heavy-traffic limit theorems in which $\rho \approx 1$, see Iglehart and Whitt (1970a,b). To fully utilize the above approximations, one needs to be able to analyze the resulting limit processes. One complication that arises in the setting of Theorems 12 and 13 is that, unlike the $\rho = 1$ situation in which $f(B) \overset{d}{=} |B|$, it is *not* true that $f(\alpha B + ce) \overset{d}{=} |\alpha B + ce|$. However, it turns out that one can develop, using methods to be discussed in Section 6, an analytical theory for the process $f(\alpha B + ce)$; this process is known as reflected (or regulated) Brownian motion with drift c and variance parameter α^2. Among the results available is the distribution of $f(\alpha B + ce)$: for $\alpha > 0$,

$$P\{f(\alpha B + ce)(t) \leq x\} = \Phi\left(\frac{x - ct}{\alpha t^{1/2}}\right) - \exp(2cx/\alpha^2)\Phi\left(\frac{-x - ct}{\alpha t^{1/2}}\right) \tag{5.4}$$

where

$$\Phi(x) = (2\pi)^{-1/2}\int_{-\infty}^{x} \exp(-\tfrac{1}{2}t^2)\,\mathrm{d}t$$

is the standard normal distribution function. We will now illustrate the application of this result to a GI/G/1/∞ queueing system.

Example 6. Consider the analysis of $\{W_n : n \geq 0\}$ for a GI/G/1/∞ queue in which $\lambda \approx \mu$. Suppose that we wish to study the distribution of W_n for n large, using (5.4). The idea is to set $\alpha^2 = \sigma_A^2$, $c = n^{1/2}(\mu^{-1} - \lambda^{-1})$. Then, Theorem 13 yields

$$P\{W_n \leq xn^{1/2}\} \approx P\{f((\sigma_A^2 + \sigma_S^2)^{1/2}B + [n^{1/2}(\mu^{-1} - \lambda^{-1})]e)(1) \leq x\} \tag{5.5}$$

or, more descriptively,

$$W_n \overset{\mathrm{d}}{\approx} n^{1/2}f((\sigma_A^2 + \sigma_S^2)^{1/2}B + [n^{1/2}(\mu^{-1} - \lambda^{-1})]e)(1)\,.$$

Note that the right-hand side of (5.5) can be evaluated using (5.4).

In Example 6, observe that if $c = n^{1/2}(\mu^{-1} - \lambda^{-1})$ is a large negative (large positive) number (this occurs if n is large relative to $(\mu^{-1} - \lambda^{-1})$), the approximation (5.5) is poor, since the r.v. $f(\alpha B + ce)(1)$ is then essentially a point mass at zero (infinity). Better approximations are available if $c_n \equiv n^{1/2}(\mu^{-1} - \lambda^{-1})$ is large, however. Relation (5.5) says that $n^{-1/2}W_n \overset{\mathrm{d}}{\approx} f(\alpha B + c_n e)(1)$. Proceeding formally, this suggests that

$$(\mu^{-1} - \lambda^{-1})W_n = c_n\, n^{-1/2}W_n \overset{\mathrm{d}}{\approx} c_n f(\alpha B + c_n e)(1)\,.$$

But the r.v. $c_n f(\alpha B + c_n e)(1)$ is equivalent to

$$f(\alpha c_n B + c_n^2 e)(1) \overset{\mathrm{d}}{=} f(\alpha B(c_n^2\,\cdot\,) + e(c_n^2\,\cdot\,))(1) = f(\alpha B + e)(c_n^2)$$

(the first equality relies on the scaling behavior of Brownian motion: $\gamma B(\cdot) = B(\gamma^2\cdot))$. Note that if $(\mu^{-1} - \lambda^{-1})$ is negative and c_n is large, this argument yields the approximation $|\mu^{-1} - \lambda^{-1}|W_n \overset{\mathrm{d}}{\approx} f(\alpha B - e)(\infty)$. Turning to (5.4) and letting $t \to \infty$ it turns out that $f(\alpha B - e)(t) \xrightarrow{\mathrm{w}} \frac{1}{2}\alpha^2 \exp(1)$ as $t \to \infty$, where $\exp(1)$ is an exponential r.v. with parameter 1. Thus, we approximate $|\mu^{-1} - \lambda^{-1}|W_n$ via $\frac{1}{2}\alpha^2 \exp(1)$. If, on the other hand, $(\mu^{-1} - \lambda^{-1})$ is positive and c_n is large, $|\mu^{-1} - \lambda^{-1}|W_n \approx f(\alpha B + e)(c_n^2)$. Because of the positive drift of $\alpha B + e$,

$$f(\alpha B + e)(c_n^2) \overset{\mathrm{d}}{\approx} (\alpha B + e)(c_n^2) \overset{\mathrm{d}}{=} c_n(\alpha B(1) + c_n e)(1)$$

i.e.

$$n^{-1/2}W_n - c_n e(1) = n^{1/2}(n^{-1}W_n - (\mu^{-1} - \lambda^{-1})) \overset{\mathrm{d}}{\approx} \alpha B(1)\,.$$

We summarize the above discussion with the approximations shown in Table 1.

Table 1

Parameter region	Approximating distribution for W_n
$(\mu^{-1} - \lambda^{-1})$ small, $n^{1/2}(\mu^{-1} - \lambda^{-1})$ large and negative	$\frac{(\sigma_A^2 + \sigma_S^2)}{2(\mu^{-1} - \lambda^{-1})}\exp(1)$
n large, $(\mu^{-1} - \lambda^{-1})$ small, $n^{1/2}(\mu^{-1} - \lambda^{-1})$ small to moderate	$n^{1/2}f((\sigma_A^2 + \sigma_S^2)^{1/2}B + [(\mu^{-1} - \lambda^{-1})n^{1/2}]\,e)(1)$
$(\mu^{-1} - \lambda^{-1})$ small, $n^{1/2}(\mu^{-1} - \lambda^{-1})$ large and positive	$n(\mu^{-1} - \lambda^{-1}) + n^{1/2}(\sigma_A^2 + \sigma_S^2)^{1/2}B(1)$

Except for the topmost entry in the table, the approximating distributions cited above are given by Theorem 13. For the entry concerning the situation in which $n^{1/2}(\mu^{-1} - \lambda^{-1})$ is large and negative, Prohorov (1963) proved the appropriate result (see Kingman, 1961, for a related theorem).

Theorem 14. *Assume* (5.3)(i)–(v) *and suppose that* $n^{1/2}(\mu_n^{-1} - \lambda_n^{-1}) \to -\infty$. *Then*

$$|\mu_n^{-1} - \lambda_n^{-1}|W_n \xrightarrow{\mathrm{w}} \tfrac{1}{2}(\sigma_A^2 + \sigma_S^2)\exp(1) \quad \text{as } n \to \infty\,.$$

Theorem 14 suggests that as $\rho \nearrow 1$, $EW_\rho \sim \frac{1}{2}\lambda(\sigma_A^2 + \sigma_S^2)/(1-\rho)$ where W_ρ is the steady-state waiting time associated with the ρth queue. In fact, this heavy-traffic approximation is an upper bound on EW for all ρ; see Marshall (1968).

Theorem 15. *In a* GI/G/1/∞ *queue with* $\rho < 1$, $EW \leq \frac{1}{2}\lambda(\sigma_A^2 + \sigma_S^2)/(1-\rho)$.

To conclude this section, we note that further details on the structure of reflected Brownian motion $f(\alpha B + ce)$ may be found in Abate and Whitt (1987a,b). It also turns out that the results of this section are valid under much more general hypotheses than those stated above. In particular, the results hold even when the inter-arrival and service time sequences are dependent.

This is important, since typically the arrival stream to a queue is the superposition of departure processes from other queues; such an arrival stream yields dependent inter-arrival times. To deal with dependent sequences, it is enough to observe that the above results essentially require only that the arrival and service streams satisfy a joint FCLT: There exists λ^{-1}, μ^{-1} and Σ such that

$$n^{1/2}\Big(n^{-1}T_{[nt]} - t\lambda^{-1}, n^{-1}\sum_{j=1}^{[nt]} V_j - t\mu^{-1}\Big) \xrightarrow{\text{w}} \Sigma^{1/2}B(t) \tag{5.6}$$

in $D_{\mathbb{R}^2}[0, \infty)$. Using (5.6), analogues to Theorems 7, 8, 9, 10, and 11 may be obtained. Analogues to Theorems 12, 13, and 14 depend on assuming a double-indexed version of (5.6).

Finaly, Rosenkrantz (1978) has studied the question of how rapidly the distribution of $n^{-1/2}W_n$ converges to that of reflected Brownian motion, when the queue is in heavy-traffic ($\rho = 1$). By using the strong approximation theorem of Section 4, he shows that the rate is $\mathrm{O}(\log n/n^{1/2})$.

6. Background on diffusion processes

In this section, we briefly describe several important elements of the general theory for diffusion processes. We are particularly interested here in describing how the theory of diffusion processes is intimately connected to the study of certain partial differential equations (PDE's). It is that connection to PDE's that makes diffusion processes computationally and analytically attractive.

To develop the connection with PDE's, we review briefly the different modeling approaches that give rise to diffusion processes:

(1) A diffusion process may be directly postulated as an appropriate model (e.g. Brownian motion as a model of molecular motion).

(2) A diffusion process may arise as a limit occurring in the study of a stochastic model (e.g. reflected Brownian motion arises as a limit in the study of queues in heavy traffic).

(3) One may postulate that the stochastic process X driving a model has certain infinitesimal characteristics. Specifically, if X is $\mathbb{R}^d$-valued one specifies functions $\mu : \mathbb{R}^d \to \mathbb{R}^d$, $\sigma^2 : \mathbb{R}^d \to \mathbb{R}^{d+d}$ called the *infinitesimal drift* and *infinitesimal covariance* functions, respectively. These functions μ, σ^2 are related to X by requiring that

$$E\{X(t+h)|X(u)\colon u \leq t\} = \mu(X(t))h + \mathrm{o}(h)\,,$$

$$\mathrm{cov}\{X(t+h)|X(u)\colon u \leq t\} = \sigma^2(X(t))h + \mathrm{o}(h)\,.$$

If $\sigma^2(\cdot)$ is positive definite and X is additionally postulated to be Markov with continuous paths, then the resulting process (if it exists) must be a diffusion process.

