Chapter 2

Random Walks in Applications

The random walks we have considered in Chapter 1 are easy to think about, because they have a relatively simple structure. However, the random walks are abstract, so that they may seem disconnected from reality. But that is not so!

Even though the random walks are abstract, they play a fundamental role in many applications. Many stochastic processes in applied probability models are very closely related to random walks. Indeed, we are able to obtain many stochastic-process limits for stochastic processes of interest in applied probability models directly from established probability limits for random walks, using the continuous-mapping approach.

To elaborate on this important point, we now give three examples of stochastic processes closely related to random walks. The examples involve stock prices, the Kolmogorov-Smirnov test statistic and a queueing model for a buffer in a switch. In the final section we discuss the engineering significance of the queueing model and the (heavy-traffic) stochastic-process limits.

2.1. Stock Prices

In some applications, random walks apply very directly. A good example is finance, which often can be regarded as yet another game of chance; see *A Random Walk Down Wall Street* by Malkiel (1996).

Indeed, we might model the price of a stock over time as a random walk; i.e., the position S_n can be the price in time period n. However, it is common to consider a refinement of the direct random-walk model, because the magnitude of any change is usually considered to be approximately proportional to the price.

A popular alternative model that captures that property is obtained by letting the price in period n be the *product* of the price in period $n-1$ and a *random multiplier* Y_n; i.e., if Z_n is the price in period n, then we have

$$Z_n = Z_{n-1} Y_n, \quad n \geq 1 . \tag{1.1}$$

That in turn implies that

$$Z_n = Z_0 (Y_1 \times \cdots \times Y_n), \quad n \geq 1 . \tag{1.2}$$

Just as for random walks, for tractability we often assume that the successive random multipliers $Y_n : n \geq 1$, are IID. Hence, if we take logarithms, then we obtain

$$log(Z_n) = log(Z_0) + S_n, \quad n \geq 0 ,$$

where $\{S_n : n \geq 0\}$ is a random walk, defined as in (3.4), with steps $X_n \equiv log(Y_n), \quad n \geq 1$ that are IID. With this multiplicative framework, the *logarithms of successive prices constitute an initial position plus a random walk.* Approximations for random walks thus produce direct approximations for the logarithms of the prices.

It is natural to consider limits for the stock prices, in which the duration of the discrete time periods decreases in the limit, so that we can obtain convergence of the sequence of discrete-time price processes to a continuous-time limit, representing the evolution of the stock price in continuous time. To do so, we need to change the random multipliers as we change n. We thus define a sequence of price models indexed by n. We let Z_k^n and Y_k^n denote the price and multiplier, respectively, in period k in model n. For each n, we assume that the sequence of multipliers $\{Y_k^n : k \geq 1\}$ is IID. Since the periods are shrinking as $n \to \infty$, we want $Y_k^n \to 1 \quad \text{as} \quad n \to \infty$. The general idea is to have

$$E[log(Y_k^n)] \approx m/n \quad \text{and} \quad Var[log(Y_k^n)] \approx \sigma^2/n .$$

We let the initial price be independent of n; i.e., we let $Z_0^n \equiv Z_0$ for all n.

Thus, we incorporate the scaling within the partial sums for each n. We make further assumptions so that

$$\mathbf{S}_n(t) \equiv S_{\lfloor nt \rfloor}^n \Rightarrow \sigma \mathbf{B}(t) + mt \quad \text{as} \quad n \to \infty \tag{1.3}$$

for each $t > 0$, where $\mathbf{B}$ is standard Brownian motion. Given (1.3), we obtain

$$log(\mathbf{Z}_n(t)) \equiv log(Z^n_{\lfloor nt \rfloor}) = log(Z_0) + S^n_{\lfloor nt \rfloor} \Rightarrow log(Z_0) + \sigma\mathbf{B}(t) + mt \ ,$$

so that

$$\mathbf{Z}_n(t) \equiv Z^n_{\lfloor nt \rfloor} \Rightarrow \mathbf{Z}(t) \equiv Z_0 exp(\sigma\mathbf{B}(t) + mt) \ ; \tag{1.4}$$

i.e., the price process converges in distribution as $n \to \infty$ to the stochastic process $\{\mathbf{Z}(t) : t \geq 0\}$, which is called *geometric Brownian motion.*

Geometric Brownian motion tends to inherit the tractability of Brownian motion. Since the moment generating function of a standard normal random variable is

$$\psi(\theta) \equiv E[exp(\theta N(0,1))] = exp(\theta^2/2) \ ,$$

the k^{th} moment of geometric Brownian motion for any k can be expressed explicitly as

$$E[\mathbf{Z}(t)^k] = E[(Z_0)^k] exp(kmt + k^2t^2\sigma^2/2) \ . \tag{1.5}$$

See Section 10.4 of Ross (1993) for an introduction to the application of geometric Brownian motion to finance, including a derivation of the Black-Scholes option pricing formula.

The analysis so far is based on the assumption that the random-walk steps $X^n_k \equiv log(Y^n_k)$ are IID with finite mean and variance. However, even though the steps must be finite, the volatility of the stock market has led people to consider alternative models. If we drop the finite-mean or finite-variance assumption, then we can still obtain a suitable continuous-time approximation, but it is likely to be a geometric stable Lévy motion (obtained by replacing the Brownian motion by a stable Lévy motion in the exponential representation in (1.4)). Even other limits are possible when the steps come from a double sequence $\{\{X^n_k : k \geq 1\} : n \geq 1\}$. When we consider models for volatile prices, we should be ready to see stochastic-process limits with jumps. For further discussion, see Embrechts, Klüppelberg and Mikosch (1997), especially Section 7.6.

In addition to illustrating how random walks can be applied, this example illustrates that we sometimes need to consider double sequences of random variables, such as $\{\{X^n_k : k \geq 1\} : n \geq 1\}$, in order to obtain the stochastic-process limit we want.

2.2. The Kolmogorov-Smirnov Statistic

For our second random-walk application, let us return to the empirical cdf's considered in Example 1.1.1 in Section 1.1.3. What we want to see now is a stochastic-process limit for the difference between the empirical cdf and the underlying cdf, explaining the statistical regularity we saw in Figure 1.8. The appropriate limit process is the *Brownian bridge* $\mathbf{B}_0$, which is just Brownian motion $\mathbf{B}$ over the interval $[0,1]$ conditioned to be 0 at the right endpoint $t = 1$.

Recall that the applied goal is to develop a statistical test to determine whether or not data from an unknown source can be regarded as an independent sample from a candidate cdf F. The idea is to base the test on the "difference" between the candidate cdf and the empirical cdf. We determine whether or not the observed difference is significantly greater than the difference for an independent sample from the candidate cdf F is likely to be. The problem, then, is to characterize the probability distribution of the difference between a cdf and the associated empirical cdf obtained from an independent sample. Interestingly, even here, random walks can play an important role.

Hence, let F be an arbitrary continuous candidate cdf and let F_n be the associated empirical cdf based on an independent sample of size n from F. A convenient test statistic, called the *Kolmogorov-Smirnov statistic*, can be based on the limit

$$D_n \equiv \sqrt{n} \sup_{t \in \mathbb{R}} \{|F_n(t) - F(t)|\} \Rightarrow sup(|\, \mathbf{B}_0 \,|) \quad \text{as} \quad n \to \infty \ , \tag{2.1}$$

where $\mathbf{B}_0$ is the Brownian bridge, which can be represented as

$$\mathbf{B}_0(t) = \mathbf{B}(t) - t\mathbf{B}(1), \quad 0 \le t \le 1, \tag{2.2}$$

$$sup(|\, \mathbf{B}_0 \,|) \equiv \sup_{0 \le t \le 1} \{|\, \mathbf{B}_0(t) \,|\}$$

and

$$P(sup(|\, \mathbf{B}_0 \,|) > x) = 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 x^2}, \quad x > 0 \ . \tag{2.3}$$

Notice that the limit in (2.1) is independent of the cdf F (assuming only that the cdf F is continuous). The candidate cdf F could be the uniform cdf in Example 1.1.1, a normal cdf, a Pareto cdf or a stable cdf. In particular, the limit process here is unaffected by the cdf F having a heavy tail.

In practice, we would compute the Kolmogorov-Smirnov statistic D_n in (2.1) for the empirical cdf associated with the data from the unknown source and the candidate cdf F. We then compute, using (2.3), the approximate probability of observing a value as large or larger than the observed value of the Kolmogorov-Smirnov statistic, under the assumption that the empirical cdf does in fact come from an independent sample from F. If that probability is very small, then we would reject the hypothesis that the data come from an independent sample from F.

