# Chapter 6

# Unmatched Jumps in the Limit Process

## 6.1. Introduction

As illustrated by the random walks with Pareto steps in Section 1.4 and the workload process with Pareto inputs in Section 2.3, it can be important to consider stochastic-process limits in which the limit process has jumps, i.e., has discontinuous sample paths. The jumps observed in the plots in Chapter 1 correspond to exceptionally large increments in the plotted sequences, i.e., large steps in the simulated random walk and large inputs of required work in the simulated workload process of the queue. Thus, in the associated stochastic-process limit, the jumps in the limit process are *matched* by corresponding jumps in the converging processes. However, there are related situations in which the jumps in the limit process are not matched by jumps in the converging processes.

Indeed, a special focus of this book is on stochastic-process limits with unmatched jumps in the limit process. In the extreme case, the converging stochastic processes have continuous sample paths. Then the sample paths of the converging processes have portions with steep slope corresponding to the limiting jumps. In other cases, a single jump in the sample path of the limiting stochastic process corresponds to many small jumps in the sample path of one of the converging stochastic processes. In this chapter we give several examples showing how a stochastic-process limit with unmatched jumps in the limit process can arise. Most of these examples will be treated in detail later.

We give special attention to stochastic-process limits with unmatched

jumps in the limit process because they represent an interesting phenomenon and because they require special treatment beyond the conventional theory. In particular, as discussed in Section 3.3, whenever there are unmatched jumps in the limit process, we cannot have a stochastic-process limit in the function space $D$ with the conventional Skorohod (1956) $J_1$ topology. To establish the stochastic-process limit, we instead use the $M$ topology.

Just as in Chapter 1, we primarily draw our conclusions in this chapter by looking at pictures. By plotting initial segments of the stochastic processes for various sample sizes, we can see the stochastic-process limits emerging before our eyes. As before, the plots often do the proper scaling automatically, and thus reveal statistical regularity associated with a macroscopic view of uncertainty. The plots also show the relevance of stochastic-process limits with unmatched jumps in the limit process.

First, though, we should recognize that it is common for the limit process in a stochastic-process limit to have continuous sample paths. For example, that is true for Brownian motion, which is the canonical limiting stochastic process, occurring as the limit in Donsker's theorem, discussed in Chapters 1 and 4. In many books on stochastic-process limits, *all* the stochastic-process limits that are considered have limit processes with continuous sample paths, and there is much to consider.

Moreover, when a limit process in a stochastic-process limit does have discontinuous sample paths, the jumps in the limit process are often matched in the converging processes. We have already pointed out that only matched jumps appear in the examples in Chapter 1. Indeed, there is a substantial literature on stochastic-process limits where the limit process may have jumps and those jumps are matched in the converging processes. The extreme-value limits in Resnick (1987) and the many stochastic-process limits in Jacod and Shiryaev (1987) are all of this form.

However, even for the examples in Chapter 1 with limit processes having discontinuous sample paths, we would have stochastic-process limits with unmatched jumps in the limit process if we formed the continuous-time representation of the discrete-time process using linear interpolation, as in (2.1) in Chapter 1. We contend that the linearly interpolated processes should usually be regarded as asymptotically equivalent to the step-function versions used in Chapter 1; i.e., one sequence of scaled processes should converge if and only if the other does, and they should have the same limit process. That asymptotic eqivalence is suggested by Figure 1.13, which plots the two continuous-time representations of a random walk with uniform random steps. As the sample size $n$ increases, both versions approach Brownian motion. Indeed, as $n$ increases, the two alternative continuous-

time representations become indistinguishable.

In Section 6.2 we look at more examples of random walks, comparing the linearly interpolated continuous-time representations (which always have continuous sample paths) to the standard step-function representation for the same random-walk sample paths. Now we make this comparison for random walks approaching a limit process with discontinuous sample paths. Just as in Chapter 1, we obtain jumps in the limit process by considering random walks with steps having a heavy-tailed distribution, in particular, a Pareto distribution. As before, the plots reveal statistical regularity. The plots also show that it is natural to regard the two continuous-time representations of scaled discrete-time processes as asymptotically equivalent.

However, the unmatched jumps in the limit process for the random walks in Section 6.2 can be avoided if we use the step-function representation instead of the linearly interpolated version. Since the step-function version seems more natural anyway, the case for considering unmatched jumps in the limit process is not yet very strong. In the rest of this chapter we give examples in which stochastic-process limits with unmatched jumps in the limit process cannot be avoided.

## 6.2. Linearly Interpolated Random Walks

All the stochastic-process limits with jumps in the limit process considered in Chapter 1 produce unmatched jumps when we form the continuous-time representation of the original discrete-time process by using linear interpolation. We now want to show, by example, that it is natural to regard the linearly interpolated continuous-time representation as asymptotically equivalent to the standard step-function representation in settings where the limit process has jumps.

Given a random walk or any discrete-time process $\{S_k : k \geq 0\}$, the scaled-and-centered step-function representations are defined for each $n \geq 1$ by

$$\mathbf{S}_n(t) \equiv c_n^{-1}(S_{\lfloor nt \rfloor} - m\lfloor nt \rfloor), \quad 0 \leq t \leq 1 \ , \tag{2.1}$$

where $\lfloor x \rfloor$ is the greatest integer less than $x$ and $c_n \to \infty$ as $n \to \infty$. The associated linearly interpolated versions are

$$\tilde{\mathbf{S}}_n(t) \equiv (nt - \lfloor nt \rfloor)\mathbf{S}_n((\lfloor nt \rfloor + 1)/n) + (1 + \lfloor nt \rfloor - nt)\mathbf{S}_n(\lfloor nt \rfloor/n) \ , \tag{2.2}$$

for $0 \leq t \leq 1$. Clearly the sample paths of $\mathbf{S}_n$ in (2.1) are discontinuous for all $n$ (except in the special case in which $S_k = S_0, 1 \leq k \leq n$), while the sample paths of $\tilde{\mathbf{S}}_n$ in (2.2) are continuous for all $n$.

### 6.2.1. Asymptotic Equivalence with $M_1$

We contend that the two sequences of processes $\{\mathbf{S}_n : n \geq 0\}$ and $\{\tilde{\mathbf{S}}_n : n \geq 0\}$ in the function space $D \equiv D([0,1], \mathbb{R})$ should be *asymptotically equivalent*, i.e., if either converges in distribution as $n \to \infty$, then so should the other, and they should have the same limit. It is easy to see that the desired asymptotic equivalence holds with the $M_1$ metric. In particular, we can show that $d_{M_1}(\mathbf{S}_n, \tilde{\mathbf{S}}_n) \Rightarrow 0$ as $n \to \infty$.

**Theorem 6.2.1.** (the $M_1$ distance between the continuous-time representations) *For any discrete-time process* $\{S_k : k \geq 0\}$,

$$d_{M_1}(\mathbf{S}_n, \tilde{\mathbf{S}}_n) \leq n^{-1} \quad \textit{for all} \quad n \geq 1 \ ,$$

*for* $\mathbf{S}_n$ *in* (2.1) *and* $\tilde{\mathbf{S}}_n$ *in* (2.2).

**Proof.** For the $M_1$ metric, we can use an arbitrary parametric representation of the step-function representation $\mathbf{S}_n$. Then, for any $\epsilon > 0$, we can construct the associated parametric representation of $\tilde{\mathbf{S}}_n$ so that it agrees with the other parametric reprentation at the finitely many points in the domain $[0,1]$ mapping into the points $(k/n, \mathbf{S}_n(k/n))$ on the completed graph of $\mathbf{S}_n$ for $0 \leq k \leq n$, with the additional property that the spatial components of the two parametric representations differ by at most $n^{-1} + \epsilon$ anywhere. Since $\epsilon$ was arbitrary, we obtain the desired conclusion. ∎

We can apply Theorem 6.2.1 and the convergence-together theorem, Theorem 11.4.7, to establish the desired asymptotic equivalence with respect to convergence in distribution.

**Corollary 6.2.1.** (asymptotic equivalence of continuous-time representations) *If either* $\mathbf{S}_n \Rightarrow \mathbf{S}$ *in* $(D, M_1)$ *or* $\tilde{\mathbf{S}}_n \Rightarrow \mathbf{S}$ *in* $(D, M_1)$, *then both limits hold.*

Note that the conclusion of Theorem 6.2.1 is much stronger than the conclusion of Corollary 6.2.1. Corollary 6.2.1 concludes that $\mathbf{S}_n$, $\tilde{\mathbf{S}}_n$ and $\mathbf{S}$ all have approximately the same probability laws for all suitably large $n$, whereas Theorem 6.2.1 concludes that the individual sample paths of $\mathbf{S}_n$ and $\tilde{\mathbf{S}}_n$ are likely to be close for all suitably large $n$.

We used plots to illustrate the asymptotic equivalence of $\tilde{\mathbf{S}}_n$ and $\mathbf{S}_n$ for random walks with uniform steps, for which the limit process is Brownian motion, in Figure 1.13. That asymptotic equivalence is proved by Corollary 6.2.1. (Since the limit process has continuous sample paths, the various non-uniform Skorohod topologies are equivalent in this example.)

Now we use plots again to illustrate the asymptotic equivalence of $\tilde{\mathbf{S}}_n$ and $\mathbf{S}_n$ in random walks with jumps in the limit process. Since the asymptotic equivalence necessarily holds in the $M_1$ topology by virtue of Corollary 6.2.1, but not in the $J_1$ topology, we are presenting a case for using the $M_1$ topology.

### 6.2.2. Simulation Examples

We give three examples, all involving variants of the Pareto distribution.

**Example 6.2.1.** *Centered random walk with Pareto(p) steps.*

As in (3.5) (iii) in Section 1.3, we consider the random walk $\{S_k : k \geq 0\}$ with IID steps

$$X_k \equiv U_k^{-1/p} \tag{2.3}$$

for $U_k$ uniformly distributed on the interval $[0, 1]$. The steps then have a Pareto($p$) distribution with parameter $p$, having ccdf $F^c(t) = t^{-p}$ for $t \geq 1$. We first consider the case $1 < p \leq 2$. In that case, the steps have a finite mean $m = 1 + (p-1)^{-1}$ but infinite variance. In Figures 1.20 – 1.22, we saw that the plots of the centered random walks give evidence of jumps. The supporting FCLT (in Section 4.5) states that the step-function representations converge in distribution to a stable Lévy motion, which indeed has discontinuous sample paths.

Just as in Chapter 1, we use the statistical package $S$ to simulate and plot the initial segments of the stochastic processes. Plots of the two continuous-time representations $\mathbf{S}_n$ and $\tilde{\mathbf{S}}_n$ for the same sample paths of the random walk are given for the case $p = 1.5$ and $n = 10^j$ with $j = 1, 2, 3$ in Figure 6.1. For $n = 10$, the two continuous-time representations look quite different. Indeed, at first it may seem that they cannot be corresponding continuous-time representations of the same realized segment of the random walk, but closer examination shows that the two continuous-time representations are correct. However, for $n = 100$ and beyond, the two continuous-time representations look very similar. For larger values of $n$ such as $n = 10^4$ and beyond, the two continuous-time representations look virtually identical.

