

7

Weak Convergence: Introduction

7.0 Outline of Chapter

Up to now, we have concentrated on the convergence of $\{\theta_n\}$ or of $\{\theta^n(\cdot)\}$ to an appropriate limit set with probability one. In this chapter, we work with a weaker type of convergence. In practical applications, this weaker type of convergence most often yields the same information about the asymptotic behavior as the probability one methods. Yet the methods of proof are simpler (indeed, often substantially simpler), and the conditions are weaker and more easily verifiable. The weak convergence methods have considerable advantages when dealing with complicated problems, such as those involving correlated noise, state-dependent noise processes, decentralized or asynchronous algorithms, and discontinuities in the algorithms. If probability one convergence is still desired, starting with a weak convergence argument can allow one to “localize” the probability one proof, thereby simplifying both the argument and the conditions that are needed. For example, the weak convergence proof might tell us that the iterates spend the great bulk of the time very near some point. Then a “local” method such as that for the “linearized” algorithm in Theorem 6.1.2 can be used. The basic ideas have many applications to problems in process approximation and for getting limit theorems for sequences of random processes.

Mathematically, the basic idea of weak convergence concerns the characterization of the limits of the sequence of measures of the processes $\theta^n(\cdot)$ on the appropriate path space. In particular, one shows that the limit measures induce a process (on the path space) supported on some set of limit

trajectories of the ODE $\dot{\theta} = \bar{g}(\theta) + z$, $z \in -C(\theta)$, (or $\dot{\theta} = \bar{g}(\theta)$ for the unconstrained case). Despite this abstract formulation, one does not work with the measures in either the proofs or the applications, but with the iterate sequence itself, and the entire process of proof and applications is actually simpler than what probability one methods require. The basic ideas are applications of only an elementary part of the theory of weak convergence of probability measures.

The main convergence results for stochastic approximations are in Chapter 8. This chapter provides an introduction to the subject. Section 1 motivates the importance and the role of weak convergence methods. The ideas and developments in Section 2 are intended to illustrate some of the ideas that underlie the theory of weak convergence and to provide a kind of “behind the scene” view. They do not require any of the machinery of the general theory. They need only be skimmed for the general ideas, because stronger results will be proved in Chapter 8 with the use of the general theory, without requiring any of the explicit constructions or methods that are used in the proofs in Section 2. Despite the fact that the statements of the theorems are more limited and the proofs require more details than those that use the general theory of Section 3, they are included since they relate the weak convergence method to what was done in Chapter 5, and illustrate the role of “tightness” and the minimal requirements on the step sizes ϵ_n and the moments of the martingale difference terms δM_n . It was seen in Section 6.10 that, under broad conditions and even for the constant-step-size algorithm, if the iterate is close to a stable point or set at time n , it will stay close to it for a time *at least* of the order of e^{c/ϵ_n} for some $c > 0$.

The general theory of weak convergence is introduced in Section 3. The theorems cited there and in Section 4 are all we will require for the convergence proofs in subsequent chapters. The reader need only understand the statements and need not know their proofs. Subsection 4.1 gives criteria for verifying that the “limit” is a martingale; this important idea will be used in the proofs of Chapter 8. Subsection 4.2 gives “martingale-type” criteria to verify that a given continuous-time martingale with continuous paths is a Wiener process. A very useful perturbed test function criterion for verifying tightness (relative compactness) is stated in Subsection 4.3. The latter two results will be used in the proofs of the rates of convergence in Chapter 10. The reference [241] contains a useful intuitive discussion of the advantages and nature of weak convergence, with many graphs illustrating the convergence.

7.1 Introduction

Introductory remarks on weak vs. probability one convergence.

Chapters 5 and 6 were concerned with methods of proving the convergence of $\{\theta_n\}$ or of $\{\theta^n(\cdot)\}$ with probability one to an appropriate limit set. In the context of the actual way in which stochastic approximation algorithms are used in applications, an assertion of probability one convergence can be misleading. For example, there is usually some sort of stopping rule that tells us when to stop the iteration and to accept as the final value either the most recent iterate or some function of the iterates that were taken “shortly” before stopping. The stopping rule might simply be a limit on the number of iterations allowed, or it might be a more sophisticated rule based on an estimate of the improvement of the mean performance over the recent past, or perhaps on the “randomness” of the behavior of the recent iterations (the more “random,” the more likely that the iterate is in a neighborhood of a stationary point). Generally, at the stopping time all we know about the closeness to a limit point or set is information of a distributional type.

If the application of stochastic approximation is done via a simulation, then one can control the model so that it does not change over time (but even then there is a stopping rule). Nevertheless, the situation is different when the stochastic approximation is used to optimize a system on-line, since convergence with probability one implies that we can iterate essentially forever, and the system will remain unchanged however long the procedure is. In practical on-line applications, the step size ϵ_n is often not allowed to decrease to zero, due to considerations concerning robustness and to allow some tracking of the desired parameter as the system changes slowly over time. Then probability one convergence does not apply. Indeed, it is the general practice in signal processing applications to keep the step size bounded away from zero. In the Kiefer–Wolfowitz procedure for the minimization of a function via a “noisy” finite difference-based algorithm, the difference interval is often not allowed to decrease to zero. This creates a bias in the limit, but this bias might be preferred to the otherwise slower convergence and “noisier” behavior when the variance of the effective noise is inversely proportional to the square of a difference interval that goes to zero. Thus, even under a probability one convergence result, the iterates might converge to a point close to the minimum, but not to the minimum itself. Such biases reduce the value of a probability one convergence result.

The proofs of probability one results tend to be quite technical. They might not be too difficult when the noise terms are martingale differences, but they can be very hard for multiscale, state-dependent-noise cases or decentralized/asynchronous algorithms. To handle the technical difficulties in an application where one wishes to prove probability one convergence, one might be forced to introduce assumptions that are not called for (such

as modifications of the algorithm) or that are hard to verify.

These concerns do not eliminate the value of convergence with probability one. Convergence theorems are a guide to behavior. Although no algorithm is carried to infinity, it is still comforting to know that if the iterations are allowed to continue forever in the specified ideal environment, they will assuredly converge. However, the concerns that have been raised emphasize that methods for probability one convergence might offer less than what appears at first sight, and that methods with slightly more limited convergence goals can be just as useful, particularly if they give a lot of insight into the entire process, are technically easier, require weaker conditions, and are no less informative under the conditions that prevail in applications.

This and the next chapter will focus on convergence in a weak or distributional sense. It will turn out that the proofs are easier and conditions weaker and that we can learn nearly as much about where the iterate sequence spends its time as with probability one methods. For complicated algorithms, the proofs are substantially simpler. The methods are the natural ones if the step sizes do not decrease to zero, where probability one convergence is not pertinent. When the step sizes do go to zero, weak convergence does not preclude convergence with probability one. In fact, first proving weak convergence can simplify the ultimate proof of probability one convergence, since it allows a “localization” of the proof. Recall that the general approach has been to get the mean ODE determined by the “average dynamics,” show that the solution to the ODE tends to an appropriate limit set (or a set of stationary points if the algorithm is of the gradient descent type), and then show that the chosen limit points of the solution to the ODE are the limit points of $\{\theta_n\}$. The mean ODE is easier to derive in that there are weaker conditions and simpler proofs when weak convergence methods are used. The process $\{\theta_n\}$ can still be shown to spend nearly all of its time arbitrarily close to the same limit point or set. For example, suppose that the limit set is a just a unique asymptotically stable (in the sense of Liapunov) point $\bar{\theta}$ of the ODE. Then, once we know, via a weak convergence analysis, how to characterize the path to $\bar{\theta}$ and that $\{\theta_n\}$ spends nearly all of its time (asymptotically) in any arbitrarily small neighborhood of $\bar{\theta}$, one can use a *local analysis* to get convergence with probability one, under weaker conditions (due to the local nature of the proof) than what would be needed by a pure probability one technique. For example, the methods of Chapters 5 and 6 can be used locally, or the local large deviations methods of [63] can be used. Whether or not one follows a weak convergence proof with a probability one convergence proof, under broad conditions it can be shown that if the error $|\theta_n - \bar{\theta}|$ is small, it stays small afterwards for an average time of at least the order of e^{c/ϵ_n} for some $c > 0$.

Some basic ideas and facts from the theory of weak convergence will be discussed in the next section. The theory is a widely used tool for obtain-

ing approximation and limit theorems for sequences of stochastic processes. There is only a small amount of machinery to be learned, and this machinery has applications well beyond the needs of this book. Before discussing the ideas of the theory of weak convergence in detail, we return to the model of Chapter 5, where the noise terms are martingale differences and prove a convergence theorem under weaker conditions than used there. The proof that will be given is of a “weak convergence nature” and gives some of the flavor of weak convergence. Owing to the martingale difference property, it is quite straightforward and does not require any of the general machinery of weak convergence analysis of Sections 3 and 4. The proofs and the statements of the theorems are intended to be illustrative of some of the ideas underlying the theory of weak convergence. They are more awkward than necessary, since the tools of the general theory are avoided. However, the constructions used are of independent interest and play an important role in relating the general theory to what has been done for the probability one case, although they will not be used in applications of that general theory in the following chapters. Since more general results will be obtained in Chapter 8, the results and ideas of the next section should be skimmed for their intuitive content and insights into the types of approximations that can be used for “distributional-sense” approximation and limit theorems, and what might be required if the general theory were not available.

7.2 Martingale Difference Noise: Simple Alternative Approaches

Introductory comments and definitions. Convergence results for two simple algorithmic forms will be given in this section. The theorems are quite similar, although different methods are used for the proofs. In the second problem, there is no constraint set H , and it is assumed that $\{\theta_n\}$ is bounded in probability. These models are chosen for illustrative purposes only. The methods to be used avoid the explicit use of the machinery of weak convergence theory, but they illustrate some of the concepts. The explicit constructions used in these theorems are not necessary when the general theory is used.