Much of the theory of diffusion processes concerns mathematical difficulties that arise when a diffusion process is postulated as in (3) above. More precisely, given μ and σ^2, does there exist a unique diffusion process X (unique in distribution) which satisfies the formulas in (3)? To answer this question, suppose there exists such a process X. If f is sufficiently smooth,

$$E_x f(X(h)) = E_x\Big[f(x) + \sum_{i=1}^{d} \frac{\partial f}{\partial x_i}(x)(X_i(h) - x_i) + \tfrac{1}{2} \sum_{i,j=1}^{d} \frac{\partial^2 f}{\partial x_i \partial x_j}(\xi)(X_i(h) - x_i)(X_j(h) - x_j)\Big]$$

where ξ is on the line segment joining x and $X(h)$. Using the infinitesimal characteristics, the above Taylor expansion yields

$$E_x f(X(h)) = f(x) + (Lf)(x)h + \mathrm{o}(h) \tag{6.1}$$

where

$$(Lf)(x) = \sum_{i=1}^{d} \mu_i(x) \frac{\partial f}{\partial x_i}(x) + \tfrac{1}{2} \sum_{i,j=1}^{d} \sigma_{ij}(x) \frac{\partial^2 f}{\partial x_i \partial x_j}(x) .$$

This expansion holds for a large class of f's, call it $\mathscr{D}(L)$. One can state (6.1) differently. Let $P_t(x, \mathrm{d}y) = P_x\{X(t) \in \mathrm{d}y\}$ be the transition measure for X. Then, (6.1) asserts that

$$h^{-1}\Big[\int P_h(x, \mathrm{d}y)f(y) - f(x)\Big] \to (Lf)(x) \tag{6.2}$$

as $h \downarrow 0$, provided $f \in \mathscr{D}(L)$. Of course, the family $(P_t\colon t > 0)$ must have the *semigroup property*

$$P_{t+s}(x, A) = \int P_t(x, \mathrm{d}y)P_s(y, A) \tag{6.3}$$

and it must be *stochastic*: $P_t \geq 0$, $P_t(x, \mathbb{R}^d) = 1$. An extensive theory has been developed to determine when there exists a stochastic semigroup satisfying (6.2), for a given L and $\mathscr{D}(L)$. (A principal result here is the Hille–Yosida theorem; see Karlin and Taylor, 1981.) A positive answer to this question generally guarantees the existence of a diffusion X having infinitesimals μ and σ^2. (One also needs to check that a version of X can be constructed with continuous paths, and thus existence of (P_t) may not be quite enough to guarantee existence of a diffusion.)

Suppose now that we are given that a diffusion X exists, described by L and $\mathscr{D}(L)$. The operator L is called the *infinitesimal generator* of X and $\mathscr{D}(L)$ is called the *domain of the generator*. We shall now describe some of the problems related to X that may be found by solving PDE's involving L.

Problem 1. Let X be a real-valued diffusion. Let $T(a)=\inf\{t\geqslant 0 : X(t)=a\}$ be the first time that X takes on the value a. For fixed levels a and b $(a<b)$, let $u(x)=P_x\{T(a)<T(b)\}(a<x<b)$ be the probability that X hits level b before hitting level a. Then, $u\in\mathscr{D}(L)$ and u satisfies the differential equation

$$Lu=0\,,$$

subject to the (obvious) boundary conditions $u(a)=1$, $u(b)=0$.

Example 7. If X is a one-dimensional standard Brownian motion, it is easily verified from the stationary independent increments of X that $\mu(x)=0$ and $\sigma^2(x)=1$. The domain $\mathscr{D}(L)$ consists of all bounded functions $f:\mathbb{R}\to\mathbb{R}$ having a bounded continuous second derivative. The solution to the differential equation

$$\tfrac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2}u(x)=0$$

subject to $u(a)=1$, $u(b)=0$ is the affine function $u(x)=(b-x)/(b-a)$.

Problem 2. Let x be an interior point of a set A with a 'nice' boundary and assume that $P_y\{T(A^c)<\infty\}=1$ for all $y\in A$, where $T(A^c)=\inf\{t>0\colon X(t)\in A^c\}$. For given real-valued functions g and k, set

$$w(x)=E_x\left\{\int_0^{T(A^c)}\exp\left(-\int_0^t k(X(\tau))\,\mathrm{d}\tau\right)g(X(t))\,\mathrm{d}t\right\}.$$

Then, w satisfies the PDE,

$$(Lw)(\cdot)-k(\cdot)w(\cdot)=-g(\cdot)\quad\text{in } A$$

subject to $w\in\mathscr{D}(L)$ and $w(y)=0$ on the boundary of A.

Example 7 (continued). Let $k=0$ and $g=1$ in Problem 2, so that $w(x)=E_xT(A^c)$ is the expected amount of time required for the diffusion X to 'escape' from the set A. If X is a one-dimensional standard Brownian motion and $A=[a,b]$, then w is obtained by solving

$$\tfrac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2}w(x)=-1$$

subject to $w(a)=w(b)=0$. The solution is $w(x)=(x-a)(b-x)$.

Problem 3. For a given real-valued g, set $u(t,x)=E_xg(X(t))$. Then, u satisfies

$$\frac{\partial}{\partial t}u(t,x)=\sum_{i=1}^{d}\mu_i(x)\frac{\partial}{\partial x_i}u(t,x)+\tfrac{1}{2}\sum_{i,j=1}^{d}\sigma_{ij}^2(x)\frac{\partial^2}{\partial x_i\,\partial x_j}u(t,x) \tag{6.4}$$

subject to $u(t,\cdot)\in\mathscr{D}(L)$ (and $u(0,\cdot)=g(\cdot)$); in short-hand, (6.4) is written $u_t=Lu$. Equation (6.4) is called the *backward equation* for X, and is fundamental to the study of diffusion processes. Assume that $P(t,x,\mathrm{d}y)$ has a density $p(t,x,y)$ with respect to Lebesgue measure $\mathrm{d}y$. By formally setting $g(\cdot)=\delta_y(\cdot)$ ($\delta_y(\cdot)$ is the Dirac delta function), (6.4) yields an equation for the transition density p:

$$\frac{\partial}{\partial t}p(t,x,y)=\sum_{i=1}^{d}\mu_i(x)\frac{\partial}{\partial x_i}p(t,x,y) + \tfrac{1}{2}\sum_{i,j=1}^{d}\sigma_{ij}^2(x)\frac{\partial^2}{\partial x_i\,\partial x_j}p(t,x,y) \tag{6.5}$$

subject to $p(t,\cdot,y)\in\mathscr{D}(L)$ and $p(0,\cdot,y)=\delta_y(\cdot)$; the partial differential equation (6.5) can be re-written in the more convenient shorthand notation

$$\frac{\partial}{\partial t}p=Lp\,.$$

If X is a process that lives on all of $\mathbb{R}^d$ (so that X has no non-trivial 'boundary behavior'), one can obtain the 'adjoint' equation to (6.5):

$$\frac{\partial}{\partial t}p(t,x,y)=-\sum_{i=1}^{d}\frac{\partial}{\partial y_i}(\mu_i(y)p(t,x,y)) + \tfrac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2}{\partial y_i\,\partial y_j}(\sigma_{ij}^2(y)p(t,x,y)) \tag{6.6}$$

subject to $p(0,x,\cdot)=\delta_x(\cdot)$; (6.6) is the *forward equation* for the density p. The adjoint equation can be derived by viewing the 'backwards operator' L as an operator on an appropriately chosen space of smooth functions, and performing an 'integration by parts' to obtain the adjoint (or forward) operator L^*.

Example 8. Let X be a real-valued diffusion process for which $\mu(x)=-\mu x$ and $\sigma^2(x)=1$; this is a special case of a one dimensional *Ornstein–Uhlenbeck process*. (Note that if $\mu=0$, X is a standard Brownian motion.) Then, the backwards equation for the transition density p takes the form

$$\frac{\partial p}{\partial t}=\tfrac{1}{2}\frac{\partial^2 p}{\partial x^2}-\mu x\frac{\partial p}{\partial x}$$

subject to $p(0,x,y)=\delta_y(x)$. On the other hand, the forward equation is given by

$$\frac{\partial p}{\partial t}=\tfrac{1}{2}\frac{\partial^2 p}{\partial y^2}+\frac{\partial}{\partial y}[\mu y\cdot p]$$

subject to $p(0, x, y) = \delta_y(x)$. If $\mu = 0$, the (common) solution p is given by

$$p(t, x, y) = \varphi(t, x, y)$$

where φ is the Gaussian kernel

$$\varphi(t, x, y) = (2\pi t)^{-1/2} \exp(-\tfrac{1}{2}(y - x)^2/t) ,$$

$t > 0$. However, if $\mu \neq 0$, the solution p takes the form

$$p(t, x, y) = \varphi(\tfrac{1}{2}(1 - \mathrm{e}^{-2\mu t})/\mu, x\,\mathrm{e}^{-\mu t}, y) .$$

Problem 4. For a real-valued g and a non-negative function k, consider the expectation

$$w(t, x) = E_x\left\{\exp\left(-\int_0^t k(X(s))\,\mathrm{d}s\right) g(X(t))\right\} .$$

Then, under suitable regularity conditions on X, g, and k, it can be shown that w solves the PDE,

$$\frac{\partial}{\partial t} w = Lw - kw ,$$

subject to $w(t, \cdot) \in \mathscr{D}(L)$ and $w(0, \cdot) = g(\cdot)$.

Example 7 (continued). Let X be a one-dimensional standard Brownian motion and suppose that $k(x) = \alpha$ and $g(x) = x$. Then,

$$w(t, x) = x\,\mathrm{e}^{-\alpha t}$$

solves Problem 4.

Problem 5. Suppose that g is a real-valued function. If we view $g(x)$ as the rate at which cost accumulates in state x, then

$$u(x) = E_x \int_0^\infty \mathrm{e}^{-\alpha t} g(X(t))\,\mathrm{d}t$$

is the α-discounted cost associated with starting the diffusion X in state x. Under suitable regularity conditions, one can show that $u \in \mathscr{D}(L)$ and satisfies

$$\alpha u - Lu = g .$$

Example 7 (continued). If X is a one-dimensional standard Brownian motion and $\alpha > 0$, then the solution u to Problem 5 is given by

$$u(x) = \frac{1}{\sqrt{2\alpha}} \int_{-\infty}^{\infty} \exp(-\sqrt{2\alpha}|x - y|)g(y)\,\mathrm{d}y$$

(provided the intergral exists and is finite).

Problem 6. Suppose that X is positive recurrent so that a density $p(\cdot)$ exists such that $p(t, x, y) \to p(y)$ as $t \to \infty$ for each x. This type of limit behavior suggests that $\partial p(t, x, y)/\partial t \to 0$ as $t \to \infty$; formal substitution in (6.6) yields

$$\tfrac{1}{2} \sum_{i,j=1}^{d} \frac{\partial^2}{\partial y_i\, \partial y_j} (\sigma_{ij}^2(y)p(y)) - \sum_{i=1}^{d} \frac{\partial}{\partial y_i} (\mu_i(y)p(y)) = 0 \tag{6.7}$$

subject to $\int p(y)\,\mathrm{d}y = 1$; (6.7) can be solved to obtain the steady-state density of X. (We caution that if X has non-trivial boundary behavior (e.g. reflection), additional boundary conditions must be prescribed.) Typically, the process X becomes a stationary process when initialized according to the density p. Hence, p is often termed a *stationary density* for X.

Example 8 (continued). Suppose that X is a one-dimensional Ornstein–Uhlenbeck process with $\mu(x) = -\mu x$ and $\sigma^2(x) = 1$. Then, the stationary density p must satisfy

$$\tfrac{1}{2} \frac{\mathrm{d}^2}{\mathrm{d}y^2} p(y) + \frac{\mathrm{d}}{\mathrm{d}y} (\mu y p(y)) = 0\,.$$

If $\mu > 0$, such a solution p exists; it is given by

$$p(y) = \frac{1}{\sqrt{\pi/\mu}} \exp(-\mu y^2)\,.$$

An easily accessible treatment of the analytical theory for diffusions is given in Karlin and Taylor (1981).

