5

Convergence with Probability One: Martingale Difference Noise

5.0 Outline of Chapter

Much of the classical work in stochastic approximation dealt with the situation where the "noise" in each observation Y_n is a martingale difference, that is, where there is a function $g_n(\cdot)$ of θ such that $E[Y_n|Y_i, i < n, \theta_0] = g_n(\theta_n)$ [17, 40, 45, 47, 56, 79, 86, 132, 154, 159, 169, 181]. Then we can write $Y_n = g_n(\theta_n) + \delta M_n$, where δM_n is a martingale difference. This "martingale difference noise" model is still of considerable importance. It arises, for example, where Y_n has the form $Y_n = F_n(\theta_n, \psi_n)$ where ψ_n are mutually independent. The convergence theory is relatively easy in this case, because the noise terms can be dealt with by well-known and relatively simple probability inequalities for martingale sequences. This chapter is devoted to this martingale difference noise case. Nevertheless, the ODE, compactness, and stability techniques to be introduced are of basic importance for stochastic approximation, and will be used in subsequent chapters.

A number of definitions that will be used throughout the book are introduced in Section 1. In particular, the general "ODE" techniques used in the rest of the book are based on the analysis of continuous time interpolations of the stochastic approximation sequence. These interpolations are defined in Section 1. The general development in the book follows intuitively reasonable paths but cannot be readily understood unless the definitions of the interpolated processes are understood.

Section 2 gives a fundamental convergence theorem and shows how the stochastic approximation sequence is related to a "mean limit" ODE that

characterizes the asymptotic behavior. The Arzelà Ascoli theorem is crucial to getting the ODE since it guarantees that there will always be convergent subsequences of the set of interpolated processes. The limits of any of these subsequences will satisfy the "mean limit" ODE. The first theorem (Theorem 2.1) uses a simple constraint set to get a relatively simple proof and allows us to concentrate on the essential structure of the "ODE method"-type proofs. This constraint set is generalized in Theorem 2.3, where a method for characterizing the reflection terms is developed, which will be used throughout the book. All the results carry over to the case where the constraint set is a smooth manifold of any dimension.

The conditions used for the theorems in Section 2 are more or less classical. For example, square summability of the step sizes ϵ_n is assumed. The square summability, together with the martingale noise property and a stability argument, can be used to get a simpler proof if the algorithm is unconstrained. However, the type of proof given readily generalizes to one under much weaker conditions. The set to which the iterates converge is a limit or invariant set for the mean limit ODE. These limit or invariant sets might be too large in that the convergence can only be to a subset. Theorem 2.5 shows that the only points in the limit or invariant set that we need to consider are the "chain recurrent" points, an idea due to Benaim [6].

The conditions are weakened in Subsection 3.1, which presents the "final form" of the martingale difference noise case in terms of conditions that require the "asymptotic rates of change" of certain random sequences to be zero with probability one. These conditions are satisfied by the classical case of Section 2. They are phrased somewhat abstractly but are shown to hold under rather weak and easily verifiable conditions in Subsection 3.2. Indeed, these "growth rate" conditions seem to be nearly minimal for convergence, and they hold even for "very slowly" decreasing step sizes. The conditions have been proved to be necessary in certain cases. The essential techniques of this chapter originated in [99].

A stability method for getting convergence, when there are no *a priori* bounds on the iterates, is in Section 4. A stochastic Liapunov function method is used to prove recurrence of the iterates, and then the ODE method takes over in the final stage of the proof. This gives a more general result than one might obtain with a stability method alone and is more easily generalizable. Section 5 concerns "soft" constraints, where bounds on functionals of the iterate are introduced into the algorithm via a penalty function. The results in Section 6 on the random directions Kiefer Wolfowitz method and on the minimization of convex functions are suggestive of additional applications. Section 7 gives the proof of convergence for the "lizard learning" problem of Section 2.1 and the pattern classification problem of Section 1.1. When using stochastic approximation for function minimization, where the function has more than one local minimum, one would like to assure at least that convergence to other types of station-

ary points (such as local maxima or saddles) is impossible. One expects that the noise in the algorithm will destabilize the algorithm around these "undesirable" points. This is shown to be the case for a slightly perturbed algorithm in Section 8.