The first result (Theorem 2.1 and its corollary) depends on the fact that if a sequence of random variables converges in probability, there is always a subsequence that converges with probability one to the same limit. The second result (Theorem 2.2) depends on the fact that any sequence of random variables which is bounded in probability has a subsequence that converges in distribution to some random variable. These basic facts provide simple connections between the convergence of the sequence $\{\theta^n(\cdot)\}$ with probability one and in the weak or distributional sense. In both cases,

the technique of proof depends on the choice of appropriate subsequences. In the first case, it is shown that for any sequence $\{\theta^n(\cdot), Z^n(\cdot)\}$, there is always a subsequence to which the methods of Theorems 5.2.1 and 5.2.3 can be applied. The second method works with convergence in distribution directly and leads to a “functional” limit theorem. The reader should keep in mind that the assumptions are selected for convenience in exposing the basic ideas and that stronger results are to be obtained in Chapter 8.

Let $\{\mathcal{F}_n\}$ denote a sequence of nondecreasing σ -algebras, where \mathcal{F}_n measures at least $\{\theta_i, Y_{i-1}, i \leq n\}$, and let E_n denote the expectation conditioned on \mathcal{F}_n . Suppose that we can write $E_n Y_n = \bar{g}(\theta_n) + \beta_n$, where β_n is a small bias term. First, we work with the constrained algorithm

$$\theta_{n+1} = \Pi_H(\theta_n + \epsilon_n Y_n) = \theta_n + \epsilon_n Y_n + \epsilon_n Z_n. \quad (2.1)$$

It will be shown that the mean ODE

$$\dot{\theta} = \bar{g}(\theta) + z, \quad z(t) \in -C(\theta(t)) \quad (2.2)$$

continues to characterize the asymptotic behavior. Recall the definitions $t_n = \sum_{i=0}^{n-1} \epsilon_i$, $\theta^0(t) = \theta_n$ on $[t_n, t_{n+1})$, and $\theta^n(t) = \theta^0(t + t_n)$, where $\theta^n(t) = \theta_0$ for $t \leq -t_n$. As usual, decompose the interpolated process $\theta^n(\cdot)$ as

$$\theta^n(t) = \theta_n + \bar{G}^n(t) + M^n(t) + B^n(t) + Z^n(t), \quad (2.3)$$

where we recall that, for $t \geq 0$,

$$\begin{aligned} \bar{G}^n(t) &= \sum_{i=n}^{m(t_n+t)-1} \epsilon_i \bar{g}(\theta_i), & M^n(t) &= \sum_{i=n}^{m(t_n+t)-1} \epsilon_i \delta M_i, \\ B^n(t) &= \sum_{i=n}^{m(t_n+t)-1} \epsilon_i \beta_i, & Z^n(t) &= \sum_{i=n}^{m(t_n+t)-1} \epsilon_i Z_i, \end{aligned}$$

where $\delta M_n = Y_n - E_n Y_n$.

Theorem 2.1. *Assume the step-size condition (5.1.1), (A5.2.1)–(A5.2.3), and any of the constraint set conditions (A4.3.1), (A4.3.2), or (A4.3.3). Suppose that $E|\beta_n| \rightarrow 0$. Then, for each subsequence of $\{\theta^n(\cdot), Z^n(\cdot)\}$, there is a further subsequence (indexed by n_k) such that $\{\theta^{n_k}(\cdot), Z^{n_k}(\cdot)\}$ is equicontinuous in the extended sense with probability one (on a sequence of intervals going to infinity), and whose limits satisfy the ODE (2.2). Let there be a unique limit point $\bar{\theta}$ of (2.2), which is asymptotically stable in the sense of Liapunov. Then, for each $\mu > 0, T > 0$,*

$$\lim_n P \left\{ \sup_{t \leq T} |\theta^n(t) - \bar{\theta}| > \mu \right\} = 0. \quad (2.4a)$$

More generally, there are $\mu_n \rightarrow 0, T_n \rightarrow \infty$ such that

$$\lim_n P \left\{ \sup_{t \leq T_n} \text{distance}[\theta^n(t), L_H] \geq \mu_n \right\} = 0. \quad (2.4b)$$

In the sense of convergence in probability, the fraction of time $\theta^n(\cdot)$ spends in a δ neighborhood of L_H goes to one as $n \rightarrow \infty$.

Remarks on the theorem. According to the estimates in Sections 6.9 and 6.10, under broad conditions one can use $T_n = O(e^{c/\epsilon_n})$ in (2.4) for some $c > 0$. It will be seen that the behavior proved by both the weak convergence and the probability one methods are similar. They both show that the path essentially follows the solution to the ODE, for large n . Suppose that the path enters the domain of attraction of an asymptotically stable point $\bar{\theta}$ infinitely often. Then (ignoring some null set of paths), the probability one methods show that it will eventually converge to $\bar{\theta}$. Under the weaker conditions used for the weak convergence proofs we might not be able to prove that it will never escape. But this escape, if it ever occurs, will be a “large deviations” phenomena; i.e., it will be very rare, perhaps too rare to be of concern.

Note that we do not need a summability condition of the type $\sum_n \epsilon_n^{1+\gamma} < \infty$ for some $\gamma > 0$; only $\epsilon_n \rightarrow 0$ is needed. The corollary given after the proof shows that uniform square integrability of $\{Y_n\}$ can be replaced by uniform integrability. Further comments on the nature of the convergence results appear after Theorems 2.2 and 8.2.1.

Proof. The proof is modeled on that of Theorem 5.2.1. The main idea is the careful choice of subsequence. By the fact that there is a $0 \leq K_1 < \infty$ such that $\sup_n E|Y_n|^2 \leq K_1$, the inequality (4.1.4) implies that for $T > 0$ and $\mu > 0$,

$$\begin{aligned} P \left\{ \sup_{t \leq T} |M^n(t)| \geq \mu \right\} &\leq \frac{E \left| \sum_{i=n}^{m(t_n+T)-1} \epsilon_i \delta M_i \right|^2}{\mu^2} \\ &\leq \frac{K_1 \sum_{i=n}^{m(t_n+T)-1} \epsilon_i^2}{\mu^2}. \end{aligned} \quad (2.5)$$

Next, it will be shown that, for any $T < \infty$ and $\mu > 0$,

$$\lim_n P \left\{ \sup_{t \leq T} |y^n(t)| \geq \mu \right\} = 0 \quad (2.6)$$

for $y^n(\cdot)$ being either $M^n(\cdot)$ or $B^n(\cdot)$. Since $\epsilon_n \rightarrow 0$, we have

$$\lim_n \sum_{i=n}^{m(t_n+T)} \epsilon_i^2 = 0 \quad (2.7)$$

for each $T > 0$, which yields (2.6) for $M^n(\cdot)$. Since $E|\beta_n| \rightarrow 0$,

$$E \sum_{i=n}^{m(t_n+T)} \epsilon_i |\beta_i| \rightarrow 0,$$

which implies that (2.6) also holds for $B^n(\cdot)$.

By (2.6), for $M^n(\cdot)$ and $B^n(\cdot)$ (or by the proof of the following corollary in the case where uniform integrability of $\{Y_n\}$ replaces uniform square integrability),

$$\theta^n(t) = \theta_n + \sum_{i=n}^{m(t_n+t)-1} \epsilon_i \bar{g}(\theta_i) + Z^n(t) + \kappa^n(t), \quad (2.8)$$

where $\kappa^n(t) = M^n(t) + B^n(t)$ and for any $\mu > 0$,

$$\lim_n P \left\{ \sup_{t \leq T} |\kappa^n(t)| \geq \mu \right\} = 0.$$

By the fact that $M^n(\cdot)$ and $B^n(\cdot)$ satisfy (2.6), there are $m_k \rightarrow \infty$ and $T_k \rightarrow \infty$ such that

$$P \left\{ \sup_{t \leq T_k} |\kappa^n(t)| \geq 2^{-k} \right\} \leq 2^{-k}, \quad n \geq m_k. \quad (2.9)$$

Now, (2.9) and the Borel–Cantelli Lemma imply that for any sequence $n_k \geq m_k$,

$$\lim_k \sup_{t \leq T_k} |\kappa^{n_k}(t)| = 0, \quad (2.10)$$

with probability one. From this point on, the proof concerning equicontinuity and the mean ODE follows that of Theorems 5.2.1 or 5.2.3.

The proof of (2.4) uses a contradiction argument. Assuming that it is false, extract a suitable subsequence for which the \liminf is positive and use the previous conclusions to get a contradiction. Proceed as follows. Let $T > 0$ and (extracting another subsequence if necessary) work with the left shifted sequence $\{\theta^{n_k}(-T + \cdot), Z^{n_k}(-T + \cdot)\}$. Then use the fact that for any $\delta > 0$ the time required for the solution of (2.2) to reach and remain in $N_\delta(L_H)$ is bounded in the initial condition in H . Since T is arbitrary, this yields that the solution to the ODE on $[0, \infty)$ is in $N_\delta(L_H)$ for each $\delta > 0$. The few remaining details are left to the reader. \square

Remark on equicontinuity. In Theorem 5.2.1, the original sequence $\{\theta^n(\cdot), Z^n(\cdot)\}$ was equicontinuous in the extended sense with probability one and $\lim_n \sup_{t \leq T} |\kappa^n(t)| = 0$ with probability one. Thus, we were able to examine the convergent subsequences of $\{\theta^n(\omega, \cdot), Z^n(\omega, \cdot)\}$ for (almost all) fixed ω , with the “errors” $\kappa^n(\omega, \cdot)$ vanishing as $n \rightarrow \infty$. In the current case, we know only that each sequence $\{\theta^n(\cdot), Z^n(\cdot)\}$ has a further subsequence that is equicontinuous in the extended sense with probability one (on a sequence of time intervals increasing to the entire real line), and the errors vanish (for almost all ω) only along that subsequence. Hence, under only the conditions of Theorem 2.1, we cannot expect that for almost

all ω , any subsequence of $\{\theta^n(\omega, \cdot), Z^n(\omega, \cdot)\}$ will always have a further subsequence that converges to a solution to the mean ODE.