So far we have considered only $p = 1.5$. We now illustrate how the plots depend on $p$ for $1 < p \leq 2$. In Figure 6.2 we plot the two continuous-time representations of the random walk with Pareto($p$) steps for three values of $p$, in particular for $p = 1.1, 1.5$ and $1.9$. We do the plot for the case $n = 100$ using the *same uniform random numbers* (exploiting (2.3)). In each plot the largest steps stem from the smallest uniform random numbers. The

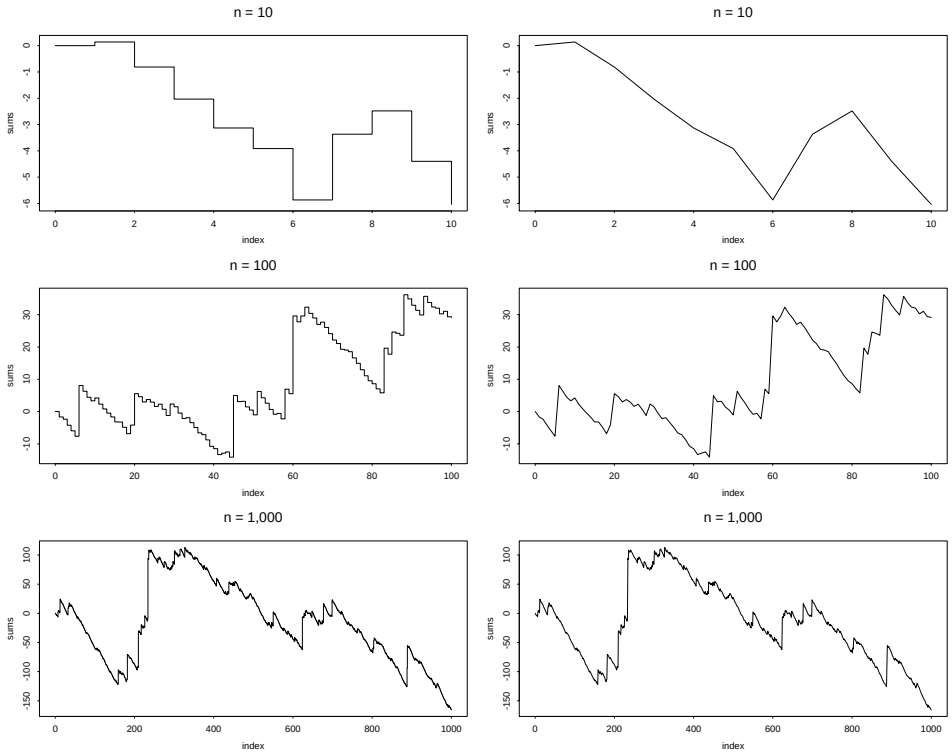



Figure 6.1: Plots of the two continuous-time representations of the centered random walk with Pareto(1.5) steps for $n = 10^j$ with $j = 1, 2, 3$. The step-function representation $\mathbf{S}_n$ in (2.1) appears on the left, while the linearly interpolated version $\tilde{\mathbf{S}}_n$ in (2.2) appears on the right.

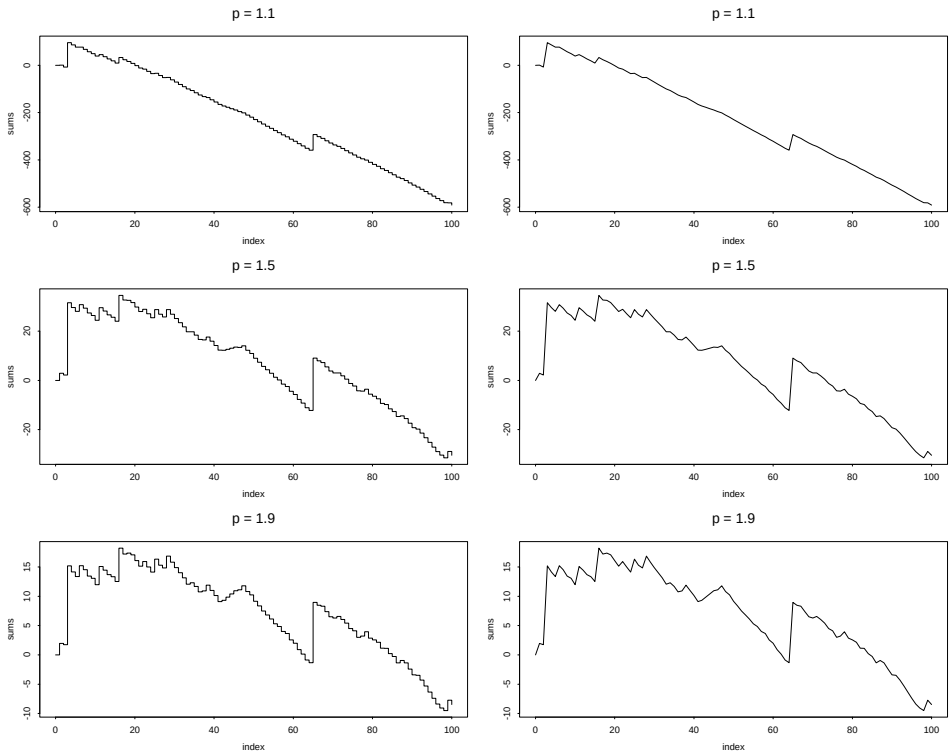
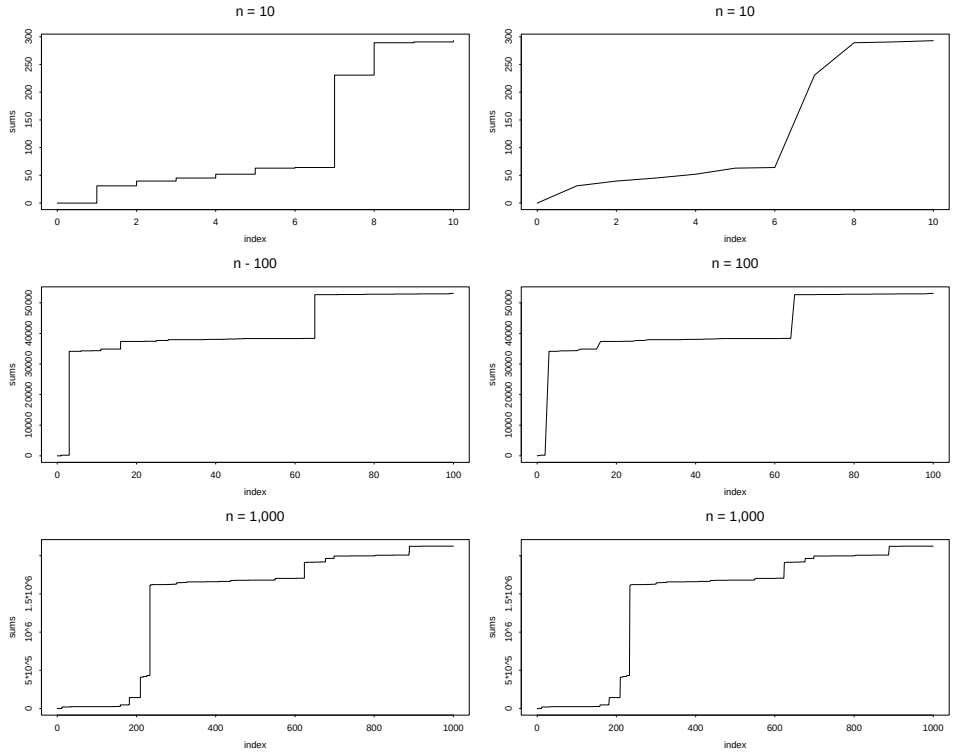



Figure 6.2: Plots of the two continuous-time representations of the centered random walk with Pareto($p$) steps with $p = 1.1, 1.5$ and $1.9$ for $n = 10^2$ based on the same uniform random numbers (using (2.3)). The step-function representation $\mathbf{S}_n$ in (2.1) appears on the left, while the linearly interpolated version $\tilde{\mathbf{S}}_n$ in (2.2) appears on the right.

three smallest uniform random numbers in this sample were $U_3 = 0.00542$, $U_{65} = 0.00836$ and $U_{16} = 0.0201$. The corresponding large steps can be seen in each case of Figure 6.2. Again, we see that the limiting stochastic process should have jumps (up). That conclusion is confirmed by considering larger and larger values of $n$. As in Figures 6.1 and 6.2, the two continuous-time representations look very similar. And the little difference we see for $n = 100$ deceases as $n$ increases.

**Example 6.2.2.** *Uncentered random walk with Pareto(0.5) steps.* In Figures 1.19, 1.25 and 1.26 we saw that the *uncentered* random walk with Pareto(0.5) steps should have stochastic-process limits with jumps in the limit process. The supporting FCLT implies convergence to another stable




Plots of the uncentered random walk with Pareto(0.5) steps for $n = 10^j$ with $j = 1, 2, 3$. The step-function representation $\mathbf{S}_n$ in (2.1) appears on the left, while the linearly interpolated version $\tilde{\mathbf{S}}_n$ in (2.2) appears on the right.

Lévy motion as $n \to \infty$ (again see Section 4.5). Moreover, such a limit holds for IID Pareto($p$) steps whenever $p \leq 1$, because then the steps have infinite mean.

Now we look at the two continuous-time representations in this setting. We now plot the two continuous-time representations $\tilde{\mathbf{S}}_n$ and $\mathbf{S}_n$ associated with the uncentered random walk with Pareto(0.5) steps for $n = 10^j$ with $j = 1, 2, 3$ in Figure 6.2.2. Again, the two continuous-time representations initially (for small $n$) look quite different, but become indistinguishable as $n$ increases. Just as in Chapter 1, even though there are jumps, we see statistical regularity associated with large $n$. Experiments with different $n$ show the self-similarity discussed before.

**Example 6.2.3.** *Centered random walk with limiting jumps up and down.*

The Pareto distributions considered above have support on the inter-

val $[1, \infty)$, so that, even with centering, the positive tail of the step-size distribution is heavy, but the negative tail of the step-size distribution is light. Consequently the limiting stochastic process in the stochastic-process limit for the random walks with Pareto steps can only have jumps up. (See Section 4.5)

We can obtain a limit process with both jumps up and jumps down if we again use (2.3) to define the steps, but we let $U_k$ be uniformly distributed on the interval $[-1, 1]$ instead of in $[0, 1]$. Then we can have both arbitrarily large negative jumps and arbitrarily large positive jumps. We call the resulting distribution a *symmetric Pareto distribution* (with parameter $p$). Since the distribution is symmetric, no centering need be done for the plots or the stochastic-process limits.

To illustrate, we make additional comparisons between the linearly interpolated continuous-time representation and the step-function continuous-time representation of the random walk, now using the symmetric Pareto($p$) steps for $p = 1.5$. The plots are shown in Figure 6.3. We plot the two continuous-time representations for $n = 10^j$ with $j = 2, 3, 4$. From the plots, it is evident that the limit process now should have jumps down as well as jumps up. Again, the two continuous-time representations look almost identical for large $n$.

## 6.3. Heavy-Tailed Renewal Processes

One common setting for stochastic-process limits with unmatched jumps in the limit process, which underlies many applications, is a heavy-tailed renewal process. Given partial sums $S_k \equiv X_1 + \cdots + X_k, k \geq 1$, from a sequence of nonnegative random variables $\{X_k : k \geq 1\}$ (without an IID assumption), the associated stochastic process $N \equiv \{N(t) : t \geq 0\}$ defined by

$$N(t) \equiv max\{k \geq 0 : S_k \leq t\}, \quad t \geq 0, \tag{3.1}$$

where $S_0 \equiv 0$, is called a *stochastic counting process*. When the random variables $X_k$ are IID, the counting process is called a *renewal counting process* or just a *renewal process*.

### 6.3.1. Inverse Processes

Roughly speaking (we will be more precise in Chapter 13), the stochastic processes $\{S_k : k \geq 1\}$ and $N \equiv \{N(t) : t \geq 0\}$ can be regarded as *inverses*

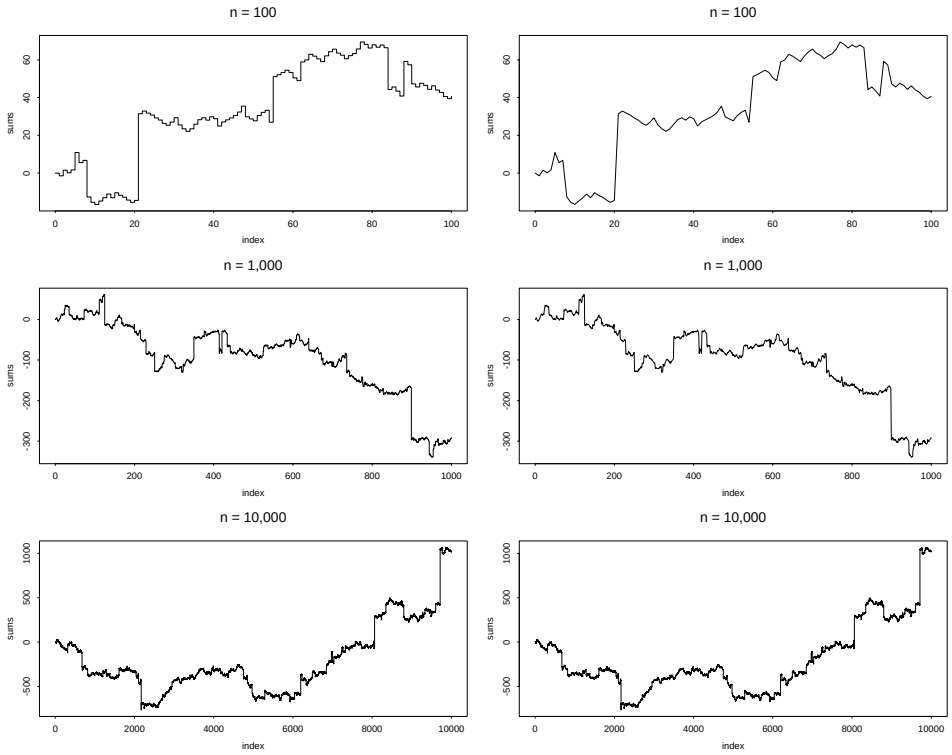



Figure 6.3: Plots of the two continuous-time representations of the random walk with symmetric Pareto(1.5) steps for $n = 10^j$ with $j = 2, 3, 4$. The step-function representation $\mathbf{S}_n$ in (2.1) appears on the left, while the linearly interpolated version $\tilde{\mathbf{S}}_n$ in (2.2) appears on the right.

of each other, without imposing the IID condition, because

$$S_k \le t \quad \text{if and only if} \quad N(t) \ge k. \tag{3.2}$$

The $M_1$ topology is convenient for relating limits for partial sums to associated limits for the counting processes, because the $M_1$-topology definition makes it easy to exploit the inverse relation in the continuous-mapping approach.

