Chapter 15

The Spaces E and F

15.1. Introduction

In this final chapter we introduce new spaces of functions larger than D. These new spaces of functions are intended to serve as spaces of sample paths for new stochastic processes that are limits for sequences of appropriately scaled stochastic processes that have significant fluctuations in two different time scales. We have in mind applications to queueing models of communication networks and manufacturing systems, but there are many other possible applications, e.g., to stochastic models of earthquake dynamics, cancer growth or stock prices.

Here is how this chapter is organized: To provide motivation for the new function spaces, we start in Section 15.2 by discussing three important time scales for the performance of queues: the performance time scale, the service time scale and the failure time scale. When the failure time scale falls between a shorter service time scale and a longer performance time scale, stochastic-process limits with scaling of space and time may provide insight into the impact of the failures upon performance. However, the failures in an intermediate time scale can lead to more complicated oscillations in buffer content, which require a new framework for the stochastic-process limits. We discuss these oscillations and their impact in Section 15.3.

In order to have stochastic-process limits with greater oscillations, we need functions spaces larger than D. In Section 15.4 we define the space E and specify conditions under which the Hausdorff metric inducing the analog of the M_2 topology on E is well defined. In Section 15.5 we develop alternative characterizations of convergence in (E, M_2). In Section 15.6 we give an example of convergence of stochastic processes in (E, M_2), where there is no convergence in D. The example is closely related to extreme-

value limits to extremal processes. We show how previous limits in that context can be viewed in a new way.

In Section 15.7 we define the space F and introduce the analog of the M_1 metric. We conclude in Section 15.8 by discussing queueing applications. We obtain a heavy-traffic stochastic-process limit describing a queue that experiences long rare failures, where the failures are more complicated than the service interruptions considered previously.

The present chapter is a brief introduction, identifying a direction for future research. The goal is to establish analogs for the spaces E and F, to the extent possible, of all the results for the space D in earlier chapters. There could even be another book!

15.2. Three Time Scales

In this section we discuss three important time scales for the performance of queues. A source of motivation for the queueing models is the desire to understand and control the performance of evolving communication networks. In some communication networks, such as Internet Protocol (IP) networks, there is evidence that performance degradation occurs, not only because of periods of exceptionally high user demand, but also because of various kinds of system failures (or by the combination of the two phenomena). One possible model of this phenomenon is a complex queueing system (network of queues) subject to occasional failures. This system alternates between periods of being "in control," and "out of control," where only some portion of the network may be experiencing difficulties when the system is out of control. The net-input process of packets into the network nodes (with buffers or queues) changes when the system changes state from in control to out of control, and so on. In this context, our goal is to obtain a suitable framework for establishing heavy-traffic stochastic-process limits for buffer-content stochastic processes that reveal the performance degradation caused by the system failures.

Similar problems arise in manufacturing systems. As above, a key factor in system performance is often system failures. We anticipate a customary randomness in the production process, e.g., associated with setups when changing a machine from one function to another, but occasional larger disruptions may be caused by system failures. A factory also alternates between periods of being "in control" and "out of control." The factory may be out of control when a critical machine breaks down or when a shipment of essential parts fails to arrive when scheduled. The system failures may

cause work-in-process inventories to temporarily build up to high levels at various work centers.

Both the communication network and the manufacturing system can be modeled as a network of queues, with the queues containing the bits or packets to be transmitted or the work-in-process that needs further processing. In that context, our goal is to capture the impact of system failures upon performance through appropriate stochastic-process limits. As an initial abstraction, consider a single queue. The system failures may cause significant modification in the input or the service capacity. For example, failure at that queue might cause an interruption of service. A substantial backlog can build up if input keeps arriving during the service interruption. Similarly, failures elsewhere might cause sudden decrease or increase of input. These failure modes can have a dramatic impact on system performance, if the failures cannot be corrected quickly. Obviously it is desirable to eliminate the failures whenever possible, but to effectively manage the system it is also important to understand and control the system when occasional failures do occur.

To understand the impact of system failures upon the performance of queues, it is useful to examine the relevant time scales. We draw attention to three different times scales for the performance of queues:

$$\begin{aligned} &\text{(i) the performance time scale} \\ &\text{(ii) the service time scale} \\ &\text{(iii) the failure time scale.} \end{aligned} \tag{2.1}$$

The *performance time scale* is the time scale of concern when judging system performance; the *service time scale* is the time scale of individual service times; the *failure time scale* is the time scale of failure durations.

We are primarily concerned with the case in which the performance time scale is much longer than the service time scale. For example, in communication networks, the performance time scale is often rooted in human perception. We often want to ensure satisfactory performance in the "human" time scale of seconds. In contrast, the service time scale corresponds to packet transmission times, which may be in milliseconds or less, depending upon the transmission rate. Similarly, in a make-to-order production facility, the performance time scale may be associated with product delivery times, i.e., the interval between a customer order and the product delivery, which might be measured in days or weeks. In contrast, the service time scale, determined by the production times on individual machines, might be measured in minutes.

Earthquake dynamics would seem to be a very different phenomenon, because the performance time scale is usually shorter than the "service" time scale: As before, our concern might be rooted in human perception, and thus may be measured in the time scale of seconds, minutes or days, whereas key factors in the system dynamics (which serve as analogs of the service times in a queue) may occur in the much longer time scale of months or years, or even longer.

When the time scale of interest in performance evaluation of a queueing system is much longer than the time scale of individual service times, it is likely that asymptotic analysis can provide useful insight. We may be able to usefully describe the macroscopic behavior of the system in the longer time scale of interest for performance by establishing heavy-traffic limits for stochastic processes after appropriately scaling space and time. Hopefully, the macroscopic description will capture the essential features of the microscopic model, while discarding inessential detail.

We want to include system failures in this framework. Assuming that the performance time scale is much greater than the service time scale, there are five possibilities for the failure time scale:

(i) failure time scale $>$ performance time scale

(ii) failure time scale $\approx$ performance time scale

(iii) performance time scale $>$ failure time scale $>$ service time scale

(iv) failure time scale $\approx$ service time scale

(v) failure time scale $<$ service time scale (2.2)

Here we are primarily concerned with the common case (iii), but we are also interested in cases (ii) and (iv). From a performance-analysis perspective, cases (i) and (v) are relatively trivial. In case (i), the failures totally dominate, in which case attention should be concentrated there. In case (v), failures tend to be have little consequence, so that they probably can be ignored.

In cases (ii), (iii) and (iv), it is worth carefully studying the impact of failures upon performance. In case (iv), the failures can be treated as random perturbations of the service times, so that the failures can usually be analyzed by making minor modifications of models that do not account for the failures.

We are particularly interested in case (iii) in (2.2). We then say that the failures occur in an *intermediate time scale*. When the failure time scale falls between a shorter service time scale and a longer performance time scale, it is useful to consider heavy-traffic stochastic-process limits with time scaling.

In such a limit, the failure durations may be asymptotically negligible in the performance time scale, while the overall performance impact of the failures may still be significant, being much greater than can be captured by inflating the mean and variance of service times in case (iv).

In fact, we have already considered examples illustrating failures occuring in an intermediate time scale: The examples were the queues with rare long service interruptions, discussed in Sections 6.5 and 14.7. In those models we assumed that the times between successive failures are in the long performance time scale, while the failure durations occur in an intermediate time scale, shorter than the performance time scale but longer than the service time scale. With appropriate definitions and scaling, heavy-traffic limits can be established in which failures occasionally occur in the (performance) time scale of the limit process. Since the failure durations are in the intermediate time scale, the failure durations are asymptotically negligible in the limit, so that failures occur instantaneously at single time points in the limit. However, the failures cannot be disregarded altogether. When the failure occurs, service stops while the input keeps arriving, so that a large backlog can quickly build up. With appropriate scaling, the failure leads to a sudden jump up in the limit process representing the asymptotic queue length or workload. The jump up represents the stronger consequence of the failure when the failure time scale is longer than the service time scale.

To be more explicit, we review the scaling that was used for the heavy-traffic stochastic-process limits for queues with rare long service interruptions. (See Section 5.5 for a general discussion of heavy-traffic scaling.) For the standard heavy-traffic limit for the single-server queue with unlimited waiting room, time is scaled by multiplying by n while space is scaled by dividing by $\sqrt{n}$, where n typically corresponds to $(1-\rho)^{-2}$, with ρ being the traffic intensity. When the traffic intensity is allowed to increase with n in this way, the queue length or waiting time unscaled tends to grow without bound. Since the steady-state means are of order $(1-\rho)^{-1}$ as $\rho \to 1$, the appropriate space scaling to obtain a nondegenerate limit is to divide by $\sqrt{n} = (1-\rho)^{-1}$. It turns out that the associated time scaling is to multiply by $n = (1-\rho)^{-2}$.

To capture the impact of failures in the way described, we let the time between failures be of order n and the failure durations be of order $\sqrt{n}$. In particular, assuming that a system failure corresponds to a service interruption, the queue buildup will be by the number of arrivals during the service interruption. Assuming that the service times are of order 1, the arrival rate is also order 1. Thus the queue buildup in a service interruption whose duration is of order $\sqrt{n}$ will also be of order $\sqrt{n}$. Hence the failures are

represented in a content limit process by occasional jumps up. In performance analysis, these jumps quantitatively describe the performance impact of the failures. The jumps reveal sudden performance degradation from the perspective of the long performance time scale.