Definition. A sequence $\{Y_n\}$ of vector-valued random variables is said to be *uniformly integrable* if

$$\sup_n E|Y_n|I_B \rightarrow 0 \quad \text{as } P\{B\} \rightarrow 0,$$

where B is a measurable set. This is equivalent to the property

$$\lim_{K \rightarrow \infty} \sup_n E|Y_n|I_{\{|Y_n| \geq K\}} = 0.$$

Remark on uniform integrability. A nice aspect of the weak convergence approach is that the uniform integrability of $\{Y_n\}$ is enough to assure that the limit processes are continuous without using the “reflection” character of the Z_n terms, as required in the proof of Theorem 5.2.3.

Corollary. *The conclusions of the theorem continue to hold if the square integrability of $\{Y_n\}$ is replaced by uniform integrability.*

Proof of the corollary. Assume the uniform integrability condition in lieu of square integrability. The only problem is the verification of (2.6) for $y^n(\cdot) = M^n(\cdot)$. For $K > 0$, let $I_K(v)$ denote the indicator function of the set $\{v \in \mathbb{R}^r : |v| \geq K\}$. Define the truncated sequence $\{Y_{n,K}\}$ by $Y_{n,K} = Y_n(1 - I_K(Y_n))$. Then $Y_n = Y_{n,K} + Y_n I_K(Y_n)$. Define $\delta M_{n,K}$ and $\delta \kappa_{n,K}$ by

$$\begin{aligned} \delta M_n &= (Y_n - E_n Y_n) = [Y_{n,K} - E_n Y_{n,K}] + [Y_n I_K(Y_n) - E_n Y_n I_K(Y_n)] \\ &\equiv \delta M_{n,K} + \delta \kappa_{n,K}. \end{aligned}$$

The uniform integrability of $\{Y_n\}$ and Jensen’s inequality (4.1.11) imply that

$$\lim_{K \rightarrow \infty} \sup_n E[|Y_n I_K(Y_n)| + |E_n Y_n I_K(Y_n)|] \leq 2 \lim_{K \rightarrow \infty} \sup_n E|Y_n|I_K(Y_n) = 0. \quad (2.11)$$

Equation (2.11) and the definition of $\delta \kappa_{i,K}$ imply that

$$\lim_{K \rightarrow \infty} \sup_n E \sum_{i=n}^{m(t_n+T)} \epsilon_i |\delta \kappa_{i,K}| = 0. \quad (2.12)$$

For $\mu > 0$ and $T > 0$, we can write

$$\begin{aligned} & \limsup_n P \left\{ \sup_{t \leq T} |M^n(t)| \geq \mu \right\} \\ & \leq \limsup_n P \left\{ \sup_{t \leq T} \left| \sum_{i=n}^{m(t_n+t)-1} \epsilon_i \delta M_{i,K} \right| \geq \mu/2 \right\} \\ & \quad + \limsup_n P \left\{ \sup_{t \leq T} \left| \sum_{i=n}^{m(t_n+t)-1} \epsilon_i \delta \kappa_{i,K} \right| \geq \mu/2 \right\}. \end{aligned} \quad (2.13)$$

Now, given $\nu > 0$, there is a $0 \leq K < \infty$ such that (2.12) implies that the last term on the right side in (2.13) is less than ν . Next, (2.5) holds with $\delta M_{n,K}$ (that is bounded by $2K$) replacing δM_n , where $K_1 = 4K^2$. These facts imply (2.6) for $y^n(\cdot) = M^n(\cdot)$. The rest of the details are as in the theorem. \square

Remarks on extensions to correlated noise. The conditions required in Theorem 2.1 and its corollary showed some of the possibilities inherent in the weak convergence method, since we required only $\epsilon_n \rightarrow 0$, uniform integrability of $\{Y_n\}$, and $E|\beta_n| \rightarrow 0$ (the latter condition will be weakened in Chapter 8). For more general algorithms, where there are noise processes such as the sequence $\{\xi_n\}$ appearing in Chapter 6, some additional averaging is needed. In Theorem 6.1.1, the condition (A6.1.3) was used to average out the noise. To use (A6.1.3), it was necessary to show that the paths of $\theta^n(\cdot)$ were asymptotically continuous with probability one, so that the time varying θ_n could be replaced by a fixed value of θ over small time intervals. The analog of that approach in the present weak convergence context involves showing that for each positive T and μ ,

$$\lim_{\Delta \rightarrow 0} \limsup_n P \left\{ \max_{j\Delta \leq T} \max_{0 \leq t \leq \Delta} |\theta^n(j\Delta + t) - \theta^n(j\Delta)| \geq \mu \right\} = 0. \quad (2.14)$$

This condition does not imply asymptotic continuity of $\{\theta^n(\cdot)\}$ with probability one, and hence it is weaker than what was needed in Chapter 6. Indeed, (2.14) is implied by the uniform integrability of $\{Y_n\}$. Thus, for the correlated noise case and under uniform integrability, one could redo Theorem 2.1 by replacing condition (A6.1.3) by the weaker condition obtained by deleting the $\sup_{j \geq n}$ inside the probability there. With analogous adjustments to assumptions (A6.1.4)–(A6.1.7), such an approach can be carried out. But the method to be used in the next chapter is preferable in general because it is simpler to use, requires weaker conditions, and is more versatile.

An alternative approach. In the next theorem, $\tilde{\theta}^0(\cdot)$ denotes the *piecewise linear* interpolation of $\{\theta_n\}$ with interpolation interval $\{[t_n, t_{n+1})\}$,

and $\tilde{\theta}^n(\cdot)$ the left shift by t_n . Let $C^r[0, \infty)$ denote the space of continuous \mathbb{R}^r -valued functions on the time interval $[0, \infty)$ with the local sup norm metric (i.e., a sequence converges if it converges uniformly on each bounded time interval). The convergence assertion (2.15) is in terms of the convergence of a sequence of expectations of bounded and continuous functionals. This is actually equivalent to the types of convergence assertions given in Theorem 2.1, as can be seen by suitable choices of the function $F(\cdot)$. The form (2.15) of the convergence assertion is typical of the conclusions of weak convergence theory. The methods to be used in Chapter 8 work with the original piecewise constant interpolated processes $\theta^n(\cdot)$ and do not require the piecewise linear interpolation.

Theorem 2.2. *Assume the step-size condition (5.1.1) and that $\{Y_n\}$ is uniformly integrable. Drop the constraint set H and let $\{\theta_n\}$ be bounded in probability. Let $E_n Y_n = \bar{g}(\theta_n) + \beta_n$, where $E|\beta_n| \rightarrow 0$ and $\bar{g}(\cdot)$ is bounded and continuous. Suppose that the solution to the ODE is unique (going either forward or backward in time) for each initial condition, and that the limit set L , over all initial conditions, is bounded. Then, for each subsequence of $\{\tilde{\theta}^n(\cdot)\}$ there is a further subsequence (indexed by n_k) and a process $\theta(\cdot)$ that satisfies $\dot{\theta} = \bar{g}(\theta)$ such that $\{\tilde{\theta}^{n_k}(\cdot)\}$ converges in distribution to $\theta(\cdot)$ in the sense that for any bounded and continuous real-valued function $F(\cdot)$ on $C^r[0, \infty)$*

$$EF(\tilde{\theta}^{n_k}(\cdot)) \xrightarrow{k} EF(\theta(\cdot)). \quad (2.15)$$

For almost all ω , $\theta(t, \omega)$ takes values in an invariant set of the ODE. Also, (2.4b) holds when L_H is replaced by the invariant set. If the limit set is simply a point, $\bar{\theta}$, then $EF(\tilde{\theta}^{n_k}(\cdot)) \rightarrow F(\bar{\theta}(\cdot))$, where $\bar{\theta}(t) = \bar{\theta}$.

Remark on the convergence assertion and the limit points. Analogously to the conclusions of Theorem 2.1, the theorem says that for large n , the paths of $\tilde{\theta}^n(\cdot)$ are essentially concentrated on the set of limit trajectories of the ODE $\dot{\theta} = \bar{g}(\theta)$. This can be seen as follows. Let L denote the largest bounded invariant set of the ODE. For $y(\cdot) \in C^r[0, \infty)$ and any positive T , define the function

$$\tilde{F}_T(y(\cdot)) = \sup_{t \leq T} \text{distance}[y(t), L],$$

where $\text{distance}[y, L] = \min_{u \in L} |y - u|$. The function $\tilde{F}_T(\cdot)$ is continuous on $C^r[0, \infty)$. Then the theorem says that for each subsequence there is a further subsequence (indexed by n_k) such that $E\tilde{F}_T(\tilde{\theta}^{n_k}(\cdot)) \rightarrow E\tilde{F}_T(\theta(\cdot)) = 0$, where the limit is zero since the value of $\tilde{F}_T(\cdot)$ on the paths of the limit process is zero with probability one. Thus, the sup over $t \in [0, T]$ of the distance between the original sequence $\tilde{\theta}^n(t)$ and L goes to zero in

probability as $n \rightarrow \infty$. Indeed, the same result holds if T is replaced by $T_n \rightarrow \infty$ slowly enough. Thus, there are $T_n \rightarrow \infty$ such that for any $\mu > 0$,

$$\lim_n P \left\{ \sup_{t \leq T_n} \text{distance}[\tilde{\theta}^n(t), L] \geq \mu \right\} = 0.$$

We note the key role to be played by the estimate (2.14). This estimate implies the “tightness” condition, which will be basic to the results of Chapter 8 and is guaranteed by the uniform integrability of $\{Y_n\}$.