5.1 Truncated Algorithms: Introduction

To develop the basic concepts behind the convergence theory in a reasonably intuitive way, we will first work with a relatively simple form and then systematically generalize it.

An important issue in applications of stochastic approximation concerns the procedure to follow if the iterates become too large. Practical algorithms tend to deal with this problem via appropriate adjustments to the basic algorithm, but these are often ignored in the mathematical developments, which tend to allow unbounded iterate sequences and put various "stability" conditions on the problem. However, even if these stability conditions do hold in practice, samples of the iterate sequence might get large enough to cause concern. The appropriate procedures to follow when the parameter value becomes large is, of course, dependent on the particular problem and the form of the algorithm that has been chosen, and it is unfortunate that there are no perfectly general rules to which one can appeal. Nevertheless, the useful parameter values in properly parameterized practical problems are usually confined by constraints of physics or economics to some compact set. This might be given by hard physical constraint that requires that, say, a dosage be less than a certain number of milligrams or a temperature set point in a computer simulation of a chemical process be less than 200°C. There are also implicit bounds in most problems. If θ_n is the set point temperature in a chemical processor and it reaches the temperature at the interior of the sun, or if the cost of setting the parameter at θ_n reaches the U.S. gross national product, then something is very likely wrong with the model or with the algorithm or with both. The models used in simulations are often inaccurate representations of physical reality at excessive values of the parameter (or of the noise), and so a mathematical development that does not carefully account for the changes in the model as the parameter (and the noise) values go to infinity might well be assuming much more than is justified. The possibility of excessive values of θ_n is a problem unique to computer simulations, because any algorithm that is used on a physical process would be carefully controlled.

Excessively large values of θ_n might simply be a consequence of poor choices for the algorithm structure. For example, instability can be caused by values of ϵ_n that are too large or values of finite difference intervals that are too small. The path must be checked for undesirable behavior, whether or not there are hard constraints. If the algorithm appears to be unstable,

then one could reduce the step size and restart at an appropriate point or even reduce the size of the constraint set. The path behavior might suggest a better algorithm or a better way of estimating derivatives. Conversely, if the path moves too slowly, we might wish to increase the step sizes. If the problem is based on a simulation, one might need to use a cruder model, with perhaps fewer parameters and a more restricted constraint set, to get a rough estimate of the location of the important values of the parameters. Even hard constraints are often somewhat "flexible," in that they might be intended as rough guides of the bounds, so that if the iterate sequence "hugs" a bounding surface, one might try to slowly increase the bounds, or perhaps to test the behavior via another simulation. In practice, there is generally an upper bound, beyond which the user will not allow the iterate sequence to go. At this point, either the iterate will be truncated in some way by the rules of the algorithm or there will be external intervention.

Much of the book is concerned with projected or truncated algorithms, where the iterate θ_n is confined to some bounded set, because this is a common practice in applications. Allowing unboundedness can lead to needless mathematical complications because some sort of stability must be shown or otherwise assumed, with perhaps artificial assumptions introduced on the behavior at large parameter values, and it generally adds little to the understanding of practical algorithms.

Many practical variations of the constraints can be used if the user believes they will speed convergence. For example, if the iterate leaves the constraint set, then the projection need not be done immediately. One can wait several iterates. Also, larger step sizes can be used near the boundary, if desired.

Throughout the book, the step size sequence will satisfy the fundamental condition

$$\sum_{n=0}^{\infty} \epsilon_n = \infty, \ \epsilon_n \geq 0, \ \epsilon_n \to 0, \quad \text{for } n \geq 0; \quad \epsilon_n = 0, \ \text{for } n < 0. \tag{1.1}$$

When *random* ϵ_n are used, it will always be supposed that (1.1) holds *with probability one*. Let $Y_n = (Y_{n,1}, \ldots, Y_{n,r})$ denote the $\mathbb{R}^r$-valued "observation" at time n, with the real-valued components $Y_{n,i}$.