To establish such limits with jumps for queues with service interruptions, we need to exploit the M_1 topology on D, because the limiting jump is approached gradually as a consequence of many arrivals during the service interruption. Thus the system failures serve as a major source of motivation for considering the function space D with the M_1 topology.

15.3. More Complicated Oscillations

Our purpose now is to go beyond the space (D, M_1). We want to be able to treat failures that cause more complicated oscillations in the stochastic process of interest. On D, the M_2 topology is useful because it allows the converging functions to have quite general fluctuations in the neighborhood of a limiting discontinuity.

Example 15.3.1. *The use of the M_2 topology on D.* Let the limit function be $x = I_{[1,2]}$ in $D([0,2], \mathbb{R})$. Then $x_n \to x$ in (D, M_2) as $n \to \infty$, but not for any of the other Skorohod topologies if

$$\begin{aligned} x_n(0) &= x_n(1 - 3n^{-1}) = 0, \\ x_n(1 - 2n^{-1}) &= 3/4,\ x_n(1 - n^{-1}) = 1/5, \\ x_n(1) &= 7/8,\ x_n(1 + n^{-1}) = 1/16, \\ x_n(1 + 2n^{-1}) &= x_n(2) = 1, \end{aligned} \tag{3.1}$$

with x_n defined by linear interpolation elsewhere. ∎

Example 15.3.1 shows the advantage of the space (D, M_2) over the space (D, M_1), but Example 15.3.1 also shows two shortcomings of the space (D, M_2): First, the nature of the fluctuations in x_n in the neighborhood of $t = 1$ are not evident from the limit function x; we only know that the process x_n is in the neighborhood of the interval $[0, 1]$ for $t \in (1 - \epsilon, 1 + \epsilon)$ for all sufficiently small ϵ. We might want functions to be defined so that the limit shows that x_n goes up from 0 to 3/4, then down to 1/5, then up to 7/8, then down to 1/16, and finally up to 1 in the neighborhood of time 1.

A second shortcoming of the space (D, M_2) is the requirement that all functions assume values at discontinuity points between the left and right

limits there. Of course, we actually have required more; we have required that the functions actually be right-continuous. However, the space D with the M_1 or M_2 topology is unchanged if we only require that the functions have left and right limits everywhere, provided that the function value at each discontinuity point falls between the left and right limit, because the completed graph is then the same as for the right-continuous version.

Now we want to allow for more general fluctuations. If functions x_n do have significant fluctuations in the neighborhood of some point t, then the functions x_n might well visit regions outside the interval $[x(t-) \wedge x(t+), x(t-) \vee x(t+)]$ in the neighborhood of t. And we might well want to say that x_n converges as $n \to \infty$ in that situation.

To illustrate, consider the following example.

Example 15.3.2. *The need for spaces larger than D.* With the same limit function $x = I_{[1,2]}$ in Example 15.3.1, the functions x_n might be defined, instead of by (3.1), by

$$\begin{aligned} x_n(0) &= x_n(1-3n^{-1}) = 0, \\ x_n(1-2n^{-1}) &= 2,\ x_n(1-n^{-1}) = -1, \\ x_n(1) &= 3,\ x_n(1+n^{-1}) = 1/16, \\ x_n(1+2n^{-1}) &= x_n(2) = 1, \end{aligned} \tag{3.2}$$

with x_n again defined by linear interpolation elsewhere. In this case, $x_n \not\to x$ in (D, M_2) as $n \to \infty$. However, we might want to say that x_n does actually converge to a limit as $n \to \infty$. It is natural that the graph of the limit be the graph Γ_x of x augmented by the vertical line segment $[-1,3] \times \{1\}$. Then the graph shows the range of points visited by x_n. In addition, we might want the limit to reflect the fact that x_n goes up from 0 to 2, then down to -1, then up to 3, then down to $1/16$, and finally up to 1 in the neighborhood of time 1. ∎

Example 15.3.3. *Another example requiring a larger space.* Another simple example is the case in which $x_n = I_{[1+n^{-1},1+2n^{-1})}$ in $D([0,2],\mathbb{R})$. This is the classic example in which x_n converges pointwise for all t to x with $x(t) = 0$, $0 \le t \le 2$, but x_n fails to converge in any of the Skorohod topologies. However, we might well want to say that x_n does in fact converge, and that it converges to a limit that captures the fluctuation experienced by x_n. Clearly, the graph of the limit should be the graph of x, i.e., $\{0\} \times [0,T]$, augmented by the vertical line segment $[0,1] \times \{1\}$. ∎

The first new space we introduce, the space E, allows for extra excursions at individual time points. We construct an element of $E([0,T],\mathbb{R})$ by augmenting a function x in $D([0,T],\mathbb{R})$ by adding vertical line segments to the graph of x at these specially designated excursion times. We do this in such a way that the graphs remain compact subsets of $\mathbb{R}^{k+1}$. Then, paralleling (D,M_2), we induce a topology (which we call the M_2 topology) on E by using the Hausdorff metric on the resulting set of graphs.

We also construct another space of functions, F, in order to capture more information about the fluctuations of the converging functions in the limit. One important source of motivation for the space F is our desire to apply reflection maps to establish heavy-traffic stochastic-process limits for queueing processes with limits in one of these larger spaces. Thus it is helpful to reconsider Example 13.5.2, which shows that the standard one-dimensional reflection map is not continuous on D with the M_2 topology.

Example 15.3.4. *The reflection map is not continuous on* (D,M_2). Recall that Example 13.5.2 shows that the one-sided one-dimensional reflection map $\phi : D \to D$ defined in Section 13.5 is not continuous with the M_2 topology. That example has $x = -I_{[1,2]}$ in $D([0,2],\mathbb{R})$,

$$x_n(0) = x_n(1-3n^{-1}) = x(1-n^{-1}) = 0$$

and

$$x_n(1-2n^{-1}) = x_n(1) = x_n(2) = -1$$

with x_n defined by linear interpolation elsewhere. Then $x_n \to x$ in (D,M_2), but $\phi(x_n) \not\to \phi(x)$ in (D,M_2), where $\phi(x)(t) = 0$ for all t. This example fails to provide a counterexample in (D,M_1) because then $x_n \not\to x$.

However, we might well want to have a topology allowing us to say that $x_n \to \hat{x}$ and $\phi(x_n) \to \hat{y}$, where $\hat{x}$ and $\hat{y}$ are different from x and $\phi(x)$ above, being elements of F instead of elements of D. Indeed, for this example, it is natural to say that $\phi(x_n)$ converges to a limit in (E,M_2), which coincides with the zero-function $\phi(x)$ augmented by the vertical line $[0,1]\times\{1\}$. Our new space allows us to reach that conclusion.

For that purpose, we want the initial limit $\hat{x}$ to reflect the parametric representations that can be used to justify M_2 convergence $x_n \to x$. In particular, when $r(s) = 1$, $u(s)$ goes from 1 to 0, to 1 and back to 0, in order to match the fluctuation in x_n. Assuming that we keep track of the fluctuation detail in $\hat{x}$, the associated limit $\phi(\hat{x})$ should have a graph equal to the zero-function $\phi(x)$ augmented by the vertical line $[0,1]\times\{1\}$. When we introduce another space with an appropriate topology, we will be able

to say that $x_n \to \hat{x}$. Then the reflection map will be continuous, so that we will have $\phi(x_n) \to \phi(\hat{x})$. Moreover, since this new topology will be stronger than the M_2 topology, we will be able to deduce that $\phi(x_n)$ converges to the graph of $\phi(\hat{x})$ in E, as desired. ∎

The need for a new space can also be explained by the fact that the reflection map is not even well defined on E. In order to define the reflection of an element of E, we need to know the order in which the points in a vertical segment are visited.

Example 15.3.5. *Maximal and minimal reflection maps on E.* To see that the reflection map is not well defined on E, consider the function $\tilde{x}$ in E obtained from $x = I_{[0,1)}$ in $D([0,2], \mathbb{R})$ with graph

$$\Gamma_x = (\{1\} \times [0,1)) \cup (\{0\} \times [1,2]) \tag{3.3}$$

augmented by the vertical line $[-2,3] \times \{1\}$, i.e., with graph

$$\Gamma_{\tilde{x}} = (\{1\} \times [0,1)) \cup ([-2,3] \times \{1\}) \cup (\{0\} \times [1,2]) \ . \tag{3.4}$$

A *maximal reflection map* ϕ^+ can be defined by assuming that, at time 1, $\tilde{x}$ first goes from 1 down to -2, then goes up to 3 and finally goes down to the right limit 0. The graph of $\phi^+(\tilde{x})$ is

$$\Gamma_{\psi^+(\tilde{x})} = (\{1\} \times [0,1)) \cup ([0,5] \times \{1\}) \cup (\{2\} \times (1,2]) \ . \tag{3.5}$$

On the other hand, a *minimal reflection map* ϕ^- can be defined by assuming that $\tilde{x}$ first goes up to 3, then down to -2 and then up to 1. The graph of $\phi^-(\tilde{x})$ is

$$\Gamma_{\psi^-(\tilde{x})} = (\{1\} \times [0,1)) \cup ([0,3] \times \{1\}) \cup (\{2\} \times (1,2]) \ . \tag{3.6}$$

Note that the maximum value of $\phi^+(\tilde{x})$ is 5, while the maximum value of $\phi^-(\tilde{x})$ is 3. Thus, to be well defined, the reflection map needs to know the order of the points visited during the excursions. ∎

To faithfully represent the fluctuations at excursions we introduce a new space F based on parametric representations. In particular, we consider parametric representations of the graphs of the elements of E. We say that two parametric representations of $\Gamma_{\tilde{x}}$ for $\tilde{x} \in E$ are equivalent if they visit the same points in the same order. We let F be the space of equivalence classes with respect to that equivalence relation. The space F is larger than

E because there are many different elements of F with the same graph. We give F an analog of the M_1 metric. The space (D, M_1) itself is homeomorphic to a subset of F, in which the graphs correspond to functions in D and the parametric representations are monotone in the order put on the graphs in Sections 3.3 and 12.3.