In most operations research applications, diffusions are not obtained via an infinitesimal characterization of the process; rather, as illustrated in Section 5, they are usually obtained as limit processes that are functionals of Brownian motion. In order to use the PDE's cited above, one needs to calculate L for such processes. We now indicate how to do this for reflected Brownian motion $f(\alpha B + ce)$; the tool used is, however, quite general.

Let X be a reflected Brownian motion starting at $x \geq 0$. Then, for $t \geq 0$,

$$X(t) = \tilde{f}(\alpha B + x + ce)(t)\,,$$

where

$$\tilde{f}(y)(t) = y(t) - \min\{y(s) \wedge 0 : 0 \leq s \leq t\}\,.$$

Hence, we may represent X as

$$X(t) = x + \alpha B(t) + ct - V(t)\,,$$

where

$$V(t) = \min\{(x + \alpha B(s) + cs) \wedge 0 : 0 \leqslant s \leqslant t\}\,.$$

The process V has two important characteristics. Firstly, V is non-decreasing; hence, it may be used as an integrator. Secondly, the points of increase occur only at those t for which $X(t)$ vanishes (i.e., $\int_0^\infty I(X(s) > 0)V(\mathrm{d}s) = 0$).

We are going to use Itô's formula (see Chapter 2 of this handbook) to derive the backward equation for the diffusion X. Consider a function $\eta(t, x)$ on $[0, T] \times [0, \infty)$ which is twice continuously differentiable on its domain. We will apply an Itô-type formula to the process $\nu(t, X(t))$, where $\nu(t, x) = \eta(T - t, x)$. Note that the stochastic differential of the process $\nu(t, X(t))$ is given by

$$\begin{aligned}\mathrm{d}\nu(t, X(t)) = {} & \frac{\partial \nu}{\partial t}(t, X(t))\,\mathrm{d}t + \frac{\partial \nu}{\partial x}(t, X(t))\,\mathrm{d}X(t) \\ & + \tfrac{1}{2}\frac{\partial^2 \nu}{\partial x^2}(t, X(t))(\mathrm{d}X(t))^2\,.\end{aligned}$$

Clearly, $\mathrm{d}X(t) = \alpha\,\mathrm{d}B(t) + c\,\mathrm{d}t - \mathrm{d}V(t)$. On the other hand, the Itô calculus asserts that $(\mathrm{d}X(t))^2 = \alpha^2\,\mathrm{d}t$. Hence,

$$\begin{aligned}\mathrm{d}\nu(t, X(t)) = {} & \left\{\frac{\partial \nu}{\partial t}(t, X(t)) + c\,\frac{\partial \nu}{\partial x}(t, X(t)) + \tfrac{1}{2}\alpha^2\,\frac{\partial^2 \nu}{\partial x^2}(t, X(t))\right\}\mathrm{d}t \\ & + \alpha\,\frac{\partial \nu}{\partial x}(t, X(t))\,\mathrm{d}B(t) - \frac{\partial \nu}{\partial x}(t, X(t))\,\mathrm{d}V(t)\,. \qquad (6.8)\end{aligned}$$

Suppose now that we choose the function η so that the bracketed term in (6.8) disappears, i.e., choose η to satisfy the PDE,

$$\frac{\partial \eta}{\partial t}(t, x) = c\,\frac{\partial \eta}{\partial x}(t, x) + \tfrac{1}{2}\alpha^2\,\frac{\partial^2}{\partial x^2}\,\eta(t, x)\,. \qquad (6.9)$$

We can simplify (6.8) further by observing that since V increases only when X is zero, we have

$$\frac{\partial \nu}{\partial x}(t, X(t))\,\mathrm{d}V(t) = \frac{\partial \nu}{\partial x}(t, 0)\,\mathrm{d}V(t)\,. \qquad (6.10)$$

Hence, (6.10) vanishes if we require that η satisfy the boundary condition

$$\frac{\partial}{\partial x}\,\eta(t, 0) = 0\,. \qquad (6.11)$$

We may therefore conclude that if η satisfies (6.9) and (6.11), then

$$\mathrm{d}\nu(t, X(t)) = \alpha \frac{\partial \nu}{\partial x}(t, X(t))\,\mathrm{d}B(t)$$

and thus

$$\nu(T, X(T)) - \nu(0, X(0)) = \alpha \int_0^T \frac{\partial \nu}{\partial x}(t, X(t))\,\mathrm{d}B(t)\,.$$

But the right-hand side is a martingale (see Chapter 2 of this handbook) so we conclude that

$$E_x \nu(T, X(t)) = \nu(0, x)$$

i.e.,

$$E_x \eta(0, X(T)) = \eta(T, x)\,.$$

If we set $g(x) = \eta(0, x)$, we find that $\eta(T, x) = E_x g(X(T))$, provided that η satisfies (6.9) and (6.11).

A glance at (6.4) shows that η is the solution to the backwards equation corresponding to the diffusion $X = f(\alpha B + ce)$. Thus, the infinitesimal generator L of X is given by the second-order differential operator

$$L = c\frac{\partial}{\partial x} + \tfrac{1}{2}\alpha^2 \frac{\partial^2}{\partial x^2}\,;$$

the domain $\mathscr{D}(L)$ of the reflected Brownian motion X includes bounded functions h that possess a bounded continuous second derivative and satisfy $(\partial/\partial x)h(0) = 0$.

Hence, Itô's formula can be applied to determine both the infinitesimal generator and the 'boundary conditions' that characterize a diffusion process.

7. Diffusion approximations for an open network of queues in heavy traffic

In Section 5, we showed how a single-server queue in heavy traffic can be approximated by a one-dimensional reflecting Brownian motion. In this section, we discuss how to approximate an open network of d queueing stations in heavy traffic by a d-dimensional diffusion process.

Consider a network of d queueing stations. Each of the d stations consists of a single work-conserving server that serves customers in the order in which they arrive. By work conserving, we mean that the server never goes idle when facing customers in its queue. We assume that the customer routing within the network is Markovian. Let P be the routing matrix for the network (i.e., P_{ij}

represents the probability that a customer completing service at station i goes immediately to station j; $1 - \Sigma_{j=1}^{d} P_{ij}$ is the probability that a customer released from station i leaves the network). To simplify our following discussion, we assume that $P_{ii} = 0$ for $1 \leqslant i \leqslant d$. We further require that P be irreducible and that $(I - P)^{-1}$ exist (i.e., P is substochastic).

Let $U_i = \{U_i(m) \colon m \geqslant 1\}$ be the sequence of exogenous customer arrivals to the ith station; the $U_i(m)$'s are assumed to be i.i.d. with mean $1/\lambda_i$ and (finite) coefficient of variation a_i^2 (i.e., $\operatorname{var} U_i(m) = \lambda_i^{-2} a_i^2$). For $1 \leqslant i \leqslant d$, let $V_i = \{V_i(m) \colon m \geqslant 1\}$ be a sequence of i.i.d. unit mean r.v.'s for which $b_i^2 = \operatorname{var} V_i(m) < \infty$; the r.v. $V_i(m)$ can be viewed as the service requirement of the mth customer to be served at station i. If $\mu_j(n)$ is the service rate at the jth station in the nth approximating network, then $V_j(m)/\mu_j(n)$ is the actual duration of service for the mth customer at the jth station in the nth network.

Let $Q_n(0)$ be a random d-vector having non-negative integer-valued components. The ith component is to be interpreted as the number of customers that are sitting in the queue of the ith station waiting to be served at time $t = 0$ (in the nth system).

We require that the $2d$ inter-arrival and service requirement sequences, as well as the customer routing dynamics and $Q_n(0)$, be mutually independent. Such a queueing network as we have described is termed a *generalized Jackson network*. (Note that if the $U_i(m)$'s and $V_i(m)$'s are exponential r.v.'s, we obtain a standard Jackson network.)

The theory of weak convergence for queueing networks is largely based on the development of convenient representations for the corresponding queue-length processes. By applying the continuous mapping principle, random time change theory, and converging-together techniques to the representation, one arrives at a diffusion limit for the network. To describe the representation that is useful here, we let

$$A_i(t) = \max\left\{m \geqslant 0 \colon \sum_{j=1}^{m} U_i(j) \leqslant t\right\}$$

be the process which counts the number of exogenous arrivals to the ith station over the interval $[0, t]$. We further let

$$C_i(t) = \max\left\{m \geqslant 0 \colon \sum_{j=1}^{m} V_i(j) \leqslant t\right\};$$

note that $C_{ni}(t) = C_i(\mu_i(n)t)$ is a process which counts the number of customers served at the ith station during the first t units of busy time (in the nth network). Let $Q_n(t) = (Q_{n1}(t), \ldots, Q_{nd}(t))$ be the vector queue-length process in which $Q_{ni}(t)$ represents the number of customers at the ith station at time t in the nth approximating network. Note that

$$B_{ni}(t) = \int_0^t I(Q_{ni}(s) > 0)\, \mathrm{d}s \tag{7.1}$$

is the amount of time that the ith station is busy over $[0, t]$. Then, $C_{ni}(B_{ni}(t))$ is the total number of customers served at the ith station (in the nth system) over $[0, t]$.

Suppose that $R_i(m)$ is a random vector which routes the mth customer completing service at the ith station. More precisely, $R_i(m) = e_j$ (e_j is the jth unit vector) if and only if the mth customer to be served at the ith station is routed immediately, upon completing service at station i, to station j. Then,

$$\sum_{j=1}^{C_{ni}(B_{ni}(t))} R_i(j)$$

is a vector in which the kth component equals the number of customers routed from station i to k over $[0, t]$. Let $C_n(B_n(t)) = (C_{n1}(B_{n1}(t)), C_{n2}(B_{n2}(t)), \ldots, C_{nd}(B_{nd}(t))$ and $A(t) = (A_1(t), \ldots, A_d(t))$. Clearly,

$$Q_n(t) = Q_n(0) + A(t) + \sum_{i=1}^{d} \sum_{j=1}^{C_{ni}(B_{ni}(t))} R_i(j) - C_n(B_n(t)) . \tag{7.2}$$

By arguing path-by-path, it is straightforward to show that the queue-length process Q_n (together with the busy time processes $B_{n1}, \ldots, B_{nd}$) is the unique solution of the coupled system (7.1)–(7.2). However, we shall shortly describe an alternative representation for Q_n that is more convenient in terms of implementing weak convergence arguments of the type mentioned earlier in this chapter.