Many of the proofs are based on the ideas in [99]. To facilitate understanding of these ideas, in Section 2 we start with conditions that are stronger than needed, and weaken them subsequently. The basic interpolations and time scalings will also be used in the subsequent chapters. In Theorem 2.1, we let the ith component of the state θ_n be confined to the interval $[a_i, b_i]$, where $-\infty < a_i < b_i < \infty$. Then the algorithm is

$$\theta_{n+1,i} = \Pi_{[a_i,b_i]}\left[\theta_{n,i} + \epsilon_n Y_{n,i}\right], \quad i = 1, \ldots, r. \tag{1.2}$$

We will write this in vector notation as

$$\theta_{n+1} = \Pi_H \left[\theta_n + \epsilon_n Y_n\right], \tag{1.3}$$

where Π_H is the projection onto the constraint set $H = \{\theta : a_i \leq \theta^i \leq b_i\}$. Define the *projection* or "correction" term Z_n by writing (1.3) as

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n + \epsilon_n Z_n. \tag{1.4}$$

Thus $\epsilon_n Z_n = \theta_{n+1} - \theta_n - \epsilon_n Y_n$; it is the vector of shortest Euclidean length needed to take $\theta_n + \epsilon_n Y_n$ back to the constraint set H if it is not in H.

To get a geometric feeling for the Z_n terms, refer to Figures 1.1 and 1.2. In situations such as Figure 1.1, where only one component is being truncated, Z_n points inward and is orthogonal to the boundary at θ_{n+1}. If more than one component needs to be truncated, as in Figure 1.2, Z_n again points inward but toward the corner, and it is proportional to a convex combination of the inward normals at the faces that border on that corner. In both cases, $Z_n \in -C(\theta_{n+1})$, where the cone $C(\theta)$ determined by the outer normals to the active constraint at θ was defined in Section 4.3.

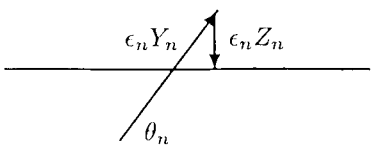

Figure 1.1. A projection with one violated constraint.

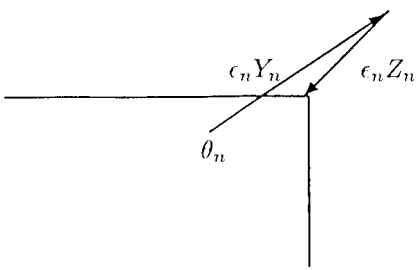

Figure 1.2. A projection with two violated constraints.

Martingale difference noise. In this chapter we will suppose that there are measurable functions $g_n(\cdot)$ of θ and random variables β_n such that Y_n can be decomposed as

$$Y_n = g_n(\theta_n) + \delta M_n + \beta_n, \quad \delta M_n = Y_n - E\left[Y_n | \theta_0, Y_i, i < n\right]. \tag{1.5}$$

The sequence $\{\beta_n\}$ will be "asymptotically negligible" in a sense to be defined. The sequence $\{\delta M_n\}$ is a martingale difference (with respect to the sequence of σ-algebras $\mathcal{F}_n$ generated by $\{\theta_0, Y_i, i < n\}$). The martingale difference assumption was used in the earliest work in stochastic approximation [17, 36, 40, 42, 45, 79, 132, 154]. Our proofs exploit the powerful ideas of the ODE methods stemming from the work of Ljung [119, 120] and Kushner [93, 99, 102]. In many of the applications of the Robbins–Monro or Kiefer–Wolfowitz algorithms, Y_n has the form $Y_n = F_n(\theta_n, \psi_n) + \beta_n$ where $\{\psi_n\}$ is a sequence of mutually independent random variables, $\{F_n(\cdot)\}$ is a sequence of measurable functions, $\beta_n \to 0$ and $E[F_n(\theta_n, \psi_n)|\theta_n = \theta] = g_n(\theta)$. For the Kiefer–Wolfowitz algorithm (see (1.3.1)–(1.3.4)), β_n represents the finite difference bias. The function $g_n(\cdot)$ might or might not depend on n. In the classical works on stochastic approximation, there was no n-dependence. The n-dependence occurs when the successive iterations are on different components of θ, the experimental procedure varies with n, or variance reduction methods are used, and so on. In the introductory result (Theorem 2.1), it will be supposed that $g_n(\cdot)$ is independent of n to simplify the development.