As usual, good judgement is needed in the interpretation of the statistical analysis. When the sample size n is not large, we might be unable to reject the hypothesis that the data is an independent sample from a cdf F for more than one candidate cdf F. On the other hand, with genuine data (not a simulation directly from the cdf F), for any candidate cdf F, we are likely to be able to reject the hypothesis that the data is an independent sample from F for all n sufficiently large. Our concern here, though, is to justify the limit (2.1).

So, how do random walks enter in? Random walks appear in two ways. First, the empirical cdf $F_n(t)$ as a function of n itself is a minor modification of a random walk. In particular,

$$nF_n(t) = \sum_{k=1}^{n} I_{(-\infty,t]}(X_k) \ ,$$

where $I_A(x)$ is the *indicator function* of the set A, with $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ otherwise. Thus, for each t, $nF_n(t)$ is the sum of the n IID Bernoulli random variables $I_{(-\infty,t]}(X_k)$, $1 \le k \le n$, and is thus a random walk.

Note that the Bernoulli random variable $I_{(-\infty,t]}(X_k)$ has mean $F(t)$ and variance $F(t)F^c(t)$. Hence we can apply the SLLN and the CLT to deduce that

$$F_n(t) \to F(t) \quad w.p.1 \quad \text{as} \quad n \to \infty$$

and

$$\sqrt{n}(F_n(t) - F(t)) \Rightarrow N(0, F(t)F^c(t)) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty \tag{2.4}$$

for each $t \in \mathbb{R}$. Note that we have to multiply the difference by $\sqrt{n}$ in (2.4) in order to get a nondegenerate limit. That explains the multiplicative factor $\sqrt{n}$ in (2.1).

Paralleling the way we obtained stochastic-process limits for random walks in Section 1.2, we can go from the limit in (2.4) to the limit in (2.1) by extending the limit in (2.4) to a stochastic-process limit in the function space D. We can establish the desired stochastic-process limit in D in two steps: first, by reducing the case of a general continuous cdf F to the case of the uniform cdf (i.e., the cdf of the uniform distribution on $[0,1]$) and, second, by treating the case of the uniform cdf. Random walks can play a key role in the second step.

To carry out the first step, we show that the distribution of D_n in (2.1) is independent of the continuous cdf F. For that purpose, let $U_k, 1 \le k \le n$, be uniform random variables (on $[0,1]$) and let G_n be the associated empirical cdf. Recall from equation (3.7) in Section 1.3.3 that

$$F^{\leftarrow}(U_k) \le t \quad \text{if and only if} \quad U_k \le F(t) \ ,$$

so that $F^{\leftarrow}(U_k) \stackrel{\mathrm{d}}{=} X_k, 1 \le k \le n$, and

$$\{G_n(F(t)) : t \in \mathbb{R}\} \stackrel{\mathrm{d}}{=} \{F_n(t) : t \in \mathbb{R}\} \ .$$

Hence,

$$D_n \equiv \sqrt{n} \sup_{t \in \mathbb{R}} \{|F_n(t) - F(t)|\} \stackrel{\mathrm{d}}{=} \sqrt{n} \sup_{t \in \mathbb{R}} \{|G_n(F(t)) - F(t)|\} \ .$$

Moreover, since F is a continuous cdf, F maps $\mathbb{R}$ into the interval $(0,1)$ plus possibly $\{0\}$ and $\{1\}$. Since $P(U = 0) = P(U = 1) = 0$ for a uniform random variable U, we have

$$D_n \stackrel{\mathrm{d}}{=} \sqrt{n} \sup_{0 \le t \le 1} \{|G_n(t) - t|\} \ , \tag{2.5}$$

which of course is the special case for a uniform cdf.

Now we turn to the second step, carrying out the analysis for the special case of a uniform cdf, i.e., starting from (2.5). To make a connection to random walks, we exploit a well known property of Poisson processes. We start by focusing on the uniform order statistics: Let $U_k^{(n)}$ be the k^{th} order statistic associated with n IID uniform random variables; i.e., $U_k^{(n)}$ is the k^{th} smallest of the uniform random numbers. It is not difficult to see that the supremum in the expression for D_n in (2.5) must occur at one of the jumps in G_n (either the left or right limit) and these jumps occur at the random times $U_k^{(n)}$. Since each jump of D_n in (2.5) has magnitude $1/\sqrt{n}$,

$$| \ D_n - \sqrt{n}(\max_{1 \le k \le n} \{| \ U_k^{(n)} - k/n \ |\}) \ | \le \ 1/\sqrt{n} \ . \tag{2.6}$$

Now we can make the desired connection to random walks: It turns out that

$$(U_1^{(n)}, \ldots, U_n^{(n)}) \stackrel{\mathrm{d}}{=} (S_1/S_{n+1}, \ldots, S_n/S_{n+1}) \ , \tag{2.7}$$

where

$$S_k \equiv X_1 + \cdots + X_k, \quad 1 \le k \le n+1 \ ,$$

with $X_k, 1 \le k \le n+1$, being IID exponential random variables with mean 1. To justify relation (2.7), consider a Poisson process and let the k^{th} point be located at S_k (Which makes the intervals between points IID exponential random variables). It is well known, and easy to verify, that the first n points of the Poisson process are distributed in the interval $(0, S_{n+1})$ as the n uniform order statistics over the interval $(0, S_{n+1})$; e.g., see p. 223 of Ross (1993). When we divide by S_{n+1} we obtain the uniform order statistics over the interval $(0,1)$, just as in the left side of (2.7).

With the connection to random walks established, we can apply Donsker's FCLT for the random walk $\{S_k : k \ge 0\}$ to establish the limit (2.1). In rough outline, here is the argument:

$$\begin{aligned} D_n &\approx \sqrt{n} \max_{1 \le k \le n} \{| \ (S_k/S_{n+1}) - (k/n) \ |\} \\ &\approx (n/S_{n+1}) \max_{1 \le k \le n} \{| \ (S_k - k)/\sqrt{n} - (k/n)(S_{n+1} - n)/\sqrt{n} \ |\} \ . \end{aligned} \tag{2.8}$$

Since $n/S_{n+1} \to 1 \quad \text{as} \quad n \to \infty$ and $(S_{n+1} - S_n)/\sqrt{n} \to 0 \quad \text{as} \quad n \to \infty$, we have

$$D_n \approx \sup_{0 \le t \le 1} \{| \ (S_{\lfloor nt \rfloor} - \lfloor nt \rfloor)/\sqrt{n} - (\lfloor nt \rfloor/n)(S_n - n)/\sqrt{n} \ |\} \ . \tag{2.9}$$

To make the rough argument rigorous, and obtain (2.9), we repeatedly apply an important tool – the convergence-together theorem – which states that $X_n \Rightarrow X$ whenever $Y_n \Rightarrow X$ and $d(X_n, Y_n) \Rightarrow 0$, where d is an appropriate distance on the function space D; see Theorem 11.4.7.

Since the functions $\psi_1 : D \to D$ and $\psi_2 : D \to \mathbb{R}$, defined by

$$\psi_1(x)(t) \equiv x(t) - tx(1), \quad 0 \le t \le 1 \ , \tag{2.10}$$

and

$$\psi_2(x) \equiv \sup_{0 \le t \le 1} \{|x(t)|\} \tag{2.11}$$

are continuous, from (2.9) we obtain the desired limit

$$D_n \Rightarrow \sup_{0 \le t \le 1} \{|\mathbf{B}(t) - t\mathbf{B}(1)|\} \ . \tag{2.12}$$

Finally, it is possible to show that relations (2.2) and (2.3) hold.

The argument here follows Breiman (1968, pp. 283–290). Details can be found there, in Karlin and Taylor (1980, p. 343) or in Billingsley (1968, pp. 64, 83, 103, 141). See Pollard (1984) and Shorack and Wellner (1986) for further development. See Borodin and Salminen (1996) for more properties of Brownian motion.

Historically, the derivation of the limit in (2.1) is important because it provided a major impetus for the development of the general theory of stochastic-process limits; see the papers by Doob (1949) and Donsker (1951, 1952), and subsequent books such as Billingsley (1968).

2.3. A Queueing Model for a Buffer in a Switch

Another important application of random walks is to queueing models. We will be exploiting the connection between random walks and queueing models throughout the queueing chapters. We only try to convey the main idea now.

To illustrate the connection between random walks and queues, we consider a discrete-time queueing model of data in a buffer of a switch or router in a packet communication network.

Let W_k represent the workload (or buffer content, which may be measured in bits) at the end of period k. During period k there is a random input V_k and a deterministic constant output μ (corresponding to the available bandwidth) provided that there is content to process or transmit. We assume that the successive inputs V_k are IID, although that is not strictly necessary to obtain the stochastic-process limits.