Proof. The theorem remains true if, for any $T > 0$, $F(\cdot)$ depends on the values of its argument only at times $t \leq T$. Both $F(\cdot)$ and T will be fixed henceforth. Recall that a compact set in $C^r[0, T]$ is a set of equicontinuous functions. Let $y(\cdot)$ denote the canonical element of $C^r[0, T]$. For any $\nu > 0$ and compact set $C_0 \subset C^r[0, T]$, there is a $\Delta > 0$ and a real-valued continuous function $F_\Delta(\cdot)$ on $C^r[0, T]$ that depends on $y(\cdot)$ only at times $\{i\Delta, i\Delta \leq T\}$ such that

$$|F(y(\cdot)) - F_\Delta(y(\cdot))| \leq \nu, \quad y(\cdot) \in C_0.$$

We can write

$$\tilde{\theta}^n(t) = \theta_n + \sum_{i=n}^{m(t_n+t)-1} \epsilon_i \bar{g}(\theta_i) + \kappa^n(t), \quad (2.16)$$

where $\kappa^n(t) = M^n(t) + B^n(t)$. Since $E|\beta_n| \rightarrow 0$, $\lim_n E \max_{t \leq T} |B^n(t)| = 0$. By the martingale property, the uniform integrability, (4.1.5) (applied to the sums of $\epsilon_i \delta M_{i,K}$ in the corollary to Theorem 2.1) and the estimates (2.13), for each $T > 0$ we have

$$\lim_n E \max_{t \leq T} |M^n(t)| = 0.$$

Since for each $\mu > 0$,

$$\lim_n P \left\{ \sup_{t \leq T} |\kappa^n(t)| \geq \mu \right\} = 0, \quad (2.17a)$$

we have

$$\lim_{K \rightarrow \infty} \sup_n P \left\{ \sup_{t \leq T} |\tilde{\theta}^n(t)| \geq K \right\} = 0. \quad (2.17b)$$

Equation (2.17a) and the representation (2.16) imply that (2.14) holds.

Equations (2.14) and (2.17b) imply that for each $\nu > 0$ there is a compact set $C_\nu \subset C^r[0, T]$ such that for each n , $\tilde{\theta}^n(\cdot) \in C_\nu$ with probability greater than $1 - \nu$. [This is also implied directly by the uniform integrability of $\{Y_n\}$.] Thus we need only show that there is a subsequence n_k and a process $\theta(\cdot)$ satisfying $\dot{\theta} = \bar{g}(\theta)$ and taking values in the largest bounded invariant set of this ODE such that for any $\Delta > 0$,

$$EF_\Delta(\tilde{\theta}^{n_k}(\cdot)) \rightarrow EF_\Delta(\theta(\cdot)),$$

where the bounded and continuous real-valued function $F_\Delta(\cdot)$ depends only on the values of the argument at times $\{i\Delta, i\Delta \leq T\}$. [We note that a key point in the general theory is to show that, with a high probability (not depending on n), the paths of $\tilde{\theta}^n(\cdot)$ are confined to a compact set in the path space. The general theory also uses limits taken along convergent subsequences to help characterize the limits of the original sequence.]

By the fact that $\{\theta_n\}$ is bounded in probability, there is a subsequence n_k and a random variable $\theta(0)$ (on some probability space) such that $\{\theta_{n_k}\}$ converges in distribution to $\theta(0)$. Let \mathcal{T} denote the positive rational numbers. By (2.17b) and the diagonal method, we can take a further subsequence $\{m_k\}$ such that $\{\tilde{\theta}^{m_k}(t), t \in \mathcal{T}\}$ converges in distribution, and denote the limit (on some probability space) by $\{\theta(t), t \in \mathcal{T}\}$. By the representation (2.16), the boundedness of $\bar{g}(\cdot)$ and the fact that (2.17a) holds for each T , there is a version of the limit that is continuous on \mathcal{T} with probability one. Hence, we can suppose that $\theta(\cdot)$ is defined for all $t \geq 0$ and is continuous.

By (2.14), (2.16), and (2.17a), we can write

$$\theta(n\delta) = \theta(0) + \sum_{i=0}^{n-1} \delta \bar{g}(\theta(i\delta)) + \kappa_\delta(n\delta), \quad (2.18)$$

where $\lim_\delta P\{\sup_{t \leq T} |\kappa_\delta(t)| \geq \mu\} = 0$ for each $\mu > 0, T > 0$. By the continuity of $\theta(\cdot)$ we see that it must satisfy the ODE $\dot{\theta} = \bar{g}(\theta)$. By the uniqueness to the solution of the ODE for each initial condition and the boundedness of $\bar{g}(\cdot)$, the limit does not depend on the chosen further subsequence $\{m_k\}$ and the original subsequence can be used. Now (2.15) clearly holds for $F_\Delta(\cdot)$.

We need only show that with probability one the paths of $\theta(\cdot)$ take values in the largest bounded invariant set of the ODE, and we will sketch the details. We have worked on the time interval $[0, \infty)$. Follow the same procedure on the interval $(-\infty, \infty)$ by replacing \mathcal{T} by $\mathcal{T}_1 = \mathcal{T} \cup (-\mathcal{T})$. Then (extracting a further subsequence $\{m_k\}$ if necessary), $\{\tilde{\theta}^{m_k}(t), t \in \mathcal{T}_1\}$ converges in distribution to a process $\theta(t), t \in \mathcal{T}_1$. As above, it can be assumed that $\theta(t)$ is defined and continuous on $(-\infty, \infty)$ and satisfies the ODE. Again, by the uniqueness of the solution to the ODE for each initial condition, the further subsequence indexed by m_k is not needed, and one can use the original subsequence.

Next, note that the boundedness in probability of the sequence $\{\theta_n\}$ implies that for any $\mu > 0$, there is a $K_\mu < \infty$ such that if θ is any limit in distribution of a subsequence of $\{\theta_n\}$, then

$$P\{|\theta| \geq K_\mu\} \leq \mu. \quad (2.19)$$

Thus, for each $\mu > 0$, we can suppose that $|\theta(t)| \leq K_\mu$ for each t with probability $\geq 1 - \mu$. Now, this fact and the stability property of the limit set L implies that the solution is bounded and lies in L for all t . \square

7.3 Weak Convergence

7.3.1 Definitions

Convergence in distribution. Let $\{A_n\}$ be a sequence of \mathbb{R}^k -valued random variables on a common probability space (Ω, P, \mathcal{F}) , with $(a_{n,i}, i = 1, \dots, k)$ being the real-valued components of A_n . Let P_n denote the measures on the Borel sets of \mathbb{R}^k determined by A_n , and let $x = (x_1, \dots, x_k)$ denote the canonical variable in \mathbb{R}^k . If there is an \mathbb{R}^k -valued random variable A with real-valued components (a_1, \dots, a_k) such that

$$P\{a_{n,1} < \alpha_1, \dots, a_{n,k} < \alpha_k\} \rightarrow P\{a_1 < \alpha_1, \dots, a_k < \alpha_k\} \quad (3.1)$$

for each $\alpha = (\alpha_1, \dots, \alpha_k) \in \mathbb{R}^k$ at which the right side of (3.1) is continuous, then we say that A_n *converges to A in distribution*. Let P_A denote the measure on the Borel sets of \mathbb{R}^k determined by A . An equivalent definition [34] is that

$$EF(A_n) = \int F(x) dP_n(x) \rightarrow EF(A) = \int F(x) dP_A(x) \quad (3.2)$$

for each bounded and continuous real-valued function $F(\cdot)$ on \mathbb{R}^k . We say that the sequence $\{P_n\}$ is *tight* or *bounded in probability* if

$$\lim_{K \rightarrow \infty} \sup_n P_n \{(-\infty, -K] \cup [K, \infty)\} = \lim_{K \rightarrow \infty} \sup_n P \{|A_n| \geq K\} = 0. \quad (3.3a)$$

For real- or vector-valued random variables, the term *mass preserving* is sometimes used in lieu of *tight*. An equivalent definition of boundedness in probability is: Let $|A_n| < \infty$ with probability one for each n and for each small $\mu > 0$, let there be finite M_μ and K_μ such that

$$P \{|A_n| \geq K_\mu\} \leq \mu, \quad \text{for } n \geq M_\mu. \quad (3.3b)$$

Given a sequence of random variables $\{A_n\}$ with values in \mathbb{R}^k (or more generally, in any complete separable metric space), tightness is a necessary and sufficient condition that any subsequence has a further subsequence that converges in distribution [34]. Convergence in distribution is also called *weak convergence*.

The notion of convergence in distribution extends, via the general theory of weak convergence, to sequences of random variables that take values in more abstract spaces than \mathbb{R}^k . The extension provides a powerful methodology for the approximation of random processes and for obtaining useful limit theorems for sequences of random processes, such as our $\theta^n(\cdot)$.

The following example is one of the classical illustrations of weak convergence. Let $\{\xi_n\}$ be a sequence of real-valued random variables that are mutually independent and identically distributed, with mean zero and unit

variance. Then, by the classical central limit theorem $\sum_{i=1}^n \xi_i/\sqrt{n}$ converges in distribution to a normally distributed random variable with zero mean and unit variance. Now, define $q(t) = \max\{i : i/n \leq t\}$ and define the process with piecewise constant paths

$$W_n(t) = \sum_{i=0}^{q(t)-1} \xi_i/\sqrt{n}. \quad (3.4)$$

Then the central limit theorem tells us that $W_n(t)$ converges in distribution to a normally distributed random variable with mean zero and variance t . For an integer k , let $0 = t_0 < t_1, \dots, t_{k+1}$ be real numbers, and let $W(\cdot)$ be a real-valued Wiener process with unit variance parameter. Then, by the multivariate central limit theorem [34], the set $\{W_n(t_{i+1}) - W_n(t_i), i \leq k\}$ converges in distribution to $\{W(t_{i+1}) - W(t_i), i \leq k\}$. It is natural to ask whether $W_n(\cdot)$ converges to $W(\cdot)$ in a stronger sense. For example, will the distribution of the first passage time for $W_n(\cdot)$ defined by $\min\{t : W_n(t) \geq 1\}$ converge in distribution to the first passage time for $W(\cdot)$ defined by $\min\{t : W(t) \geq 1\}$? Will the maximum $\max\{W_n(t) : t \leq 1\}$ converge in distribution to $\max\{W(t) : t \leq 1\}$ and similarly for other useful functionals? In general, we would like to know the class of functionals $F(\cdot)$ for which $F(W_n(\cdot))$ converges in distribution to $F(W(\cdot))$. Donsker's Theorem states that this convergence occurs for a large class of functionals [25, 68].