Definitions: Interpolated time scale and processes. The definitions and interpolations introduced in this section will be used heavily throughout the book. They are basic to the ODE method, and facilitate the effective exploitation of the time scale differences between the iterate process and the driving noise process. The ODE method uses a continuous time interpolation of the $\{\theta_n\}$ sequence. A natural time scale for the interpolation is defined in terms of the step size sequence. Define $t_0 = 0$ and $t_n = \sum_{i=0}^{n-1} \epsilon_i$. For $t \geq 0$, let $m(t)$ denote the unique value of n such that $t_n \leq t < t_{n+1}$. For $t < 0$, set $m(t) = 0$. Define the *continuous time interpolation* $\theta^0(\cdot)$ on $(-\infty, \infty)$ by $\theta^0(t) = \theta_0$ for $t \leq 0$, and for $t \geq 0$,

$$\theta^0(t) = \theta_n, \quad \text{for } t_n \leq t < t_{n+1}. \tag{1.6}$$

For later use, define the sequence of *shifted* processes $\theta^n(\cdot)$ by

$$\theta^n(t) = \theta^0(t_n + t), \quad t \in (-\infty, \infty). \tag{1.7}$$

Figures 1.3 and 1.4 illustrate the functions $m(\cdot)$, $m(t_n + \cdot)$, and interpolations $\theta^0(\cdot)$, and $\theta^n(\cdot)$.

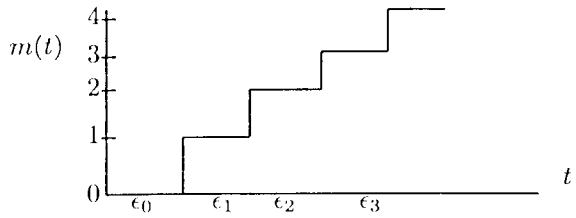

Figure 1.3a. The function $m(\cdot)$.

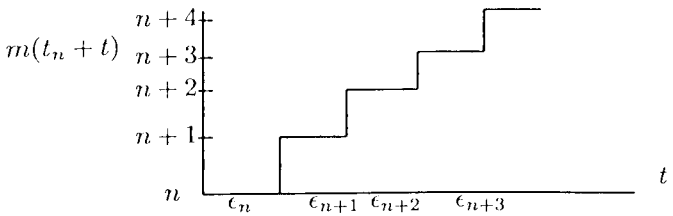

Figure 1.3b. The function $m(t_n + t)$.

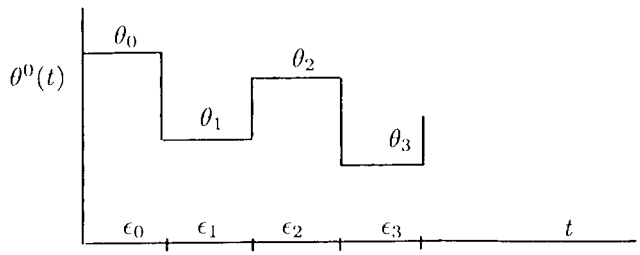

Figure 1.4a. The function $\theta^0(\cdot)$.