Before proceeding further, we need to derive conditions on the network which force each of the stations into heavy traffic. We let $\lambda = (\lambda_1, \ldots, \lambda_d)$ be the vector of exogenous arrival rates and $\mu(n) = (\mu_1(n), \ldots, \mu_d(n))$ be the vector of service rates. Then, $\lambda(I - P)^{-1}$ is the vector of 'effective arrival rates', in which the ith component is the rate at which customers arrive to station i (from both within and outside the network). Suppose that

$$n^{1/2}(\mu(n) - \mu) \rightarrow c \tag{7.3}$$

as $n \rightarrow \infty$, where the vector μ is such that $\mu_i = (\lambda(I - P)^{-1})_i$ for $1 \leq i \leq d$. Hence, each station is effectively experiencing a traffic intensity of 1, and is therefore in heavy traffic.

The alternative representation for Q_n is obtained by first centering the cumulative processes appearing in (7.2). The centering constants are suggested by (7.3) and the following FCLT's:

$$n^{1/2}(n^{-1}A_i(nt) - \lambda_i t) \xrightarrow{w} B_{1i}(\lambda_i a_i^2 t) , \tag{7.4}$$

$$n^{1/2}(n^{-1}C_i(nt) - t) \xrightarrow{w} B_{2i}(b_i^2 t) , \tag{7.5}$$

$$n^{1/2}\left(n^{-1} \sum_{j=1}^{[nt]} R_i(j) - e_i P\right) \xrightarrow{w} A_i^{1/2} B_i(t) , \tag{7.6}$$

where $\Lambda_i = (\Lambda_i(j, k){:}\ \ 1 \leqslant j,\ \ k \leqslant d)$ and $\Lambda_i(j, k) = -P_{ij}P_{ij}(j \neq k)$ and $\Lambda_i(j, j) = P_{ij}(1 - P_{ij})$. FCLT's (7.4) and (7.5) are just the FCLT's for renewal processes described in Section 5, whereas (7.6) is the multivariate version of Donsker's theorem. Let $(\mu(n)B_n)(t) = (\mu_1(n)B_{n1}(t), \ldots, \mu_d(n)B_{nd}(t))$. Then,

$$Q_n(t) = Z_n(t) + Y_n(t)(I - P) \tag{7.7}$$

where

$$Z_n(t) = Q_n(0) + (\lambda + \mu(n)P - \mu(n))t + \xi_n(t)\,,$$

$$\begin{aligned}\xi_n(t) = [A(t) - \lambda t] &+ \left[\sum_{i=1}^{d} \sum_{k=1}^{C_{ni}(B_{ni}(t))} R_i(k) - C_n(B_n(t))P\right] \\ &+ [C_n(B_n(t))P - (\mu(n)B_n)(t)P] \\ &- [C_n(B_n(t)) - (\mu(n)B_n)(t)]\,,\end{aligned}$$

$$Y_n(t) = \mu(n)t - (\mu(n)B_n)(t)\,.$$

It turns out that the process Y_n appearing in (7.7) has the following properties:

(A1) Y_n is non-decreasing (component-by-component) with $Y_n(0) = 0$.

(A2) $Y_{nj}(t)$ increases only at times t for which $Q_{nj}(t) = 0$ (i.e., $\int_0^\infty I(Q_{nj}(t) > 0)Y_{nj}(\mathrm{d}t) = 0$).

Futhermore, the pair of processes (Q_n, B_n) satisfies (7.1)–(7.2) if and only if (Q_n, Y_n) satisfies (7.7), where Y_n has properties (A1) and (A2). Hence, we may view (7.7), (A1) and (A2) as an alternative characterization of Q_n.

Let $D^0_{\mathbb{R}^d}[0, \infty) = \{x \in D_{\mathbb{R}^d}[0, \infty){:}\ x(0) \geqslant 0\}$. Then, given $z \in D^0_{\mathbb{R}^d}[0, \infty)$, there exists a unique y such that:

(B1) $q = z + y(I - P)$.

(B2) y is non-decreasing with $y(0) = 0$.

(B3) y_j increases only at times t when $q_j(t) = 0$, $1 \leqslant j \leqslant d$.

Since y is uniquely defined (for each z), we may write $y = h(z)$. If we let $f(z) = z + h(z)(I - P)$, we can rewrite (7.7), (A1) and (A2) as asserting that $Q_n = f(Z_n)$. The representation $Q_n = f(Z_n)$ plays a central role in the heavy traffic analysis of open networks. The function h is known as the *regulator map* and $f(z)$ is termed the *regulated version* of z. These mappings were first studied in Harrison and Reiman (1981). The function h has the following properties:

(C1) The function f is (suitably) continuous.

(C2) For every $t > 0$, the restriction of $f(z)$ to $[0, t]$ depends only on the restriction of z to $[0, t]$.

(C3) Let $T \geqslant 0$. Define $\tilde{z}(t) = q(T) + z(T + t) - z(T)$, $\tilde{f}(z)(t) = f(z(T + t))$. Then, $\tilde{f}(z) = f(\tilde{z})$.

Given (C1), weak convergence results for Q_n can be obtained by suitably approximating Z_n. An important (and typically difficult) step in dealing with Z_n is to show that

$$n^{-1}B_{ni}(nt)\overset{\mathrm{w}}{\rightarrow} t \tag{7.8}$$

in $D_{\mathbb{R}}[0,\infty)$ as $n\rightarrow\infty$ (i.e., each server is asymptotically busy 100% of the time). With (7.8) in hand, (7.4)–(7.6), in conjunction with a random time change argument, permits one to show that if there exists $Q(0)$ such that

$$n^{-1/2}Q_n(0)\overset{\mathrm{w}}{\rightarrow} Q(0)\,, \tag{7.9}$$

then

$$Z^n(t)\overset{\Delta}{=}n^{-1/2}Z_n(nt)\overset{\mathrm{w}}{\rightarrow} Q(0)+c(I-P)t+\Gamma^{1/2}B(t) \tag{7.10}$$

as $n\rightarrow\infty$. The process B appearing in (7.10) is a d-dimensional standard Brownian motion and the covariance matrix Γ is given by

$$\begin{aligned}\Gamma_{jk} &= [\mu_j b_j^2+\lambda_j a_j^2]\delta_{jk}-\mu_j b_j^2 P_{jk}-\mu_k b_k^2 P_{kj}\\ &\quad+\sum_{l=1}^{d}\mu_l P_{lj}[\delta_{jk}-P_{lk}+b_l^2 P_{lk}]\,.\end{aligned} \tag{7.11}$$

By applying the continuous mapping principle to the function f and the process Z^n, one obtains the following diffusion approximation for open queueing networks in heavy traffic; see Reiman (1984) and Chen and Mandelbaum (1988) for further details.

Theorem 16. *Assume* (7.3) *and* (7.9). *Let* $Q^n(t)=Q_n(nt)/\sqrt{n}$ *and set* $Z(t)=Q(0)+c(I-P)t+\Gamma^{1/2}B(t)$. *Then*

$$Q^n=f(Z^n)\overset{\mathrm{w}}{\rightarrow}f(Z)\overset{\Delta}{=}Q$$

in $D_{\mathbb{R}^d}[0,\infty)$ *as* $n\rightarrow\infty$.

To use the approximation suggested by Theorem 16 in a practical setting, we need a queueing network in which each station is in heavy traffic. By this, we mean that the difference $|(\lambda(I-P)^{-1})_i/\mu_i-1|$ ought to be of order ε for some small ε $(1\leq i\leq d)$. (Think of ε as $n^{-1/2}$ in our limit theorem.) Then, the diffusion limit Q describes the fluctuations of order $1/\varepsilon$ experienced by the queueing network over time scales of order $1/\varepsilon^2$.

As in Section 5, the reflected Brownian motion Q (also known as regulated Brownian motion) turns out to be a diffusion process. This basically is a consequence of properties (C1) and (C2) of the map f, together with the independent increments of B. The term reflection is used because the process Q can be viewed as 'reflecting' in the direction of the ith row of $I-P$ whenever the ith component of Q is zero.

The diffusion limit Q inherits much of the qualitative structure of the queueing network. For example, we note that if $c > 0$, then (7.3) guarantees that each station can serve customers (slightly) faster than they arrive, so that the network ought to be stable. The following result, due to Harrison and Williams (1987), gives the diffusion analogue.

Theorem 17. *The diffusion process Q has a stationary probability distribution π if and only if $c_i > 0$ for $1 \leq i \leq d$. Furthermore, any stationary probability distribution for Q is necessarily unique.*

Recall that if the inter-arrival and service time distributions are exponential, the queueing network studied here is a Jackson network and the stationary distribution (when it exists) is known to be of product form. This product form structure also manifests itself in the diffusion limit Q. We say that a probability distribution π on $\mathbb{R}^d$ is of *product form* if

$$\pi(\mathrm{d}x_1 \times \cdots \times \mathrm{d}x_d) = \prod_{i=1}^{d} p_i(x_i)\,\mathrm{d}x_i\,.$$

The next result is also due to Harrison and Williams (1987).

Theorem 18. *The diffusion process Q has a stationary probability distribution π of product form if and only if $c_i > 0$ for $1 \leq i \leq d$ and*

$$2\Gamma_{jk} = -(P_{kj}\Gamma_{kk} + P_{jk}\Gamma_{jj}) \tag{7.12}$$

for $j \neq k$. Furthermore, if a product form stationary probability distribution π exists,

$$p_i(x) = \eta_i \exp(-\eta_i x) I(x \geq 0)$$

where $\eta_i = 2\mu_i c_i / \Gamma_{ii}$.

In addition, Harrison and Williams (1989) show that the notion of quasireversibility for queueing networks extends, in a natural way, to the Brownian limit. Roughly speaking, a Brownian model of a queueing station is quasireversible if and only if the departure process has the same distribution as the arrival process. It is believed that a network of quasireversible Brownian queueing stations must necessarily have a stationary distribution of product form.

We note that in the Jackson network setting, $a_i^2 = b_i^2 = 1$ and (7.12) is easily verified from (7.11). In general, the stationary probability distribution π cannot be calculated in closed form analytically. Nevertheless, the Itô-type argument of Section 6 (used there to justify the backwards equation for one-dimensional reflected Brownian motion) can be used to obtain an equation that typically characterizes the stationary distribution π. In preparation for the statement of the result, let $F_k = \{x \in \mathbb{R}^d : x_k = 0\}$ be the kth face of the non-negative d-dimensional orthant, and let L and D be the differential operators defined by

$$L = \tfrac{1}{2} \sum_{i,j=1}^{d} \Gamma_{ij} \frac{\partial^2}{\partial x_i \, \partial x_j} + \sum_{i=1}^{d} \theta_i \frac{\partial}{\partial x_i}$$

and

$$D_i = \mu_i \left(\frac{\partial}{\partial x_i} - \sum_{j \neq i} P_{ij} \frac{\partial}{\partial x_j} \right)$$

where $\theta = c(I - P)$. Finally, let C^2 be the space of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that f is twice continuously differentiable and such that all partials up to order 2 are bounded on $\{x : x_i \geq 0, 1 \leq i \leq d\}$.