Now, let us consider the following extension, where the ξ_n are as given above. For given real-valued $U(0)$ and $\Delta > 0$, define real-valued random variables U_n^Δ by $U_0^\Delta = U(0)$ and for $n \geq 0$,

$$U_{n+1}^\Delta = U_n^\Delta + \Delta g(U_n^\Delta) + \sqrt{\Delta} \xi_n,$$

where $g(\cdot)$ is a continuous function. Define the interpolated process $U^\Delta(\cdot)$ by: $U^\Delta(t) = U_n^\Delta$ on $[n\Delta, (n+1)\Delta)$. Then in what sense will $U^\Delta(\cdot)$ converge to the process defined by the stochastic differential equation

$$dU = g(U)dt + dW?$$

We expect that this stochastic differential equation is the “natural” limit of $U^\Delta(\cdot)$.

More challenging questions arise when the random variables ξ_n are correlated. The central limit theorem and the laws of large numbers are very useful for the approximation of random variables, which are the “sums” of many small effects whose mutual dependence is “local,” by the simpler normally distributed random variable or by some constant, respectively. The theory of weak convergence is concerned with analogous questions when the random variables are replaced by random processes as in the above examples. There are two main steps analogous to what is done for proving

the central limit theorem: First show that there are appropriately convergent subsequences and then identify the limits. The condition (3.3) says that, neglecting a set of small probability for each n (small, uniformly in n), the values of random variables A_n are confined to some compact set. There will be an analogous condition when random processes replace random variables.

The path spaces $D(-\infty, \infty)$ and $D[0, \infty)$. The processes $\theta^n(\cdot)$, $M^n(\cdot)$, and $Z^n(\cdot)$ used in Chapters 5 and 6 and Section 1 were piecewise constant and had small discontinuities (of the order of the step size) at the “jump times.” However, when using the Arzelà–Ascoli Theorem to justify the extraction of subsequences that converge uniformly on bounded time intervals to continuous limits, we checked equicontinuity by looking at the processes at the jump times. Equivalently, we used the extended definition of equicontinuity (4.2.2) and the extended Arzelà–Ascoli Theorem 4.2.2. This procedure is obviously equivalent to working with the piecewise linear interpolations. We could continue to work with piecewise linear interpolations or the extended definition of equicontinuity. However, from a technical point of view it turns out to be easier to use a path space that allows discontinuities, in particular because the verification of the extension of the concept of tightness will be simpler.

The statement of Theorem 2.2 used the path space $C^r[0, \infty)$ of the piecewise linear interpolations of θ_n . The applications of weak convergence theory commonly use the space of paths that are right continuous and have limits from the left, with a topology known as the *Skorohod topology*. This topology is weaker than the topology of uniform convergence on bounded time intervals. The key advantage of this weaker topology is that it is easier to prove the *functional* analog of the tightness condition (3.3) for the various processes of interest. This will be more apparent in Chapters 10 and 11, where a more sophisticated form of the theory is used to get the rates of convergence. Since the limit processes $\theta(\cdot)$ will have continuous paths in our applications, the strength of the assertions of the theorem is the same no matter which topology is used.

Let $D(-\infty, \infty)$ (resp., $D[0, \infty)$) denote the space of real-valued functions on the interval $(-\infty, \infty)$ (resp., on $[0, \infty)$) that are right continuous and have left-hand limits, with the Skorohod topology used, and $D^k(-\infty, \infty)$ (resp., $D^k[0, \infty)$) its k -fold product. The exact definition of the Skorohod topology is somewhat technical and not essential for our purposes. It is given at the end of the chapter, but is not explicitly used in subsequent chapters. Full descriptions and treatments can be found in [25, 68]. We note the following properties here. Let $f_n(\cdot)$ be a sequence in $D(-\infty, \infty)$. Then the convergence of $f_n(\cdot)$ to a continuous function $f(\cdot)$ in the Skorohod topology is equivalent to *convergence uniformly on each bounded time interval*. Under the Skorohod topology, $D(-\infty, \infty)$ is a complete and separable metric space. Since we will later use the Skorohod topology to prove

rate of convergence results, for consistency we will use it from this point on. Loosely speaking, the Skorohod topology is an extension of the topology of uniform convergence on bounded time intervals in the sense that a “local” small (n, t) -dependent stretching or contraction of the time scale is allowed, the purpose of which is to facilitate dealing with “nice” discontinuities that do not disappear in the limit.

Definition of weak convergence. Let B be a metric space. In our applications, it will be either \mathbb{R}^k or one of the product path spaces $D^k(-\infty, \infty)$ or $D^k[0, \infty)$ for an appropriate integer k . Let \mathcal{B} denote the minimal σ -algebra induced on B by the topology. Let $\{A_n, n < \infty\}$ and A be B -valued random variables defined on a probability space (Ω, P, \mathcal{F}) , and suppose that P_n and P_A are the probability measures on (B, \mathcal{B}) determined by A_n and A , respectively. We say that P_n *converges weakly* to P_A if (3.2) holds for all bounded, real-valued, and continuous functions on B , and write the convergence as $P_n \Rightarrow P_A$. Equivalently, with a convenient abuse of terminology, we say that A_n *converges weakly* to A or that A is the *weak limit* or *weak sense limit* of $\{A_n\}$, and write $A_n \Rightarrow A$. These ways of expressing weak convergence will be used interchangeably.

A set $\{A_n\}$ of random variables with values in B is said to be *tight* if for each $\delta > 0$ there is a compact set $B_\delta \subset \mathcal{B}$ such that

$$\sup_n P\{A_n \notin B_\delta\} \leq \delta. \quad (3.5)$$

To prove tightness of a sequence of \mathbb{R}^k -valued processes, it is enough to prove tightness of the sequence of each of the k components. A set $\{A_n\}$ of B -valued random variables is said to be *relatively compact* if each subsequence contains a further subsequence that converges weakly.

7.3.2 Basic Convergence Theorems

A basic result, Prohorov’s Theorem, is given next.

Theorem 3.1. [25, Theorems 6.1 and 6.2]. *If $\{A_n\}$ is tight, then it is relatively compact (i.e., it contains a weakly convergent subsequence). If B is complete and separable, tightness is equivalent to relative compactness.*

Theorem 3.2. [25, Theorem 5.1] *Let $A_n \Rightarrow A$. Let $F(\cdot)$ be a real-valued bounded and measurable function on B that is continuous with probability one under the measure P_A . Then $EF(A_n) \rightarrow EF(A)$.*

Tightness in $D(-\infty, \infty)$ and $D[0, \infty)$. An advantage to working with the path space $D(-\infty, \infty)$ in lieu of $C(-\infty, \infty)$ or $C[0, \infty)$ is that it is easier to prove tightness in $D(-\infty, \infty)$. Let $A_n(\cdot)$ be processes with paths in $D(-\infty, \infty)$. The following criteria for tightness will be easy to apply to

our problems. Let \mathcal{F}_t^n denote the σ -algebra generated by $\{A_n(s), s \leq t\}$, and let τ denote an \mathcal{F}_t^n -stopping time.

Theorem 3.3. [[68, Theorem 8.6, Chapter 3], [118, Theorem 2.7b]] *Let $\{A_n(\cdot)\}$ be a sequence of processes that have paths in $D(-\infty, \infty)$. Suppose that for each $\delta > 0$ and each t in a dense set in $(-\infty, \infty)$, there is a compact set $K_{\delta,t}$ in \mathbb{R} such that*

$$\inf_n P\{A_n(t) \in K_{\delta,t}\} \geq 1 - \delta \quad (3.6)$$

and for each positive T ,

$$\lim_{\delta \downarrow 0} \limsup_n \sup_{|\tau| \leq T} \sup_{s \leq \delta} E \min[|A_n(\tau + s) - A_n(\tau)|, 1] = 0. \quad (3.7)$$

Then $\{A_n(\cdot)\}$ is tight in $D(-\infty, \infty)$. If the interval $[0, \infty)$ is used, tightness holds if $|\tau| \leq T$ is replaced by $0 \leq \tau \leq T$.

Remarks on tightness and the limit process. Let the piecewise constant interpolations $\theta^n(\cdot)$ and $Z^n(\cdot)$ be defined on $(-\infty, \infty)$ until further notice. Note that representation (2.16) and estimate (2.17) imply the tightness of $\{\theta^n(\cdot)\}$ in $D^r(-\infty, \infty)$. If a compact constraint set H is used, then (3.6) holds. For the problems in Chapters 5 and 6, the fact that the extended Arzelà–Ascoli Theorem was applicable (with probability one) to $\{\theta^n(\cdot), Z^n(\cdot)\}$ implies (3.7) for these processes. Thus the tightness criterion is always satisfied under the conditions used in Chapters 5 and 6. It is clear that (3.7) does not imply the continuity of the paths of either $A_n(\cdot)$ or any weak sense limit $A(\cdot)$. Indeed, (3.7) holds if $\{A_n(\cdot)\}$ is a sequence of continuous-time Markov chains on a compact state space S with uniformly bounded and time independent transition rates. Then it can be shown that any weak sense limit process is also a continuous-time Markov chain with values in S and time independent transition functions.

Suppose that a sequence of processes $\{A^n(\cdot)\}$ is tight in $D^r(-\infty, \infty)$ and that on each interval $[-T, T]$ the size of the maximum discontinuity goes to zero in probability as $n \rightarrow \infty$. Then any weak sense limit process must have continuous paths with probability one.