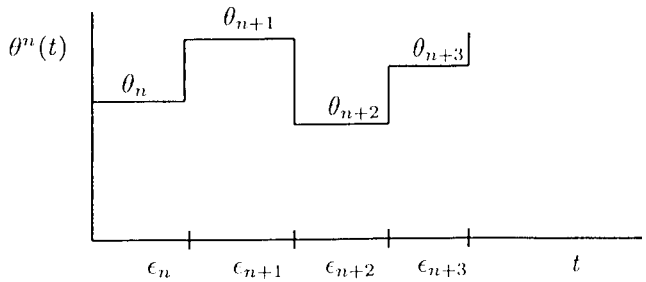

Figure 1.4b. The function $\theta^n(\cdot)$.

Let $Z_i = 0$ and $Y_i = 0$ for $i < 0$. Define $Z^0(t) = 0$ for $t \leq 0$ and

$$Z^0(t) = \sum_{i=0}^{m(t)-1} \epsilon_i Z_i, \quad t \geq 0.$$

Define $Z^n(\cdot)$ by

$$\begin{aligned} Z^n(t) &= Z^0(t_n + t) - Z^0(t_n) = \sum_{i=n}^{m(t_n+t)-1} \epsilon_i Z_i, \quad t \geq 0, \\ Z^n(t) &= -\sum_{i=m(t_n+t)}^{n-1} \epsilon_i Z_i, \quad t < 0. \end{aligned} \tag{1.8}$$

Define $Y^n(\cdot), M^n(\cdot)$, and $B^n(\cdot)$ analogously to $Z^n(\cdot)$ but using Y_i, δM_i, and β_i, resp., in lieu of Z_i. By the definitions (recall that $m(t_n) = n$)

$$\theta^n(t) = \theta_n + \sum_{i=n}^{m(t_n+t)-1} \epsilon_i [Y_i + Z_i] = \theta_n + Y^n(t) + Z^n(t), \quad t \geq 0, \tag{1.9a}$$

$$\theta^n(t) = \theta_n - \sum_{i=m(t_n+t)}^{n-1} \epsilon_i [Y_i + Z_i] = \theta_n + Y^n(t) + Z^n(t), \quad t < 0. \tag{1.9b}$$

For simplicity, we always write the algorithm as (1.9a), *whether* t *is positive or negative, with the understanding that it is to be interpreted as* (1.9b) *if* $t < 0$. All the above interpolation formulas will be used heavily in the sequel.

Note that the time origin of the "shifted" processes $\theta^n(\cdot)$ and $Z^n(\cdot)$ is time t_n for the original processes, the interpolated time at the nth iteration. The step sizes ϵ_n used in the interpolation are natural intervals for the continuous time interpolation. Their use allows us to exploit the time scale differences between the mean terms and the noise terms under quite general conditions. We are concerned with the behavior of the tail of the sequence $\{\theta_n\}$. Since this is equivalent to the behavior of $\theta^n(\cdot)$ over any finite interval for large n, a very effective method (introduced in [99]) of dealing with the tails works with these shifted processes $\theta^n(\cdot)$.

Note on piecewise linear vs. piecewise constant interpolations. The basic stochastic approximation (1.3) is defined as a discrete time process. We have defined the continuous time interpolations $\theta^n(\cdot), Z^n(\cdot)$ to be piecewise constant with interpolation intervals ϵ_n. We could have defined the interpolations to be piecewise linear in the obvious way by simply interpolating linearly between the "break" or "jump" points $\{t_n\}$. Nevertheless, there are some notational advantages to the piecewise constant interpolation. In the proofs in this chapter and Chapter 6, it is shown that for almost

sample paths the set $\{\theta^n(\omega,\cdot)\}$ is equi- (actually Lipschitz) continuous the extended sense (see Theorem 4.2.2). Thus, the set of piecewise linear terpolations is equicontinuous.

5.2 The ODE Method: A Basic Convergence Theorem

5.2.1 *Assumptions and the Main Convergence Theorem*

One way or another, all methods of analysis need to show that the "tail" effects of the noise vanish. This "tail" behavior is essentially due to the martingale difference property and the fact that the step sizes ϵ_n decrease to zero as $n \to \infty$.