Moreover, it is not possible to use the standard $J_1$ topology to establish limits of scaled versions of the counting processes, because the $J_1$ topolgy requires all jumps in the limit process to be matched in the converging stochastic processes. The difficulty with the $J_1$ topology on $D$ can easily be seen when the random variables $X_k$ are strictly positive. Then the counting process $N$ increases in unit jumps, and scaled versions of the counting process, such as

$$\mathbf{N}_n(t) \equiv c_n^{-1}(N(nt) - m^{-1}nt), \quad t \ge 0, \tag{3.3}$$

where $c_n \to \infty$, have jumps of magnitude $1/c_n$, which are asymptotically negligible as $n \to \infty$ . Hence, if $\mathbf{N}_n$ in (3.3) is ever to converge as $n \to \infty$ to a limiting stochastic process with discontinuous sample paths, then we must have unmatched jumps in the limit process. Then we need the $M_1$ topology on $D$.

What is not so obvious, however, is that $\mathbf{N}_n$ will ever converge to a limiting stochastic process with discontinuous sample paths. However, such limits can indeed occur. Here is how: A long interrenewal time creates a long interval between jumps up in the renewal process. The long interrenewal time appears horizontally rather than vertically, not directly causing a jump. However, during such an interval, the scaled process in (3.3) will decrease linearly at rate $n/mc_n$, due to the translation term not being compensated for by any jumps up. When $n/c_n \to \infty$ (the usual case), the slope approaches $-\infty$. When the interrenewal times are long enough, these portions of the sample path with steep slope down can lead to jumps *down* in the limit process.

A good way to see how jumps can appear in the limit process for $\mathbf{N}_n$ is to see how limits for $\mathbf{N}_n$ in (3.3) are related to associated limits for $\mathbf{S}_n$ in (2.1) when both scaled processes are constructed from the same underlying process $\{S_k : k \ge 0\}$. A striking result from the continuous-mapping approach to stochastic-process limits (to be developed in Chapter 13) is an equivalence between stochastic-process limits for partial sums and associated counting processes, exploiting the $M_1$ topology (but not requiring any

direct independence or common-distribution assumption). As a consequence of Corollary 13.8.1, we have the following result:

**Theorem 6.3.1.** (FCLT equivalence for counting processes and associated partial sums) *Suppose that* $0 < m < \infty$, $c_n \to \infty$, $n/c_n \to \infty$ *and* $\mathbf{S}(0) = 0$. *Then*

$$\mathbf{S}_n \Rightarrow \mathbf{S} \quad in \quad (D, M_1) \tag{3.4}$$

*for* $\mathbf{S}_n$ *in* (2.1) *if and only if*

$$\mathbf{N}_n \Rightarrow \mathbf{N} \quad in \quad (D, M_1) \tag{3.5}$$

*for* $\mathbf{N}_n$ *in* (3.3), *in which case*

$$(\mathbf{S}_n, \mathbf{N}_n) \Rightarrow (\mathbf{S}, \mathbf{N}) \quad in \quad (D^2, WM_1) \ , \tag{3.6}$$

*where the limit processes are related by*

$$\mathbf{N}(t) \equiv (m^{-1}\mathbf{S} \circ m^{-1}\mathbf{e})(t) \equiv m^{-1}\mathbf{S}(m^{-1}t), \quad t \geq 0 \ , \tag{3.7}$$

*or, equivalently,*

$$\mathbf{S}(t) = (m\mathbf{N} \circ m\mathbf{e})(t) \equiv m\mathbf{N}(mt), \quad t \geq 0 \ , \tag{3.8}$$

*where* $\mathbf{e}(t) = t, \quad t \geq 0$.

Thus, *whenever the limit process* $\mathbf{S}$ *in* (3.4) *has discontinuous sample paths, the limit process* $\mathbf{N}$ *in* (3.5) *necessarily has discontinuous sample paths as well.* Moreover, $\mathbf{S}$ has only jumps up (down) if and only if $\mathbf{N}$ has only jumps down (up). Whenever $\mathbf{S}$ and $\mathbf{N}$ have discontinuous sample paths, the $M_1$ topology is needed to express the limit for $\mathbf{N}_n$ in (3.5). In contrast, the limit for $\mathbf{S}_n$ in (3.4) can hold in $(D, J_1)$.

### 6.3.2. The Special Case with $m = 1$

The close relation between the limit processes $\mathbf{S}$ and $\mathbf{N}$ in (3.4) – (3.8) is easy to understand and visualize when we consider plots for the special case of strictly positive steps $X_k$ with translation scaling constant $m = 1$. Note that the limit process $\mathbf{N}$ in (3.7) becomes simply $-\mathbf{S}$ when $m = 1$.

Also note that we can always scale so that $m = 1$ without loss of generality: For any given sequence $\{X_k : k \geq 0\}$, when we multiply $X_k$ by $m$ for all $k$, we replace $\mathbf{S}_n$ by $m\mathbf{S}_n$ and $\mathbf{N}_n$ by $\mathbf{N}_n \circ m^{-1}\mathbf{e}$. Hence, the limits $\mathbf{S}$ and $\mathbf{N}$ are replaced by $m\mathbf{S}$ and $\mathbf{N} \circ m^{-1}\mathbf{e}$, respectively.

Hence, suppose that $m = 1$. A useful observation, then, is that $N(S_k) = k$ for all $k$. (We use the assumption that the variables $X_k$ are strictly positive.) With that in mind, note that we can plot $N(t) - t$ versus $t$, again using the statistical package $S$, by plotting the points $(0, 0)$, $(S_k, N(S_k) - 1 - S_k)$ and $(S_k, N(S_k) - S_k)$ in the plane $\mathbb{R}^2$ and then performing linear interpolation between successive points.

Roughly speaking, then, we can plot $N(t) - t$ versus $t$ by plotting $N(S_k) - S_k$ versus $S_k$. On the other hand, when we plot the centered random walk $\{S_k - k : k \geq 0\}$, we plot $(S_k - k)$ versus $k$. Since $N(S_k) = k$, we have

$$N(S_k) - S_k = k - S_k = -(S_k - k) \ .$$

Thus, the second component of the pair $(S_k, N(S_k) - S_k)$ is just minus 1 times the second component of the pair $(k, S_k - k)$. Thus, the plot of $N(t) - t$ versus $t$ should be very close to the plot of $-(S_k - k)$ versus $k$. The major difference is in the first component: For the renewal process, the first component is $S_k$; for the random walk, the first component is $k$. However, since $n^{-1}S_n \to 1$ as $n \to \infty$ by the SLLN, that difference between these two first components disappears as $n \to \infty$.

**Example 6.3.1.** *Centered renewal processes with Pareto(p) steps for* $1 < p < 2$. By now, we are well acquainted with a situation in which the limit for $\mathbf{S}_n$ in (3.4) holds and the limit process $\mathbf{S}$ has discontinuous sample paths: That occurs when the underlying process $\{S_k : k \geq 0\}$ is a random walk with IID Pareto($p$) steps for $1 < p < 2$. Then the limit (3.4) holds with $m = 1 + (p-1)^{-1}$ and $\mathbf{S}$ being a stable Lévy motion, which has discontinuous sample paths. The discontinuous sample paths are clearly revealed for the case $p = 1.5$ in Figures 1.20 – 1.22 and 6.1.

To make the relationship clear, we consider the case $m = 1$. We obtain $m = 1$ in our example with IID Pareto(1.5) steps by dividing the steps by 3; i.e., we let $X_k \equiv U_k^{-2/3}/3$. For this example with Pareto(1.5) steps having ccdf decay rate $p = 3/2$ and mean 1, we plot both the centered renewal process ($N(t) - t$ versus $t$) and minus 1 times the centered random walk ($-(S_k - k)$ versus $k$). We plot both sample paths, putting the centered renewal process on the left, for the cases $n = 10^j$ with $j = 1, 2, 3$ in Figure 6.4. We plot three possible representations of each for $n = 10^4$ in Figure 6.5. (We plot the centered random walk directly; i.e., we do not use either of the continuous-time represenations.)

For small $n$, the sample paths of the two centered processes look quite different, but as $n$ increases, the sample paths begin to look alike. The

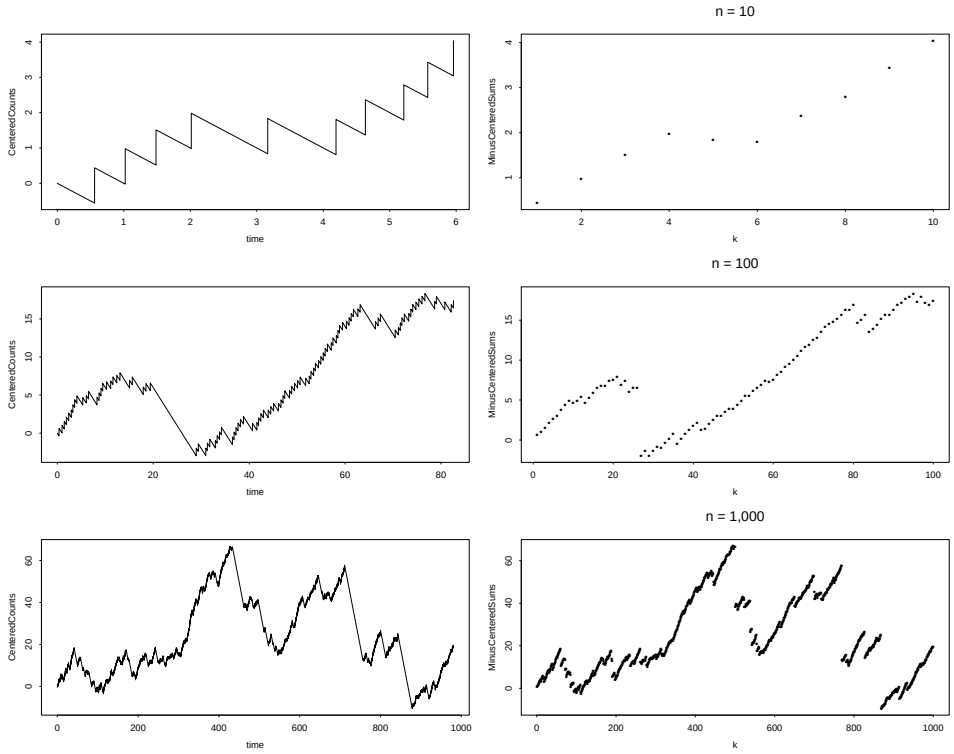



Figure 6.4: Plots of the centered renewal process (on the left) and minus 1 times the centered random walk (on the right) for Pareto(1.5) steps with mean $m = 1$ and $n = 10^j$ for $j = 1, 2, 3$.

jumps in the centered random walk plot are matched with portions of the centered-renewal-process plot with very steep slope. As $n$ increases, the slopes in the portions of the centered-renewal-process plots corresponding to the random-walk jumps tend to get steeper and steeper, approaching the jump itself.