More formally, we assume that the successive workloads can be defined recursively by

$$W_k \equiv min\{K, \, max\{0, \, W_{k-1} + V_k - \mu\}\}, \quad k \geq 1 \ , \tag{3.1}$$

where the initial workload is W_0 and the buffer capacity is K. The *maximum* appears in (3.1) because the workload is never allowed to become negative; the output (up to μ) occurs only when there is content to emit. The *minimum* appears in (3.1) because the workload is not allowed to exceed the capacity K at the end of any period; we assume that input that would make the workload exceed K at the end of the period is lost.

The workload process $\{W_k : k \geq 1\}$ specified by the recursion (3.1) is quite elementary. Since the inputs V_k are assumed to be IID, the stochastic process $\{W_k\}$ is a discrete-time Markov process. If, in addition, we assume that the inputs V_k take values in a discrete set $\{ck : k \geq 0\}$ for some constant c (which is not a practical restriction), we can regard the stochastic process $\{W_k\}$ as a discrete-time Markov chain (DTMC). Since the state space of the DTMC $\{W_k\}$ is one-dimensional, the finite state space will usually not be prohibitively large. Thus, it is straightforward to exploit numerical methods for DTMC's, as in Kemeny and Snell (1960) and Stewart (1994), to describe the behavior of the workload process.

Nevertheless, we are interested in establishing stochastic-process limits for the workload process. In the present context, we are interested in seeing how the distribution of the inputs V_k affects the workload process. We can use heavy-traffic stochastic-process limit to produce simple formulas describing the performance. (We start giving the details in Chapter 5.) Those simple formulas provide insight that can be gained only with difficulty from a numerical algorithm for Markov chains.

We also are interested in the heavy-traffic stochastic-process limits to illustrate what can be done more generally. The heavy-traffic stochastic-process limits can be established for more complicated models, for which exact performance analysis is difficult, if not impossible. Since the heavy-traffic stochastic-process limits strip away unessential details, they reveal the key features determining the performance of the queueing system.

Now we want to see the statistical regularity associated with the workload process for large n. We could just plot the workload process for various candidate input processes $\{V_k : k \geq 1\}$ and parameters K and μ. However, the situation here is more complicated than for the the random walks we considered previously. We can simply plot the workload process and let the plotter automatically do the scaling for us, but it is not possible to automatically see the desired statistical regularity. For the queueing model, we need to do some analysis to determine how to do the proper scaling in order to achieve the desired statistical regularity. (That is worth verifying.)

2.3.1. Deriving the Proper Scaling

It turns out that stochastic-process limits for the workload process are intimately related to stochastic-process limits for the random walk $\{S_k : k \geq 0\}$ with steps

$$X_k \equiv V_k - \mu \ ,$$

but notice that in general this random walk is not centered. The random walk is only centered in the special case in which the input rate $E[V_k]$ exactly matches the potential output rate μ. However, to have a well-behaved system, we want the long-run potential output rate to exceed the long-run input rate.

In queueing applications we often characterize the system load by the *traffic intensity*, which is the rate in divided by the potential rate out. Here the traffic intensity is

$$\rho \equiv EV_1/\mu \ .$$

With an infinite-capacity buffer, we need $\rho < 1$ in order for the system to be stable (not blow up in the limit as $t \to \infty$).

We are able to obtain stochastic-process limits for the workload process by applying the continuous-mapping approach, starting from stochastic-process limits for the centered version of the random walk $\{S_k : k \geq 0\}$. However, to do so when $EX_k \neq 0$, we need to consider a sequence of models indexed by n to achieve the appropriate scaling. In the n^{th} model, we let $X_{n,k}$ be the random-walk step X_k, and we let $EX_{n,k} \to 0$ as $n \to \infty$.

There is considerable freedom in the construction of a sequence of models, but from an applied perspective, it suffices to do something simple: We can keep a fixed input process $\{V_k : k \geq 1\}$, but we need to make the output rate μ and the buffer capacity K depend upon n. Let W_k^n denote the workload at the end of period k in model n. Following this plan, for model n the recursion (3.1) becomes

$$W_k^n \equiv min\{K_n,\ max\{0,\ W_{k-1}^n + V_k - \mu_n\}\}, \quad k \geq 1 \ , \tag{3.2}$$

where K_n and μ_n are the buffer capacity and constant potential one-period output in model n, respectively.

The problem now is to choose the sequences $\{K_n : n \geq 1\}$ and $\{\mu_n : n \geq 1\}$ so that we obtain a nondegenerate limit for an appropriately scaled version of the workload processes $\{W_k^n : k \geq 0\}$. If we choose these sequence of constants appropriately, then the plotter can do the scaling of the workload processes automatically.

Let $S_k^v \equiv V_1 + \cdots + V_k$ for $k \geq 1$ with $S_0^v \equiv 0$. The starting point is a FCLT for the random walk $\{S_k^v : k \geq 0\}$. Suppose that the mean $E[V_k]$ is finite, and let it equal m_v. Then the natural FCLT takes the form

$$\mathbf{S}_n^v \Rightarrow \mathbf{S}^v \quad \text{in} \quad D \quad \text{as} \quad n \to \infty \ , \tag{3.3}$$

where

$$\mathbf{S}_n^v(t) \equiv n^{-H}(S_{\lfloor nt \rfloor}^v - m_v \lfloor nt \rfloor), \quad 0 \leq t \leq 1 \ , \tag{3.4}$$

the exponent H in the space scaling is a constant satisfying $0 < H < 1$ and $\mathbf{S}^v$ is the limit process. The common case has $H = 1/2$ and $\mathbf{S}^v = \sigma\mathbf{B}$, where

$\mathbf{B}$ is standard Brownian motion. However, as seen for the random walks, if V_k has infinite variance, then we have $1/2 < H < 1$ and the limit process $\mathbf{S}^v$ is a stable Lévy motion (which has discontinuous sample paths). We elaborate on the case with $1/2 < H < 1$ in Section 4.5.

It turns out that a scaled version of the workload process $\{W_k^n : k \geq 0\}$ can be represented directly as the image of a two-sided reflection map applied to a scaled version of the uncentered random walk $\{S_k^n : k \geq 1\}$ with steps $V_k - \mu_n$. In particular,

$$\mathbf{W}_n = \phi_K(\mathbf{S}_n) \quad \text{for all} \quad n \geq 1 \ , \tag{3.5}$$

where

$$\mathbf{W}_n(t) \equiv n^{-H} W_{\lfloor nt \rfloor}^n, \quad 0 \leq t \leq 1 \ , \tag{3.6}$$

$$\mathbf{S}_n(t) \equiv n^{-H} S_{\lfloor nt \rfloor}^n, \quad 0 \leq t \leq 1 \ , \tag{3.7}$$

and $\phi_K : D \to D$ is the *two-sided reflection map.*

In fact, it is a challenge to even define the two-sided reflection map, which we may think of as serving as the continuous-time analog of (3.1) or (3.2); that is done in Sections 5.2 and 14.8; alternatively, see p. 22 of Harrison (1985). Consistent with intuition, it turns out that the two-sided reflection map ϕ_K is continuous on the function space D with appropriate definitions, so that we can apply the continuous-mapping approach with a limit for $\mathbf{S}_n$ in (3.7) to establish the desired limit for $\mathbf{W}_n$. But now we just want to determine how to do the plotting.

The next step is to relate the assumed limit for $\mathbf{S}_n^v$ to the required limit for $\mathbf{S}_n$. For that purpose, note from (3.4) and (3.7) that

$$\mathbf{S}_n(t) = \mathbf{S}_n^v(t) - n^{-H}(\mu_n - m_v)\lfloor nt \rfloor \ .$$

Hence we have the stochastic-process limit

$$\mathbf{S}_n \Rightarrow \mathbf{S} \quad \text{as} \quad n \to \infty \ , \tag{3.8}$$

where

$$\mathbf{S}(t) \equiv \mathbf{S}^v(t) - mt, \quad 0 \leq t \leq 1 \ , \tag{3.9}$$

if and only if

$$n^{-H}(\mu_n - m_v)\lfloor nt \rfloor \to mt \quad \text{as} \quad n \to \infty$$

for each $t > 0$ or, equivalently,

$$(\mu_n - m_v)n^{1-H} \to m \quad \text{as} \quad n \to \infty \ . \tag{3.10}$$

In addition, because of the space scaling by n^H in $\mathbf{S}_n$, we need to let

$$K_n \equiv n^H K \ . \tag{3.11}$$

Given the scaling in both (3.10) and (3.11), we are able to obtain the FCLT

$$\mathbf{W}_n \Rightarrow \mathbf{W} \equiv \phi_K(\mathbf{S}) \ , \tag{3.12}$$

where $\mathbf{W}_n$ is given in (3.6), $\mathbf{S}$ is given in (3.9) and ϕ_K is the two-sided reflection map.