Definition. Recall that L_H denotes the set of limit points of the mean limit ODE (2.1) in H, over all initial conditions where z is the minimum force needed to keep the solution in H. By invariant set, we always mean a two-sided invariant set; that is, if $x \in I$, an invariant set in H, then there is a path of the ODE in H on the time interval $(-\infty, \infty)$ that goes through x at time 0. If there is a constraint set, then the set of limit points might be smaller than the largest two-sided invariant set.

Definitions. Let E_n denote the expectation conditioned on the σ-algebra $\mathcal{F}_n$, generated by $\{\theta_0, Y_i, i < n\}$. When it is needed, the definition will be changed to a larger σ-algebra.

A set $A \subset H$ is said to be *locally asymptotically stable in the sense of Liapunov* for the ODE

$$\dot{\theta} = \bar{g}(\theta) + z, \quad z \in -C(\theta), \tag{2.1}$$

if for each $\delta > 0$ there is $\delta_1 > 0$ such that all trajectories starting in $N_{\delta_1}(A)$ never leave $N_\delta(A)$ and ultimately stay in $N_{\delta_1}(A)$.

Assumptions. The assumptions listed here, which will be used in Theorem 2.1, are more or less classical except for the generality of the possible limit set for the mean limit ODE and the use of a constraint set. All the conditions will be weakened in subsequent theorems. The proof is more complicated than the "minimal" convergence proof, since the algorithm is not necessarily of the gradient descent form and we do not insist that there be a unique limit point, but allow the algorithm to have a possibly complicated asymptotic behavior. Also, the proof introduces decompositions and interpolations that will be used in the sequel, as well as the basic idea of the ODE method for the probability one convergence. Condition (A2.2) simply sets up the notation, where β_n satisfies (A2.5). Condition (A2.6)

(and (A2.6′) as well) is intended to describe the limits of θ_n in terms of the limit points of the ODE. The motivating example is where the set L_H^1 in (A2.6) is either empty or a set of unstable or marginally stable limit points, and the set of remaining limit points, A_H, is asymptotically stable in the sense of Liapunov. The condition is not needed in the "gradient descent" case, and stronger results are obtained without it in Subsection 2.2, where it is shown that under essentially the other conditions of this subsection, the process converges to the subset of the limit points consisting of "chain recurrent" points, a natural set.

(A2.1) $\sup_n E|Y_n|^2 < \infty$.

(A2.2) There is a measurable function $\bar{g}(\cdot)$ of θ and random variables β_n such that

$$E_n Y_n = E\left[Y_n | \theta_0, Y_i, i < n\right] = g(\theta_n) + \beta_n.$$

(A2.3) $\bar{g}(\cdot)$ is continuous.

(A2.4) $\sum_i \epsilon_i^2 < \infty$.

(A2.5) $\sum_i \epsilon_i |\beta_i| < \infty$ w.p.1.

The following condition will sometimes be used.

(A2.6) Let L_H^1 be a subset of L_H and A_H a set that is locally asymptotically stable in the sense of Liapunov. Suppose that for any initial condition not in L_H^1 the trajectory of (2.1) goes to A_H.

For later use in this chapter, we will restate (A2.6) when a differential inclusion replaces the ODE.

(A2.6′) Let L_H now denote the (compact) set of limit points of the differential inclusion

$$\dot{\theta} \in G(\theta) + z, \quad z(t) \in -C(\theta(t)),$$

over all initial conditions in H, where z is the minimum force needed to keep the solution in H. Let L_H^1 be a subset of L_H and A_H a set that is locally asymptotically stable in the sense of Liapunov. Suppose that for any initial condition not in L_H^1 the trajectory goes to A_H.

Suppose that there is a continuously differentiable real-valued function $f(\cdot)$ such that $\bar{g}(\cdot) = -f_\theta(\cdot)$. Then the points in L_H are the stationary points, called S_H; they satisfy the stationarity condition

$$\bar{g}(\theta) + z = 0 \quad \text{for almost all } t, \quad z \in -C(\theta). \tag{2.2}$$

The set of stationary points can be divided into disjoint compact and connected subsets $S_i, i = 0, \ldots$ The following unrestrictive condition is needed.