It is natural to wonder how the plots look as the decay rate $p$ changes within the interval (1,2), which is the set of values yielding a finite mean but an infinite variance. We know that for smaller $p$ the jumps are likely to be larger. To see what happens, we plot three realizations each of the centered renewal process and minus 1 times the centered random walk for Pareto steps having decay rates $p = 7/4$ and $p = 5/4$ (normalized as before to have mean 1) for $n = 10^4$ in Figures 6.6 and 6.7. From Figures 6.5 – 6.7, we see that the required space scaling decreases, the two irregular paths

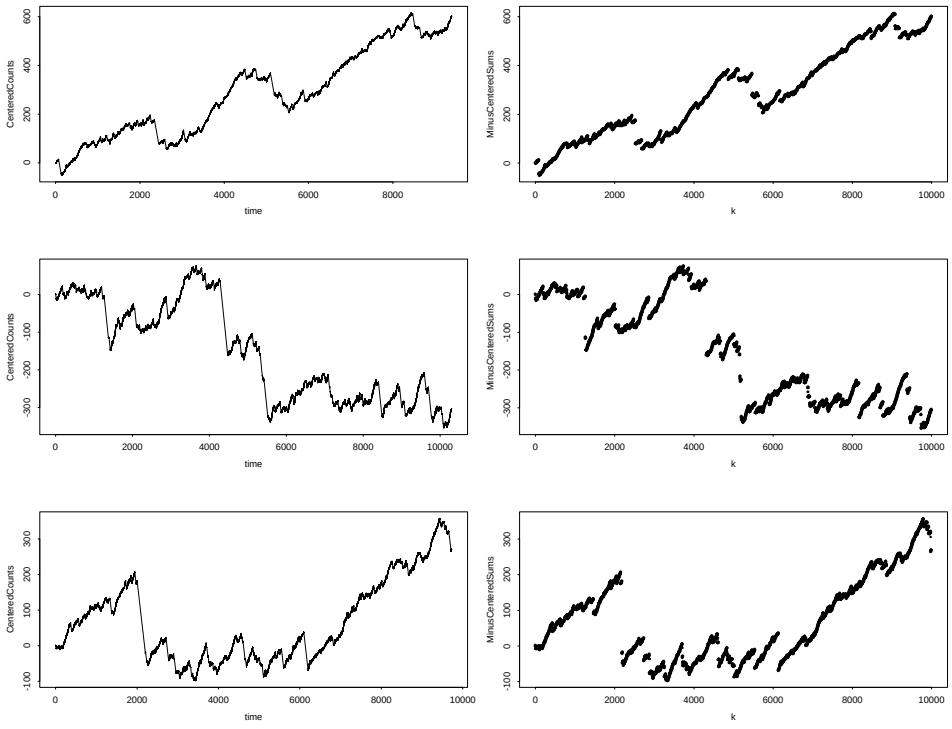



Figure 6.5: Plots of three independent realizations of the centered renewal process (on the left) and minus 1 times the centered random walk (on the right) for Pareto(1.5) steps with mean $m = 1$ and $n = 10^4$.

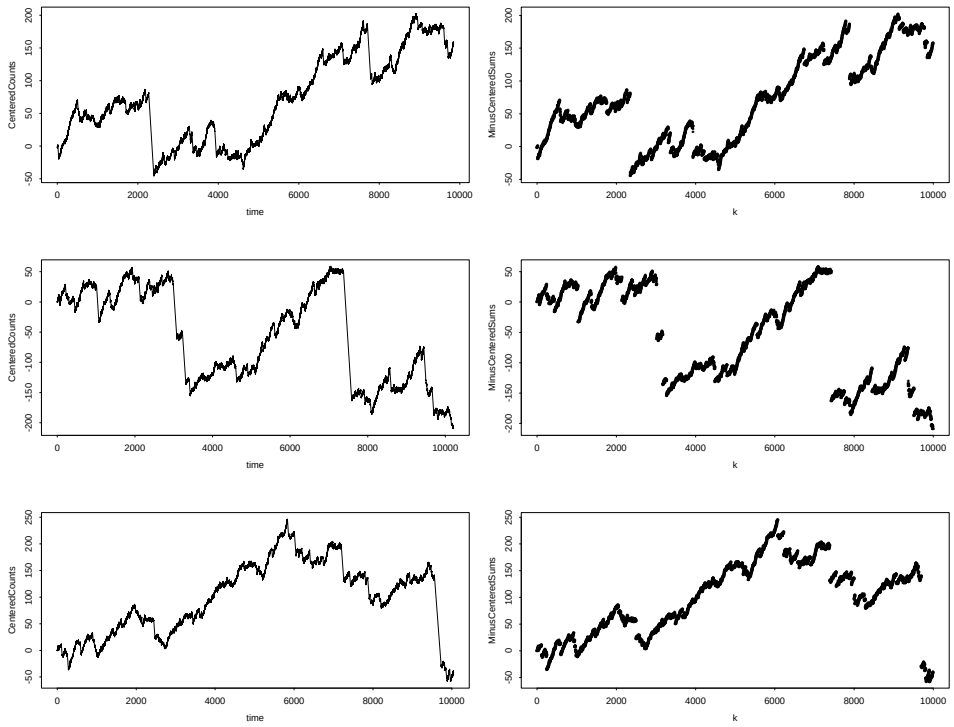



Figure 6.6: Plots of three independent realizations of the centered renewal process (on the left) and minus 1 times the centered random walk (on the right) associated with Pareto($p$) steps in (2.3) with $p = 7/4$, $m = 1$ and $n = 10^4$.

become closer, and the slopes in the renewal-process plot become steeper, as $p$ increases from 5/4 to 3/2 to 7/4. For $p = 5/4$, we need larger $n$ to see steeper slopes. However, in all cases we can see that there should be unmatched jumps in the limit process. ■

For the Pareto-step random walk plots in Figures 6.4 – 6.7, we not only have $-\mathbf{S}_n \Rightarrow -\mathbf{S}$ and $\mathbf{N}_n \Rightarrow -\mathbf{S}$, but also the realizations of $\mathbf{N}_n$ and $-\mathbf{S}_n$ are becoming close to each other as $n \to \infty$. Such asymptotic equivalence follows from Theorem 6.3.1 by virtue of Theorem 11.4.8. Recall that we can start with any translation scaling constant $m$ and rescale to $m = 1$.

**Corollary 6.3.1.** (asymptotic equivalence) *If, in addition to the assump-*

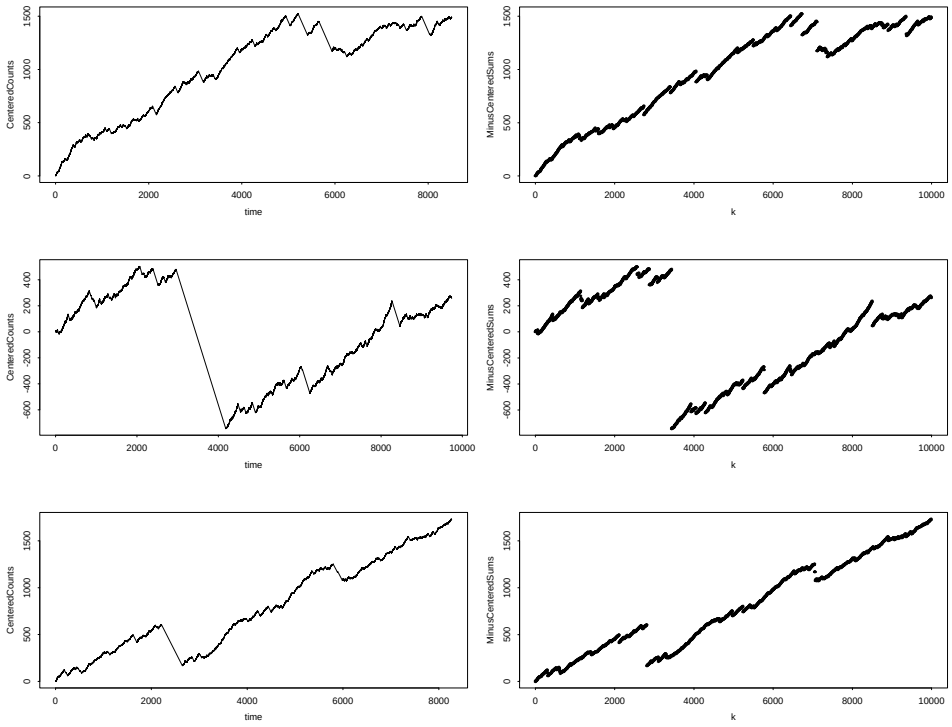



Figure 6.7: Plots of three independent realizations of the centered renewal process (on the left) and minus 1 times the centered random walk (on the right) associated with Pareto($p$) steps in (2.3) with $p = 5/4$, $m = 1$ and $n = 10^4$.

*tions of Theorem 6.3.1, the limit* $\mathbf{S}_n \Rightarrow \mathbf{S}$ *in* (3.4) *holds and* $m = 1$, *then*

$$d_{M_1}(\mathbf{N}_n, -\mathbf{S}_n) \Rightarrow 0 \ .$$

To summarize, properly scaled versions (with centering) of a renewal process (or, more generally, any counting process) are intimately connected with associated scaled versions (with centering) of random walks, so that FCLTs for random walks imply associated FCLTs for the scaled renewal process (and vice versa), provided that we use the $M_1$ topology. When the limit process for the random walk has discontinuous sample paths, so does the limit process for the renewal process, which necessarily produces unmatched jumps. We state specific FCLTs for renewal processes in Section 7.3.

## 6.4. A Queue with Heavy-Tailed Distributions

Closely paralleling the heavy-tailed renewal process just considered, heavy-traffic limits for the queue-length process in standard queueing models routinely produce stochastic-process limits with unmatched jumps in the limit process when the service times or interarrival times have heavy-tailed distributions (again meaning with infinite variance). In fact, renewal processes enter in directly, because the customer arrival process in the queueing model is a stochastic counting process, which is a renewal process when the interarrival times are IID.

We start by observing that jumps in the limit process associated with stochastic-process limits for the queue-length process almost always are unmatched jumps. That is easy to see when all the interarrival times and service times are strictly positive. (That is the case w.p.1 when the interarrival times and service times come from sequences of random variables with distributions assigning 0 probability to 0.) Then the queue length (i.e., the number of customers in the system) makes changes in unit steps. Thus, any jumps in the limit process associated with a stochastic-process limit for a sequence of queue-length processes with space scaling, where we divide by $c_n$ with $c_n \to \infty$ as $n \to \infty$, must be unmatched jumps.

The real issue, then, is to show that jumps can appear in stochastic-process limits for the queue-length process. The stochastic-process limits we have in mind occur in a heavy-traffic setting, as in Section 2.3.

### 6.4.1. The Standard Single-Server Queue

To be specific, we consider a single-server queue with unlimited waiting room and the first-come first-served service discipline. (We will discuss this model further in Chapter 9. The model can be specified by a sequence of ordered pairs of nonnegative random variables $\{(U_k, V_k) : k \geq 1\}$. The variable $U_k$ represents the *interarrival time* between customers $k$ and $k-1$, with $U_1$ being the arrival time of the first customer, while the variable $V_k$ represents the *service time* of the customer $k$. The *arrival time* of the customer $k$ is thus

$$T_k \equiv U_1 + \cdots + U_k, \quad k \geq 1 \ , \tag{4.1}$$

and the *departure time* of the customer $k$ is

$$D_k \equiv T_k + W_k + V_k, \quad k \geq 1 \ , \tag{4.2}$$

where $W_k$ is the *waiting time* (before beginning service) of customer $k$. The waiting times can be defined recursively by

$$W_k \equiv [W_{k-1} + V_{k-1} - U_k]^+, \quad k \geq 2 \ , \tag{4.3}$$

where $[x]^+ \equiv max\{x, 0\}$ and $W_1 \equiv 0$. (We have assumed that the system starts empty; that of course is not critical.)

We can now define associated continuous-time processes. The counting processes are defined just as in (3.1). The *arrival (counting) process* $\{A(t) : t \geq 0\}$ is defined by

$$A(t) \equiv max\{k \geq 0 : T_k \leq t\}, \quad t \geq 0 \ , \tag{4.4}$$

the *departure (counting) process* $\{D(t) : t \geq 0\}$ is defined by

$$D(t) \equiv max\{k \geq 0 : D_k \leq t\}, \quad t \geq 0 \ , \tag{4.5}$$

and the *queue-length process* $\{Q(t) : t \geq 0\}$ is defined by

$$Q(t) \equiv A(t) - D(t), \quad t \geq 0 \ . \tag{4.6}$$

Here the queue length is the number in system, including the customer in service, if any.

The standard single-server queue that we consider now is closely related to the infinite-capacity version of the discrete-time fluid queue model considered in Section 2.3. Indeed, the recursive definition for the waiting times in (4.3) is essentially the same as the recursive definition for the workloads

in (3.1) of Section 2.3 in the special case in which the waiting space is unlimited, i.e., when $K = \infty$. For the fluid queue model, we saw that the behavior of the workload process is intimately connected to the behavior of an associated random walk, and that heavy-tailed inputs lead directly to jumps in the limit process for appropriately scaled workload processes. The same is true for the waiting times here, as we will show in Section 9.2.

### 6.4.2. Heavy-Traffic Limits

Thus, just as in Section 2.3, we consider a sequence of models indexed by $n$ in order to obtain interesting stochastic-process limits for stable queueing systems. We can achieve such a framework conveniently by scaling a single model. We use a superscript $n$ to index the new quantities constructed in the $n^{th}$ model.