The upshot is that we obtain the desired stochastic-process limit for the workload process, and the plotter can automatically do the appropriate scaling, if we let

$$\mu_n \equiv m_v + m/n^{1-H} \quad \text{and} \quad K_n \equiv n^H K \tag{3.13}$$

for any fixed m with $0 \le m < \infty$ and K with $0 < K \le \infty$, where H with $0 < H < 1$ is the scaling exponent appearing in (3.4).

At this point, it is appropriate to pause and reflect upon the significance of the scaling in (3.13). First note that time scaling by n (replacing t by nt) and space scaling by n^H (dividing by n^H) is determined by the FCLT in (3.3). Then the output rate and buffer size should satisfy (3.13). Note that the actual buffer capacity K_n in system n must increase, indeed go to infinity, as n increases. Also note that the output rate μ_n approaches m_v as n increases, so that the traffic intensity ρ_n approaches 1 as n increases. Specifically,

$$\rho_n \equiv \frac{E[V_1]}{\mu_n} = \frac{m_v}{m_v + mn^{-(1-H)}} = 1 - (m/m_v)n^{-(1-H)} + o(n^{-(1-H)})$$

as $n \to \infty$.

The obvious application of the stochastic-process limit in (3.12) is to generate approximations. The direct application of (3.12) is

$$\{n^{-1/\alpha} W^n_{\lfloor nt \rfloor} : t \ge 0\} \approx \{\mathbf{W}(t) : t \ge 0\} \ , \tag{3.14}$$

where here $\approx$ means *approximately equal to in distribution.* Equivalently, by unscaling, we obtain the associated approximation (in distribution)

$$\{W^n_k : k \ge 0\} \approx \{n^{1/\alpha}\mathbf{W}(k/n) : k \ge 0\} \ . \tag{3.15}$$

Approximations such as (3.15), which are obtained directly from stochastic-process limits, may afterwards be refined by making modifications to meet other criteria, e.g., to match exact expressions known in special cases. Indeed, it is often possible to make refinements that remain asymptotically correct in the heavy-traffic limit, e.g., by including the traffic intensity ρ, which converges to 1 in the limit.

Often the initial goal in support of engineering applications is to develop a suitable approximation. Then heuristic approaches are perfectly acceptable, with convenience and accuracy being the criteria to judge the worth of alternative candidates. Even with such a pragmatic engineering approach, the stochastic-process limits are useful, because they generate initial candidate approximations, often capturing essential features, because the limit often is able to strip away unessential details. Moreover, the limits establish important theoretical reference points, demonstrating asymptotic correctness in certain limiting regimes.

2.3.2. Simulation Examples

Let us now look at two examples.

Example 2.3.1. *Workloads with exponential inputs.*

First let $\{V_k : k \geq 1\}$ be a sequence of IID exponential random variables with mean 1. Then the FCLT in (3.3) holds with $H = 1/2$ and $\mathbf{S}$ being standard Brownian motion $\mathbf{B}$. Thus, from (3.13), the appropriate scaling here is

$$\mu_n \equiv 1 + m/\sqrt{n} \quad \text{and} \quad K_n \equiv \sqrt{n}K \ . \tag{3.16}$$

To illustrate, we again perform simulations. Due to the recursive definition in (3.2), we can construct and plot the successive workloads just as easily as we constructed and plotted the random walks before. Paralleling our previous plots of random walks, we now plot the first n workloads, using the scaling in (3.16). In Figure 2.1 we plot the first n workloads for the case $H = 1/2$, $m = 1$ and $K = 0.5$ for $n = 10^j$ for $j = 1, \ldots, 4$. To supplement Figure 2.1, we show six independent replications for the case $n = 10^4$ in Figure 2.2.

What we see, as n becomes sufficiently large, is standard Brownian motion with drift $-m = -1$ modified by reflecting barriers at 0 and 0.5. Of course, just as for the random-walk plots before, the units on the axes are for the original queueing model. For example, for $n = 10^4$, the buffer capacity is $K_n = 0.5\sqrt{n} = 50$, so that the actual buffer content ranges from 0 to 50, even though the reflected Brownian motion ranges from 0 to 0.5. Similarly, for $n = 10^4$, the traffic intensity is $\rho_n = (1 + n^{-1/2})^{-1} = (1.01)^{-1} \approx 0.9901$ even though the Brownian motion has drift -1.

Unlike in the previous random-walk plots, the units on the vertical axes in Figure 2.2 are the same for all six plots. That happens because, in all six cases, the workload process takes values ranging from 0 to 50. The upper limit is 50 because for $n = 10^4$ the upper barrier in the queue is $0.5\sqrt{n} = 50$. The clipping at the upper barrier occurs because of overflows.

The traffic intensity 0.99 in Figure 2.2 is admittedly quite high. If we focus instead upon $n = 100$ or $n = 25$, then the traffic intensity is not so extreme, in particular, then $\rho_n = (1 + n^{-1/2})^{-1} = (1.1)^{-1} \approx 0.91$ or $(1.2)^{-1} \approx 0.83$.

In Figures 2.1 and 2.2 we see statistical regularity, just as in the early random-walk plots. Just as in the pairs of figures, (Figures 1.3 and 1.4) and (Figures 1.21 and 1.22), the plots for $n = 10^6$ look just like the plots for $n = 10^4$ when we ignore the units on the axes. The plots show that there should be a stochastic-process limit as $n \to \infty$. The plots demonstrate that a reflected Brownian motion approximation is appropriate with these parameters.

Moreover, our analysis of the stochastic processes to determine the appropriate scaling shows how we can obtain the stochastic-process limits. Indeed, we obtain the supporting stochastic-process limits for the workload process directly from the established stochastic-process limits for the random walks. In order to make the connection between the random walk and the workload process, we are constrained to use the scaling in (3.16). With that scaling, the plotter directly reveals the statistical regularity. ■

Example 2.3.2. *Workloads with Pareto*(3/2) *inputs.*

For our second example, we assume that the inputs V_k have a Pareto(p) distribution with finite mean but infinite variance. In particular, we let

$$V_k \equiv U_k^{-1/p} \quad \text{for} \quad p = 3/2 \; , \tag{3.17}$$

just as in case (*iii*) of (3.5) in Section 1.3.3, which makes the distribution Pareto(p) for $p = 3/2$. Since $H = p^{-1}$ for $p = 3/2$, we need to use different scaling than we did in Example 2.3.1. In particular, instead of (3.16), we now use

$$\mu_n \equiv 1 + m/n^{1/3} \quad \text{and} \quad K_n \equiv n^{2/3} K \; , \tag{3.18}$$

with $m = 1$ and $K = 0.5$ just as before.

Since the scaling in (3.18) is different from the scaling in (3.16), for any given triple (m, K, n), the buffer size K_n is now larger, while the output rate differs more from the input rate. Assuming that $m > 0$, the traffic intensity

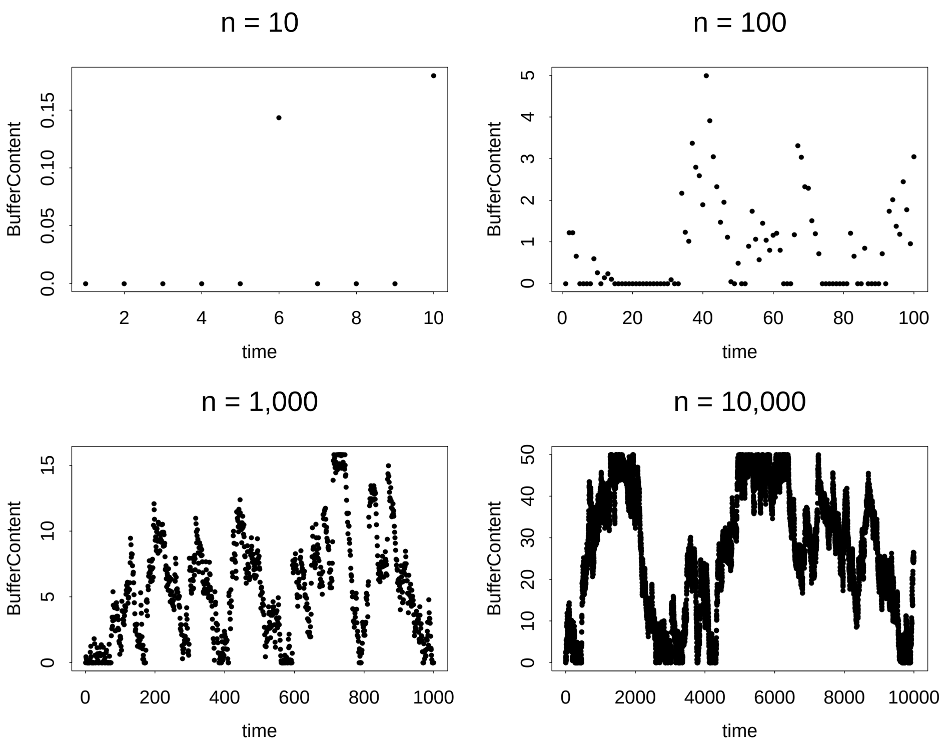

Figure 2.1: Possible realizations of the first n steps of the workload process $\{W_k^n : k \geq 0\}$ with IID exponential inputs having mean 1 for $n = 10^j$ with $j = 1, \ldots, 4$. The scaling is as in (3.16) with $m = 1$ and $K = 0.5$.