The spaces (E, M_2) and (F, M_1) allow us to obtain a satisfactory treatment of reflection. First, working with F instead of E, we avoid the ambiguity about the order points are visited in Example 15.3.5. In particular, the reflection map ϕ is well defined and continuous on (F, M_1). Thus, starting from convergence $\hat{x}_n \to \hat{x}$ in (F, M_1), we obtain $\phi(\hat{x}_n) \to \phi(\hat{x})$in (F, M_1). Since convergence in (F, M_1) implies convergence in (E, M_2) for the graphs of the elements of F, we obtain as a consequence associated convergence in (E, M_2) for the graphs of $\phi(\hat{x}_n)$ and $\phi(\hat{x})$. In particular, we obtain the proposed limit in (E, M_2) in Example 15.3.4.

The example above is for the one-dimensional reflection map. We can also treat the multidimensional reflection map. Paralleling Chapter 14, we need to make assumptions about the domain and the range: First for the domain, it suffices to assume that the limit $\hat{x}$ belongs to the space F_1, the subset of F in which the excursions and discontinuities occur only in one coordinate at a time. Next for the range, we need the product topology: First, we can use the space (F, WM_1); then we can go to (E, WM_2). In other words, convergence of functions in the domain (F, WM_1), where the limit belongs to F_1 will imply convergence of the graphs of the reflections in the range (E, WM_2).

In summary, we introduce new function spaces E and F to supplement the familiar ones C and D considered before. Hence, we have four spaces of functions, each larger than the one before:

C – *Continuous* functions
D – *Discontinuous* functions
E – functions with extra *Excursions*
F – functions with extra excursions that faithfully model *Fluctuations*

We omit most proofs in this chapter, which are similar to proofs for (D, M_2) and (D, M_1). A more extensive discussion of the spaces E and F is planned for the Internet Supplement.

15.4. The Space E

Our general approach to defining the new function spaces E and F is to base the definitions on the M_2 and M_1 topologies used on D. Now we use the graphs and parametric representations, not only to define the topologies, but also to define the functions themselves.

The space E is larger than D because the functions are allowed to have extra excursions, which occur at single time points. We initially represent $E \equiv E([0,T],\mathbb{R}^k)$ as the space of *excursion triples*

$$(x, S, \{I(t) : t \in S\}) \ , \tag{4.1}$$

where $x \in D \equiv D([0,T],\mathbb{R}^k)$, the space of all right-continuous $\mathbb{R}^k$-valued functions on $[0,T]$ with left limits everywhere, S is a countable set with

$$Disc(x) \subseteq S \subseteq [0,T] \ , \tag{4.2}$$

and, for each $t \in S$, $I(t)$ is a compact subset of $\mathbb{R}^k$ with at least two points such that

$$x(t), x(t-) \in I(t) \quad \text{for all} \quad t \in S \ . \tag{4.3}$$

We call the function x in D the *base function*. We call S the set of *excursion times* or the set of *discontinuity points* of the new function. We want to allow the new function to make excursions where the base function x is continuous, so $Disc(x)$ may be a proper subset of S. We call the set $I(t)$ the *set of excursion values*; $I(t)$ is the set of values assumed by the new function at time t for $t \in S$. By requiring that $I(t)$ contain at least two points, we ensure that $\{x(t)\} \overset{\subset}{\neq} I(t)$ when $t \in Disc(x)^c$. Hence, each $t \in S$ corresponds to some genuine form of discontinuity.

Associated with the excursion triple $(x, S, \{I(t), t \in S\})$ is the set-valued function

$$\tilde{x}(t) \equiv \begin{cases} I(t), & t \in S \\ \{x(t)\}, & t \notin S \ . \end{cases} \tag{4.4}$$

Associated with the set-valued function $\tilde{x}$ is its graph

$$\Gamma_{\tilde{x}} \equiv \{(z,t) \in \mathbb{R}^k \times [0,T] : z \in \tilde{x}(t)\} \ . \tag{4.5}$$

Clearly, it is possible to construct the excursion triple $(x, S, \{I(t), t \in S\})$, the set-valued function $\tilde{x}$ and the graph $\Gamma_{\tilde{x}}$, starting from any one of the three. Hence these are three equivalent representations for the elements of the space E. For brevity, we will let elements of E be denoted by $\tilde{x}$.

Unlike the graphs Γ_x and G_x for $x \in D$ defined in Section 12.3, the set $I(t)$ appearing in the graph $\Gamma_{\tilde{x}}$ in (4.5) need not be a segment. Indeed it need not even be connected. We will impose further restrictions below.

Paralleling our treatment of (D, M_2), we propose making E a separable metric space by using the Hausdorff metric on the space of graphs $\Gamma_{\tilde{x}}$ for $\tilde{x} \in E$. For that purpose, we want to be sure that each graph is a compact subset of $\mathbb{R}^{k+1}$. Without additional assumptions, we need not have that property.

Example 15.4.1. *The graphs need not be compact.* To see the need for additional assumptions in order to guarantee that $\Gamma_{\tilde{x}}$ in (4.5) is compact, consider the triple $\{x, S, \{I(t) : t \in S\}\}$ with $x(t) = 0$, $0 \le t \le 1$, S the set of rational numbers strictly less than $1/2$, and $I(t) = [0, 1]$ for all $t \in S$. Then $\Gamma_{\tilde{x}}$ is a dense subset of $[0, 1] \times [0, 1/2)$, which is bounded but not closed, and thus not compact. ∎

A simple way to ensure that $\Gamma_{\tilde{x}}$ is compact is to require that the set S of excursion times be a finite subset of $[0, T]$, and we think of that being an important case for applications. However, that assumption is unappealing because then D would no longer be a proper subset of E. Thus we want to allow countable sets of excursion times. We can allow countable subsets if we impose a restriction. We give equivalent characterizations of the condition below.

When we talk about convergence of sets, the sets will always be compact subsets of $\mathbb{R}^k$, and we use the Hausdorff metric, as defined in (5.2) in Section 11.5; i.e., for compact sets A_1, A_2,

$$m(A_1, A_2) \equiv \mu(A_1, A_2) \vee \mu(A_2, A_1) \tag{4.6}$$

where

$$\mu(A_1, A_2) \equiv \sup_{x \in A_1} \{\|x - A_2\|\} \tag{4.7}$$

and

$$\|x - A\| = \|A - x\| \equiv \inf_{y \in A} \{\|x - y\|\} \ . \tag{4.8}$$

For $A \subseteq \mathbb{R}^k$, let $\delta(A)$ be the *diameter* of A, i.e.,

$$\delta(A) \equiv \sup_{x, y \in A} \{\|x - y\|\} \ . \tag{4.9}$$

Theorem 15.4.1. (equivalent conditions) *The following conditions for elements of E are equivalent:*

(a) For each $\epsilon > 0$, there are only finitely many t for which $\delta(I(t)) > \epsilon$.

(b) For each $t \in [0,T)$, $\tilde{x}$ has a single-point right limit

$$\tilde{x}(t+) \equiv \lim_{s \downarrow t} \tilde{x}(s) = \{x(t)\} \tag{4.10}$$

and, for each $(0,T]$, $\tilde{x}$ has a single-point left limit

$$\tilde{x}(t-) \equiv \lim_{s \uparrow t} \tilde{x}(s) = \{x(t-)\} \ . \tag{4.11}$$

Theorem 15.4.2. (conditions to make the graphs compact) *Under the conditions in Theorem* 15.4.1*, the graph $\Gamma_{\tilde{x}}$ in* (4.5) *is a compact subset of $\mathbb{R}^{k+1}$.*

Henceforth let E be the set of graphs $\Gamma_{\tilde{x}}$ for which the conditions of Theorem 15.4.1 hold. Then, paralleling (D, M_2), we endow E with the Hausdorff metric, i.e.,

$$m(\tilde{x}_1, \tilde{x}_2) \equiv m(\Gamma_{\tilde{x}_1}, \Gamma_{\tilde{x}_2}) \ , \tag{4.12}$$

where m is the Hausdorff metric on the space of compact subsets of $\mathbb{R}^{k+1}$, as in (4.6). We call the topology induced by m on E the M_2 topology. Since the graphs are compact subsets of $\mathbb{R}^{k+1}$, we have the following result.