We start with a single sequence $\{(U_k, V_k) : k \geq 1\}$. Note that we have made no stochastic assumptions so far. The key assumption is a FCLT for the random walks, in particular,

$$(\mathbf{S}_n^u, \mathbf{S}_n^v) \Rightarrow (\mathbf{S}^u, \mathbf{S}^v) \quad \text{in} \quad (D^2, WM_1) \ , \tag{4.7}$$

where

$$\mathbf{S}_n^u \equiv c_n^{-1}(\sum_{i=1}^{\lfloor nt \rfloor} U_i - \lfloor nt \rfloor)$$

and

$$\mathbf{S}_n^v \equiv c_n^{-1}(\sum_{i=1}^{\lfloor nt \rfloor} V_i - \lfloor nt \rfloor) \ .$$

The standard stochastic assumption to obtain (4.7) is for $\{U_k\}$ and $\{V_k\}$ to be independent sequences of IID random variables with

$$EV_k = EU_k = 1 \quad \text{for all} \quad k \geq 1 \ . \tag{4.8}$$

and other regularity conditions (finite variances to get convergence to Brownian motion or asymptotic power tails to get convergence to stable Lévy motions).

Paralleling the scaling in (3.13) in Section 2.3, we form the $n^{th}$ model by letting

$$U_k^n \equiv b_n U_k \quad \text{and} \quad V_k^n \equiv V_k, \quad k \geq 1 \ , \tag{4.9}$$

where

$$b_n \equiv 1 + mc_n/n \quad \text{for} \quad n \geq 1 \ . \tag{4.10}$$

We assume that $c_n/n \downarrow 0$ as $n \to \infty$, so that $b_n \downarrow 1$ as $n \to \infty$. The scaling in (4.9) is a simple deterministic scaling of time in the arrival process; i.e., the arrival process in model $n$ is

$$A^n(t) \equiv A(b_n^{-1}t), \quad t \geq 0 \ ,$$

for $b_n$ in (4.10).

We now form scaled stochastic processes associated with the sequence of models by letting

$$\mathbf{W}_n(t) \equiv c_n^{-1} W^n_{\lfloor nt \rfloor}, \tag{4.11}$$

and

$$\mathbf{Q}_n(t) \equiv c_n^{-1} Q^n(nt), \quad t \geq 0 \ . \tag{4.12}$$

We now state the heavy-traffic stochastic-process limit, which follows from Theorems 9.3.3, 9.3.4 and 11.4.8. As before, for $x \in D$, let $Disc(x)$ be the set of discontinuities of $x$.

**Theorem 6.4.1.** (heavy-traffic limit for the waiting times and queue lengths) *Suppose that the stochastic-process limit in* (4.7) *holds and the scaling in* (4.9) *holds with* $c_n \to \infty$ *and* $c_n/n \to 0$. *Suppose that almost surely the sets* $Disc(\mathbf{S}^u)$ *and* $Disc(\mathbf{S}^v)$ *have empty intersection and*

$$P(\mathbf{S}^u(0) = 0) = P(\mathbf{S}^v(0) = 0) = 1 \ .$$

*Then*

$$\mathbf{W}_n \Rightarrow \mathbf{W} \equiv \phi(\mathbf{S}^v - \mathbf{S}^u - m\mathbf{e}) \quad in \quad (D, M_1) \ , \tag{4.13}$$

*where* $\phi$ *is the one-sided reflection map in* (5.4) *in Section* 3.5,

$$(\mathbf{W}_n, \mathbf{Q}_n) \Rightarrow (\mathbf{W}, \mathbf{W}) \quad in \quad (D^2, WM_1) \tag{4.14}$$

*and*

$$d_{M_1}(\mathbf{W}_n, \mathbf{Q}_n) \Rightarrow 0 \ . \tag{4.15}$$

We now explain why the limit process $\mathbf{Q}$ for the scaled queue-length processes can have jumps. Starting from (4.6), we have

$$\mathbf{Q}_n = \mathbf{A}_n - \mathbf{D}_n \ , \tag{4.16}$$

where

$$\mathbf{A}_n(t) \equiv c_n^{-1}(A^n(nt) - nt), \quad t \geq 0 \tag{4.17}$$

and

$$\mathbf{D}_n(t) \equiv c_n^{-1}(D^n(nt) - nt), \quad t \geq 0 \ . \tag{4.18}$$

Just as for the renewal processes in the previous section, an especially long service time (interarrival time) can cause a period of steep linear slope down in $\mathbf{D}_n$ ($\mathbf{A}_n$), which can correspond to jumps down in the associated limit process. The jump down from $\mathbf{D}_n$ ($\mathbf{A}_n$) corresponds to a jump up (down) in the limit process for $\mathbf{Q}_n$.

### 6.4.3. Simulation Examples

What we intend to do now is simulate and plot the waiting-time and queue-length processes under various assumptions on the interarrival-time and service-time distributions. Just as with the empirical cdf in Example 1.1.1 and the renewal process in Section 6.3, when we plot the queue-length process we need to plot a portion of a continuous-time process. Just as in the two previous cases, we can plot the queue-length process with the statistical package $S$, exploiting underlying random sequences. Here the relevant underlying random sequences are the arrival times $\{T_k\}$ and the departure times $\{D_k\}$, defined recursively above in (4.1) and (4.2).

Since the plotting procedure is less obvious now, we specify it in detail. We first form two dimensional vectors by appending a $+1$ to each arrival time and a $-1$ to each departure time. (Instead of the arrival time $T_n$, we have the vector $(T_n, 1)$; instead of the departure time $D_n$, we have the vector $(D_n, -1)$.) We then combine all the vectors (creating a matrix) and sort on the first component. The new first components are thus the successive times of any change in the queue length (arrival or departure). We then form the successive cumulative sums of the second components, which converts the second components into the queue lengths at the times of change. We could just plot the queue lengths at the successive times of change, but we go further to plot the full continuous-time queue-length process. We can plot by linear interpolation, if we include each queue length value twice, at the jump when the value is first attained and just before the next jump. (This method inserts a vertical line at each jump.)

We now give an $S$ program to read in the first $n$ interarrival times, service times and waiting times and plot the queue-length process over the time interval that these $n$ customers are in the system (ignoring all subsequent arrivals). At the end of the time interval the system is necessarily empty. Our construction thus gives an odd end effect, but it can be truncated. Indeed, in our plots below we do truncate (at the expected time of the $n^{th}$ arrival).

Here is the $S$ function:

```
QueueLength <- function(U, V, W) {
QueueLength <- vector("numeric", 2*length(U) + 1)
T <- cumsum(U)                          #construct arrival times
D <- T + W + V                          # departure times
TT <- cbind(T, +1)                      #append +1 to arrivals
DD <- cbind(D, -1)                      #append -1 to deps.
m <- rbind(TT, DD)                      #merge into one matrix
msort <- m[sort.list(m[, 1]),]          #sort on first comp.
time1 <- msort[, c(1)]                  #extract change times
QLchg <- msort[, c(2)]                  #queue length changes
QL1 <- cumsum(QLchg)                    #successive q. lths.
time2 <- c(0, time1, time1)             #times for lin.interp.
time <- sort(time2)
n <- length(time1)                      #q. lths. for lin. int.
QL <- c(0, QL1)
for (k in seq(n)) {
QueueLength[[2 * k - 1]] <- QL[[k]]
QueueLength[[2 * k]] <- QL[[k]] }
QueueLength[2 * n + 1] <- QL[n + 1]
plot(time, QueueLength, type = "l")     #do the plotting
}
```

We now consider a few examples. We use the *Kendall notation* to describe the model: $X/Y/c$ specifies a model with $c$ servers, arrival process of type $X$ and service process of type $Y$. For either $X$ or $Y$, $GI$ denotes an IID sequence with a general distribution, while $M$ (for Markov) denotes (in addition) the exponential distribution. We use $P_p$ for the Pareto distribution with parameter $p$.

**Example 6.4.1.** *The M/M/1 Queue.*

We first consider the standard M/M/1 queue. Thus, here we assume that the interarrival times and service times come from mutually independent sequences of IID exponentially distributed random variables. It suffices to specify the means of the interarrival time and the service time. Using the scaling in equations (4.9) and (4.10), we need to specify the constant $m$ and the space-scaling sequence $\{c_n : n \geq 1\}$.

At this point, we know what to do: There are no heavy-tailed distributions, so we should let $c_n = \sqrt{n}$. We also let $m = 1$. Thus, we fully specify

the sequence of $M/M/1$ models indexed by $n$ by letting

$$EU_k^n = 1 + 1/\sqrt{n} \quad \text{and} \quad EV_k^n = 1 \quad \text{for all} \quad k \quad \text{and} \quad n \ . \tag{4.19}$$

With that choice, the plotter can do the appropriate scaling automatically.

We are primarily interested in the queue-length process, but we also plot the waiting times, because it is instructive to compare the plotted queue-length process to the plotted waiting times. Hence, we plot both the waiting times of the first $n$ customers (linearly interpolated) and the queue-length process over the time interval $[0, nEU_1^n]$ for the cases $n = 10^j$ with $j = 1, 2, 3$ in Figure 6.8.

For small $n$, the queue-length process looks very different from the waiting time sequence, but as $n$ increases, the sample path of the queue length process becomes very similar to the sample path of the waiting times, except possibly for the final portion, where the queue length experiences some of the end effect. To confirm what we see in Figure 6.8, we plot three possible realizations of the waiting times and the queue lengths for $n = 10^4$ in Figure 6.9.

From our experience so far, we should know what to expect: The plots are approaching plots of reflected Brownian motion with drift $-1$ (which does not have any jumps). Now the conditions and conclusions of Theorem 6.4.1 hold with $c_n = \sqrt{n}$ and $\mathbf{W} = \phi(\sigma\mathbf{B} - m\mathbf{e})$, where $\mathbf{B}$ is standard Brownian motion, $\mathbf{e}$ is the identity map, $\phi : D \to D$ is the one-sided reflection map and $\sigma^2 = Var(U_1) + Var(V_1) = 2$. We apply Donsker's theorem – Theorem 4.3.2.

Moreover, the plots show that the distance between the two scaled processes is indeed asymptotically negligible. Since the limit process here has continuous sample paths, we can express this asymptotic equivalence using the uniform norm over $[0, 1]$:

$$\| \mathbf{W}_n - \mathbf{Q}_n \| \Rightarrow 0 \quad \text{as} \quad n \to \infty \ . \tag{4.20}$$

■

**Example 6.4.2.** *The $M/P_{1.5}/1$ Queue.*

We now modify the previous example by letting the service-time distribution be Pareto($p$) with $p = 1.5$ and mean 1. (In the framework of Section 1.3.3, we can use $3^{-1}U^{-2/3}$, where $U$ is uniform on the interval $[0, 1]$, which has ccdf $F^c(t) = (3t)^{-3/2}$ for $t \geq 1$.) With this heavy-tailed service-time distribtuion, we must scale space differently, because the space scaling in the

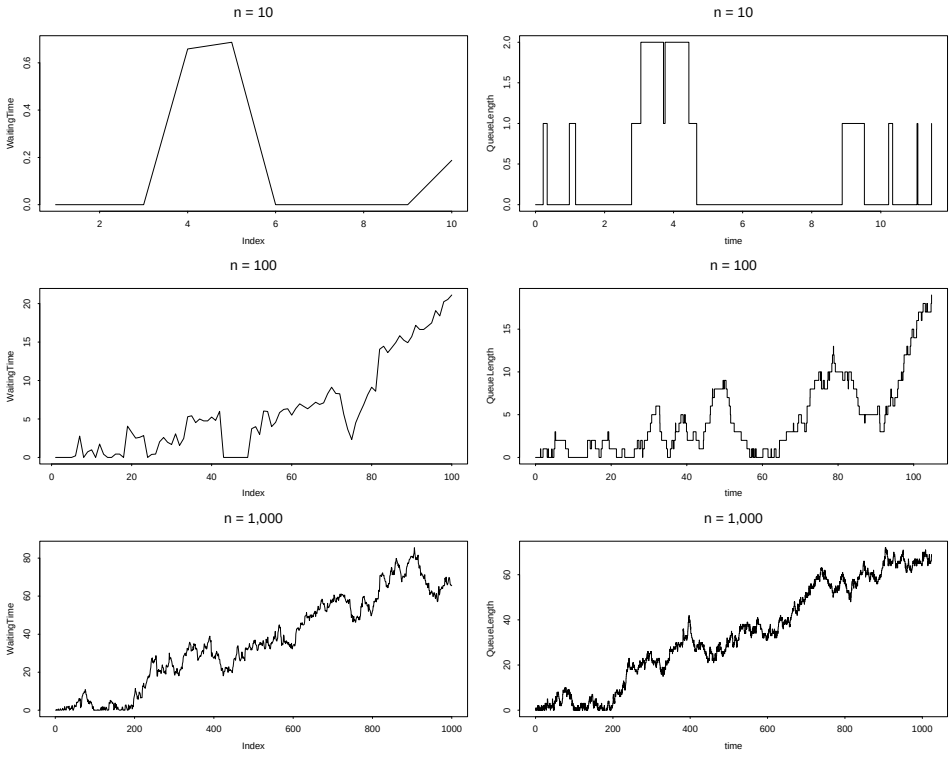



Figure 6.8: Plots of the waiting times of the first $n$ arrivals (on the left) and the queue-length process over the interval $[0, nEU_1^n]$ (on the right) in the M/M/1 queue with scaling in (4.19) for $n = 10^j$ with $j = 1, 2, 3$.