Suppose that $\{\theta^n(\cdot), Z^n(\cdot)\}$ (or, otherwise said $\{P_n\}$) is tight. Let $(\theta(\cdot), Z(\cdot))$ denote the weak sense limit of a weakly convergent subsequence. The next question concerns the characterization of the process $(\theta(\cdot), Z(\cdot))$. It will be shown that the weak sense limit process is characterized as solutions to the mean ODE. In other words, any limit measure is concentrated on a set of paths that satisfy the ODE, with $Z(\cdot)$ being the reflection term. In particular, the limit measure is concentrated on a set of paths that are limit trajectories of the ODE, as $t \rightarrow \infty$.

Skorohod representation and “probability one” convergence. In the general discussion of weak convergence and in Theorems 3.1–3.3, the

sequence $\{A_n\}$ was defined on some given probability space (Ω, P, \mathcal{F}) . Since weak convergence works with measures P_n induced on the range space of the sequence $\{A_n\}$, the actual probability space itself is unimportant, and one can select it for convenience. For purely analytical purposes, it is often helpful to be able to suppose that the convergence is with probability one rather than in the weak sense, since it enables us to work with paths directly and simplifies parts of the proofs. It turns out that the probability space can be chosen such that the weak convergence “implies” convergence with probability one. This basic result is known as the *Skorohod representation*.

Theorem 3.4. [[68, Chapter 3, Theorem 1.8], [222, Theorem 3.1]] *Let B be a complete and separable metric space with metric $d(\cdot, \cdot)$, and let $A_n \Rightarrow A$ for B -valued random variables A_n and A . Then there is a probability space $(\tilde{\Omega}, \tilde{\mathcal{B}}, \tilde{P})$ with associated B -valued random variables \tilde{A}_n and \tilde{A} defined on it such that for each set $D \in \mathcal{B}$,*

$$\tilde{P}\{\tilde{A}_n \in D\} = P\{A_n \in D\}, \quad \tilde{P}\{\tilde{A} \in D\} = P\{A \in D\}, \quad (3.8)$$

and

$$d(\tilde{A}_n, \tilde{A}) \rightarrow 0 \quad \text{with probability one.} \quad (3.9)$$

The choice of the probability space in the theorem is known as the *Skorohod representation*. Its use facilitates proofs without changing the distributions of the quantities of interest. In the rest of the book, it will be supposed where convenient in the proofs, and without loss of generality, that the probability space has been selected so that weak convergence is “equivalent to” convergence with probability one.

Note that we have started with a range space B with a σ -algebra \mathcal{B} , and measures P_n, P_A defined on it, but with only weak convergence $P_n \Rightarrow P_A$. The Skorohod representation constructs a single probability space with B -valued random variables \tilde{A}_n and \tilde{A} defined on it, where \tilde{A}_n (resp., \tilde{A}) determine the measure \tilde{P}_n (resp., \tilde{P}_A), on the range space (B, \mathcal{B}) , and where the convergence is with probability one. For notational simplicity, when the Skorohod representation is used in the sequel, the *tilde* notation will generally be omitted.

Define $A_n = (\theta^n(\cdot), Z^n(\cdot))$, which takes values in $D^{2r}(-\infty, \infty)$. If it has been shown that the sequence $(\theta^n(\cdot), Z^n(\cdot))$ converges weakly to some $D^{2r}(-\infty, \infty)$ -valued random variable $(\theta(\cdot), Z(\cdot))$, where $(\theta(\cdot), Z(\cdot))$ have continuous paths with probability one, then by the Skorohod representation (see Theorem 3.4), it can be supposed in the proof of the characterization of $(\theta(\cdot), Z(\cdot))$ that the convergence is with probability one uniformly on bounded time intervals, provided that the conclusions of the theorem remain in terms of weak convergence. In particular, the use of the Skorohod representation itself does not imply that the original sequence θ_n (or $\theta^n(\cdot)$) converges with probability one.

A simple example of Skorohod representation. Let $\{Y_n\}$ be a sequence of real-valued random variables that converges in distribution to a random variable Y , and let $F_n(\cdot)$ (resp., $F(\cdot)$) be the distribution function of Y_n (resp., of Y). Suppose for simplicity that each of the distribution functions is strictly monotonically increasing. The sequence $\{Y_n\}$ might not converge to Y with probability one. In fact, Y_n might not even be defined on the same probability space. But there are random variables \tilde{Y}_n and \tilde{Y} such that each of the pairs Y_n and \tilde{Y}_n (as well as Y and \tilde{Y}) have the same distribution, and \tilde{Y}_n converges to \tilde{Y} with probability one.

The construction is as follows. Let the probability space be $(\tilde{\Omega}, \tilde{\mathcal{B}}, \tilde{P})$ where $\tilde{\Omega} = [0, 1]$, $\tilde{\mathcal{B}}$ is the collection of Borel sets on $[0, 1]$, and \tilde{P} is the Lebesgue measure. For $\tilde{\omega} \in [0, 1]$ define $\tilde{Y}_n(\tilde{\omega}) = F_n^{-1}(\tilde{\omega})$ and $\tilde{Y}(\tilde{\omega}) = F^{-1}(\tilde{\omega})$. By the construction and the uniform distribution on $[0, 1]$, $P\{\tilde{Y}_n \leq a\} = F_n(a)$ for all a . Thus \tilde{Y}_n (resp., \tilde{Y}) has the distribution function $F_n(\cdot)$ (resp., $F(\cdot)$). Furthermore the uniform convergence of $F_n(\cdot)$ to $F(\cdot)$ and the strict monotonicity imply that $F_n^{-1}(\cdot)$ also converges pointwise to $F^{-1}(\cdot)$. This is equivalent to the convergence of $\tilde{Y}_n \rightarrow \tilde{Y}$ for all $\tilde{\omega}$. This is an easy example. In the more general case, where Y_n is replaced by a random process and $\{Y_n\}$ is tight, the limit of any weakly convergent subsequence is not so easily characterized. Then the Skorohod representation can be quite helpful in the analysis.

Return to the central limit theorem discussed in connection with (3.4). The theory of weak convergence tells us that the *process* $W_n(\cdot)$ constructed in (3.4) converges weakly to the Wiener process with unit variance parameter. This result gives us more information on the distributions of real-valued functionals of the paths of $W_n(\cdot)$ for large n than can be obtained by the classical central limit theorem alone, which is confined to working with values at a finite number of fixed points and not with the entire process; see [25, 68] for the details and a full development of the general theory and other examples. For the basic background, effective methods for dealing with wide-bandwidth noise-driven processes or discrete time processes with correlated driving noise, including many applications of the theory to approximation and limit problems arising in applications to control, communication and signal processing theory, as well as to various stochastic approximation-type problems, consult [127].

Some auxiliary results. The following theorems will simplify the analysis. Theorem 3.5 shows the fundamental role of uniform integrability in establishing the Lipschitz continuity of the paths of the weak sense limit processes and generalizes the corollary to Theorem 2.1. Further details of the proof are in Theorem 8.2.1. Sometimes one can show that a sequence of processes can be approximated in some sense by one that can be shown to be tight and for which the weak sense limit can be exhibited. This is dealt with in Theorem 3.6. The proof of Theorem 3.6 follows from the definition

of weak convergence; the details are left to the reader.

Theorem 3.5. *Let $\{Y_i^n; n \geq 0, i \geq 0\}$ be a sequence of real-valued random variables that is uniformly integrable, and let ϵ_i^n be non-negative numbers that satisfy*

$$\sum_{i=0}^{\infty} \epsilon_i^n = \infty, \quad \text{for all } n \text{ and}$$

$$\limsup_n \limsup_i \epsilon_i^n = 0.$$

Define $\tau_k^n = \sum_{i=0}^{k-1} \epsilon_i^n$ and the processes $X^n(\cdot)$ on $D^r[0, \infty)$ by

$$X^n(t) = \sum_{i=0}^{k-1} \epsilon_i^n Y_i^n \quad \text{on } [\tau_k^n, \tau_{k+1}^n).$$

Then $\{X^n(\cdot)\}$ is tight, and all weak sense limit processes have Lipschitz continuous paths with probability one. If $E|Y_i^n| \rightarrow 0$ as n and i go to infinity, then the weak sense limit process is identically zero. The analogous results hold for $D^r(-\infty, \infty)$.

Theorem 3.6. *Let the processes $X^n(\cdot)$ have paths in $D^r[0, \infty)$ with probability one. Suppose that for each $1 > \rho > 0$ and $T > 0$ there is a process $X^{n,\rho,T}(\cdot)$ with paths in $D^r[0, \infty)$ with probability one such that*

$$P \left\{ \sup_{t \leq T} |X^{n,\rho,T}(t) - X^n(t)| \geq \rho \right\} \leq \rho.$$

If $\{X^{n,\rho,T}(\cdot), n \geq 0\}$ is tight for each ρ and T , then $\{X^n(\cdot)\}$ is tight. If $\{X^{n,\rho,T}(\cdot), n \geq 0\}$ converges weakly to a process $X(\cdot)$ that does not depend on (ρ, T) , then the original sequence converges weakly to $X(\cdot)$. Suppose that for each $1 > \rho > 0$ and $T > 0$, $\{X^{n,\rho,T}(\cdot), n \geq 0\}$ converges weakly to a process $X^{\rho,T}(\cdot)$, and that there is a process $X(\cdot)$ such that the measures of $X^{\rho,T}(\cdot)$ and $X(\cdot)$ on the interval $[0, T]$ are equal, except on a set whose probability goes to zero as $\rho \rightarrow 0$. Then $\{X^n(\cdot)\}$ converges weakly to $X(\cdot)$. The analogous result holds for processes with paths in $D^r(-\infty, \infty)$.

7.4 Martingale Limit Processes and the Wiener Process

7.4.1 Verifying that a Process Is a Martingale

The criteria for tightness in Theorem 3.3 will enable us to show that for any subsequence of the shifted stochastic approximation processes $\{\theta^n(\cdot), Z^n(\cdot)\}$, there is always a further subsequence that converges weakly.