Theorem 15.4.3. (metric property) *The space (E, m) is a separable metric space.*

So far, we have defined the metric m on the space $E([0,T], \mathbb{R}^k)$ with the compact domain $[0,T]$. We extend to non-compact domains just as was done for D. We say that $\tilde{x}_n \to \tilde{x}$ in $E(I, \mathbb{R}^k)$ for an interval I in $\mathbb{R}$ if $\tilde{x}_n \to \tilde{x}$ for the restrictions in $E([t_1, t_2], \mathbb{R}^k)$ for all $t_1, t_2 \in I$ with $t_1 < t_2$ and $t_1, t_2 \notin S$.

We also define a stronger metric than the Hausdorff metric m in (4.12) on $E \equiv E([0,T], \mathbb{R}^k)$, which we call the *uniform metric*, namely,

$$m^*(\tilde{x}_1, \tilde{x}_2) = \sup_{0 \le t \le T} m(\tilde{x}_1(t), \tilde{x}_2(t)) \ , \tag{4.13}$$

where m is again the Hausdorff metric, here applied to compact subsets of $\mathbb{R}^k$. The following comparison is not difficult.

Theorem 15.4.4. (comparison with the uniform metric) *For any* $\tilde{x}_1, \tilde{x}_2 \in E$,

$$m(\tilde{x}_1, \tilde{x}_2) \leq m^*(\tilde{x}_1, \tilde{x}_2)$$

for m *in* (4.12) *and* m^* *in* (4.13).

As with the metrics inducing the U and M_2 topologies on D, m^* is complete but not separable, while m is separable but not complete.

Example 15.4.2. *The metric* m *on* E *is not complete.* To see that the Hausdorff metric m on E is not complete, let

$$x_n = \sum_{k=1}^{n-1} I_{[2k/2n,(2k+1)/2n)}, \quad n \geq 2 \ . \tag{4.14}$$

Then $x_n \in D([0,1], \mathbb{R})$,

$$m(\Gamma_{x_n}, \Gamma_{x_m}) \to 0 \quad \text{as} \quad n, m \to \infty \tag{4.15}$$

$$m(\Gamma_{x_n}, [0,1] \times [0,1]) \to 0 \quad \text{as} \quad n \to \infty \ , \tag{4.16}$$

but the limit is not in E. Hence x_n does not converge to a limit in E. ∎

We now consider approximations of functions in E by piecewise-constant functions. Let E_c be the subset of *piecewise-constant functions* in E, i.e., the set of functions for which S is finite, x is constant between successive points in S and $I(t)$ is finite valued for all $t \in S$. Rational-valued piecewise-constant functions with $S = \{jT/k, 1 \leq j \leq k\}$, which are elements of E_c, form a countable dense subset of (E, m). The following result parallels Theorem 12.2.2.

Theorem 15.4.5. (approximation by piecewise-constant functions) *If* $\tilde{x} \in (E, m)$, *then for all* $\epsilon > 0$ *there exists* $\tilde{x}_c \in E_c$ *such that* $m^*(\tilde{x}, \tilde{x}_c) < \epsilon$ *for* m^* *in* (4.13).

Even though the excursion sets $I(t)$ associated with the functions $\tilde{x}$ in E must be compact because of the conditions imposed in Theorem 15.4.1, so far they can be very general. In particular, the excursion sets $I(t)$ need not be connected. It may well be of interest to consider the space E with disconnected excursions, which the framework above allows, but for the applications we have in mind, we want to consider graphs that are connected sets. We will make the stronger assumption that the graphs are *path-connected*,

i.e., each pair of points can be joined by a path – a continuous map from $[0,1]$ into the graph; e.g., see Section V.5 of Dugundji (1966). In fact, we will assume that the graph can be represented as the image of parametric representations.

We say that $(u,r):[0,1]\to\mathbb{R}^{k+1}$ is a *strong parametric representation* of $\Gamma_{\tilde{x}}$ or

$\tilde{x}$ in E if (u,r) is a continuous function from $[0,1]$ *onto* $\Gamma_{\tilde{x}}$ such that r is nondecreasing. We say that $\tilde{x}$ and $\Gamma_{\tilde{x}}$ are *strongly connected* if there exists a strong parametric representation of $\Gamma_{\tilde{x}}$.

Similarly, we say that (u,r) is a *weak parametric representation* of $\Gamma_{\tilde{x}}$ or $\tilde{x}$ in E, (u,r) is a continuous function from $[0,1]$ into $\Gamma_{\tilde{x}}$ such that r is nondecreasing, $r(0)\in\tilde{x}(0)$ and $r(1)\in\tilde{x}(T)$. We say that $\tilde{x}$ and Γ_x are *weakly connected* if there exists a weak parametric representation of Γ_x and if the union of $(u(s),r(s))$ over all s, $0\le s\le 1$, and all weak parametric representations of $\Gamma_{\tilde{x}}$ is $\Gamma_{\tilde{x}}$ itself.

Let E_{st} and E_{wk} represent the subsets of strongly connected and weakly connected functions $\tilde{x}$ in E. When the range of the functions is $\mathbb{R}$, $E_{st}=E_{wk}$ and E_{st} is a subset of E in which $I(t)$ is a closed bounded interval for each $t\in S$.

As in (3.1) and (3.2) in Section 12.3, for $a,b\in\mathbb{R}^k$, let $[a,b]$ and $[[a,b]]$ be the standard and product segments in $\mathbb{R}^k$. It is easy to identify (D,SM_2) and (D,WM_2) as subsets of (E_{st},m) and (E_{wk},m), respectively.

Theorem 15.4.6. (when $\tilde{x}$ reduces to x) *Consider an element $\tilde{x}$ in E. Suppose that $S=Disc(x)$.*

(a) If $\tilde{x}\in E_{st}$ and

$$I(t)=[x(t-),x(t)] \quad \textit{for all} \quad t\in S\,, \tag{4.17}$$

then

$$\Gamma_{\tilde{x}}=\Gamma_x \tag{4.18}$$

for the thin graph Γ_x in (3.3) *of Section* 12.3.

(b) If $\tilde{x}\in E_{wk}$ and

$$I(t)=[[x(t-),x(t)]] \quad \textit{for all} \quad t\in S\,, \tag{4.19}$$

then

$$\Gamma_{\tilde{x}}=G_x \tag{4.20}$$

for the thick graph G_x in (3.4) *Section* 12.3.

Let D_{st} and D_{wk} be the subsets of all $\tilde{x}$ in E_{st} and E_{wk}, respectively, satisfying the conditions of Theorem 15.4.6 (a) and (b). Since the SM_2 and WM_2 topologies on D can be defined by the Hausdorff metric on the graphs Γ_x and G_x, the following corollary is immediate.

Corollary 15.4.1. (identifying the spaces (D, SM_2) and (D, WM_2)) *The space (D, SM_2) is homeomorphic to the subset D_{st} in (E_{st}, m), while the space (D, WM_2) is homeomorphic to the subset D_{wk} in (E_{wk}, m).*

15.5. Characterizations of M_2 Convergence in E

Paralleling Section 12.11, we now want to develop alternative characterizations of M_2 convergence on E. For simplicity, we consider only real-valued functions. Thus there is no need to distinguish between the SM_2 and WM_2 topologies.

We consider M_2 convergence on $E \equiv E_{st} \equiv E_{st}([0,T], \mathbb{R})$, assuming that the conditions of Theorem 15.4.1 hold. By considering E_{st}, we are assuming that the functions are strongly connected (i.e., there exist parametric representations onto the graphs.) Thus the excursion sets $I(t)$ for t in the set S of excursion times are all closed bounded intervals.

Given that the functions $\tilde{x}$ in E are all strongly connected, it is natural to consider characterizing M_2 convergence in terms of parametric representations. For $\tilde{x}_1$, $\tilde{x}_2 \in E$, let

$$d_{s,2}(\tilde{x}_1, \tilde{x}_2) = \inf_{\substack{(u_i, r_i) \in \Pi_{s,2}(\tilde{x}_i) \\ i=1,2}} \{\|u_1 - u_2\| \vee \|r_1 - r_2\|\} , \tag{5.1}$$

where $\Pi_{s,2}(\tilde{x}_i)$ is the set of all strong parametric representations (u_i, r_i) of $\tilde{x}_i \in E$. As in Section 12.11, it turns out that $d_{s,2}(\tilde{x}_n, \tilde{x}) \to 0$ if and only if $m(\tilde{x}_n, \tilde{x}) \to 0$. (We will state a general equivalence theorem below.) However, as before, $d_{s,2}$ in (5.1) is not a metric on E. That is shown by Example 12.11.1.

We next introduce a characterization corresponding to local uniform convergence of the set functions. For this purpose, let the δ-neighborhood of $\tilde{x}$ at t be

$$N_\delta(\tilde{x})(t) = \cup_{0 \vee (t-\delta) \le s \le (t+\delta) \wedge T} \tilde{x}(s) . \tag{5.2}$$

(Since N_δ includes only perturbations horizontally, it is not actually the δ neighborhood in the metric m.)