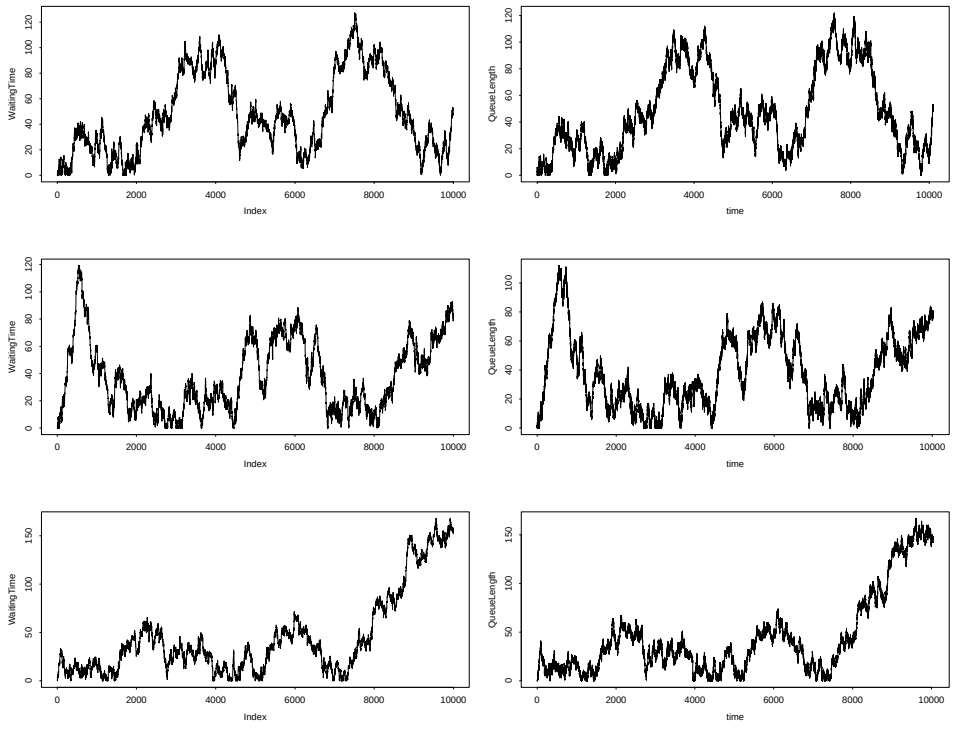



Figure 6.9: Three possible realizations of the waiting times of the first $n$ arrivals (on the left) and the queue-length process over the interval $[0, nEU_1^n]$ (on the right) in the M/M/1 queue with scaling in (4.19) for $n = 10^4$.

FCLT for the random walk involves $c_n = n^{2/3}$ instead of $c_n = n^{1/2}$. Hence, instead of the scaling in (4.19), we now use

$$EU_k^n = 1 + n^{-1/3} \quad \text{and} \quad EV_k^n = 1 \quad \text{for all} \quad k \quad \text{and} \quad n\,. \tag{4.21}$$

The new scaling makes the traffic intensity $\rho_n$ smaller than in Example 6.4.1 for any given $n$. For example, for $n = 10,000$, before we had $\rho_n = 1/1.01 \approx 0.990$, while now we have $\rho_n \approx 1/1.046 \approx 0.956$.

We plot three possible realizations of the waiting times of the first $n$ customers (on the bottom or left) and queue-length process over the interval $[0, nEU_k^n]$ (on the top or right) for $n = 10^4$, in Figure 6.10. The first two plots look much like the $M/M/1$ plots in Figure 6.9 except now we can see upward jumps. But the third plot is very different!

There is now much more variability in the sample paths because of the possibility of the occasional very large jumps. The range of values is exceptionally small in case 2 and exceptionally large in case 3. The possibility of exceptionally large jumps produces large variations from plot to plot, as we saw for the random walks in Figure 1.21.

When we look at the third plots closely, it is not evident that the waiting-time and queue-length plots are for the same sample path. For instance, the second big jump in the waiting times occurs at about index 3100, whereas the corresponding second steep incline in the queue-length path begins at about time 4100. However, upon reflection, we see that these actually are consistent, because the waiting time of the customer having the second large service time is about 1000. Since the arrival rate is 1, that customer arrives at about time 3100. Hence that customer enters service, and begins occupying the server, at about time 4100. Thus the queue length should start building up at about time 4100, as it does.

The upward jumps are less sharp for the queue-length process, which we know actually increases by unit jumps, but the asymptotic behavior is evident from the plots. In this case, we are seeing a reflected stable Lévy motion with drift $-1$, which has discontinuous sample paths, instead of a reflected Brownian motion. Again we can explain the statistical regularity we see by Theorem 6.4.1. However, now the scaling involves $c_n = n^{2/3}$.

By Theorems 4.5.2 and 4.5.3, the limit process is $\mathbf{W} \equiv \phi(\sigma \mathbf{S}^v - \mathbf{e}) \equiv \sigma\phi(\mathbf{S}^v - \sigma^{-1}\mathbf{e}$, where $\sigma = 1/3C_\alpha^{2/3}$ for $C_\alpha$ in (5.14) of Section 4.5.1, $\mathbf{S}^v$ is a centered $\alpha$-stable Lévy motion with $\mathbf{S}^v(1) \stackrel{\mathrm{d}}{=} S_\alpha(1,1,0)$ and $\alpha = 3/2$. (Its steady-state distribution is given in Section 8.5.2.) Again, it is evident that the two scaled processes $\mathbf{W}_n$ and $\mathbf{Q}_n$ should now be asymptotically equivalent. ∎

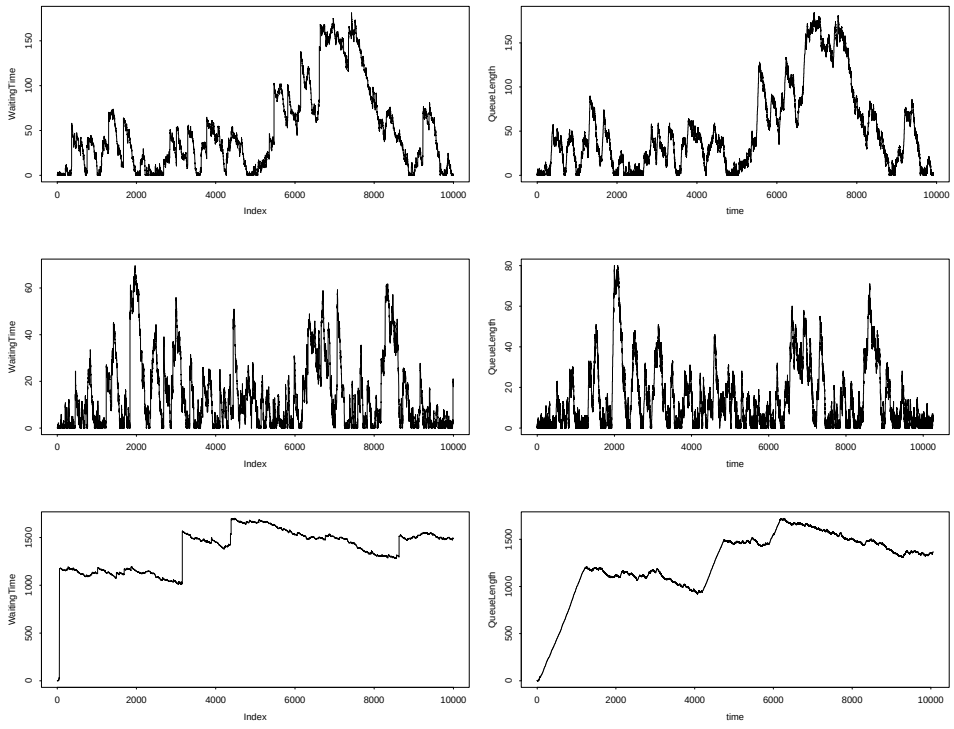



Figure 6.10: Three possible realizations of the waiting times of the first $n$ arrivals (on the left) and the queue-length process over the interval $[0, nEU_1^n]$ (on the right) in the $M/P_{1.5}/1$ queue with the scaling in (4.21) for $n = 10^4$.

**Example 6.4.3.** *The $P_{1.5}/M/1$ Queue.*

It is evident that a heavy-tailed service-time distribution should cause greater congestion, but it may not be evident that a heavy-tailed interarrival-time distribution can as well, because extra long interarrival times only serve to empty out the queue. However, heavy-tailed interarrival-time distributions can cause congestion as well. The reason is that, for given fixed mean, the occasionally exceptionally long interarrival times must be compensated for in the distribution by shorter interarrival times, and these shorter interarrival times lead to bursts of arrivals and thus increased queue lengths.

We illustrate by considering the $P_{1.5}/M/1$ queue, which has IID Pareto(1.5) interarrival times and IID exponential service times. This model is the *dual* of the model in Example 6.4.2, with the role of the interarrival times and service times switched (adjusted by scaling, so that the expected interarrival times are bigger than the expected service times in both cases).

In Figure 6.11 we plot three possible realizations of the waiting times of the first $n$ arrivals (on the left) and the queue-length process over the interval $[0, nEU_1^n]$ (on the right) in the $P_{1.5}/M/1$ queue with the scaling in (4.21) for $n = 10^4$.

As in Figures 6.8 – 6.10, the queue-length plots are similar to the waiting-time plot, except possibly for the final portion of the queue-length plot, where the queue experiences its end effect. However, unlike in the previous figures, in Figure 6.11 we see evidence of jumps down.

Just as for the $M/P_{1.5}/1$ model, the heavy-traffic FCLT in Theorem 6.4.1 applies to the $P_{1.5}/M/1$ and $P_{1.5}/P_{1.5}/1$ models. Indeed, we again have the same scaling, but now the limiting reflected stable Lévy motions are different, having jumps down only for the $P_{1.5}/M/1$ model and haviing jumps both up and down for the $P_{1.5}/P_{1.5}/1$ model, instead of having jumps up only for the $M/P_{1.5}/1$ model.

For the $P_{1.5}/M/1$ model, the heavy-traffic stochastic-process limit for the workload process is $\mathbf{W}_n \Rightarrow \mathbf{W}$, where again $c_n = n^{2/3}$, but now

$$\mathbf{W} = \phi(-\sigma \mathbf{S}^u - \mathbf{e}) \stackrel{\mathrm{d}}{=} \sigma\phi(-\mathbf{S}^u - \sigma^{-1}\mathbf{e}) \ ,$$

where $\sigma = 1/3C_\alpha^{2/3}$ for $\alpha = 3/2$, just as in Example 6.4.2. Here $-\mathbf{S}^u(1) \stackrel{\mathrm{d}}{=} S_\alpha(1, -1, 0)$.

For the $P_{1.5}/P_{1.5}/1$ model, the limit process is

$$\mathbf{W} = \phi(\sigma \mathbf{S}^v - \sigma \mathbf{S}^u - \mathbf{e}) \stackrel{\mathrm{d}}{=} \sigma\phi(\mathbf{S}^v - \mathbf{S}^u - \sigma^{-1}\mathbf{e}) \ ,$$

where $\mathbf{S}^v - \mathbf{S}^u \stackrel{\mathrm{d}}{=} \mathbf{S}$ with $\mathbf{S}$ being a stable Lévy motion satisfying $\mathbf{S}(1) \stackrel{\mathrm{d}}{=} 2^{2/3} S_\alpha(1, 0, 0)$; see (5.8) – (5.11) in Section 4.5.1.