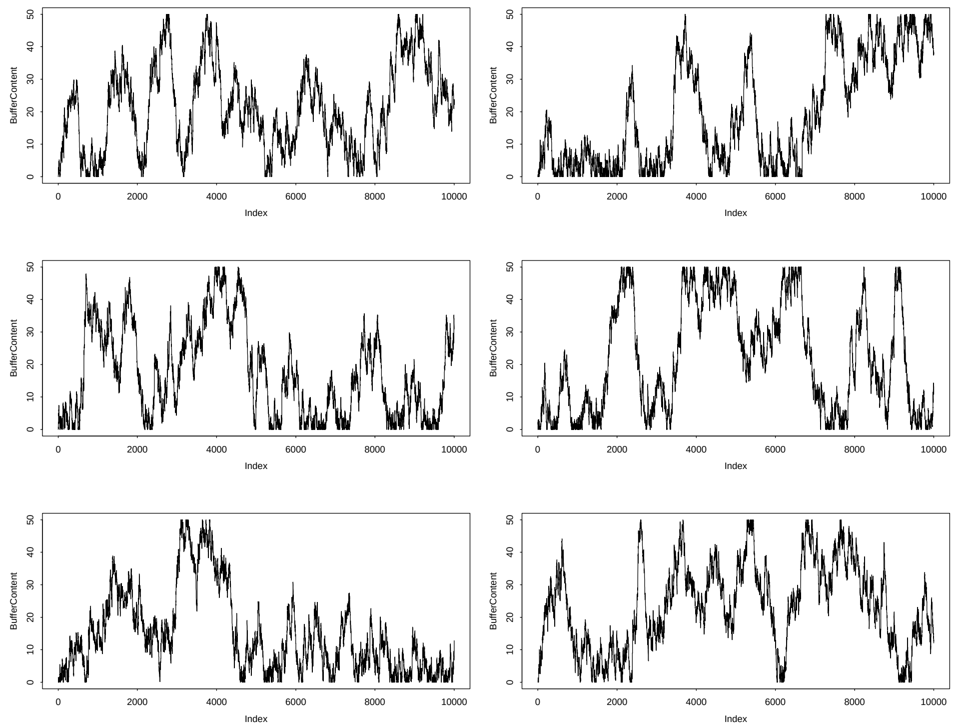

Figure 2.2: Six possible realizations of the first n steps of the workload process $\{W_k^n : k \geq 0\}$ with IID exponential inputs for $n = 10^4$. The scaling is as in (3.16) with $m = 1$ and $K = 0.5$.

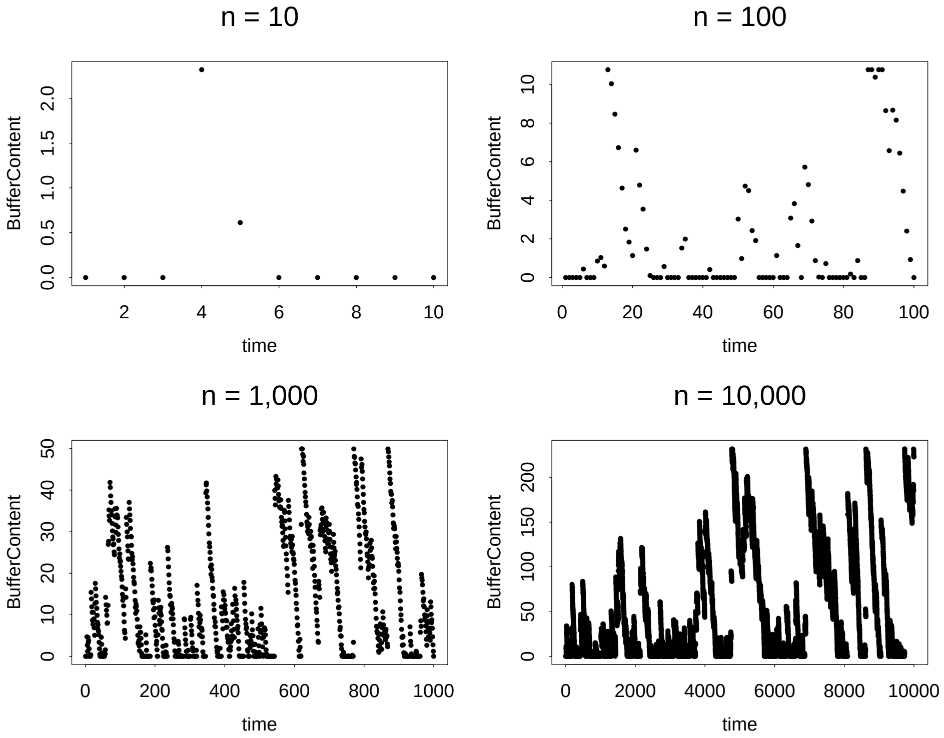

Figure 2.3: Possible realizations of the first n steps of the workload process $\{W_k^n : k \geq 0\}$ with IID Pareto(p) inputs having $p = 3/2$, mean 3 and infinite variance for $n = 10^j$ with $j = 1, \ldots, 4$. The scaling is as in (3.18) with $m = 1$ and $K = 0.5$.

in model n is now lower. That suggests that as H increases the heavy-traffic approximations may perform better at lower traffic intensities.

We plot the first n workloads, using the scaling in (3.18), for $n = 10^j$ for $j = 1, \ldots, 4$ in Figure 2.3 for the case $m = 1$ and $K = 0.5$. What we see, as n becomes sufficiently large, is a stable Lévy motion with drift $-m = -1$ modified by reflecting barriers at 0 and 0.5. To supplement Figure 2.3, we show six independent replications for the case $n = 10^4$ in Figure 2.4. As before, the plots for $n = 10^6$ look just like the plots for $n = 10^4$ if we ignore the units on the axes. Just as in Figures 1.20–1.22 for the corresponding random walk, the plots here have jumps. ∎

In summary, the workload process $\{W_k\}$ in the queueing model is intimately related to the random walk $\{S_k\}$ with steps being the net inputs

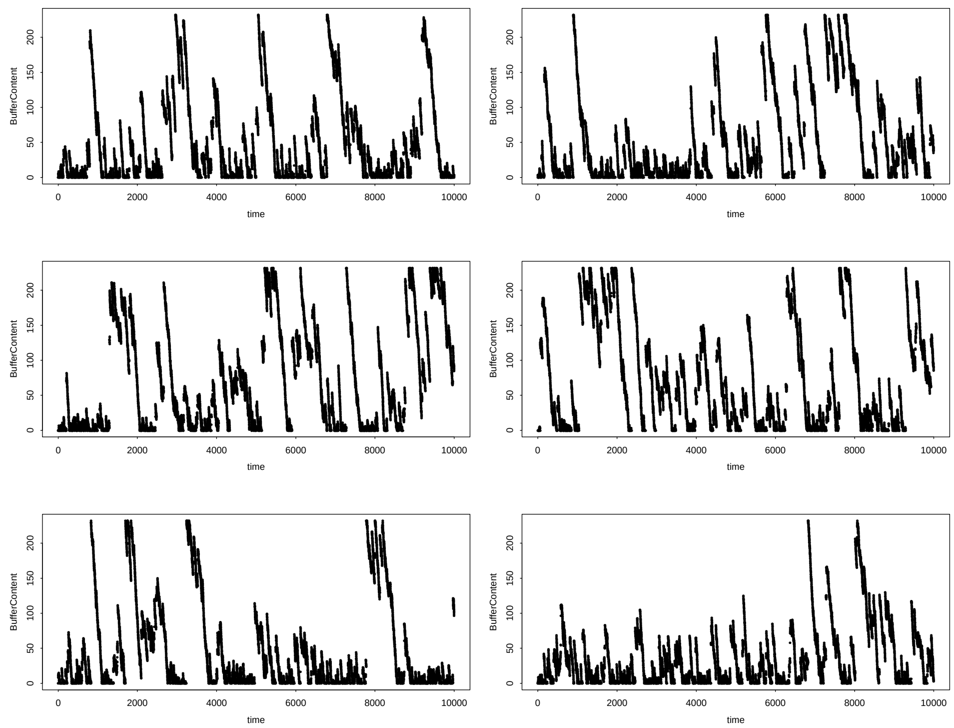

Figure 2.4: Six possible realizations of the first n steps of the workload process $\{W_k^n : k \geq 0\}$ with IID Pareto(p) inputs having $p = 3/2$, mean 3 and infinite variance for $n = 10^4$. The scaling is as in (3.18) with $m = 1$ and $K = 0.5$.