Theorem 19. *Suppose π is a stationary probability distribution for Q. Then, there exist finite measures $\nu_1, \ldots, \nu_d$ on $F_1 \ldots, F_d$ such that for each $f \in C^2$,*

$$\int_{\mathbb{R}^d} (Lf)(x) \pi(\mathrm{d}x) + \tfrac{1}{2} \sum_{i=1}^{d} \int_{F_i} (D_i f)(x) \nu_i(\mathrm{d}x) = 0 . \tag{7.13}$$

See Harrison and Williams (1987) for details. The equation (7.13) is known as the *basic adjoint relationship* satisfied by π. Given the importance of the stationary distribution π from an application viewpoint, significant activity is currently underway to solve (7.13) numerically. A recent paper by Harrison and Nguyen (1989) specializes its numerical treatment to $d = 2$. It shows that in the two station setting, the numerical approximations obtained from Theorem 19 give good results, in comparison with simulation, even for moderate traffic intensities.

For the class of open queueing networks that we have described above, it turns out that sojourn times can also be approximated (in distribution) by appropriate diffusion limits. To describe the notion of sojourn time, let $h = (h_1, \ldots, h_d)$ be a d-vector having non-negative integer-valued co-ordinates, and fix a station j. (Let us adopt the convention that station 0 corresponds to the world external to the network.) Suppose that the routing matrix P is such that it is possible for a customer, upon entering j and before returning to it, to follow a route in which station k is visited precisely h_k times (note that we must necessarily have $h_j = 1$); such an h is termed a *j-accessible visit vector*. A customer who follows such a route is said to *follow h upon entering j*. The time it takes a specified customer to follow h is called its *sojourn time along h*. For a j-accessible h, let $D_n(j, h, t)$ be the sojourn time (in the nth approximating network) along h of the first customer that follows h upon entering j after time t.

Before stating the limit theorem for $D_n(j, h, t)$, recall that Theorem 16 implies that the queue-lengths at each of the d stations (in the nth approximating network) is of order $n^{1/2}$ in magnitude. Since each station has first come/first serve queueing discipline, it is evident that the amount of time that a customer spends in the queue of station j is roughly Q_j/μ_j, where Q_j is the queue-length at the instant at which the customer joins the queue at station j. (We obtain the approximation Q_j/μ_j by noting that customers at station j are served (asymptotically) at rate μ_j.) Hence, the sojourn time of a customer in the nth system is of order $n^{1/2}$ in magnitude. Since time is measured in units of order n in the nth network (see Theorem 16), this implies that a customer visits all the stations along its route instantaneously in the diffusion limit. Hence, in the diffusion time scale, queues and workloads seen by a customer arriving to station j freeze while the customer follows h. As a consequence, the limit process for sojourn times has the property that the order in which stations are visited along a route does not affect the diffusion limit of the route's travel time. Furthermore, the diffusion limits of $D_n(j, h, t)$ and $D_n(k, h, t)$ are identical, provided that h is both j-accessible and k-accessible; see Reiman (1984) and Chen and Mandelbaum (1988) for details.

Theorem 20. *Assume* (7.3) *and* (7.9). *Let* $Q(t) = (Q_1(t), \ldots, Q_d(t))$ *be the limit process specified by Theorem* 16. *Set* $D^n(j, h, t) = D_n(j, h, nt)/n^{1/2}$. *Then,*

$$D^n(j, h, t) \xrightarrow{\mathrm{w}} \sum_{i=1}^{d} h_i Q_i(t)/\mu_i$$

in $D_{\mathbb{R}}[0, \infty)$ *as* $n \to \infty$.

We conclude this section with a brief discussion of some extensions of the open queueing model that has been described above. Firstly, (7.3) can be relaxed to deal with networks in which certain stations are not asymptotically in heavy traffic. (In other words, there exists at least one station i for which $\mu_i \neq (\lambda(I - P)^{-1})_i$.) This generalization has been fully explored in Chen and Mandelbaum (1988). Basically, the stations of such a network can be classified into one of three types: balanced, non-bottleneck, and strict bottleneck stations. At a balanced station, the 'effective arrival rate' (both internal and external arrivals) is roughly identical to the service rate, and the station is in heavy traffic. On the other hand, at a non-bottleneck station, the effective arrival rate is less than the service rate, so that the station is in light traffic. As a consequence, the queue length at a non-bottleneck station is basically bounded (in a stochastic sense) as a function of n. Thus, in the diffusion re-scaling in which the spatial variables are measured in units of $n^{1/2}$, the non-bottlenecks are essentially drained of customers instantaneously, after which the corresponding queue-length processes are effectively zero. Hence, the diffusion limit gives no information about the queue lengths at the non-bottleneck stations. In some sense, the queueing network can then be reduced to an analysis of the remaining balanced and strict bottleneck stations. This is an example of what is known, in the literature, as 'state space collapse'. This should not be interpreted as suggesting that the non-bottlenecks play no role in the diffusion limit for, in fact, they do. For example, the diffusion limit needs to keep track of how customers move between stations in the reduced network consisting of balanced and strict bottleneck stations. The routing matrix of the full network (including non-bottleneck stations) must be analyzed in order to calculate the submatrix which characterizes flows within the reduced network.

The behavior of the network at strict bottlenecks is comparable to that of a queue in which the arrival rate is strictly greater than the service rate. As we saw in Section 5, oversaturated queues have very simple behavior. In particular, the queue lengths grow linearly in time. When appropriately centered to reflect the linear trend, the queue-length process exhibits fluctuations of order $n^{1/2}$. These fluctuations are contributed by a non-reflecting Brownian motion term (from the non-bottleneck stations) as well as a reflecting Brownian motion term (to handle input from the balanced stations).

Thus, the components of the vector-valued diffusion limit display three different types of behavior. Components corresponding to non-bottlenecks vanish identically, whereas balanced components exhibit reflection at the boundary of the non-negative orthant. On the other hand, the strict bottlenecks display no non-trivial boundary behavior (other than that induced from the balanced stations). This is because the diffusion limit merely characterizes the non-zero fluctuations about the linear drift term, as the queue-lengths drift to infinity. These fluctuations are of arbitrary sign.

We turn now to discussion of a second generalization of the queueing networks considered in this section. Suppose that the network serves two types of customers, one type of which has a pre-emptive resume priority over the other at each station of the network. In this setting, the high priority customers are unaffected by the low priority customers. Hence, in heavy traffic, the high priority customers race through the network relative to the low priority customers. Thus, the queue-length processes (at the various stations) of the high priority customers are negligible compared to those of the low priority customers. As a consequence, in the diffusion limit, the components corresponding to high priority customers vanish identically. However, the effect of the high priority customers is manifested in the limit processes obtained for the low priority customers; see Johnson (1983) for further details. It is also possible to obtain diffusion approximations for multiple customer type networks, in which the routing is 'feedforward' (i.e., the stations can be numbered so that the route of each customer type is an increasing sequence). The queue discipline at each station j is defined by a partition of the customer types into subsets H_j and L_j. Customers of type $i \in H_j$ have pre-emptive resume priority over those of type $k \in L_j$. Within each subset H_j and L_j, the queue discipline is first come/first served. As in the previous priority queue that was described, the queue-lengths of the high priority customers at a particular station vanish in the diffusion limit. On the other hand, the queue-lengths corresponding to lower priority customers at a particular station are represented by fixed multiples of a single limiting marginal process corresponding to total queue length at that station. Thus, the diffusion limit for such a d station queueing network (with multiple customer types) is basically a d dimensional process. This is another example of 'state space collapse' (i.e., the diffusion limit is d-dimensional, whereas the exact model records the number of customers of each type at each station and is therefore higher dimensional); this limit theorem is described in Peterson (1985).

For single station open systems, several other diffusion approximations have been studied. Johnson (1983) considers a single station model with two or more customer types under first come/first serve and processor sharing queue disciplines. A heavy traffic diffusion limit is obtained for the total queue-length process; it is then shown that a fixed fraction of the limit process corresponds to customers of a particular type. A similar result is obtained in Reiman (1983). The heavy traffic behavior of a two queue system in which customers join the shortest queue is shown to converge to a limit process with equal fractions of customers at each of the two stations. Thus, in the limit, only the total number of customers in the system varies stochastically. A further example of 'state space collapse' appears in Reiman (1988), where a multi-class feedback queue with round-robin service discipline is studied.

We conclude this discussion of generalizations of Theorem 16 by pointing out that the i.i.d. assumptions on the inter-arrival times, service times, and routing vectors are unnecessary to the basic argument. As suggested by (7.4)–(7.6), all that is really needed is that the various input processes satisfy a joint FCLT. This point of view is stressed in the modeling approach described in Harrison and Williams (1987) and Harrison (1985).

8. Diffusion approximations for a closed network of queues in heavy traffic

In this section, we describe the closed network analogs of the results given in Section 7 for open networks. We start by describing the basic model. As in the open case, we consider a network of d queueing stations, in which each station has an infinite waiting room and a single work-conserving server that serves customers in the order in which they arrive. The customer routing between stations is assumed to be Markovian and is described by a stochastic matrix P (called the routing matrix); we require that P be irreducible and that $P_{ii} = 0$ for $1 \leqslant i \leqslant d$. As in Section 7, let $V_i = \{V_i(m): m \geqslant 1\}$ be a sequence of i.i.d. unit mean r.v.'s for which $b_i^2 = \text{var } V_i(m) < \infty$. If $\mu_j(n)$ is the service rate at the jth station in the nth network, then $V_j(n)/\mu_j(n)$ is the actual duration of service for the mth customer served at the jth station in the nth network.

We assume that the nth approximating network contains precisely $[n^{1/2}]$ customers. Thus, the parameter n that indexes the system here has a physically tangible meaning (unlike the open case); it characterizes the total number of customers in the closed network. Let $Q_n(0) = (Q_{n1}(0), \ldots, Q_{nd}(0))$ be a random d-vector having non-negative integer-valued components for which $Q_{n1}(0) + \cdots + Q_{nd}(0) = [n^{1/2}]$. The ith component is to be interpreted as the number of customers that are sitting in the queue of the ith station waiting to be served at time $t = 0$ (in the nth system). We require that $V_1, \ldots, V_d$, $Q_n(0)$ be mutually independent. The network that we have just described is termed a *generalized closed Jackson network.*

To obtain heavy traffic behavior at each of the d stations, we require that the service rate $\mu_i(n)$ be approximately proportional to π_i, where $\pi = (\pi_1, \ldots, \pi_d)$ is the unique stationary probability vector associated with P (i.e., $\pi P = \pi$). Thus, heavy traffic is obtained when the service rate is roughly proportional to the relative expected number of visits that a customer makes to a given station. Stated in mathematical terms, we require that

$$n^{1/2}(\mu(n) - \mu) \to c \tag{8.1}$$

as $n \to \infty$, where $\pi_i/\mu_i = d$ (a constant). Since the diffusion approximation will be obtained by speeding up time by a factor of n and re-scaling space by a factor of $n^{1/2}$, we can re-express the conditions for heavy traffic in the following way. A diffusion limit will be a reasonable approximation to a closed queue if the total number of customers m is large and the service rate μ_i at the ith station has the property that

$$\max_{1 \leq i \leq d} (\pi_i/\mu_i) - \min_{1 \leq i \leq d} (\pi_i/\mu_i)$$

is roughly of order $1/m$. Then, the diffusion limit will describe fluctuations, in the queueing network, of order m on the time scale of order m^2.