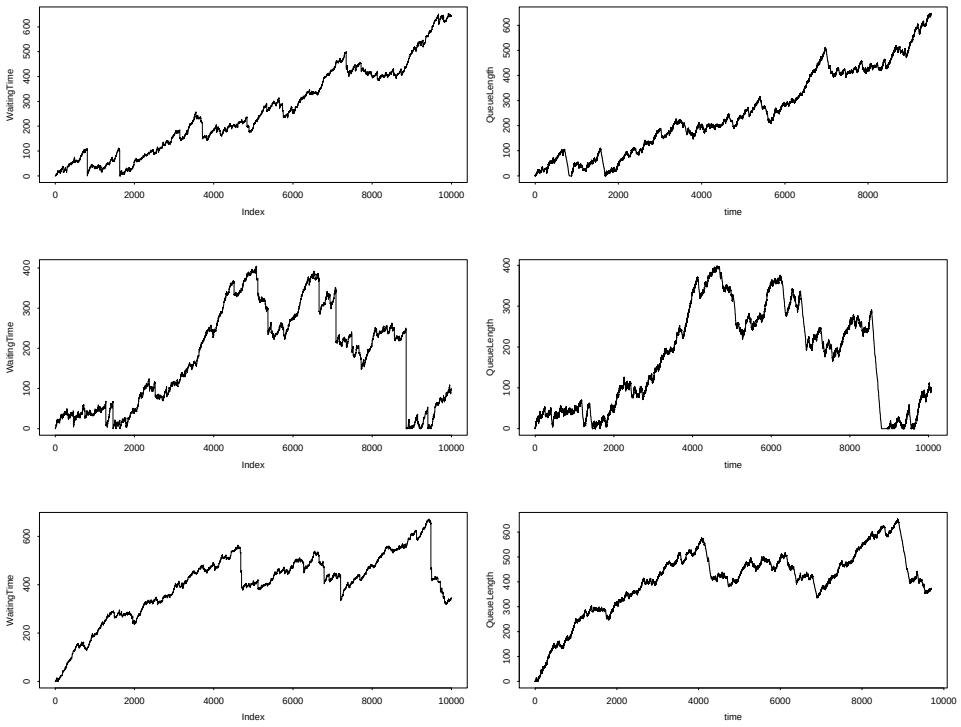



Figure 6.11: Three possible realizations of the waiting times of the first $n = 10^4$ arrivals (on the left) and the queue-length process over the interval $[0, nEU_1^n]$ (on the right) in the $P_{1.5}/M/1$ queue with the scaling in (4.21) for $n = 10^4$.

We should not be fooled by the jumps down for the $P_{1.5}/M/1$ model. Of course, the jumps down do constitute reductions in congestion, but elsewhere in the plot the sample path is rising, so that the range of values experienced can be substantial. Indeed, that is demonstrated by the heavy-traffic FCLT, which has space scaling by $n^{2/3}$, just as for the $M/P_{1.5}/1$ model in Example 6.4.2. ∎

## 6.5. Rare Long Service Interruptions

The queueing example just considered illustrates a common cause of congestion in queues: stochastic vartiability in the interarrival times and service times. However, congestion in queues can occur for other reasons: For example, the servers may be subject to breakdown and failure, causing service interruptions. In manufacturing systems, service interruptions due to machine failures or the unavailability of parts are often the dominant sources of congestion. With evolving communication networks, there is debate about whether the most important source of congestion is the uncertain burstiness of customer input or the uncertain failure of system elements. The biggest problems tend to occur when both happen together.

We can better understand the impact of service interruptions upon performance if we develop a probability model and establish appropriate stochastic-process limits. One such model, considered by Kella and Whitt (1990), is a queue with rare long service interruptions. The queue can be a standard single-server queue with unlimited waiting space, the first-come first-served service discipline and random arrivals and service times, as considered in the previous section. We can supplement that model by allowing random service interruptions. The interruptions can be triggered by queueing events; e.g., they could occur only when the queue becomes empty. Or they can occur exogenously. We will consider the case in which they occur exogenously.

Specifically, we will assume that the availability of the server is characterized by an alternating renewal process; i.e., there are alternating periods in which the server is available (up) or unavailable (down). For tractability, we assume that the up and down times come from mutually independent sequences of IID positive random variables with finite means and variances.

A revealing stochastic-process limit can be obtained by considering the queue in a heavy-traffic limit, in which the load is allowed to approach the critical value for stability. If the interruptions remain unchanged, then the service interruptions alter the conventional heavy-traffic limit with a reflected Brownian motion limit process only by increasing the traffic intensity and increasing the variance parameter of the Brownian motion, both of

which cause increased congestion. However, we obtain a different nondegenerate limit, which is consistent with many applications, if we let the intervals between interruptions and the durations of the interruptions increase in the limit. If we let these quantities increase appropriately, with the duration of an interruption being asymptotically negligible compared to the time between interruptions, then we can obtain a revealing nondegenerate limit.

In particular, an interesting limiting regime has the random up times be of order $n$ and the random down times be of order $\sqrt{n}$ as a function of the number $n$ of customers being considered. Then, with the customary scaling of time by $n$ and space by $\sqrt{n}$, the scaled up times become of order 1 and the scaled down times become of order $1/\sqrt{n}$. That makes the scaled down times asymptotically negligible. Thus, after scaling, the service interruptions occur in the limit according to a stochastic point process, with a finite positive expected number of interruptions in a finite time interval.

Since the scaled durations of the service interruptions are aymptotically negligible, the service interruptions occur instantaneously in the limit. Nevertheless, the service interruptions can have a significant spatial impact, because the number of arrivals during the order $\sqrt{n}$ down time is also of order $\sqrt{n}$. Thus, after scaling space by $\sqrt{n}$, the input during the down time causes a random jump of order 1 in the scaled queueing process at each interruption time.

The proposed scaling, with up times of order $n$ and down times of order $\sqrt{n}$, thus produces random jumps of order-1 size, spaced at random order-1 intervals. In the limit, the proportion of time that the server is unavailable because of interruption is asymptotically negligible. Nevertheless, the asymptotic impact of the interruptions can be dramatic. With this limit, it is possible to compare the effects of the service interruptions (which appear in the limit process as jumps) to the customary stochastic fluctuations. Depending on the specific parameter settings, one or the other may dominate. In Section 14.7, following Kella and Whitt (1990) and Chen and Whitt (1993), we consider networks of queues with rare long service interruptions.

When we consider limits for sequences of queue-length stochastic processes affected by rare long interruptions of the kind just described, the jumps in the limit process are typically not matched in the converging scaled queue-length processes. In the queueing system, arrivals usually are coming one at a time. During a service interruption, service stops, but the arrivals keep coming. Thus the queue length process increases by many unit steps during such periods. After scaling time and space, the $n^{th}$ scaled queue-length process increases more rapidly (due to the time scaling) but by smaller asymptotically negligible amounts (due to the space scaling). Thus

the resulting limit is a stochastic-process limit with unmatched jumps in the limit process.

In the rest of this subsection we illustrate the kind of limiting behavior provided by rare long service interruptions. To do so, we simplify the model: Even though service interruptions represent a different source of congestion than variability in customer demand, we often can represent service interruptions within the framework of a standard queueing model. We can simply include the interruption in the service time of one of the customers. Specifically, we can redefine the service-time distribution: The new service-time distribution becomes a mixture: With probability $p$, the new service time is the sum of an original service time and the interruption duration; with probability $1-p$, the new service time reduces to an original service time. We then choose the probability $p$ to match the probability that a customer is the first customer to experience a service interruption. If the timing of service interruptions needs to be modeled very precisely, then we can think of interruptions as special high-priority customers that preempt regular customers (in line or in service), but the simple model above often suffices

We have in mind rare long service interruptions occuring randomly, but to illustrate the interruption phenomenon, we let the interruptions occur in a fixed manner in our example below.

**Example 6.5.1.** *The $M/M/1$ queue with two fixed service interruptions.*

We construct a simple example to illustrate the kind of limit behavior associated with rare long service interruptions. Specifically, we consider the $M/M/1$ queue with the heavy-traffic scaling in (4.19), just as in Example 6.4.1, except that now we let customers number $n/4$ and $3n/4$ have service times of $2\sqrt{n}$ and $\sqrt{n}$, respectively, as a function of $n$. These special service times are introduced to represent interruptions that occur approximately at times $t/4$ and $3t/4$ in the scaled processes plotted over the interval $[0,1]$. (By the SLLN, the scaled arrival time of customer number $n/4$ approaches $t/4$ as $n \to \infty$.) Note that the spacings between the interruptions is indeed order $n$, while the durations of the interruptions (as captured by the special service times) are of order $\sqrt{n}$, as specified above.

We plot the waiting times of the first $n$ customers and the queue-length process for the time interval $[0, nEU_1^n]$, the expected time for the $n$ customers to arrive, for $n = 10^j$ with $j = 2, 3, 4$ in Figure 6.12. In Figure 6.12 the impact of the interruptions is clearer for the waiting times than for the queue lengths, especially for smaller $n$. For the queue-length process, the portion of the plot corresponding to the jump gets steeper as $n$ increases.

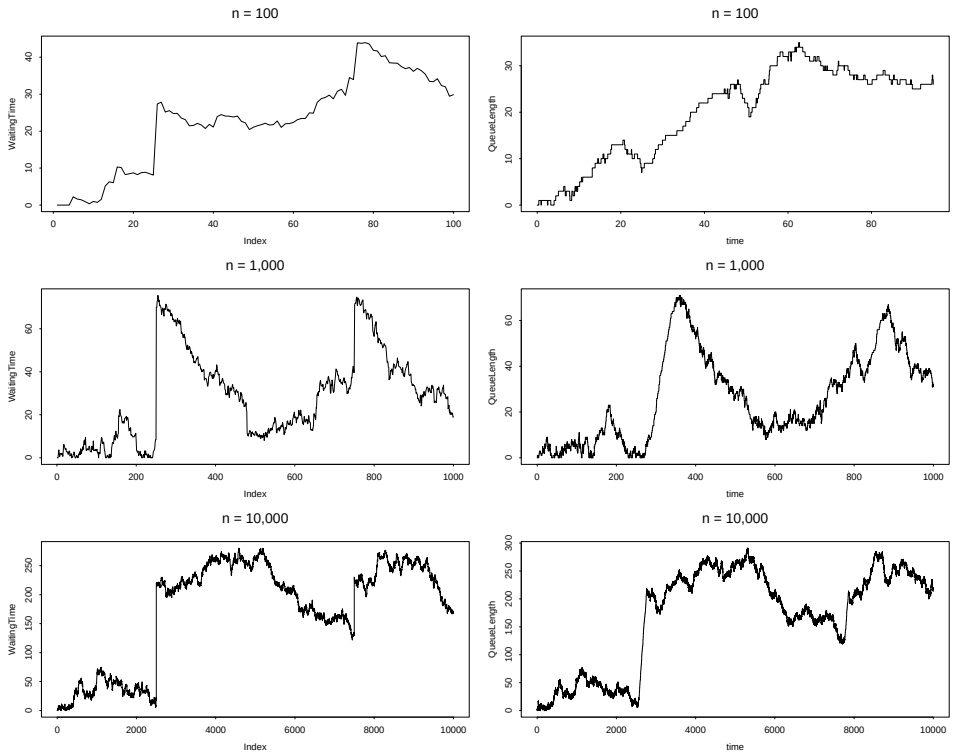



Figure 6.12: Plots of the waiting times of the first $n$ arrivals (on the left) and the queue-length process over the interval $[0, nEU_1^n]$ (on the right) for in the M/M/1 queue with scaling in (4.19) and service interruptions of length $2\sqrt{n}$ and $\sqrt{n}$ associated with customers $n/4$ and $3n/4$ for $n = 10^j$ with $j = 2, 3, 4$.

As before, we see that the queue-length and waiting-time plots coalesce as $n$ increases. Now both scaled processes approach reflected Brownian motion with drift $-1$, modified by jumps of size 2 at time $t = 1/4$ and of size 1 at time $t = 3/4$. For the scaled queue-length process, the limit process must have unmatched jumps. ∎

**Example 6.5.2.** *The $P_{1.5}/M/1$ queue with two fixed service interruptions.*

Now, as in Example 6.4.3 we consider the $P_{1.5}/M/1$ queue with heavy-traffic scaling in (4.21), modified by having customers number $n/4$ and $3n/4$ experience interruptions. We choose the $P_{1.5}/M/1$ model instead of the $M/P_{1.5}/1$ model, because it naturally (without the interruptions) produces jumps down instead of up. Thus, it will be easier to recognize the new jumps up caused by the service interruptions.