$V_k - \mu$ each period. With appropriate scaling, as in (3.13), which includes the queue being in heavy traffic, stochastic-process limits for a sequence of appropriately scaled workload processes can be obtained directly from associated stochastic-process limits for the underlying random walk.

Moreover, the limit process for the workload process is just the limit process for the random walk modified by having two reflecting barriers. Thus, the workload process in the queue exhibits the same statistical regularity for large sample sizes that we saw for the random walk. Indeed, the random walk is the source of that statistical regularity.

Just as for the random walks, the form of the statistical regularity may lead to the limit process for the workload process having discontinuous sample paths.

2.4. Engineering Significance

In the previous section, we saw that queueing models are closely related to random walks. With the proper (heavy-traffic) scaling, the same forms of statistical regularity that hold for random walks also hold for the workload process in the queueing model. But does it matter? Are there important engineering consequences?

To support an affirmative answer, in this final section we discuss the engineering significance of heavy-traffic stochastic-process limits for queues. First, in Section 2.4.1, we discuss buffer sizing in a switch or router in a communication network. Then, in Section 2.4.2, we discuss scheduling service with multiple sources, as occurs in manufacturing when scheduling production of multiple products on a single machine with setup costs or setup times for switching.

2.4.1. Buffer Sizing

The buffer (waiting space) in a network switch or router tends to be expensive to provide, so that economy dictates it be as small as possible. On the other hand, we want very few lost packets due to buffer overflow.

Queueing models are ideally suited to determine an appropriate buffer size. Let $L(K)$ be the long-run proportion of packets lost as a function of the buffer size K. We might specify a maximum allowable proportion of lost packets, ϵ. Given the function L, we then choose the buffer size K to satisfy the buffer-sizing equation

$$L(K) = \epsilon \ . \tag{4.1}$$

Classical queueing analysis, using standard models such as in Example 2.3.1, shows that $L(K)$ decays exponentially in K; specifically, L tends to have an *exponential tail*, satisfying

$$L(K) \sim \alpha e^{-\eta K} \quad \text{as} \quad K \to \infty \tag{4.2}$$

for asymptotic constants α and η depending upon the model details. (As in (4.6), $\sim$ means asymptotic equivalence. See Remark 5.4.1 for further discussion about asymptotics.)

It is natural to exploit the exponential tail asymptotics for L in (4.2) to generate the approximation

$$L(K) \approx \alpha e^{-\eta K} \tag{4.3}$$

for all K not too small. We then choose K to satisfy the *exponential buffer-sizing equation*

$$\alpha e^{-\eta K} = \epsilon \; , \tag{4.4}$$

from which we deduce that the target buffer size K^* should be

$$K^* = \eta^{-1} \log{(\alpha/\epsilon)} \; . \tag{4.5}$$

This analysis shows that the target buffer size should be directly proportional to η^{-1} and $\log \alpha$, and inversely proportional to $\log \epsilon$. It remains to determine appropriate values for the three constants η, α and ϵ, but the general relationships are clear. For example, if $\epsilon = 10^{-j}$, then K^* is proportional to the exponent j, which means that the cost of improving performance (as measured by the increase in buffer size K^* required to make ϵ significantly smaller) tends to be small.

So far, we have yet to exploit heavy-traffic limits. Heavy-traffic limits can play an important role because it actually is difficult to establish the exponential tail asymptotics in (4.2) directly for realistic models. As a first step toward analytic tractability, we may approximate the loss function $L(K)$ by the tail probability $P(W(\infty) > K)$, where $W(\infty)$ is the steady-state workload in the corresponding queue with unlimited waiting space. Experience indicates that the asymptotic form for $L(K)$ tends to be the same as the asymptotic form for the tail probability $P(W(\infty) > K)$ (sometimes with different asymptotic constants). From an applied point of view, we are not too concerned about great accuracy in this step, because the queueing model is crude (e.g., it ignores congestion controls) and the loss proportion $L(K)$ itself is only a rough performance indicator.

As a second step, we approximate $W(\infty)$ in the tail probability $P(W(\infty) > K)$ by the steady-state limit of the approximating process obtained from the heavy-traffic stochastic-process limit. For standard models, the approximating process is reflected Brownian motion, as in Example 2.3.1. Since the steady-state distribution of reflected Brownian motion with one-sided reflection is exponential (see Section 5.7), the heavy-traffic limit provides strong support for the approximations in (4.3)–(4.5) and helps identify approximate values for the asymptotic constants η and α. (The heavy-traffic limits also can generate approximations directly for the loss proportion $L(K)$; e.g., see Section 5.7.) The robustness of heavy-traffic limits (discussed in Chapters 4 and 5) suggests that the analysis should be insensitive to fine system details.

However, the story is not over! Traffic measurements from communication networks present a very different view of the world: These traffic measurements have shown that the traffic carried on these networks is remarkably bursty and complex, exhibiting features such as heavy-tailed probability distributions, strong positive dependence and self-similarity; e.g., see Leland et al. (1994), Garrett and Willinger (1994), Paxson and Floyd (1995), Willinger et al. (1995, 1997), Crovella and Bestavros (1996), Resnick (1997), Adler, Feldman and Taqqu (1998), Barford and Crovella (1998), Crovella, Bestavros and Taqqu (1998), Willinger and Paxson (1998), Park and Willinger (2000), Krishnamurthy and Rexford (2001) and references therein. These traffic studies suggest that different queueing models may be needed.

In particular, the presence of such traffic burstiness can significantly alter the behavior of the queue: *Alternative queueing analysis suggests alternative asymptotic forms for the function L.* Heavy-tailed probability distributions as in Example 2.3.2 lead to a different asymptotic form: When the inputs have power tails, like the Pareto inputs in Example 2.3.2, the function L tends to have a power tail as well: Instead of (4.2), we may have

$$L(K) \sim \alpha K^{-\eta} \quad \text{as} \quad K \to \infty , \tag{4.6}$$

where again α and η are positive asymptotic constants; see Remark 5.4.1.

The change from the exponential tail in (4.2) to the power tail in (4.6) are contrary to the conclusions made above about the robustness of heavy-traffic approximations. Even though the standard heavy-traffic limits are remarkably robust, there is a limit to the robustness! The traffic burstiness can cause the robustness of the standard heavy-traffic limits to break down. Just as we saw in Example 2.3.2, the burstiness can have a major impact on the workload process.

However, we can still apply heavy-traffic limits: Just as before, we can approximate $L(K)$ by $P(W(\infty) > K)$, where $W(\infty)$ is the steady-state workload in the corresponding queue with unlimited waiting space. Then we can approximate $W(\infty)$ by the steady-state limit of the approximating process obtained from a heavy-traffic limit. However, when we properly take account of the traffic burstiness, the heavy-traffic limit process is no longer reflected Brownian motion. Instead, as in Example 2.3.2, it may be a reflected stable Lévy motion, for which $P(W(\infty) > K) \sim \alpha K^{-\eta}$. (For further discussion about the power tails, see Sections 4.5, 6.4 and 8.5.) Thus, different heavy-traffic limits support the power-tail asymptotics in (4.6) and yield approximations for the asymptotic constants.

Paralleling (4.3), we can use the approximation

$$L(K) \approx \alpha K^{-\eta} \tag{4.7}$$

for K not too small. Paralleling (4.4), we use the target equation (4.1) and (4.7) to obtain the *power buffer-sizing equation*

$$\alpha K^{-\eta} = \epsilon \tag{4.8}$$

from which we deduce that the *logarithm* of the target buffer size K^* should be

$$\log K^* = \eta^{-1} \log (\alpha/\epsilon) \ . \tag{4.9}$$

In this power-tail setting, we see that the required buffer size K^* is much more responsive to the parameters η, α and ϵ: Now the logarithm $\log K^*$ is related to the parameters η, α and ϵ the way K^* was before. For example, if $\epsilon = 10^{-j}$, then the logarithm of the target buffer size K^* is proportional to j, which means that the cost of improving performance (as measured by the increase in buffer size K^* required to make ϵ significantly smaller) tends to be large.