As in the open network setting, the theory of diffusion approximation for closed systems depends critically on representing the vector queue-length process $Q_n(t) = (Q_{n1}(t), \ldots, Q_{nd}(t))$ as a regulated version of the centered input processes corresponding to the service times and routing vectors. Specifically, one represents Q_n as $Q_n = f(Z_n)$, where Z_n is a process that can be approximated by a Brownian motion and $f(z) = z + h(z)(I - P)$, where h satisfies (7.9).

To state the diffusion approximation for the above class of closed networks, we need to further assume that there exists a random d-vector $Q(0)$ such that

$$n^{-1/2}Q_n(0) \xrightarrow{w} Q(0) \tag{8.2}$$

as $n \to \infty$; the vector $Q(0)$ must necessarily have components that add to one. Then, the process Z_n (when appropriately re-scaled) can be approximated by Z, where

$$Z(t) = Q(0) + c(I - P)t + \Gamma^{1/2}B(t)\,, \tag{8.3}$$

the random elements $Q(0)$ and B appearing in (8.3) are independent. Also, B is a standard Brownian motion, c is the vector appearing in (8.1) and Γ is the covariance matrix given by

$$\Gamma_{jk} = \mu_j \delta_{jk}(1 + b_j^2) - \mu_j b_j^2 P_{jk} - \mu_k b_k P_{kj} - \sum_{l=1}^{d} \mu_l P_{lj} P_{lk}(1 - b_l^2) .$$

By applying continuous mapping ideas to the mapping f, one obtains the following result; see Chen and Mandelbaum (1988) for a complete proof.

Theorem 21. *Assume* (8.1) *and* (8.2). *Let* $Q^n(t) = Q_n(nt)/n^{1/2}$. *Then,*

$$Q^n \xrightarrow{w} f(Z) \overset{\Delta}{=} Q$$

in $D_{\mathbb{R}^d}[0, \infty)$ *as* $n \to \infty$, *where* Z *is defined by* (8.3).

We observe that since the vector Q_n is normalized by $n^{1/2}$ to obtain Q^n, it is evident that Q must be a stochastic process that lives on the simplex $S = \{x \in \mathbb{R}^d: x_i \geq 0, x_1 + \cdots + x_d = 1\}$. As a consequence, the covariance matrix Γ must be singular (since the d-dimensional Brownian motion Z must lie in a $(d-1)$ dimensional subspace). We henceforth assume that Γ is such that the $(d-1) \times (d-1)$ principal submatrices of Γ are positive definite. Because of properties (C1)–(C3) of the regulator map, it turns out that Q is (as in Section 7) a diffusion process. In fact, Q can be viewed a a regulated version of the Brownian motion Z. The regulation forces Z to 'reflect' in the direction of the ith row of $I - P$ whenever the ith component of Z is zero.

Since Q takes values in the compact set S, it seems reasonable to expect that Q possesses a stationary distribution π. This is the principal content of the following result, due to Harrison, Williams and Chen (1989).

Theorem 22. *The diffusion process Q has a unique stationary probability distribution π. The probability distribution π has a strictly positive density $p(x)$ with respect to Lebesgue measure on S.*

Since a certain subclass of the diffusion processes Q described above can be obtained as limits of Markovian Jackson networks having 'product form' stationary distributions, one hopes that the product form theory carries over to the diffusion setting, thereby permitting $p(x)$ to be calculated explicitly in certain cases. We say that p is an *exponential density* if it can be represented in the form

$$p(x) = C \prod_{i=1}^{d} \exp(-\eta_i x_i)$$

for $x \in S$ (for some constant C). The next result is also due to Harrison, Williams and Chen (1989).

Theorem 23. *The density p of the stationary distribution π is an exponential density if and only if*

$$2\Gamma_{jk} = -(P_{kj}\Gamma_{kk} + P_{jk}\Gamma_{kk}) \tag{8.4}$$

for $j \neq k$. Furthermore, if p is an exponential density, then

$$p(x) = C \prod_{i=1}^{d} \exp(-\eta_i x_i) ,$$

where

$$\eta_i = 2\mu_i c_i / \Gamma_{ii} , \quad C = \int_S \prod_{i=1}^{d} \exp(-\eta_i x_i) m(\mathrm{d}x)$$

and m is Lebesgue measure on S.

As in the open case, condition (8.4) is automatically satisfied when $b_i^2 = 1$ (i.e., the service times have the same coefficient of variation as does an exponential r.v.). Also, if (8.4) is satisfied and $\mu(n) = \mu$ for $n \geq 1$, then $c = 0$ and the stationary distribution π then evidently reduces to uniform distribution on the simplex.

We conclude our discussion of the qualitative structure of the process Q by describing the analog of the basic adjoint relationship (7.13) for closed networks; see Harrison, Williams and Chen (1989) for further details.

Theorem 24. *Let L, D_i, C^2, F_i be defined as in Section 7. Suppose π is the stationary probability distribution for Q. Then, there exist finite measures $\nu_1, \ldots, \nu_d$ on $F_1 \cap S, \ldots, F_d \cap S$ such that for each $f \in C^2$,*

$$\int_S (Lf)(x)p(x)m(\mathrm{d}x) + \tfrac{1}{2} \sum_{i=1}^{d} \int_{F_i \cap S} (D_i f)(x)\nu_i(\mathrm{d}x) = 0 . \tag{8.5}$$

The development of numerical solvers for dealing with (8.5) remains an important open problem.

Diffusion limits can also be obtained for the sojourn times that were defined in Section 7. The basic limit theorem is identical to Theorem 20 and the statement is omitted.

As for generalizations of the closed model that has been described here, a number of possibilities have been investigated in the literature. In Harrison, Williams and Chen (1989), the assumption that the service requirements at each section are i.i.d. is dropped, and replaced with a requirement that the input processes satisfy functional central limit theorems. Although no proof is offered, the paper does calculate the appropriate diffusion limit for such a model. Closed networks in which (8.1) is weakened to permit the possibility of including stations in light traffic are studied in Chen and Mandelbaum (1988).

As in the open case, the queue-length populations at the light traffic stations vanish in the diffusion limit. In addition, closed networks with multiple customer classes are considered in Chen and Mandelbaum (1988). The priority ranking of the various customer classes is assumed to be the same at each station of the network. Again, the theory is similar to that obtained in the open case; high priority customers disappear in the diffusion limit, although they do influence the structure of the limiting process associated with the lower priority customers.

9. Approximations for queues with many servers

In this section, we discuss limit theorems for queues in which each station possesses a single waiting room and a large number of servers. Customers are assigned to the first available server on a first come/first serve basis. The types of limits obtained here will typically not exhibit any of the boundary behavior that characterized the reflecting Brownian motions studied in the heavy traffic settings of Sections 5, 7, and 8. On the other hand, the limit processes that arise here need not be Markov processes. Thus, the approximations typical of the many server context are not true diffusion approximations, since the limit processes need not be diffusions. However, we choose to discuss these approximations here because of their intrinsic importance and because the ideas required to derive these limits are largely identical to those used to obtain the diffusion approximations described earlier in this chapter (see Sections 2 and 3).

We initiate this discussion by considering a single station queue with an infinite number of servers. Interesting limit behavior is obtained by sending the queue into heavy traffic. Heavy traffic, in this setting, means that the arrival rate is high, so that the expected number of busy servers is large. More precisely, consider a sequence of GI/G/∞/∞ queues constructed in the following manner. The service time sequence $V = \{V_i : i \geq 1\}$ is i.i.d. with common distribution F, whereas the inter-arrival times in the nth system are an independent sequence $U_n = \{U_{ni} : i \geq 1\}$ of i.i.d. r.v.'s in which U_{ni} can be represented as U_i/n; we assume that the system is idle at $t = 0$, for simplicity. Note that in the nth system, the inter-arrival times are re-scaled so that arrivals are occurring n times faster than in the first system, whereas the service times are not re-scaled. This is in contrast to the diffusion approximations previously discussed in this chapter, in which both inter-arrival times and service times are re-scaled simultaneously. In any case, we will establish a limit theorem for the queue-length process at the station as the parameter n tends to infinity.

The result is most transparent when the service time r.v.'s have a discrete distribution with finite support. Suppose, in particular, that F assigns probability p_i to the value x_i, for $1 \leq i \leq m$. Let $N_{ni}(t)$ be the total number of customers having received service time x_i by time t in the nth system. Because the nth inter-arrival stream is obtained from the first inter-arrival stream by speeding it up by a factor of n, we may write $N_{ni}(t) = N_i(nt)$, where $N_i(\cdot) \overset{\triangle}{=} N_{i1}(\cdot)$. Let $N(t) = N_1(t) + \cdots + N_m(t)$ be the total number of customers to arrive by time t in the first system. Then,

$$N_i(t) = \sum_{j=1}^{N(t)} I(V_j = x_i) .$$

Suppose that $\sigma^2 = \operatorname{var} U_1 < \infty$ and that $\lambda^{-1} = EU_1 > 0$. Then, the multivariate version of Donsker's theorem applies (see Theorem 4), yielding the fact that

$$\begin{aligned} n^{1/2}\Bigg(n^{-1} \sum_{i=1}^{[nt]} I(V_j = x_1) - tp_1, \ldots, n^{-1} \sum_{i=1}^{[nt]} I(V_j = x_m) - tp_m, \\ n^{-1} \sum_{i=1}^{[nt]} U_i - t\lambda^{-1}\Bigg) \end{aligned} \tag{9.1}$$

converges to a Brownian motion taking values in $\mathbb{R}^{m+1}$. Using a random time change argument similar to that used to obtain (5.1), we can substitute $t' = n^{-1}N(nt)$ into (9.1), thereby yielding the fact that

$$\begin{aligned} n^{1/2}(n^{-1}N_1(nt) - n^{-1}N(nt)p_1, \ldots, n^{-1}N_m(nt) - n^{-1}N(nt)p_m, \\ t - n^{-1}N(nt)\lambda^{-1}) \end{aligned}$$

converges to a $\mathbb{R}^{m+1}$-valued Brownian motion. The continuous mapping principle then implies that

$$n^{1/2}(n^{-1}N_{n1}(t) - \lambda p_1 t, \ldots, n^{-1}N_{nm}(t) - \lambda p_m t) \tag{9.2}$$

converges to an m-dimensional Brownian motion $B = \{B(t) = (B_1(t), \ldots, B_m(t)) \colon t \geq 0\}$. Recalling that the service times for customers corresponding to $N_{ni}(t)$ are all identical to x_i, we find that $Q_{ni}(t) = N_{ni}(t) - N_{ni}((t - x_i) \wedge 0)$, where $Q_{ni}(t)$ is the number of customers at the station at time t in the nth system that were assigned service time x_i. Consequently, $Q_n(t)$ (the total number of customers at the station at time t in the nth system) can be represented as

$$Q_n(t) = \sum_{i=1}^{m} (N_{ni}(t) - N_{ni}((t - x_i) \wedge 0)) .$$

It then follows from (9.2) that

$$n^{1/2}(n^{-1}Q_n(t) - \lambda E\{V_1 \wedge t\}) \xrightarrow{\mathrm{w}} \sum_{i=1}^{m} (B_i(t) - B_i((t - x_i) \vee 0)) . \tag{9.3}$$

Since Brownian motion has finite dimensional distributions that are Gaussian, it is evident that the same must be true of the limit process appearing in (9.3). As a consequence, the limit appearing in (9.3) is termed a *Gaussian approximation*. Gaussian processes are highly tractable, since their finite dimensional distributions are totally characterized by their mean and covariance functions. Thus, these limit processes are somewhat easier to use (when applicable) than the diffusion limits of Sections 7 and 8, since the diffusions obtained there typically require sophisticated numerical routines to calculate performance measures of interest. We also note that (9.3)'s limit process has the interesting property that for $t > \max\{x_i: 1 \leqslant i \leqslant m\}$, the marginal distribution is independent of t (i.e., the system reaches steady-state in finite time).