The next step will be to identify the limit process; in particular to show that it is a solution to the desired mean ODE, and to do this without excessive effort and under weak conditions. If the noise is not of the martingale difference type, then this step requires an averaging of the noise effects so that the “mean dynamics” appear. Suppose that $(\theta(\cdot), Z(\cdot))$ is the weak sense limit of a weakly convergent subsequence. A particularly useful way of doing both the averaging under weak conditions and identifying the limit process involves showing that $\theta(t) - \theta(0) - \int_0^t \bar{g}(\theta(s)) ds - Z(t)$ is a martingale with Lipschitz continuous paths. Recall the fact (Section 4.1) that any continuous-time martingale with Lipschitz continuous paths (with probability one) is a constant (with probability one). The Lipschitz continuity will be easy to prove. Then the martingale property implies that the expression is a constant. Since it takes the value zero at $t = 0$, the limit process satisfies the desired ODE. A convenient criterion for showing that a process is a martingale is needed, and a useful approach is suggested by the definition of a martingale in terms of conditional expectations.

Let Y be a random variable with $E|Y| < \infty$, and let $\{V(s), 0 \leq s < \infty\}$, be an arbitrary sequence of random variables. Suppose that for fixed real $t > 0$, each integer p and each set of real numbers $0 \leq s_i \leq t, i = 1, \dots, p$, and each bounded and continuous real-valued function $h(\cdot)$, we have

$$Eh(V(s_i), i \leq p)Y = 0.$$

This and the arbitrariness of $h(\cdot)$ imply that $E[Y|V(s_i), i \leq p] = 0$. The arbitrariness of p and $\{s_i, i \leq p\}$ now imply that

$$E[Y|V(s), s \leq t] = 0$$

with probability one [34]. To extend this idea, let $U(\cdot)$ be a random process with paths in $D^r[0, \infty)$ such that for all $p, h(\cdot), s_i \leq t, i \leq p$, as given above and a given real $\tau > 0$,

$$Eh(U(s_i), i \leq p) [U(t + \tau) - U(t)] = 0. \quad (4.1)$$

Then $E[U(t + \tau) - U(t)|U(s), s \leq t] = 0$. If this holds for all t and $\tau > 0$ then, by the definition of a martingale, $U(\cdot)$ is a martingale. Sometimes it is convenient to work with the following more general format whose proof follows from the preceding argument. The suggested approach is a standard and effective method for verifying that a process is a martingale.

Theorem 4.1. *Let $U(\cdot)$ be a random process with paths in $D^r[0, \infty)$, where $U(t)$ is measurable on the σ -algebra \mathcal{F}_t^V determined by $\{V(s), s \leq t\}$ for some given process $V(\cdot)$ and let $E|U(t)| < \infty$ for each t . Suppose that for each real $t \geq 0$ and $\tau \geq 0$, each integer p and each set of real numbers $s_i \leq t, i = 1, \dots, p$, and each bounded and continuous real-valued function $h(\cdot)$,*

$$Eh(V(s_i), i \leq p) [U(t + \tau) - U(t)] = 0, \quad (4.2)$$

then $U(t)$ is an \mathcal{F}_t^V -martingale.

7.4.2 The Wiener Process

One of the most important martingales in applications is the Wiener process. Theorem 4.1.2 gave a criterion for verifying that a process is a Wiener process, and we now repeat and elaborate it. Let $W(\cdot)$ be an \mathbb{R}^r -valued process with continuous paths such that $W(0) = 0$, $EW(t) = 0$, for any set of increasing real numbers $\{t_i\}$, the set $\{W(t_{i+1}) - W(t_i)\}$ is mutually independent, and the distribution of $W(t+s) - W(t)$, $s > 0$ does not depend on t . Then $W(\cdot)$ is called a vector-valued *Wiener process* or *Brownian motion*. There is a matrix Σ , called the covariance, such that $EW(t)W'(t) = \Sigma t$, and the increments are normally distributed [34].

There are other equivalent definitions; one will now be given. Let the \mathbb{R}^r -valued process $W(\cdot)$ have continuous paths and satisfy $W(0) = 0$ w.p.1. Let \mathcal{F}_t be a sequence of nondecreasing σ -algebras such that $W(t)$ is \mathcal{F}_t -measurable and let $E_{\mathcal{F}_t}[W(t+s) - W(t)] = 0$ with probability one for each t and each $s \geq 0$. Let there be a non-negative definite matrix Σ such that for each t and each $s \geq 0$

$$E_{\mathcal{F}_t} [W(t+s) - W(t)] [W(t+s) - W(t)]' = \Sigma s \text{ w.p.1.}$$

Then $W(\cdot)$ is a vector-valued Wiener process with covariance Σ , also called an \mathcal{F}_t -Wiener process [163, Volume 1, Theorem 4.1].

The criterion of Theorem 4.1 for verifying that a process is a martingale can be adapted to verify that it is a vector-valued \mathcal{F}_t -Wiener process for appropriate \mathcal{F}_t . Suppose that $W(\cdot)$ is a continuous vector-valued process with $E|W(t)|^2 < \infty$ for each t . Let $V(\cdot)$ be a random process and let \mathcal{F}_t^V be the smallest σ -algebra that measures $\{V(s), W(s), s \leq t\}$. Let $h(\cdot), p, t, \tau > 0, s_i \leq t$ be arbitrary but satisfy the conditions put on these quantities in Theorem 4.1. Suppose that

$$Eh(V(s_i), W(s_i), i \leq p) [W(t+\tau) - W(t)] = 0 \quad (4.3)$$

and that there is a non-negative definite matrix Σ such that

$$\begin{aligned} & Eh(V(s_i), W(s_i), i \leq p) \\ & \times [[W(t+\tau) - W(t)] [W(t+\tau) - W(t)]' - \Sigma\tau] = 0. \end{aligned} \quad (4.4)$$

Then $W(\cdot)$ is an \mathcal{F}_t^V -Wiener process, with covariance Σ .

Proving that (4.4) holds for the weak sense limit of a sequence of processes $\{W^n(t)\}$ might require showing that $\{|W^n(t)|^2\}$ is uniformly integrable. This can be avoided by using the following equivalent characterization.

For a matrix $\Sigma = \{\sigma_{ij}\}$, let A_Σ denote the operator acting on twice continuously differentiable real-valued functions $F(\cdot)$ on \mathbb{R}^r :

$$A_\Sigma F(w) = \frac{1}{2} \sum_{i,j} \sigma_{ij} \frac{\partial^2 F(w)}{\partial w_i \partial w_j}. \quad (4.5)$$

Theorem 4.2. *Let $F(\cdot)$ be an arbitrary continuous real-valued function on \mathbb{R}^r with compact support and whose mixed partial derivatives up to order three are continuous and bounded. Let $V(\cdot)$ be a random process. Let the \mathbb{R}^r -valued process $W(\cdot)$ have continuous paths with probability one and $\Sigma = \{\sigma_{ij}\}$ a non-negative definite symmetric matrix. Suppose that for each real $t \geq 0$ and $\tau \geq 0$, each integer p and each set of real numbers $s_i \leq t, i = 1, \dots, m$, each bounded and continuous real-valued function $h(\cdot)$,*

$$\begin{aligned} & Eh(V(s_i), W(s_i), i \leq p) \\ & \times \left[F(W(t+\tau)) - F(W(t)) - \int_t^{t+\tau} A_\Sigma F(W(u)) du \right] = 0. \end{aligned} \quad (4.6)$$

Then $W(\cdot)$ is an \mathcal{F}_t^V -Wiener process with covariance Σ , where \mathcal{F}_t^V is the smallest σ -algebra that measures $\{V(s), W(s), s \leq t\}$.

7.4.3 A Perturbed Test Function Method for Verifying Tightness and Verifying the Wiener Process

In Chapters 5 and 6, we have seen the usefulness of perturbed Liapunov functions and perturbed state methods for proving stability or for averaging correlated noise. Perturbed test function methods are also very useful for proving tightness or characterizing the limit of a weakly convergent sequence. The original perturbed test function ideas stem from the work of Blankenship and Papanicolaou [26], Papanicolaou, Stroock, and Varadhan [187], and Kurtz [117]. Kushner extended them to cover quite general non-Markovian situations and developed powerful techniques for their construction, exploitation and applications to diverse problems; see for example, [127, 132]; see also the remarks on perturbations in Subsection 6.3.1.

In the following perturbed test function theorems, ϵ_i^n are positive real numbers and $\tau_i^n = \sum_{j=0}^{i-1} \epsilon_j^n$. The \mathbb{R}^r -valued processes $X^n(\cdot)$ are constant on the intervals $[\tau_i^n, \tau_{i+1}^n)$ and are right continuous. Define $m^n(t) = \max\{i : \tau_i^n \leq t\}$. For each n , let \mathcal{F}_i^n be a sequence of nondecreasing σ -algebras such that \mathcal{F}_i^n measures at least $\{X^n(\tau_j^n), j \leq i\}$, and let E_i^n denote the expectation conditioned on \mathcal{F}_i^n . Let \mathcal{D}^n denote the class of right continuous, real-valued random functions $F(\cdot)$ that are constant on the intervals $[\tau_i^n, \tau_{i+1}^n)$, with bounded expectation for each t , and that $F(\tau_i^n)$ is \mathcal{F}_i^n -measurable. Define the operator \hat{A}^n acting on random functions $F(\cdot)$ in \mathcal{D}^n by

$$\hat{A}^n F(\tau_i^n) = \frac{E_i^n F(\tau_{i+1}^n) - F(\tau_i^n)}{\epsilon_i^n}. \quad (4.7)$$

The next theorem is an extension of Theorem 3.5.