Lemma 15.5.1. (compactness) *For each $\tilde{x} \in E$, $t \in [0,T]$ and $\delta > 0$, $N_\delta(\tilde{x})(t)$ in (5.2) is a compact subset of $\mathbb{R}^k$.*

Let

$$v(\tilde{x}_1, \tilde{x}_2, t, \delta) \equiv m(N_\delta(\tilde{x}_1)(t), \tilde{x}_2(t)) \tag{5.3}$$

where m is the Hausdorff metric in (4.6), and

$$v(\tilde{x}_1, \tilde{x}_2, \delta) \equiv \sup_{0 \le t \le T} v(\tilde{x}_1, \tilde{x}_2, t, \delta) \ . \tag{5.4}$$

An attractive feature of the function space $D \equiv D([0,T], \mathbb{R}^k)$ not shared by the larger space E is that all function values $x(t)$ for $x \in D$ are determined by the function values $x(t_k)$ for any countable dense subset $\{t_k\}$ in $[0,T]$. For $t \in [0,T)$, by the right continuity of x, $x(t) = \lim x(t_k)$ for $t_k \downarrow t$. In contrast, in E since we do not necessarily have $Disc(x) = S$, we cannot even identify the set S of excursion times from function values $\tilde{x}(t)$ for t outside S. In some settings it may be reasonable to assume that $S = Disc(x)$, in which case S can be identified from the left and right limits. For instance, in stochastic settings we might have $P(X(t-) = X(t)) = 0$ when $t \in S$, so that it may be reasonable to assume that $S = Disc(x)$.

However, even when $S = Disc(x)$, we are unable to discover the set $\tilde{x}(t)$ for $t \in S$ by observing the set-valued function $\tilde{x}$ at other time points t. Hence, in general *stochastic processes with sample paths in E are not separabile*; e.g., see p. 65 of Billingsley (1968). Thus it is natural to look for alternative representations for the functions $\tilde{x}$ that are determined by function values on any countable dense subset.

Thus, for $\tilde{x} \in E$, let the local-maximum function be defined by

$$M_{t_1,t_2}(\tilde{x}) = \sup\{z : z \in \tilde{x}(t) : t_1 \le t \le t_2\} \tag{5.5}$$

for $0 \le t_1 < t_2 \le T$. It is significant that we can also go the other way. For $\tilde{x} \in E$, if we are given $M_{t_1,t_2}(\tilde{x})$ and $M_{t_1,t_2}(-\tilde{x})$ for all t_1, t_2 in a countable dense subset A of $[0,T]$, we can reconstruct $\tilde{x}$. In particular, for $0 < t < T$,

$$\tilde{x}(t) = [a(t), b(t)] \ , \tag{5.6}$$

with $[a,a] \equiv \{a\}$, where

$$b(t) = \lim_{\substack{t_{1,k} \uparrow t \\ t_{2,k} \downarrow t}} M_{t_{1,k},t_{2,k}}(\tilde{x}) \tag{5.7}$$

and

$$-a(t) = \lim_{\substack{t_{1,k} \uparrow t \\ t_{2,k} \downarrow t}} M_{t_{1,k},t_{2,k}}(-\tilde{x}) \tag{5.8}$$

with $t_{1,k}$ and $t_{2,k}$ taken from the countable dense subset A. For $t = 0$,

$$b(t) = \lim_{t_{2,k} \downarrow t} M_{0,t_{2,k}}(\tilde{x}) \tag{5.9}$$

and

$$-a(t) = \lim_{t_{2,k} \downarrow t} M_{0,t_{2,k}}(-\tilde{x}) \ , \tag{5.10}$$

where $t_{2,k}$ is again taken from A. A similar construction holds for $t = T$. It is significant that M_2 convergence in E is also determined by the maximum function.

We are now ready to state our convergence-characterization theorem. It is an analog of Theorem 12.11.1 for (D, SM_2).

Theorem 15.5.1. (alternative characterizations of M_2 convergence in E) *The following are equivalent characterizations of convergence $\tilde{x}_n \to \tilde{x}$ in $E([0,T], \mathbb{R})$:*

(i) $m(\tilde{x}_n, \tilde{x}) \to 0$ for the metric m in (4.12) and (4.6).

(ii) $\mu(\Gamma_{\tilde{x}_n}, \Gamma_{\tilde{x}}) \to 0$ for μ in (4.7).

(iii) $d_{s,2}(\tilde{x}_n, \tilde{x}) \to 0$ for $d_{s,2}$ in (5.1); i.e. for any $\epsilon > 0$ and n sufficiently large, there exist $(u, r) \in \Pi_{s,2}(\tilde{x})$ and $(u_n, r_n) \in \Pi_{s,2}(\tilde{x}_n)$ such that $\|u_n - u\| \vee \|r_n - r\| < \epsilon$.

(iv) Given $v(\tilde{x}_1, \tilde{x}_2, \delta)$ in (5.4),

$$\lim_{\delta \downarrow 0} v(\tilde{x}_n, \tilde{x}, \delta) = 0 \ . \tag{5.11}$$

(v) For each t, $0 \le t \le T$,

$$\lim_{\delta \downarrow 0} \overline{\lim_{n \to \infty}} \ v(\tilde{x}_n, \tilde{x}, t, \delta) = 0 \tag{5.12}$$

for $v(\tilde{x}_1, \tilde{x}_2, t, \delta)$ in (5.3).

(vi) For all t_1, t_2 in a countable dense subset of $[0,T]$ with $t_1 < t_2$, including 0 and T,

$$M_{t_1,t_2}(\tilde{x}_n) \to M_{t_1,t_2}(\tilde{x}) \quad \text{in} \quad \mathbb{R} \tag{5.13}$$

and

$$M_{t_1,t_2}(-\tilde{x}_n) \to M_{t_1,t_2}(-\tilde{x}) \quad \text{in} \quad \mathbb{R} \tag{5.14}$$

for the local maximum function M_{t_1,t_2} in (5.5).

For $\tilde{x} \in E$, let $S'(\tilde{x})$ be the subset of non-jump discontinuities in $S(\tilde{x})$, i.e.,

$$S'(\tilde{x}) \equiv \{t \in S(\tilde{x}) : I(t) \neq [x(t-), x(t+)]\} . \tag{5.15}$$

We envision that in many applications $S'(\tilde{x})$ will be a finite subset and x_n will belong to D for all n. A more elementary characterization of convergence holds in that special case.

Theorem 15.5.2. (when there are only finitely many non-jump discontinuities) *Suppose that $\tilde{x} \in E$, $S'(\tilde{x})$ in (5.15) is a finite subset $\{t_1, \ldots, t_k\}$ and $x_n \in D$ for all n. Then $x_n \to \tilde{x}$ in (E, M_2) holds if and only if*

(i) $x_n \to x$ in (D, M_2) for the restrictions over each of the subintervals $[0, t_1)$, $(t_1, t_2), \ldots, (t_{k-1}, t_k)$ and $(t_k, T]$, where x is the base function of $\tilde{x}$, and

(ii) (5.12) holds for $t = t_i$ for each $t_i \in S'(\tilde{x})$.

15.6. Convergence to Extremal Processes

In this section we give a relatively simple example of a stochastic-process limit in which random elements of D converge in E to a limiting stochastic-process whose sample paths are in E but not in D.

Let $\{X_n : n \geq 1\}$ be a sequence of real-valued random variables and let $\{S_n : n \geq 0\}$ be the associated sequence of partial sums, i.e.,

$$S_n = X_1 + \cdots + X_n, \quad n \geq 1 , \tag{6.1}$$

with $S_0 = 0$. Let $\mathbf{S}_n$ and $\mathbf{X}_n$ be associated scaled random elements of $D \equiv D([0, T], \mathbb{R})$ defined by

$$\begin{aligned} \mathbf{S}_n(t) &= c_n^{-1}(S_{\lfloor nt \rfloor} - \mu nt) \\ \mathbf{X}_n(t) &= c_n^{-1} X_{\lfloor nt \rfloor}, \quad 0 \leq t \leq T , \end{aligned} \tag{6.2}$$

where $c_n \to \infty$ as $n \to \infty$. Since the maximum jump functional is continuous at continuous functions, p. 301 of Jacod and Shiryaev (1987), we have $\mathbf{X}_n \Rightarrow 0\mathbf{e}$ in D, where $\mathbf{e}$ is the identity function, or equivalently $\|\mathbf{X}_n\| \Rightarrow 0$, whenever $\mathbf{S}_n \Rightarrow \mathbf{S}$ in D and $P(\mathbf{S} \in C) = 1$.

However, we cannot conclude that $\|\mathbf{X}_n\| \Rightarrow 0$ when the limit $\mathbf{S}$ fails to have continuous sample paths. In this section we show that we can have $\mathbf{X}_n \Rightarrow \mathbf{X}$ in $E \equiv E([0, T], \mathbb{R})$ in that situation, where $\mathbf{X}$ is a random element of E and not a random element of D.

We establish our result for the case in which $\{X_n : n \geq 1\}$ is a sequence of IID nonnegative random variables with cdf F, where the complementary

cdf $F^c \equiv 1 - F$ is regularly varying with index $-\alpha$, i.e., $F^c \in \mathcal{R}(-\alpha)$, with $\alpha < 2$. That is known to be a necessary and sufficient condition for F to belong to both the domain of attraction of a non-normal stable law of index α and the maximum domain of attraction of the Frechet extreme value distribution with index α; see Section 4.5, especially Theorem 4.5.4.