A straightforward calculation shows that the covariance function of the limit appearing in (9.3) is given by

$$\begin{aligned} c(s, s+t) = {} & \lambda \int_0^s F(u)(1 - F(t+u))\,\mathrm{d}u \\ & + \sigma^2\lambda^3 \int_0^s (1 - F(t+u))(1 - F(u))\,\mathrm{d}u \end{aligned} \tag{9.4}$$

for s, $t \geqslant 0$ (note that the mean of (9.3)'s limit is identically zero). Since r.v.'s with arbitrary distribution can be approximated by discrete r.v.'s, this suggests that the above limit theorem ought to hold more generally. The following theorem is due to Borovkov (1967).

Theorem 25. *Consider a sequence of* GI/G/∞/∞ *queues constructed as described earlier in this section. Suppose that* $\lambda^{-1} = EU_1 > 0$ *and* $\sigma^2 = \mathrm{var}\ U_1 > \infty$. *If* $\mathrm{var}\ V_1 < \infty$, *then*

$$n^{1/2}(n^{-1}Q_n(t) - E\{V_1 \wedge t\}) \xrightarrow{\mathrm{w}} Q(t)$$

in $D_{\mathbb{R}^d}[0, \infty)$, *where* $\{Q(t): t \geqslant 0\}$ *is a process having Gaussian finite dimensional distributions. Furthermore,* $EQ(t) = 0$ *and its covariance function is given by* (9.4).

In contrast to the diffusion limits obtained earlier in this chapter, the distribution of the Gaussian limit $Q = \{Q(t): t \geqslant 0\}$ depends on the entire service time distribution F, not just on its mean and variance. Thus, the tail behavior of the service times has a significant impact on an infinite server queue in heavy traffic. We further note that if $F(t) = 1 - \mathrm{e}^{-\mu t}$, then the covariance function is identical to that of an Ornstein–Uhlenbeck process. As a consequence, it turns out that for a GI/M/∞/∞ queue, the limit process Q is an Ornstein-Uhlenbeck process with infinitesimal mean $-\mu x$ and infinitesimal variance $(\lambda^3\sigma^2 + \lambda)$. Thus, in the case of exponential service times, the limit is a diffusion. However, Q is typically not Markov. Glynn (1982) shows that Q is Markov if and only if $F(t) = 1 - p\,\mathrm{e}^{-\mu t}$ for $0 < p \leqslant 1$ and $\mu > 0$.

The process $Q(t)$ converges in distribution to a limit $Q(\infty)$ as $t \to \infty$. The steady-state r.v. $Q(\infty)$ is normally distributed with zero mean and variance

$$z = \frac{\lambda}{\mu}\left(1 + (\lambda^2\sigma^2 - 1)\mu \int_0^\infty (1 - F(t))^2 \, \mathrm{d}t\right),$$

where $\mu^{-1} = EV_1$. This is another confirmation of the analytical tractability of the Gaussian limit Q; clearly, $Q(\infty)$ can be used as an approximation to the long-run behavior of a $\mathrm{GI}/\mathrm{G}/\infty/\infty$ queue in heavy traffic.

Similar Gaussian approximations to Theorem 25 can be obtained for the cumulative departure process of a $\mathrm{GI}/\mathrm{G}/\infty/\infty$ queue of the type described above; see Whitt (1984) for further details. In addition, Gaussian limits can be derived for networks of infinite server stations; the relevant techniques are sketched out in Whitt (1982). Finally, it turns out that one can extend these results to finite server stations in which the number of servers increases with the arrival rate suitably rapidly. Note that the number of busy servers (according to Theorem 25) in the nth infinite server queue is approximately $n\lambda EV_1$. Hence, if the number of available servers s_n associated with the nth system grows sufficiently more rapidly than $n\lambda EV_1$, the finite server model will act asymptotically like the infinite server system. Specifically, this holds if $n^{-1/2}(s_n - n\lambda EV_1) \to \infty$ as $n \to \infty$.

We conclude this section by briefly discussing approximations for closed networks with a large number of servers at each station. Again, we start by considering the case where each of the d stations has an infinite number of servers. We assume, for concreteness, that the customer routing between stations is Markovian (although this can easily be generalized). In addition, the service time streams for each of the d stations form independent sequences of i.i.d. random variables with continuous distributions. We further assume that at $t = 0$, all the customers in the network are sitting at the first station, waiting to be served. We shall describe a limit theorem for the network by letting the number of customers n contained within the network tend to infinity. In contrast with previous approximations that we have analyzed, there will be no need to re-scale time in any way.

The key observation here is to recognize that since each station has an infinite number of servers, customers never queue for service. As a consequence, customers do not interfere with each other as they circulate through the network. Hence, each of the n paths followed by customers through the network are independent. Furthermore, it is clear that these paths are identically distributed. Let $X_i(t)$ be the station occupied by the ith customer at time t. Then, $Q_{nj}(t)$, the number of customers at the jth station at time t in the nth system, is given by

$$Q_{nj}(t) = \sum_{i=1}^{n} I(X_i(t) = j) .$$

If we let $Y_i(t) = (I(X_i(t) = 1), \ldots, I(X_i(t) = d))$, we conclude that $Q_n(t) = (Q_{n1}(t), \ldots, Q_{nd}(t))$ can be expressed as

$$Q_n(t) = \sum_{i=1}^{n} Y_i(t)$$

and hence $Q_n = \{Q_n(t): t \geqslant 0\}$ can be expressed as a sum of n i.i.d. $D_{\mathbb{R}^d}[0, \infty)$-valued random elements. Central limit theorems exist for such objects. When applied in this setting, we obtain the following limit theorem; see Glynn and Kurtz (1990) for additional details.

Theorem 26. *Let Q be the Gaussian process with covariance function identical to that of Y_1. Then,*

$$n^{1/2}(n^{-1}Q_n(t) - EY_1(t)) \xrightarrow{w} Q(t)$$

in $D_{\mathbb{R}^d}[0, \infty)$ as $n \to \infty$.

The limit process $Q = \{Q(t): t \geqslant 0\}$ is again a Gaussian process that is typically non-Markovian. Since the processes Q_n are Markov when the service times are exponential, the same property is inherited by the limit Q in that case, however.

Suppose now that the routing matrix is irreducible and that at least one station has an associated service time distribution that is spread-out (i.e., some n-fold convolution of the distribution that is spread-out (i.e., some n-fold convolution of the distribution possesses a density component). We further require that the mean μ_i^{-1} of the service time distribution for the ith station is finite for $1 \leqslant i \leqslant d$. Then, $Q_n(t) \xrightarrow{w} Q_n(\infty)$ at $t \to \infty$, for some limiting r.v. $Q_n(\infty)$. The following result gives an approximation to $Q_n(\infty)$ when the number of customers n in the network is large.

Theorem 27. *Let π be the unique stationary distribution of the routing matrix P. Set $p_i = \pi_i \mu_i^{-1} / (\sum_{j=1}^{d} \pi_j \mu_j^{-1})$ and let $p = (p_1, \ldots, p_d)$. Then,*

$$n^{1/2}(n^{-1}Q_n(\infty) - p) \xrightarrow{w} \Gamma^{1/2} N(0, I)$$

in $\mathbb{R}^d$, where $\Gamma_{ii} = p_i(1 - p_i)$ and $\Gamma_{ij} = -p_i p_j$ for $i \neq j$.

Limit theorems have also been derived for closed Jackson networks in which the number of servers at each station is large but finite. (Recall that in a Jackson network, service times are exponentially distributed.) The limit processes in this setting are typically vector-valued Ornstein–Uhlenbeck processes. The tools that are used here are somewhat different from those described earlier in this chapter. Rather than attempt to represent the queue-length process as some continuous functional of its inputs (thereby permitting one to use continuous mapping ideas), the technique that has commonly been used here is to show that the infinitesimal generator of an appropriately scaled version of the vector queue-length process converges to the infinitesimal generator of the limiting Ornstein–Uhlenbeck process. This approach is analytical, in contrast to the more probabilistic continuous mapping approach used earlier in this chapter. For two-station networks, the work of Stone (1963) on weak convergence of birth death processes can be used. For more general networks, the techniques outlined in Stroock and Varadhan (1979) have proved successful. For further details on these limit theorems, see Iglehart (1965) and Prisgrove (1987).

10. Conditional weak convergence theorems

In this section, we briefly describe an interesting class of diffusion limits that arise as approximations to the behavior of certain stochastic processes when conditioned on an appropriate rare event.

Consider, for example, the behavior of the waiting time sequence $\{W_n: n \geq 0\}$ of the single server queue GI/G/1/∞. As discussed in Section 5, the W_n's satisfy the recursion $W_{n+1} = [W_n + X_{n+1}]^+$, where $X_n = V_{n-1} - U_n$ and V_n, U_n are the nth service time and interarrival time, respectively. If we assume that the 0th customer encounters an idle server, then $W_0 = 0$. We shall be interested in the behavior of the waiting time sequence within the first busy period (i.e., over the interval $[0, T)$, where $T = \inf\{n \geq 1: W_n = 0\}$). We note that if we let $S_n = \Sigma_{i=1}^n X_i$ $(S_0 = 0)$, $W_n = S_n$ for $n < T$. Hence, we can alternatively view the problem as the study of random walk over the interval $[0, T')$, where $T' = \inf\{n \geq 1: S_n \leq 0\}$ is the time of first entry into $(-\infty, 0]$.

Consider the case where the traffic intensity of the queue is equal to one; this translates into the assumption that $EX_1 = 0$. Two particular processes, defined in terms of standard real-valued Brownian motion, play a special role in the subsequent development.