In addition, the durations of the interruptions need to be scaled differently from the scaling in Example 6.5.1. In order to be consistent with the heavy-traffic limiting behavior in Example 6.4.3, we now need to scale the durations of the interruptions by $n^{2/3}$ instead of $n^{1/2}$. In particular, now we let the service times of customers number $n/4$ and $3n/4$ be $2n^{2/3}$ and $n^{2/3}$, respectively. We plot three possible realizations of the waiting times of the first $n$ customers and the queue-length process over the time interval $[0, nEU_1^n]$, ignoring all arrivals after the first $n$, for the case $n = 10^4$ in Figure 6.13.

Just as we would expect from Figures 6.11 and 6.12, we see randomly occurring jumps down because of the $P_{1.5}$ arrival process and jumps up of magnitude 2 at time $t = 1/4$ and 1 at time $t = 3/4$. However, both kinds of jumps are much sharper for the waiting times than for the queue-length process. Hence, we evidently need larger $n$ in this case to have the queue-length plots be visually similar to the waiting-time plots. The supporting FCLTs state that both scaled processes converge to a stable Lévy motion (with jumps down only) modified by the addition of two jumps up, a jump of size 2 at $t = 1/4$ and a jump of size 1 at $t = 3/4$; again, see Sections 4.5 and 14.7. Again, for the scaled queue-length process, that limit process must have unmatched jumps. ∎

The simple models of service interruptions considered in Examples 6.5.1 and 6.5.2 are of course quite artificial. However, from these examples, we can anticipate what we will see when we use the more realistic alternating renewal process model for up and down times.

## 6.6. Time-Dependent Arrival Rates

In many service systems, congestion occurs primarily because of systematic, deterministic variations in the input rate over time. Many service systems have arrival rates that vary systematically with time, so that there are known busy periods with higher loads than average. However, everything is not known. There remains uncertainty about the actual input; there are unanticipated fluctuations about the known time-varying deterministic rates.

To better understand the behavior of queues with time-varying arrival rates, we need to focus directly on queueing models with time-varying arrival rates. Just as for stationary queueing models, it can be helpful to consider heavy-traffic limits for queues with time-varying arrival rates. With time-varying arrival rates, we still scale time, but we think of expanding time

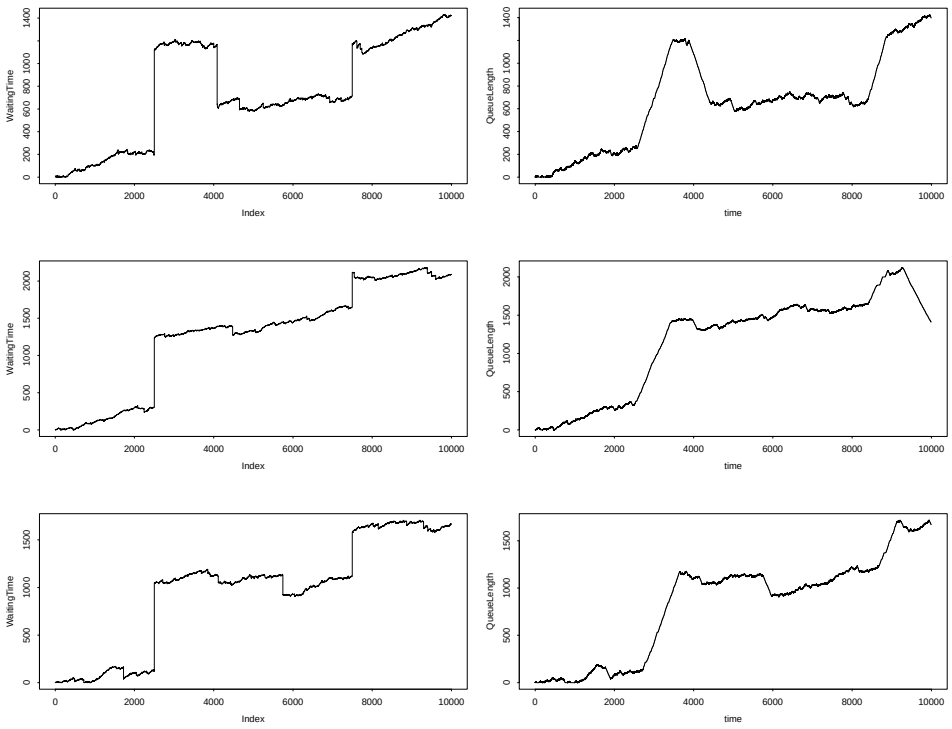



Figure 6.13: Three possible realizations of the waiting times of the first $n$ arrivals (on the left) and the queue-length process over the interval $[0, nEU_1^n]$ (on the right) in the $P_{1.5}/M/1$ queue with scaling in (4.21) and service interruptions of length $2n^{2/3}$ and $n^{2/3}$ associated with customers $n/4$ and $3n/4$ for $n = 10^4$.

immediately prior to the time of interest. We increase the overall arrival and service rate, which is tantamount to decreasing the rate of change in the arrival-rate and service-rate functions, so that temporary periods of overload or underload before the time of interest tend to persist longer and longer.

With such scaling, a law of large numbers can be established, in which the scaled queue-length process converges to a reflection of a deterministic net-input process, where the limiting deterministic net-input process satisfies an *ordinary differential equation* (ODE) driven by the original time-dependent arrival and service rates. That limit is identical to the direct deterministic ODE approximation we obtain if we ignore the stochastic aspects of the model. In the direct deterministic approximation, the net input becomes the solution an ODE driven by the time-dependent arrival and service rates; i.e., if $\lambda$ is the arrival-rate function and $\mu$ is the service-rate function, then the deterministic approximation for the queue length is the function $q$ satisfying

$$q(t) = \phi(x)(t) \equiv x(t) - \inf_{0 \le s \le t} x(s), \quad t \ge 0 \ , \tag{6.1}$$

where $\phi$ is again the one-sided reflection map, $q(0)$ is the initial queue length (assumed to satisfy $q(0) = 0$) and $x$ is the deterministic net-input function, satifying the ODE

$$\dot{x}(t) = \lambda(t) - \mu(t), \quad t \ge 0 \ . \tag{6.2}$$

When the deterministic fluctuations dominate the stochastic fluctuations, such a deterministic analysis can be very useful to describe system performance; e.g., see Oliver and Samuel (1962), Newell (1982) and Hall (1991).

However, in stochastic-process limits, we are primarily interested in going beyond the deterministic ODE limit described above. For example, Mandelbaum and Massey (1995) show that it is possible to establish a stochastic (FCLT) refinement to the deterministic ODE limit. It again can be obtained by applying the continuous-mapping approach to stochastic-process limits. In this setting, the continuous-mapping approach involves convergence preservation with nonlinear centering, and can be approached by identifying the directional derivative of the reflection map; see Chapter 6 of the Internet Supplement.

The behavior of the limit process in the stochastic-process limit depends on the deterministic function $q$. At any time, the deterministic function $q$ must be in one of three states (based on the history of the build up prior to the time of interest): overloaded, critically loaded (when the cumulative input rate is in balance with the output rate) or underloaded. (Roughly

speaking, these regimes correspond to the three cases $\rho > 1$, $\rho = 1$ and $\rho < 1$ in a stationary queueing model.)

With the usual stochastic assumptions (without any heavy-tailed distributions), the stochastic-process refinement is a diffusion process centered about the deterministic function $q$. The diffusion process corresponds to: ordinary Brownian motion when $q$ is overloaded, reflected Brownian motion when $q$ is critically loaded, and the zero function when $q$ is underloaded.

Within each region, i.e., within any interval in which the determinisitic function $q$ remains in one of its three basic states (overloaded, critically loaded or underloaded), the limiting stochastic process has continuous sample paths, but at the boundaries between different regions the limiting stochastic process can have jumps that are unmatched in the converging processes. Thus, the boundary points between different regions for the deterministic function $q$ act as phase transitions for the queueing system. Relatively abrupt changes in the queueing process can occur at these transition times. And, once again, we have a stochastic-process limit with unmatched jumps.

**Example 6.6.1.** *A shift from critically loaded to underloaded.*

We now give a simple example. In the standard situation we have in mind, the arrival-rate function is changing continuously, so that we can obtain the deterministic net-input function by solving the ODE in (6.2). However, now we consider the more elementary situation in which there is a sudden shift down in the arrival rate at one time. As in the standard situation, we let the service rate be constant (although that is not required).

We let the queue initially be critically loaded, i.e., with $\rho = 1$, and then in the middle of the time period, we reduce the arrival rate, making the model underloaded. For simplicity, we again use the $M/M/1$ queue. We let the mean service time always be 1. We actually deviate slightly from the prescription for the arrival rate: We let the mean interarrival time for the first $n/2$ customers be 1 and the mean interarrival time of the next $n/2$ customers be 2. Hence, after $n/2$ arrivals, the instantaneous traffic intensity suddenly shifts from $\rho = 1$ to $\rho = 0.5$. Of course, with this definition, the shift in arrival rate occurs at a random time instead of a deterministic time, but after scaling time by $n$, that scaled random shift time converges to $t/2$ w.p.1. Thus, what we do is essentially the same as if we let the arrival-rate shift occur exactly at time $n/2$ when we consider $n$ arrivals.

For the specified model, we plot the waiting times of the first $n$ customers and the queue-length process over the time interval $[0, n]$ for $n = 10^j$ for $j = 2, 3, 4$ in Figure 6.14. As in previous plots, the situation is somewhat

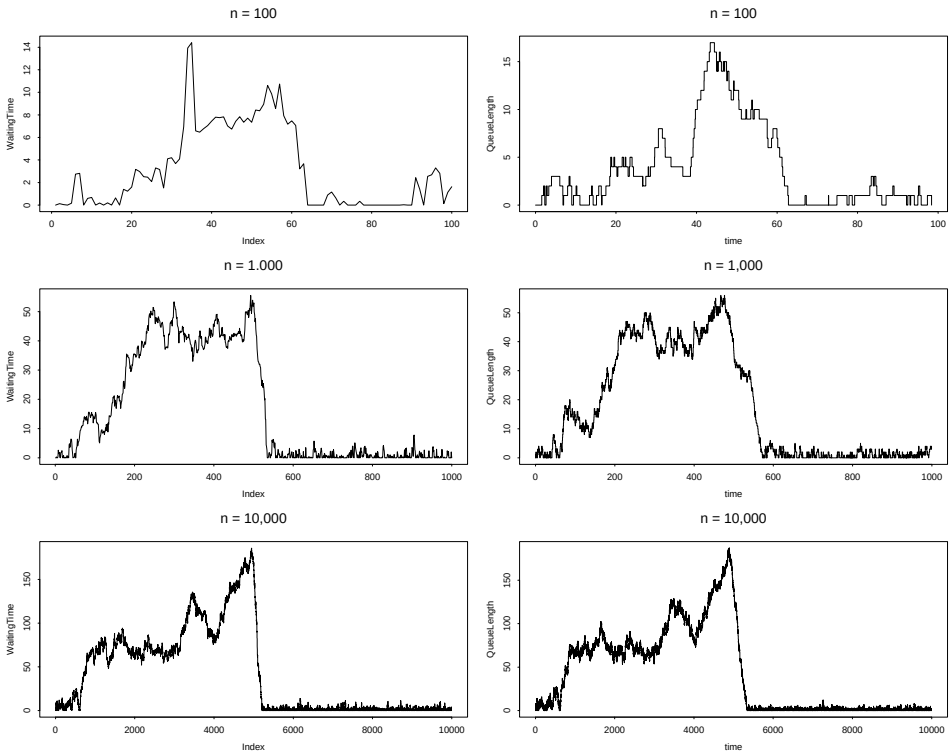



Figure 6.14: Plots of the waiting times of the first $n$ arrivals (on the left) and the queue-length process over the interval $[0, n]$ (on the right) in the $M/M/1$ queue with $\rho = 1$ for the first $n/2$ arrivals and $\rho = 1/2$ for the last $n/2$ arrivals for $n = 10^j$ with $j = 2, 3, 4$.

ambiguous for smaller $n$, but as $n$ increases, we see statistical regularity. As before, the scaled waiting-time and queue-length plots coalesce as $n$ increases. As $n$ increases, a sharp jump down is visible when the traffic intensity shifts from $\rho = 1$ to $\rho = 1/2$. As we indicated before, asymptotically, this shift for the scaled processes occurs at time $t = 1/2$.

Again, we are able to establish supporting FCLTs. Both the scaled waiting-time process and the scaled queue-length process are approaching reflecting Brownian motion over the subinterval $[0, t/2)$ and the 0 function over the subinterval $[t/2, 1]$. As in the previous examples, the scaled queue-length and waiting-time processes are asymptotically equivalent.

Thus, the limit process for the scaled queue-length process has an unmatched jump at $t = 1/2$. In this example, the limit for the waiting-time

process also has an unmatched jump at the same time.