We will use extremal processes to characterize the limit of $\mathbf{X}_n$ in E. Hence we now briefly describe extremal processes; see Resnick (1987) for more details. We can construct an extremal process associated with any cdf F on $\mathbb{R}$. Given the cdf F, we define the finite-dimensional distributions of the extremal process by

$$F_{t_1,\ldots,t_k}(x_1,\ldots,x_k) \equiv F^{t_1}\left(\bigwedge_{i=1}^{k} x_i\right) F^{t_2-t_1}\left(\bigwedge_{i=2}^{k} x_i\right)\cdots F^{t_k-t_{k-1}}(x_k)\,, \quad (6.3)$$

where

$$\bigwedge_{i=j}^{k} x_i \equiv \min\{x_i : j \le i \le k\}\ . \quad (6.4)$$

We are motivated to consider definition (6.3) because the first n successive maxima

$$\{M_k : 1 \le k \le n\}$$

have cdf $F_{1,2,\ldots,n}$; e.g.,

$$P(M_k \le x_k, M_n \le x_n) = F^k(x_k \wedge x_n)F^{n-k}(x_n)\ . \quad (6.5)$$

It is easy to see that the finite-dimensional distributions in (6.3) are consistent, so that there is a stochastic process $Y \equiv \{Y(t) : t > 0\}$, with those finite-dimensional distributions. Moreover, there is a version in D. We summarize the basic properties in the following theorem; see Resnick (1987) for a proof.

Theorem 15.6.1. (characterization of the extremal process asociated with the cdf F) *For any cdf F on $\mathbb{R}_+$, there is a stochastic process $Y \equiv \{Y(t) : t \ge 0\}$, called the extremal process associated with F, with sample paths in $D((0,\infty),\mathbb{R},J_1)$ such that*

(i) $P(t \in Disc(Y)) = 0$ *for all* $t > 0$,

(ii) Y *has nondecreasing sample paths,*

(iii) Y *is a jump Markov process with* $P(Y(t+s) \le x | Y(s) = y) = \begin{cases} F^t(\alpha), & x \ge y \\ 0, & x < y, \end{cases}$

(iv) the parameter of the exponential holding time in state x is $Q(x) \equiv -\log F(x)$; given that a jump occurs in state x, the process jumps from x to $(-\infty, y]$ with probability $1 - Q(y)/Q(x)$ if $y > x$ and 0 otherwise.

Having defined and characterized extremal processes, we can now state our limit in E.

Theorem 15.6.2. (stochastic-process limit with limit process in E) *If $\{X_n : n \geq 1\}$ is a sequence of IID nonnegative real-valued random variables with cdf $F \in \mathcal{R}(-\alpha)$, then*

$$\mathbf{X}_n \Rightarrow \mathbf{X} \quad in \quad E((0,\infty), \mathbb{R})$$

for $\mathbf{X}_n$ in (6.2) *and*

$$c_n \equiv (1/F^c)^{\leftarrow}(n) \equiv \inf\{s : (1/F^c)(s) \geq n\} = \inf\{s : F^c(s) \leq n^{-1}\} \ , \quad (6.6)$$

where $\mathbf{X}$ is characterized by the maxima $M_{t_1,t_2}(\mathbf{X})$ for $0 < t_1 < t_2$. These maxima satisfy the properties:

(i) for each $k \geq 2$ and k disjoint intervals $(t_1, t_2), (t_3, t_4), \ldots, (t_{2k-1}, t_{2k})$, the random variables $M_{t_1,t_2}(\mathbf{X})$, $M_{t_3,t_4}(\mathbf{X}), \ldots, M_{t_{2k-1},t_{2k}}(\mathbf{X})$ are mutually independent, and

(ii) for each t_1, t_2 with $0 < t_1 < t_2$,

$$M_{t_1,t_2}(\mathbf{X}) \stackrel{\mathrm{d}}{=} Y(t_2 - t_1) \ , \quad (6.7)$$

where Y is the extremal process associated with the Frechet extreme-value cdf in (5.34) *in Section* 4.5.

Proof. To carry out the proof we exploit convergence of random point measures to a Poisson random measure, as in Chapter 4 of Resnick (1987). We will briefly outline the construction. We use random point measures on $\mathbb{R}^2$. For $a \in \mathbb{R}^2$, let ϵ_a be the measure on $\mathbb{R}^2$ with

$$\epsilon_a(A) = \begin{cases} 1, & a \in A \\ 0, & a \notin A \ . \end{cases} \quad (6.8)$$

A *point measure* on $\mathbb{R}^2$ is a measure μ of the form

$$\mu \equiv \sum_{i=1}^{\infty} \epsilon_{a_i} \ , \quad (6.9)$$

where $\{a_i : i \geq 1\}$ is a sequence of points in $\mathbb{R}^2$ and

$$\mu(K) < \infty \tag{6.10}$$

for each compact subset K of $\mathbb{R}^2$. Let $M_p(\mathbb{R}^2)$ be the space of point measures on $\mathbb{R}^2$. We say that μ_n *converges vaguely* to μ in $M_p(\mathbb{R}^2)$ and write $\mu_n \to \mu$ if

$$\int f d\mu_n \to \int f d\mu \quad \text{as} \quad n \to \infty \tag{6.11}$$

for all nonnegative continuous real-valued functions with compact support. It turns out that $M_p(\mathbb{R}^2)$ with vague convergence is metrizable as a complete separable metric space; see p. 147 of Resnick (1987).

We will consider random point measures, i.e., probability measures on the space $M_p(\mathbb{R}^2)$. A Poisson random measure N with mean measure ν is a random point measure with the properties

(i) $N(A_1)$, $N(A_2), \ldots, N(A_k)$ are independent random variables for all k and all disjoint measurable subsets A_1, $A_2, \ldots, A_k$ of $\mathbb{R}^2$

(ii) for each measurable subset A of $\mathbb{R}^k$,

$$P(N(A) = k) = e^{-\nu(A)} \frac{\nu(A)^k}{k!} \tag{6.12}$$

where ν is a measure on $\mathbb{R}^2$. In our context, the random point measure limit associated with the sequence $\{X_k : k \geq 1\}$ is

$$\sum_{k=1}^{\infty} \epsilon_{\{k/n, X_k/c_n\}} \Rightarrow N \tag{6.13}$$

where c_n is again as in (6.6) and N is a Poisson random measure with mean measure ν determined by

$$\nu(A \times [x, \infty)) = \lambda(A) x^{-\alpha}, \quad x > 0 \ , \tag{6.14}$$

where λ is Lebesgue measure; the proof of convergence is shown in Resnick (1987).

We now use the Skorohod representation theorem to replace the convergence in distribution in (6.13) by convergence w.p.1 for special versions. It then follows (p. 211 of Resnick) that

$$M_{t_1,t_2}(X_n) \to Y(t_2 - t_1) \quad \text{in} \quad \mathbb{R} \tag{6.15}$$

for almost all t_1, t_2 with $0 < t_1 < t_2$, again for the special versions. We then can apply Theorem 15.5.1 to deduce that

$$\mathbf{X}_n \to \mathbf{X} \quad \text{in} \quad E((0,\infty,\mathbb{R}) \ , \tag{6.16}$$

again for the special versions. However, the w.p.1 convergence implies the desired convergence in distribution. Clearly the distributions of $\mathbf{X}_n$ and $\mathbf{X}$ are characterized by the independence and the distributions of the maxima $M_{t_1,t_2}(\mathbf{X}_n)$ and $M_{t_1,t_2}(\mathbf{X})$. ∎

Thus we have established a limit for the scaled process $\mathbf{X}_n$ in E and connected it to the convergence of successive maxima to extremal processes.

15.7. The Space F

We now use parametric representations to define the space F of functions larger than E. Our purpose is to more faithfully model the fluctuations associated with the excursions. In particular, we now want to describe the order in which the points are visited in the excursions.

Suppose that the graphs $\Gamma_{\tilde{x}}$ of elements $\tilde{x}$ of $E \equiv E([0,T],\mathbb{R}^k)$ are defined as in Section 15.4 and that the conditions of Theorem 15.4.1 hold. We will focus on the subset E_{st} of strongly connected functions in E, and call the space E.

We say that two (strong) parametric representations of a graph $\Gamma_{\tilde{x}}$ are *equivalent* if there exist nondecreasing continuous function λ_1 and λ_2 mapping $[0,1]$ onto $[0,1]$ such that

$$(u_1, r_1) \circ \lambda_1 = (u_2, r_2) \circ \lambda_2 \ . \tag{7.1}$$

A consequence of (7.1) is that, for each $t \in S$, the functions u_1 and u_2 in the two equivalent parametric representations visit all the points in $I(t)$, the same number of times and in the same order. (Staying at a value is regarded as a single visit.)

We let F be the set of *equivalence classes* of these parametric representations. We thus regard any two parametric representations that are equivalent in the sense of (7.1) as two representations of the same function in F. Let $\hat{x}$ denote an element of F and let $\Pi_s(\hat{x})$ be the set of all parametric representations of $\hat{x}$, i.e., all members of the equivalence class.