Theorem 4.3. [127, Theorems 4 and 8; Chapter 3] *For each real-valued function $F(\cdot)$ on \mathbb{R}^r with compact support and whose mixed partial deriva-*

tives up to second order are continuous, let there be a sequence of processes $F^n(\cdot) \in \mathcal{D}^n$ such that for each $\alpha > 0$ and $T > 0$,

$$\lim_n P \left\{ \sup_{s \leq T} |F^n(s) - F(X^n(s))| \geq \alpha \right\} = 0, \quad (4.8)$$

and suppose that

$$\lim_{N \rightarrow \infty} \sup_n P \left\{ \sup_{t \leq T} |X^n(t)| \geq N \right\} = 0. \quad (4.9)$$

Define γ_i^n by

$$E_i^n F^n(\tau_{i+1}^n) - F^n(\tau_i^n) = \epsilon_i^n \gamma_i^n = \epsilon_i^n \hat{A}^n F^n(\tau_i^n). \quad (4.10)$$

If $\{\gamma_i^n; n, i : \tau_i^n \leq T\}$ is uniformly integrable for each T and $F(\cdot)$, then $\{F(X^n(\cdot))\}$ is tight in $D[0, \infty)$, and $\{X^n(\cdot)\}$ is tight in $D^r[0, \infty)$. The analogous result holds on $D^r(-\infty, \infty)$. If, in addition, for each $T > 0$, $E|\gamma_i^n| \rightarrow 0$ uniformly in $\{i : \tau_i^n \leq T\}$ as $n \rightarrow \infty$, then the weak sense limit is the “zero” process.

Theorem 4.4. Let $X^n(0) = 0$, and suppose that $\{X^n(\cdot)\}$ is tight in $D^r[0, \infty)$ and that each of the weak sense limit processes has continuous paths with probability one. Let

$$\limsup_n \sup_i \epsilon_i^n = 0. \quad (4.11a)$$

Suppose that there are integers μ_i^n such that $\lim_n \inf_i \mu_i^n = \infty$ with the following properties:

(a)

$$\limsup_n \sup_i \sum_{j=i}^{i+\mu_i^n-1} \epsilon_i^n = 0, \quad \lim_n \sup_{i+\mu_i^n \geq j \geq i} \left| \frac{\epsilon_j^n - \epsilon_i^n}{\epsilon_i^n} \right| = 0; \quad (4.11b)$$

(b) for each continuous real-valued function $F(\cdot)$ on \mathbb{R}^r with compact support and whose mixed partial derivatives up to order three are continuous, and for each $T > 0$, there is an $F^n(\cdot)$ in \mathcal{D}^n such that

$$\lim_n E |F^n(t) - F(X^n(t))| = 0, \quad t \leq T, \quad (4.12)$$

$$\sup_{t \leq T} E |\hat{A}^n F^n(t)| < \infty, \quad \text{each } n; \quad (4.13)$$

(c) for the non-negative definite matrix $\Sigma = \{\sigma_{ij}\}$

$$\lim_n E \left| \frac{1}{\mu_i^n} \sum_{j=i}^{i+\mu_i^n-1} E_i^n \left[\hat{A}^n F^n(\tau_j^n) - A_\Sigma F(X^n(\tau_j^n)) \right] \right| = 0, \quad (4.14)$$

where the limit is taken on uniformly in i for $\tau_i^n \leq T$, and each $T > 0$, and A_Σ is defined in (4.5).

Then $X^n(\cdot)$ converges weakly in $D^r[0, \infty)$ to the Wiener process with covariance Σ .

Let the set $\{V^n(\cdot)\}$ also be tight in $D^r[0, \infty)$, where $V^n(\cdot)$ is constant on the intervals $[\tau_i^n, \tau_{i+1}^n)$, and $\{V^n(\cdot), X^n(\cdot)\}$ converges weakly to $(V(\cdot), X(\cdot))$. Let (4.12)–(4.14) continue to hold, where \mathcal{F}_i^n measures at least $\{X^n(\tau_j^n), V^n(\tau_j^n), j \leq i\}$. If \mathcal{F}_t^V is the smallest σ -algebra that measures $\{V(s), X(s), s \leq t\}$, then $X(\cdot)$ is an \mathcal{F}_t^V -Wiener process. The analogous result holds on $(-\infty, \infty)$.

Let there be a continuous function $\bar{g}(\cdot)$ such that the conditions hold with $A_\Sigma F(x)$ replaced by $A_\Sigma F(x) + F'_x(x)\bar{g}(x)$. Then the limit $X(\cdot)$ of any weakly convergent subsequence of $\{X^n(\cdot)\}$ can be characterized as follows: There is a Wiener process $W(\cdot)$ with covariance matrix Σ such that $X(\cdot)$ satisfies the stochastic differential equation

$$X(t) = X(0) + \int_0^t \bar{g}(X(s))ds + W(t), \quad (4.15)$$

where for each t , $\{X(s), s \leq t\}$ and $\{W(s) - W(t), s \geq t\}$ are independent.

Note on the proof. This is actually an adaptation of [127, Theorem 8, Chapter 5] to the decreasing-step-size case. In the proof, one needs to show that for each $t \geq 0$ and small $\tau > 0$, with $t + \tau \leq T$,

$$E_{m^n(t)}^n \sum_{j=m^n(t)}^{m^n(t+\tau)-1} \epsilon_j^n \left[\hat{A}^n F^n(\tau_j^n) - A_\Sigma F(X^n(\tau_j^n)) \right] \rightarrow 0$$

in mean as $n \rightarrow \infty$. By the conditions on ϵ_i^n , this is implied by (4.14).

The Skorohod topology. Let Λ_T denote the space of continuous and strictly increasing functions from the interval $[0, T]$ onto the interval $[0, T]$. The functions in this set will be “allowable time scale distortions” for the functions in $D[0, T]$. Define the metric $d_T(\cdot)$ by

$$d_T(f(\cdot), g(\cdot)) = \inf \left\{ \mu : \sup_{0 \leq s \leq T} |s - \lambda(s)| \leq \mu \text{ and } \sup_{0 \leq s \leq T} |f(s) - g(\lambda(s))| \leq \mu \text{ for some } \lambda(\cdot) \in \Lambda_T \right\}.$$

If there are $\eta_n \rightarrow 0$ such that the discontinuities of $f_n(\cdot)$ are less than η_n in magnitude and if $f_n(\cdot) \rightarrow f(\cdot)$ in $d_T(\cdot)$, then the convergence is uniform on $[0, T]$ and $f(\cdot)$ must be continuous. Because of the “time scale distortion” involved in the definition of the metric $d_T(\cdot)$, we can have (loosely speaking) convergence of a sequence of discontinuous functions, where there are only a finite number of discontinuities, where both the locations and the values of the discontinuities converge, and a type of “equicontinuity” condition holds between the discontinuities. For example, let $T > 1$ and define $f(\cdot)$

by: $f(t) = 1$ for $t < 1$ and $f(t) = 0$ for $t \geq 1$. Define the function $f_n(\cdot)$ by $f_n(t) = 1$ for $t < 1 + 1/n$ and $f_n(t) = 0$ for $t \geq 1 + 1/n$. Then $f_n(\cdot)$ converges to $f(\cdot)$ in the Skorohod topology, but not in the sup norm, and $d_T(f(\cdot), f_n(\cdot)) = 1/n$. The minimizing time scale distortion is illustrated in Figures 4.1 and 4.2.

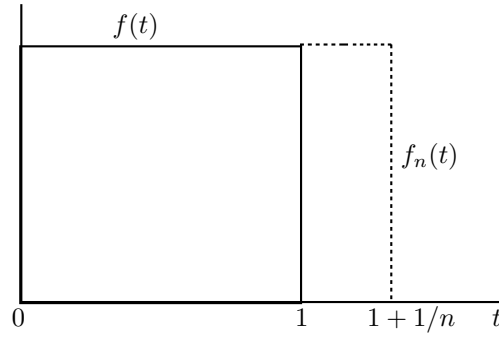


Figure 4.1. The functions $f(\cdot)$ and $f_n(\cdot)$.

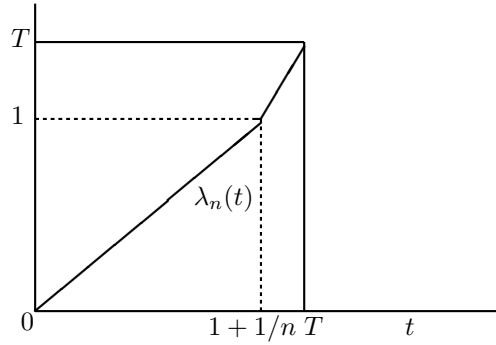


Figure 4.2. The time scale distortion $\lambda_n(\cdot)$.

Under $d_T(\cdot)$, the metric space $D[0, T]$ is separable but not complete [25, 68]. There is an equivalent metric $\hat{d}_T(\cdot)$ under which the space is both complete and separable. The $\hat{d}_T(\cdot)$ weights the “derivative” of the time scale changes $\lambda(t)$ and its deviation from t . For $\lambda(\cdot) \in \Lambda_T$, define

$$|\lambda| = \sup_{s < t < T} \left| \log \left\{ \frac{\lambda(t) - \lambda(s)}{t - s} \right\} \right|.$$

The metric $\widehat{d}_T(\cdot)$ is defined by

$$\widehat{d}_T(f(\cdot), g(\cdot)) = \inf \left\{ \mu : |\lambda| \leq \mu \text{ and } \sup_{0 \leq s \leq T} |f(s) - g(\lambda(s))| \leq \mu, \right. \\ \left. \text{for some } \lambda(\cdot) \in \Lambda_T \right\}.$$

Both $d_T(\cdot)$ and $\widehat{d}_T(\cdot)$ are referred to as *Skorohod metrics*. The topology under $\widehat{d}_T(\cdot)$ is called the *Skorohod topology*.

On the space $D[0, \infty)$, the metric for the Skorohod topology is defined by

$$\widehat{d}(f(\cdot), g(\cdot)) = \int_0^\infty e^{-t} \min(1, \widehat{d}_t(f(\cdot), g(\cdot))) dt,$$

and analogously on $D(-\infty, \infty)$. The metrics on the product spaces $D^r[0, \infty)$ and $D^r(-\infty, \infty)$ can be taken to be the sum of the metrics on the component spaces.



<http://www.springer.com/978-0-387-00894-3>

Stochastic Approximation and Recursive Algorithms and
Applications

Kushner, H.; Yin, G.

2003, XXII, 478 p., Hardcover

ISBN: 978-0-387-00894-3