Definition 4. Let $B = \{B(t): t \geq 0\}$ be a real-valued standard Brownian motion. Set $\tau_1 = \sup\{t \in [0, 1]: B(t) = 0\}$ and $\Delta_1 = 1 - \tau_1$. Then, the process $B^+ = \{B^+(t): 0 \leq t \leq 1\}$ defined by

$$B^+(t) = |B(\tau_1 + t\Delta_1)| / \Delta_1^{1/2}$$

is called (standard) *Brownian meander.*

Definition 5. Let B and τ_1 be defined as in Definition 4. Set $\tau_2 = \inf\{t \geq 1: B(t) = 0\}$ and $\Delta_2 = \tau_2 - \tau_1$. Then, the process $B_0^+ = \{B_0^+(t): 0 \leq t \leq 1\}$ defined by

$$B_0^+(t) = |B(\tau_1 + t\Delta_2)| / \Delta_2^{1/2}$$

is called (standard) *Brownian excursion.*

The processes B^+ and B_0^+ clearly have continuous paths. Furthermore, it can be shown that they are both (strong) Markov and hence diffusions. In fact, the transition density of B^+ is given by

$$P\{B^+(t) \in \mathrm{d}y \,|\, B^+(s) = x\} = g(t-s, x, y)\,\frac{|N|(y/(1-t)^{1/2})}{|N|(x/(1-s)^{1/2})}\,\mathrm{d}y$$

for $0 < s < t \leq 1$ and $x, y > 0$, where

$$g(t, x, y) = (2\pi t)^{-1/2}[\exp(-\tfrac{1}{2}(y-x)^2/t) - \exp(-\tfrac{1}{2}(y+x)^2/t)]$$

and

$$|N|(x) = (2/\pi)^{1/2}\int_0^x \exp(-\tfrac{1}{2}u^2)\,\mathrm{d}u\,.$$

On the other hand, the transition density of B_0^+ is given by

$$P\{B_0^+(t) \in \mathrm{d}y \,|\, B_0^+(s) = x\}$$

$$= g(t-s, x, y)\left(\frac{1-s}{1-t}\right)^{3/2} \cdot \frac{y\,\mathrm{e}^{-y^2/2(1-t)}}{x\,\mathrm{e}^{-x^2/2(1-s)}}\,\mathrm{d}y\,,$$

for $0 < s \leq t < 1$ and $x, y > 0$.

Returning to the waiting time sequence of the GI/G/1/∞ queue (in heavy traffic), we consider the re-scaled process

$$Z_n(t) = n^{-1/2}W_{[nt]}$$

for $0 \leq t \leq 1$. We note that $Z_n \in D_{\mathbb{R}}[0, 1]$, the function space consisting of the restrictions of functions in $D_{\mathbb{R}}[0, \infty]$ to the interval $[0, 1]$. The idea is now to study the behavior of Z_n within the first busy period. One way to accomplish this is to consider the conditional probability distribution

$$P_n(\cdot) = P\{Z_n \in \cdot \,|\, T > n\}\,.$$

This permits us to study only those paths of the waiting time sequence in which the first busy period is still in progress at time n. The following theorem is due to Iglehart (1974).

Theorem 28. *Suppose that $E|X_1|^3 < \infty$ and that X_1 is either non-lattice or integer-valued with span* 1. *Then,*

$$P_n \xrightarrow{\mathrm{w}} P$$

in $D_{\mathbb{R}}[0, 1]$ as $n \to \infty$, where $P(\cdot) = P\{B^+ \in \cdot\}$.

Hence, the behavior of the waiting time sequence, when conditioned on the first busy period still being in progress, is well approximated by that of a Brownian meander.

Brownian excursion also arises as a limit of the waiting time sequence when conditioned on the behavior of the first busy period. Specifically, let

$$P'_n(\cdot) = P\{X_n \in \cdot \mid T = n\} .$$

Hence, P'_n describes the distribution of precisely those paths of the waiting time sequence that conclude their first busy period at time n. Iglehart (1975) states that $P'_n \xrightarrow{w} P'$ as $n \to \infty$, under certain conditions on X_1, where $P'(\cdot) = P\{B_0^+ \in \cdot\}$. Note that Brownian excursion returns to zero at time $t = 1$, in accordance with the observation that the nth waiting time is zero if $T = n$. Further conditioned limit theorems of the above type may be found in Durrett (1980), Kaigh (1976) and Kao (1978).

The mathematical tools that are used to establish conditioned limit theorems are somewhat different from those described earlier in this chapter. For example, even if the unconditional approximating processes are tight, there is no guarantee that the conditioned approximations will be tight when the conditioning event has probability tending to zero. Hence, the problem of establishing tightness becomes more delicate here. As a consequence, most of the results available pertain to conditioned limit theorems for very specific classes of stochastic process (for example, random walk). Typically, the structure of the process plays an important role in the argument that is needed. However, for certain types of conditioned limit theorems, general tools are available; see, for example, Durrett (1978).

References

Abate, J. and W. Whitt (1987a). Transient behavior of regulated Brownian motion I: starting at the origin. *Adv. in Appl. Probab.* **19**, 560–598.

Abate, J. and W. Whitt (1987b). Transient behavior of regulated Brownian motion II: non-zero initial conditions. *Adv. in Appl. Probab.* **19**, 599–631.

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

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

Chen, H. and A. Mandelbaum (1988). Stochastic discrete flow networks: diffusion approximations and bottlenecks. Working Paper, Graduate School of Business, Stanford University, Stanford, CA.

Duffie, D. and P. Protter (1988). From discrete to continuous time finance: weak convergence of the financial gain process. Working Paper, Graduate School of Business, Stanford University, Stanford, CA.

Durrett, R. (1980). Conditioned limit theorems for random walks with negative drift. *Z. Wahrsch. Verw. Gebiete* **52**, 277–287.

Durrett, R. (1978). Conditioned limit theorems for some null recurrent Markov processes. *Ann. Probab.* **6**, 798–828.

Ethier, S.N. and T.C. Kurtz (1986). *Markov Processes: Characterization and Convergence*. Wiley, New York.

Glynn, P.W. (1982). On the Markov property of the $GI/G/\infty$ Gaussian limit. *Adv. in Appl. Probab.* **14**, 191–194.

Glynn, P.W. and W. Whitt (1986a). Sufficient conditions for a functional limit theorem version of $L = \lambda W$. *Queueing Systems* **1**, 279–287.

Glynn, P.W. and W. Whitt (1986b). A central limit version of $L = \lambda W$. *Queueing Systems* **2**, 191–215.

Glynn, P.W. and D.L. Iglehart (1990). Simulation output analysis using standardized time series. *Math. Oper. Res.* **15**, 1–16.

Glynn, P.W. and T.G. Kurtz (1990). Gaussian approximations for closed networks of infinite server queues. Working Paper, Department of Operations Research, Stanford University, Stanford, CA.

Harrison, J.M. (1985). *Brownian Motion and Stochastic Flow Systems*. Wiley, New York.

Harrison, J.M. and V. Nguyen (1989). The QNET method for two-moment analysis of open queueing networks. Working Paper, Graduate School of Business, Stanford University, Stanford, CA.

Harrison, J.M. and M.I. Reiman (1981). Reflected Brownian motion on an orthant, *Ann. Probability* **9**, 302–308.

Harrison, J.M. and R.J. Williams (1987). Brownian models of open queueing networks with homogeneous customer populations. *Stochast.* **22**, 77–115.

Harrison, J.M. and R.J. Williams (1989). On the quasireversibility of a multiclass Brownian service station. Working Paper, Graduate School of Business, Stanford University, Stanford, CA.

Harrison, J.M., R.J. Williams and H. Chen (1989). Brownian models of closed queueing networks with homogeneous customer populations. Working Paper, Graduate School of Business, Stanford University, Stanford, CA.

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

Iglehart, D.L. and W. Whitt (1970a). Multiple channel queues in heavy traffic, I. *Adv. in Appl. Probab.* **2**, 150–177.

Iglehart, D.L. and W. Whitt (1970b). Multiple channel queues in heavy traffic, II: sequences, networks, and batches. *Adv. in Appl. Probab.* **2**, 355–369.

Iglehart, D.L. (1974). Functional central limit theorems for random walks conditioned to stay positive. *Ann. Probab.* **2**, 608–619.

Iglehart, D.L. (1975). Conditioned limit theorems for random walks. In: M. Puri (Ed.), *Stochastic Processes and Related Topics, Vol. 1.* Academic Press, New York.

Johnson, D.P. (1983). Diffusion approximations for optimal filtering of jump processes and for queueing networks. Ph.D. Dissertation, University of Wisconsin, Madison, WI.

Kaigh, W.D. (1976). An invariance principle of random walks conditioned by a late return to zero. *Ann. Probab.* **4**, 115–121.

Kao, P. (1978). Limiting diffusion for random walks with drift conditioned to stay positive. *J. Appl. Probab.* **15**, 280–291.

Karlin, S. and H.M. Taylor (1975). *A First Course in Stochastic Processes*. Academic Press, New York.

Karlin, S. and H.M. Taylor (1981). *A Second Course in Stochastic Processes*. Academic Press, New York.

Kingman, J.F.C. (1961). The single server queue in heavy traffic. *Proc. Cambridge Philos. Soc.* **57**, 902–904.

Komlós, J., P. Major and G. Tusnády (1975). An approximation of partial sums of independent r.v.'s and the sample d.f. I. *Z. Wahrsch. Verw. Gebiete* **32**, 111–131.

Marshall, K.T. (1968). Some inequalities in queues. *Oper. Res.* **16**, 651–665.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press, Cambridge, MA.

Peterson, W.P. (1985). Diffusion approximations for networks of queues with multiple customer types. Ph.D. Dissertation, Stanford University, Stanford, CA.

Philipp, W. and W. Stout (1975). *Almost Sure Invariance Principles for Partial Sums of Weakly Dependent Random Variables*. American Mathematical Society, Providence, RI.

Prisgrove, L.A. (1987). Closed queueing networks with multiple servers: transient and steady-state approximations. Ph.D. Dissertation, Stanford University, Stanford, CA.

Prohorov, Y. (1956). Convergence of random processes and limit theorems in probability theory. *Theory Probability Appl.* **1**, 157–214.

Prohorov, Y. (1963). Transient phenomena in processes of mass service. *Litovsk. Mat. Sb.* **3**, 199–205. [In Russian.]

Reiman, M.I. (1983). Some diffusion approximations with state space collapse. *Proceedings of the International Seminar on Modeling and Performance Evaluation Methodology*. Springer-Verlag, Berlin–New York.

Reiman, M.I. (1984). Open queueing networks in heavy traffic. *Math. Oper. Res.* **9**, 441–458.

Reiman, M.I. (1988). A multiclass feedback queue in heavy traffic. *Adv. in Appl. Probab.* **20**, 179–207.

Rosenkrantz, W. (1978). On the accuracy of Kingman's heavy traffic approximation in the theory of queues. *Z. Wahrsch. Verw. Gebiete* **51**, 115–121.

Schruben, L.W. (1983). Confidence interval estimation using standardized time series. *Oper. Res.* **31**, 1090–1108.

Stone, C.J. (1963). Limit theorems for random walks, birth and death processes, and diffusion processes. *Illinois J. Math.* **7**, 638–660.

Stroock, D.W. and S.R.S. Varadhan (1979). *Multidimensional Diffusion Processes*. Springer-Verlag, New York.

Whitt, W. (1982). On the heavy-traffic limit theorem for GI/G/∞ queues. *Adv. in Appl. Probab.* **14**, 171–190.

Whitt, W. (1984). Departures from a queue with many busy servers. *Math. Oper. Res.* **9**, 534–544.

Yamada, K. (1984). Diffusion approximations for storage processes with general release rules. *Math. Oper. Res.* **9**, 459–470.