Paralleling the metric d_s inducing the M_1 topology on D in equation (3.7) in Section 12.3, let d_s be defined on $F \equiv F([0,T],\mathbb{R}^k)$ by

$$d_s(\hat{x}_1, \hat{x}_2) = \inf_{\substack{(u_i,r_i)\in\Pi_s(\hat{x}_i) \\ i=1,2}} \{\|u_1 - u_2\| \vee \|r_1 - r_2\|\} \ . \tag{7.2}$$

We call the topology on F induced by the metric d_s in (7.2) the M_1 topology.

Paralleling Theorem 12.3.1, we can conclude that d_s in (7.2) is a bonafide metric.

Theorem 15.7.1. *The space (F, d_s) for d_s in (7.2) is a separable metric space.*

To prove Theorem 15.7.1, we use the following lemma, which closely parallels Lemma 12.3.2.

Lemma 15.7.1. *For any $\hat{x}_1$, $\hat{x}_2 \in F$, $(u_1, r_1) \in \Pi_s(\hat{x}_1)$ and $\epsilon > 0$, it is possible to find $(u_2, r_2) \in \Pi_s(\hat{x}_2)$ such that*

$$\|u_1 - u_2\| \vee \|r_1 - r_2\| \leq d_s(\hat{x}_1, \hat{x}_2) + \epsilon \ . \tag{7.3}$$

Given an element $\hat{x}$ in F, let $\gamma(\hat{x})$ denote the associated element of E. (Note that $\gamma : F \to E$ is a many-to-one map.) Let $\Gamma_{\hat{x}} \equiv \Gamma_{\gamma(\hat{x})}$ be the graph of $\gamma(\hat{x})$ or just $\hat{x}$. It is easy to relate convergence $\hat{x}_n \to \hat{x}$ in (F, M_1) to convergence $\gamma(\hat{x}_n) \to \gamma(\hat{x})$ in (E, M_2) because both modes of convergence have been characterized by parametric representations. Clearly, M_1 convergence in F implies M_2 convergence in E, but not conversely.

Theorem 15.7.2. (relating the metrics $d_{s,2}$ and d_s) *For $\tilde{x}_1$, $\tilde{x}_2 \in F$,*

$$d_{s,2}(\gamma(\hat{x}_1), \gamma(\hat{x}_2)) \leq d_s(\hat{x}_1, \hat{x}_2) \ ,$$

where $d_{s,2}$ and d_s are defined in (5.1) and (7.2). Hence, if $\hat{x}_n \to \hat{x}$ in (F, M_1), then $\gamma(\hat{x}_n) \to \gamma(\hat{x})$ in (E_{st}, SM_2).

To put (F, d_s) into perspective, recall that for D with the metric d_s in (3.7) of Section 12.3 inducing the SM_1 topology, all parametric representations were required to be nondecreasing, using an order introduced on the graphs. Thus, all parametric representations of x in D are equivalent parametric representations. In contrast, here with the more general functions in F, there is in general no one natural order to consider on the graphs.

Paralleling Theorem 15.4.6 and Corollary 15.4.1, we can identify (D, SM_1) as a subset of (F, M_1).

Theorem 15.7.3. *Let D' be the subset of functions $\hat{x}$ in F for which $\Gamma_{\hat{x}}$ is the graph of a function in D, i.e., for which*

$$I(t) = [x(t-), x(t+)] \quad \textit{in} \quad \mathbb{R}^k \tag{7.4}$$

for all $t \in S$, and for which the parametric representation is monotone in the order on the graphs defined in Section 12.3. *Then (D', d_s) for d_s in* (7.2) *is homeomorphic to (D, SM_1). Indeed,*

$$d_s(\hat{x}_1, \hat{x}_2) = d_s(\gamma(\hat{x}_1), \gamma(\hat{x}_2)) \tag{7.5}$$

where $d_s(\gamma(\hat{x}_1), \gamma(\hat{x}_2))$ is interpreted as the metric on D in (3.7) *of Section* 12.3.

We now give an example to show how F allows us to establish new limits. We will show that we can have $x_n \to x$ in (D, M_2), $x_n \not\to x$ in (D, M_1) and $x_n \to \hat{x}$ in (F, M_1). An important point is that the new M_1 limit $\hat{x}$ in F is not equivalent to the M_2 limit x in D.

Example 15.7.1. *Convergence in (F, M_1) where convergence in (D, M_1) fails.* Consider the functions $x = I_{[1,2]}$ and

$$\begin{aligned} x_n(0) &= x_n(1) = x_n(1 + 2n^{-1}) = 0 \\ x_n(1 + n^{-1}) &= x_n(1 + 3n^{-1}) = x_n(2) = 1 \ , \end{aligned}$$

with x_n defined by linear interpolation elsewhere. Note that $x_n \to x$ in (D, M_2), but $x_n \not\to x$ in (D, M_1). However, $\hat{x}_n \to \hat{x}$ in (F, M_1), where $\hat{x}$ is different from x. The convergence $\hat{x}_n \to \hat{x}$ in F implies that $\gamma(\hat{x}_n) \to \gamma(\hat{x})$ in (E, M_2), which in this case is equivalent to $x_n \to x$ in (D, M_2). ∎

15.8. Queueing Applications

In this final section we discuss queueing applications of the spaces E and F. First, for queueing networks, the key result is the following analog of results in Chapter 14.

Theorem 15.8.1. (the multidimensional reflection map on F) *The multidimensional reflection map R defined in Section 14.2 is well defined and measurable as a map from (F, M_1) to (F, M_1). Moreover, R is continuous at all $\hat{x}$ in F_1, the subset of functions in F with discontinuities or excursions in only one coordinate at a time, provided that the range is endowed with the product topology. Hence, if $\hat{x}_n \to \hat{x}$ in (F, M_1), where $\hat{x} \in F_1$, then $\gamma(R(\hat{x}_n)) \to \gamma(R(\hat{x}))$ in (E_{wk}, WM_2).*

A natural way to establish stochastic-process limits in F and E is to start from stronger stochastic-process limits in D and then abandon some

of the detail. In particular, we can start with a stochastic-process limit in (D, SM_1) that simultaneously describes the asymptotic behavior in two different time scales. We can then obtain a limit in (F, M_1) when we focus on only the longer time scale. From the perspective of the longer time scale, what happens in the shorter time scale happens instantaneously. We keep some of the original detail when we go from a limit in (D, SM_1) to a limit in (F, M_1), because the order in which the points are visited in the shorter time scale at each one-time excursion is preserved.

To illustrate, we consider a single infinite-capacity fluid queue that alternates between periods of being "in control" and periods of being "out of control." When the system is in control, the net-input process is "normal;" when the system is out of control, the net-input process is "exceptional." The model is a generalization of the infinite-capacity version of the fluid queue in Chapter 5. The alternating periods in which the system is in and out of control generalize the up and down times for the queueing network with service interruptions in Section 14.7, because here we allow more complicated behavior during the down times.

As in Section 14.7, let $\{(U_k, D_k) : k \geq 1\}$ be the sequence of successive up and down times. We assume that these random variables are strictly positive and that $\sum_{i=1}^{\infty} U_k = \infty$ w.p.1. The system is up or in control during the intervals $[T_k, T_k + U_k)$ and down or out of control during the intervals $[T_k + U_k, T_{k+1})$, where $T_k \equiv \sum_{i=1}^{k}(U_i + D_i)$, $k \geq 1$, with $T_0 \equiv 0$.

Let X_k^u and X_k^u be potential net-input processes in effect during the k^{th} up period and down period, respectively. We regard X_k^u and X_k^d as random elements of $D \equiv D([0,\infty), \mathbb{R})$ for each k. However, X_k^u is only realized over the interval $[0, U_k)$, and X_k^u is only realized over the interval $[0, D_k)$. Specifically, the overall net-input process is the process X in D defined by

$$X(t) = \begin{cases} X_{k+1}^u(t - T_k) + X(T_k), & T_k \leq t < T_k + U_k \ , \\ X_{k+1}^d(t - T_k - U_k) + X(T_k + U_k), & T_k + U_k \leq t < T_{k+1} \ , \end{cases} \tag{8.1}$$

for $t \geq 0$.

Just as in equations (2.5) – (2.7) in Section 5.2, the associated workload or buffer-content process, starting out empty, is defined by

$$W(t) \equiv \phi(X)(t) \equiv X(t) - \inf_{0 \leq s \leq t} X(s), \quad t \geq 0 \ . \tag{8.2}$$

We want to establish a heavy-traffic stochastic-process limit for a sequence of these fluid-queue models. In preparation for that heavy-traffic limit, we need to show how to identify the limit process in F: We need to

show how the net-input random element of D approaches a random element of F when the down times decrease to 0, while keeping the toital net inputs over the down intervals unchanged. What happens during the down time is unchanged; it just happens more quickly as the down time decreases.

To formalize that limit, suppose that we are given a fluid queue model as specified above. We now create a sequence of models by altering the variables D_k and the processes X_k^d for all k. Specifically, we let

$$D_{n,k} \equiv c_n^{-1} D_k \tag{8.3}$$

for all k, where $c_n \to \infty$ as $n \to \infty$. To keep the associated net-input processes during these down times unchanged except for time scaling, we let

$$X_{n,k}^d(t) \equiv X_k^d(c_n t), \quad t \geq 0 \ , \tag{8.4}$$

which implies that

$$X_{n,k}^d(pD_{n,k}) = X_k^d(pD_k) \quad \text{for} \quad 0 \leq p \leq 1 \ .$$

With this special construction, it is easy to see that the original element of D approaches an associated element of F as $n \to \infty$.