And that is not the end! The story is still not over. There are other possibilities: There are different forms of traffic burstiness. In Example 2.3.2 we focused on heavy-tailed distributions for IID inputs, but the traffic measurements also reveal strong dependence. The strong dependence observed in traffic measurements leads to considering fractional-Brownian-motion models of the input, which produce another asymptotic form for the function L; see Sections 4.6, 7.2 and 8.7. Unlike both the exponential tail in (4.2) and the power tail in (4.5), we may have a *Weibull tail*

$$L(K) \sim \alpha e^{-\eta K^{\gamma}} \quad \text{as} \quad K \to \infty \tag{4.10}$$

for positive constants α, η and γ, where $0 < \gamma < 1$; see (8.10) in Section 8.8. The available asymptotic results actually show that

$$P(W(\infty) > K) \sim \alpha K^{-\beta} e^{-\eta K^{\gamma}} \quad \text{as} \quad K \to \infty$$

for asymptotic constants η, α and β, where $W(\infty)$ is the steady-state of reflected fractional Brownian motion. Thus, the asymptotic results do not directly establish the asymptotic relation in (4.10), but they suggest the rough approximation

$$L(K) \approx \alpha e^{-\eta K^{\gamma}} \tag{4.11}$$

for all K not too small and the associated *Weibull buffer-sizing equation*

$$\alpha e^{-\eta K^{\gamma}} = \epsilon \ , \tag{4.12}$$

from which we deduce that the γ^{th} *power* of the target buffer size K^* should be

$$K^{*\gamma} = \eta^{-1} \log{(\alpha/\epsilon)} \ . \tag{4.13}$$

In (4.13) the γ^{th} power of K^* is related to the parameters α, η and ϵ the way K^* was in (4.5) and $\log K^*$ was in (4.9). Thus, consistent with the intermediate asymptotics in (4.10), since $0 < \gamma < 1$, we have the intermediate buffer requirements in (4.13).

Unfortunately, it is not yet clear which models are most appropriate. Evidence indicates that it depends on the context; e.g., see Heyman and Lakshman (1996, 2000), Ryu and Elwalid (1996), Grossglauser and Bolot (1999), Park and Willinger (2000), Guerin et al. (2000) and Mikosch et al. (2001). Consistent with observations by Sriram and Whitt (1986), long-term variability has relatively little impact on queueing performance when the buffers are small, but can be dramatic when the buffers are large.

Direct traffic measurements are difficult to interpret because they describe the carried traffic, not the offered traffic, and may be strongly influenced by congestion controls such as the Transmission Control Protocol (TCP); see Section 5.2 of Krishnamurthy and Rexford (2001) and Arvidsson and Karlsson (1999). Moreover, the networks and the dominant applications keep changing. For models of TCP, see Padhye et al. (2000), Bu and Towsley (2001), and references therein.

From an engineering perspective, it may be appropriate to ignore congestion controls when developing models for capacity planning. We may wish to provide sufficient capacity so that we usually meet the *offered load* (the original customer demand). When the system is heavily loaded, the controls slow down the stream of packets. From a careful analysis of traffic measurements, we may be able to reconstruct the intended flow. (For further discussion about offered-load models, see Remark 10.3.1.) However, heavy-traffic limits can also describe the performance with congestion-controlled sources, as shown by Das and Srikant (2000).

Our goal in this discussion, and more generally in the book, is not to draw engineering conclusions, but to describe an approach to engineering problems: Heavy-traffic limits yield simple approximations that can be used in engineering applications involving queues. Moreover, nonstandard heavy-traffic limits can capture the nonstandard features observed in network traffic. The simple analysis above shows that the consequences of the model choice can be dramatic, making order-of-magnitude differences in the predicted buffer requirements.

When the analysis indicates that very large buffers are required, instead of actually providing very large buffers, we may conclude that buffers are relatively ineffective for improving performance. Instead of providing very large buffers, we may choose to increase the available bandwidth (processing rate), introduce scheduling to reduce the impact of heavy users upon others, or regulate the source inputs (see Example 9.8.1). Indeed, all of these approaches are commonly used in practice. It is common to share the bandwidth among sources using a "fair queueing" discipline. Fair queueing disciplines are variants of the head-of-line processor-sharing discipline, which gives each of several active sources a guaranteed share of the available bandwidth. See Demers, Keshav and Shenker (1989), Greenberg and Madras (1992), Parekh and Gallager (1993, 1994), Anantharam (1999) and Borst, Boxma and Jelenković (2000).

Many other issues remain to be considered: First, given any particular asymptotic form, it remains to estimate the asymptotic constants. Second, it remains to determine how the queueing system scales with increasing load. Third, it may be more appropriate to consider the transient or time-dependent performance measures instead of the customary steady-state performance measures. Fourth, it may be necessary to consider more than a single queue in order to capture network effects. Finally, it may be necessary to create appropriate controls, e.g., for scheduling and routing. Fortunately, for all these problems, and others, heavy-traffic stochastic-process limits can come to our aid.

2.4.2. Scheduling Service for Multiple Sources

In this final subsection we discuss *the engineering significance of the time-and-space scaling* that occurs in heavy-traffic limits for queues. The heavy-traffic scaling was already discussed in Section 2.3; now we want to point out its importance for system control.

We start by extending the queueing model in Section 2.3: Now we assume that there are inputs each time period from m separate sources. We let each source have its own infinite-capacity buffer, and assume that the work in each buffer is served in order of arrival, but otherwise we leave open the order of service provided to the different sources. As before, we can think of there being a single server, but now the server has to switch from queue to queue in order to perform the service, with there being a setup cost or a setup time to do the switching.

We initially assume that the server can switch from queue to queue instantaneously (within each discrete time period), but we assume that there are switchover costs for switching. To provide motivation for switching, we also assume that there are source-dependent holding costs for the workloads. To specify a concrete optimization problem, let W_k^i denote the source-i workload in its buffer at the end of period k and let $S_k^{i,j}$ be the number of switches from queue i to queue j in the first k periods. Let the total cost incurred in the first k periods be the sum of the total holding cost and the total switching cost, i.e.,

$$C_k \equiv H_k + S_k \ ,$$

where

$$H_k \equiv \sum_{i=1}^{m} \sum_{j=1}^{k} h_i W_j^i$$

and

$$S_k \equiv \sum_{i=1}^{m} \sum_{j=1}^{m} c_{i,j} S_k^{i,j} \ ,$$

where h_i is the source-i holding cost per period and $c_{i,j}$ is the switching cost per switch from source i to source j. Our goal then may be to choose a switching policy that minimizes the long-run average expected cost

$$\bar{C} \equiv \lim_{k \to \infty} k^{-1} E[C_k] \ .$$

This is a difficult control problem, even under the regularity condition that the inputs come from m independent sequences of IID random variables with finite means m_v^i. Under that regularity condition, the problem can be formulated as a *Markov sequential decision process*; e.g., see Puterman (1994): The state at the beginning of period $k+1$ is the workload vector $(W_k^1, \ldots, W_k^m)$ and the location of the server at the end of period k. An action is a specification of the sequence of queues visited and the allocation of the available processing per period, μ, during those visits. Both the state and action spaces are uncountably infinite, but we could make reasonable simplifying assumptions to make them finite.

To learn how we might approach the optimization problem, it is helpful to consider a simple scheduling policy: A *polling* policy serves the queues to exhaustion in a fixed cyclic order, with the server starting each period where it stopped the period before. We assume that the server keeps working until either its per-period capacity μ is exhausted or all the queues are empty.

There is a large literature on polling models; see Takagi (1986) and Boxma and Takagi (1992). For classical polling models, there are analytic solutions, which can be solved numerically. For those models, numerical transform inversion is remarkably effective; see Choudhury and Whitt (1996). However, analytical tractability is soon lost as model complexity increases, so there is a need for approximations.

The polling policy is said to be a *work-conserving service policy*, because the server continues serving as long as there is work in the system yet to be done (and service capacity yet to provide). An elementary, but important, observation is that the total workload process for any work-conserving policy is identical to the workload process with a single shared infinite-capacity buffer. Consequently, the heavy-traffic limit described in Section 2.3 in the special case of an infinite buffer ($K = \infty$) also holds for the total-workload process with polling; i.e., with the FCLT for the cumulative inputs in (3.3) and the heavy-traffic scaling in (3.10), we have the heavy-traffic limit for the scaled total-workload processes in (3.12), with the two-sided reflection map ϕ_K replaced by the one-sided reflection map. Given the space scaling by n^H and the time scaling by n, where $0 < H < 1$, the unscaled total workload at any time in the n^{th} system is of order n^H and changes significantly over time intervals having length of order n.

The key observation is that the time scales are very different for the individual workloads at the source buffers. First, the individual workloads are bounded above by the total workload. Hence the unscaled individual workloads are also of order n^H. Clearly, the mean inputs must satisfy the relation

$$m_v = m_{v,1} + \cdots + m_{v,m} \ .$$

Assuming that $0 < m_{v,i} < m_v$ for all i, we see that *each source by itself is not in heavy traffic when the server is dedicated to it:* With the heavy-traffic scaling in (3.10), the total traffic intensity approaches 1, i.e.,

$$\rho_n \equiv m_v/\mu_n \uparrow 1 \quad \text{as} \quad n \to \infty \ ,$$

but the instantaneous traffic intensity for source i when the server is devoted to it converges to a limit less than 1, i.e.,

$$\rho_{n,i} \equiv m_{v,i}/\mu_n \uparrow m_{v,i}/m_v \equiv \rho_i^* < 1 \ .$$

Since each source alone is not in heavy-traffic when the server is working on that source, the net output is at a constant positive rate when service is being provided, even in the heavy-traffic limit. Thus the server processes the order n^H unscaled work there in order n^H time, by the law of large numbers (see Section 5.3).

The upshot is that the unscaled individual workloads change significantly in order n^H time whenever the server is devoted to them, and the server cycles through the m queues in order n^H time, whereas the unscaled total workload changes significantly in order n time. Since $H < 1$, in the heavy-traffic limit the individual workloads change on a faster time scale. Thus, in the heavy-traffic limit we obtain a *separation of time scales*: When we consider the evolution of the individual workload processes in a short time scale, we can act as if the total workload is fixed.

Remark 2.4.1. *The classic setting: NCD Markov chains.* The separation of time scales in the polling model is somewhat surprising, because it occurs in the heavy-traffic limit. In other settings, a separation of time scales is more evident. With computers and communication networks, the relevant time scale for users is typically seconds, while the relevant time scale for system transactions is typically milliseconds. For those systems, engineers know that time scales are important.

There is a long tradition of treating different time scales in stochastic models using nearly-completely-decomposable (NCD) Markov chains; see Courtois (1977). With a NCD Markov chain, the state space can be decomposed into subsets such that most of the transitions occur between states in the same subset, and only rarely does the chain move from one subset to another. In a long time scale, the chain tends to move from one local steady-state regime to another, so that the long-run steady-state distribution is an appropriate average of the local steady-state distributions.

However, different behavior can occur if the chain does not approach steady-state locally within a subset. For example, that occurs in an infinite-capacity queue in a slowly changing environment when the queue is unstable in some environment states. Heavy-traffic limits for such queues were established by Choudhury, Mandelbaum, Reiman and Whitt (1997). Even though the queue content may ultimately approach a unique steady-state distribution, the local instability may cause significant fluctuations in an intermediate time scale. The transient behavior of the heavy-traffic limit process captures this behavior over the intermediate time scale. ∎

For the polling model, the separation of time scales suggests that in the heavy-traffic limit, given the fixed scaled total workload $\mathbf{W}_n(t) = w$, in the neighborhood of time t the vector of scaled individual workloads $(\mathbf{W}_n^1(t), \ldots, \mathbf{W}_n^m(t))$ rapidly traverses a deterministic piecewise-linear trajectory through points $(w^1, \ldots, w^m)$ in the hyperplane in $\mathbb{R}^m$ with $w^1 + \cdots + w^m = w$. For example, with three identical sources served in numerical cyclic order, the path is piecewise-linear, passing through the vertices $(2w/3, w/3, 0)$, $(0, 2w/3, w/3)$ and $(w/3, 0, 2w/3)$, corresponding to the instants the server is about to start service on sources 1, 2 and 3, respectively. In general, identifying the vertices is somewhat complicated, but the experience of each source is clear: it builds up to its peak workload at constant rate and then returns to emptiness at constant rate. And it does this many times before the total workload changes significantly. Hence at any given time its level can be regarded as uniformly distributed over its range.

As a consequence, we anticipate a *heavy-traffic averaging principle*: We should have a limit for the average of functions of the scaled individual workloads; i.e., for any $s, h > 0$ and any continuous real-valued function f,

$$h^{-1} \int_s^{s+h} f(\mathbf{W}_n^i(t))dt \Rightarrow h^{-1} \int_s^{s+h} (\int_0^1 f(a_i u \mathbf{W}(t))du)dt \ , \qquad (4.14)$$

where a_i is a constant satisfying $0 < a_i \le 1$ for $1 \le i \le m$. In words, the time-average of the scaled individual-source workload process over the time interval $[s, s+h]$ approaches the corresponding time-average of a proportional space-average of the limit $\mathbf{W}$ for the scaled total workload process. (For other instances of the averaging principle, see Anisimov (1993) and Freidlin and Wentzell (1993).)

This heavy-traffic averaging principle was rigorously established for the case of two queues by Coffman, Puhalskii and Reiman (1995) for a slightly different model in the Brownian case, with $H = 1/2$ and $\mathbf{W}$ reflected Brownian motion. They also determined the space-scaling constants a_i appearing in (4.14) for m sources: They showed that

$$a_i = \frac{\rho_i^*(1 - \rho_i^*)}{\sum_{1 \le j < k \le m} \rho_j^* \rho_k^*} \ , \qquad (4.15)$$

where ρ_i^* is the limiting source-i traffic intensity, i.e., $\rho_i^* \equiv m_{v,i}/m_v$ for our model. The upper limits a_i depend only on the means $m_{v,j}$, $1 \le j \le m$. For $m = 2$, $a_i = 1$; for m identical sources, $a_i = 2/m$. The variability affects the limit in (4.14) only through the scaling and the one-dimensional limit process $\mathbf{W}$.

Coffman, Puhalskii and Reiman (1998) also considered the two-queue polling model with unscaled switchover times. Even though the switchover times are asymptotically negligible in the heavy-traffic scaling, they have a significant impact because the relative amount of switching increases as the total workload decreases. Coffman, Puhalskii and Reiman (1998) show that the heavy-traffic averaging principle is still valid with switchover times, with the scaled total workload processes converging to a Bessel diffusion process, which has state-dependent drift of the form $-a + b/x$ for positive constants a and b. (For additional heavy-traffic limits for polling models, see van der Mei and Levy (1997) and van der Mei (2000).)

Even though the polling models have yet to be analyzed for nonstandard scaling, with $H \neq 1/2$ and $\mathbf{W}$ not a diffusion process, it is evident that the heavy-traffic averaging principle still applies. We can anticipate that the other forms of variability (associated with heavy tails and strong dependence) affect the heavy-traffic limit only through the limit process $\mathbf{W}$.

The separation of time scales provides a way to attack complicated service control problems such as the one formulated at the beginning of this subsection. Even if all the desired supporting mathematics cannot be established, the heavy-traffic limits provide a useful perspective for approximately solving these problems. The heavy-traffic averaging principle reduces the dimension of the state-space in the control problem. It provides a form of *state-space collapse*; see Reiman (1984b), Harrison and van Mieghem (1997), Bramson (1998) and Williams (1998b). It lets us focus on the single process that is the heavy-traffic limit for the scaled total-workload process. For natural classes of service policies, we can express the local cost rate associated with a fixed total workload and then determine an expression for the long-run average total cost as a function of the controls that produces a tractable optimization problem. In the more challenging cases it may be necessary to apply numerical methods to solve the optimization problem, as in Kushner and Dupuis (2000).

By now, there has been substantial work on this heavy-traffic approach to scheduling, yielding excellent results. We do not try to tell the story here; instead we refer to Reiman and Wein (1998), Markowitz, Reiman and Wein (2000), Markowitz and Wein (2001) and Kushner (2001).

For these more complicated control problems, there are many open technical problems: It remains to establish the heavy-traffic averaging principle in more complicated settings and it remains to show that the derived policies are indeed asymptotically optimal in the heavy-traffic limit. Markowitz et al. (2000, 2001) restrict attention to dynamic cyclic policies in which each source is served once per cycle in the same fixed order. It is easy to construct examples in which larger classes of policies are needed: With three sources, it may be necessary to serve one source more frequently; e.g., the cycle $(1,2,1,3)$ may be much better than either $(1,2,3)$ or $(1,3,2)$.

Nevertheless, the practical value of the heavy-traffic approach is well established: Numerical comparisons have shown that the policies generated from the heuristic heavy-traffic analysis perform well for systems under normal loading. Moreover, the heavy-traffic analysis produces important insight about the control problem, as illustrated by concluding remarks on p. 268 of Markowitz and Wein (2001) about the way model features – setups, due dates and product mix – affect the structure of policies. And there is opportunity for further work along these lines.

Heavy-traffic analysis has also been applied to other queueing control problems. We have discussed the scheduling of service for multiple sources by a single server. We may instead have to schedule and route input from multiple sources to several possible servers; see Bell and Williams (2001), Harrison and Lopez (1999) and references therein. More generally, we may have multiclass processing networks; see Harrison (1988, 2000, 2001a,b), Kumar (2000) and references therein.

In conclusion, the successful application of heavy-traffic analysis to these classic operations-research stochastic scheduling problems provides ample evidence that heavy-traffic stochastic-process limits for queues have engineering significance.