Lemma 15.8.1. (decreasing down times) *Consider a fluid-queue model as specified above. Suppose that we construct a sequence of fluid-queue models indexed by n, where the models change only by* (8.3) *and* (8.4) *with $c_n \to \infty$ as $n \to \infty$. Then $X_n \to \hat{X}$ in (F, M_1).*

We are now ready to state the heavy-traffic stochastic-process limit for the fluid-queue model. We start with a more-detailed limit in D, and obtain the limit in F by applying Lemma 15.8.1 to the limit process in D.

For the heavy-traffic limit, we start with a sequence of fluid-queue models indexed by n. In this initial sequence of models, we let the initial up and down times be independent of n. We then introduce the following scaled random elements:

$$\begin{aligned}
U_{n,k} &\equiv nU_k, \quad k \geq 1, \\
D_{n,k} &\equiv n^{\beta} D_k, \quad k \geq 1, \\
\mathbf{X}_{n,k}^u(t) &\equiv n^{-H} X_{n,k}^u(nt), \quad t \geq 0, \\
\mathbf{X}_{n,k}^d(t) &\equiv n^{-H} X_{n,k}^u(n^{\beta} t), \quad t \geq 0, \\
\mathbf{X}_n(t) &\equiv n^{-H} X_n(nt), \quad t \geq 0, \\
\mathbf{W}_n(t) &\equiv n^{-H} W_n(nt), \quad t \geq 0 \ ,
\end{aligned} \tag{8.5}$$

where

$$0 < \beta < 1 \quad \text{and} \quad H > 0 \ . \tag{8.6}$$

Note that in (8.5) the time scaling of U_k, $X^u_{n,k}$, $X_n(t)$ and $W_n(t)$ in (8.5) is all by n, while the time scaling of D_k and $X^d_{n,k}$ is by only n^β, where $0 < \beta < 1$. Thus the durations of the down periods are asymptotically negligible compared to the durations of the uptimes in the limit as $n \to \infty$.

Here is the result:

Theorem 15.8.2. (heavy-traffic limit in F) *Consider a sequence of fluid-queue models with the scaling in* (8.5). *Suppose that*

$$\begin{aligned} &\{(U_{n,k}, D_{n,k}, \mathbf{X}^u_{n,k}, \mathbf{X}^d_{n,k}) : k \geq 1\} \\ &\quad \Rightarrow \{(U'_k, D'_k, \mathbf{X}^u_k, \mathbf{X}^d_k) : k \geq 1\} \end{aligned} \tag{8.7}$$

in $(\mathbb{R} \times \mathbb{R} \times D \times D)^\infty$ *as* $n \to \infty$, *where* $U'_k > 0$ *for all* k *and* $\sum_{i=1}^\infty U'_i = \infty$ *w.p.1. Then*

$$(\mathbf{X}_n, \mathbf{W}_n) \Rightarrow (\hat{\mathbf{X}}, \hat{\mathbf{W}}) \quad \textit{in} \quad (F, M_1) \times (F, M_1)$$

with the product topology on the product space $(F, M_1) \times (F, M_1)$, *where* $\hat{\mathbf{X}}$ *is the limit in* F *obtained in Lemma* 15.8.1 *applied to the limiting fluid queue model with up and down times* (U'_k, D'_k) *and potential net-input processes* $\mathbf{X}^u_k$ *and* $\mathbf{X}^d_k$ *in* (8.7) *and* $\hat{\mathbf{W}} = \phi(\hat{\mathbf{X}})$. *Consequently,*

$$\mathbf{W}_n \Rightarrow \gamma(\hat{\mathbf{W}}) \quad \textit{in} \quad (E, M_2) \ .$$

A standard sufficient condition for condition (8.7) in Theorem 15.8.2 is to have the random elements $(U_{n,k}, D_{n,k}, \mathbf{X}^u_{n,k}, \mathbf{X}^d_{n,k})$ of $\mathbb{R} \times \mathbb{R} \times D \times D$ be IID for $k \geq 1$ and to have

$$(U_{n,1}, D_{n,1}, \mathbf{X}^u_{n,1}, \mathbf{X}^d_{n,1}) \quad \Rightarrow \quad (U'_1, D'_1, \mathbf{X}^u_1, \mathbf{X}^d_1) \tag{8.8}$$

in $\mathbb{R} \times \mathbb{R} \times D \times D$. In the standard case, we have $H = 1/2$ and $\mathbf{X}^u_1$ Brownian motion with drift down, but we have seen that there are other possibilities. In this framework, the complexity of the limit process is largely determined by the limit process $\mathbf{X}^d_1$ describing the asymptotic behavior of the net-input process during down times. We conclude by giving an example in which the limiting stochastic process, regarded as an element of F or E, is tractable.

Example 15.8.1. *The Poisson-excursion limit process.* A special case of interest in the IID cycle framework above occurs when the limit $\mathbf{X}^u_k$ in (8.7)

is deterministic, e.g., $\mathbf{X}_k^u(t) = ct$, $t \geq 0$. A further special case is to have $c = 0$. Then the process is identically zero except for the excursions, so all the structure is contained in the excursions.

Suppose that we have this special structure; i.e., suppose that $\mathbf{X}_k^u(t) = 0$, $t \geq 0$. If, in addition, U_k' is exponentially distributed, then in the limit the excursions occur according to a Poisson process. We call such a process a *Poisson-excursion process.*

To consider a simple special case of a Poisson-excursion process, suppose that, for each k, the process $\mathbf{X}_{n,k}^d$ corresponds to a partial sum of three random variables $Z_{n,k,1}$, $Z_{n,k,2}$, $Z_{n,k,3}$ evenly spaced in the interval $[0, n^\beta D_k]$, i.e., occurring at times $n^\beta D_k/4$, $2n^\beta D_k/4$ and $3n^\beta D_k/4$. For example, we could have deterministic down times of the form $D_k = 4$ for all k, which implies that $D_{n,k} = 4n^\beta$ for all k.

Consistent with (8.7), we assume that

$$n^{-H}(Z_{n,k,1}, Z_{n,k,2}, Z_{n,k,3}) \Rightarrow (Z_{k,1}, Z_{k,2}, Z_{k,3}) \quad \text{as} \quad n \to \infty \ . \tag{8.9}$$

Then the base process $\mathbf{X}$ associated with the graph of the limit $\hat{\mathbf{X}}$ is

$$\mathbf{X}(t) = \sum_{i=1}^{N(t)} Y_i, \quad t \geq 0 \ , \tag{8.10}$$

where $\{N(t) : t \geq 0\}$ is a Poisson process with rate $1/EU_1'$ and $\{Y_i\}$ is an IID sequence with $Y_i \stackrel{\mathrm{d}}{=} \sum_{j=1}^3 Z_{i,j}$.

The excursions in more detail are described by the successive partial sums $\{\sum_{j=1}^n Z_{i,j} : 1 \leq n \leq 3\}$; i.e., the first excursion occuring at time U_1' goes from 0 to Z_1, then to $Z_1 + Z_2$, and finally to $Z_1 + Z_2 + Z_3$. Note that we have not yet made any restrictive assumptions on the joint distribution of $(Z_{n,k,1}, Z_{n,k,2}, Z_{n,k,3})$, so that the limit $(Z_{k,1}, Z_{k,2}, Z_{k,3})$ can have an arbitrary distribution. However, if in addition, $Z_{k,1}$, $Z_{k,2}$ and $Z_{k,3}$ are IID, then we can characterize the limiting distribution of $\gamma(\hat{\mathbf{W}})$, the random element of E representing the buffer content. The base process W can be represented as

$$W(t) = Q_{3N(t)}, \quad t \geq 0 \ , \tag{8.11}$$

where $\{N(t) : t \geq 0\}$ is the Poisson process counting limiting excursions and Q_k is the queue-content in period k of a discrete-time queue with IID net inputs distributed as Z_1, i.e., where $\{Q_k : k \geq 0\}$ satisfies the Lindley equation

$$Q_k = \max\{Q_{k-1} + Z_k, 0\}, \quad k \geq 1 \ , \tag{8.12}$$

with $Q_0 = 0$. Consequently, $W(t) \Rightarrow W$ as $t \to \infty$, where the limiting workload W is distributed as Q with $Q_k \Rightarrow Q$ as $k \to \infty$.

The additional excursions in $\hat{\mathbf{W}}$ occur according to the Poisson process N. Given that the k^{th} excursion occurs at time t, we can easily describe it in terms of the four random variables $\mathbf{W}(t-)$, $Z_{k,1}$, $Z_{k,2}$ and $Z_{k,3}$: The process moves from $A_0 \equiv \mathbf{W}(t-)$ to $A_1 \equiv max\{0, A_0 + Z_{k,1}\}$, then to $A_2 \equiv max\{0, A_1 + Z_{k,2}\}$ and finally to $\mathbf{W}(t) = A_3 \equiv max\{0, A_2 + Z_{k,3}\}$. For the process in F, these three steps all occur at the time t. The parametric representation moves continuously from A_0 to A_1, then to A_2 and A_3. The limiting workload then remains constant until the next excursion. ∎