(A2.7) Let $\bar{g}(\cdot) = -f_\theta(\cdot)$ for continuously differentiable real-valued $f(\cdot)$. Then $f(\cdot)$ is constant on each S_i.

If $f(\cdot)$ and the $q_i(\cdot)$ in (A4.3.2) (which define the constraint set) are twice continuously differentiable, then (A2.7) holds.

Comment on equality constraints and smooth manifolds. The equality constrained problem and the case where the constraint set H is a smooth manifold in $\mathbb{R}^{r-1}$ are covered by the results of the book. A convenient alternative approach that works directly on the manifold and effectively avoids the reflection terms can be seen from the following comments. The reader can fill in the explicit conditions that are needed. Suppose that the constraint set H is a smooth manifold. The algorithm $\theta_{n+1} = \Pi_H(\theta_n + \epsilon_n Y_n)$ can be written as

$$\theta_{n+1} = \theta_n + \epsilon_n \gamma(\theta_n) Y_n + \epsilon_n \beta_n,$$

where $\gamma(\cdot)$ is a smooth function and $\epsilon_n \gamma(\theta_n) Y_n$ is the projection of $\epsilon_n Y_n$ onto the orthogonal complement of the normal hyperplane (or line, depending on the case) to H at the point θ_n, and $\epsilon_n \beta_n$ represents the "error." Under reasonable conditions on the smoothness and on the sequence $\{Y_n\}$, the sequences $\{\gamma(\theta_n) Y_n, \beta_n\}$ will satisfy the conditions required by the $\{Y_n, \beta_n\}$ in the theorems. The mean limit ODE will be $\dot{\theta} = \gamma(\theta)\bar{g}(\theta)$. Similar comments hold when the ODE is replaced by a differential inclusion, for the correlated noise case of Chapter 6 and the various weak convergence cases of Chapters 7 and 8. The results can be extended to the case where H is the intersection of the $\mathbb{R}^{r-1}$-dimensional manifold defined by (A4.3.3) and a set satisfying (A4.3.2) or (A4.3.1).

Theorem 2.1. *Let* (1.1) *and* (A2.1)–(A2.5) *hold for algorithm* (1.3). *Then there is a set* N *of probability zero such that for* $\omega \notin N$, *the set of functions* $\{\theta^n(\omega, \cdot), Z^n(\omega, \cdot), n < \infty\}$ *is equicontinuous. Let* $(\theta(\omega, \cdot), Z(\omega, \cdot))$ *denote the limit of some convergent subsequence. Then this pair satisfies the projected ODE* (2.1), *and* $\{\theta_n(\omega)\}$ *converges to some invariant set of the ODE in* H. *If the constraint set is dropped, but* $\{\theta_n\}$ *is bounded with probability one, then for almost all* ω, *the limits* $\theta(\omega, \cdot)$ *of convergent subsequences of* $\{\theta^n(\omega, \cdot)\}$ *are trajectories of*

$$\dot{\theta} = \bar{g}(\theta) \tag{2.3}$$

in some bounded invariant set and $\{\theta_n(\omega)\}$ *converges to this invariant set. Let* p_n *be integer-valued functions of* ω, *not necessarily being stopping times or even measurable, but that go to infinity with probability one. Then the conclusions concerning the limits of* $\{\theta^n(\cdot)\}$ *hold with* p_n *replacing* n. *If* $\bar{\theta}$ *is an asymptotically stable point of* (2.1) *and* θ_n *is in some compact set in the domain of attraction of* $\bar{\theta}$ *infinitely often with probability* $\geq \rho$, *then* $\theta_n \to \bar{\theta}$ *with at least probability* ρ.

Assume (A2.6). *Then the limit points are in* $L_H^1 \cup A_H$ *with probability one.*

Suppose that (A2.7) *holds. Then, for almost all* ω, $\{\theta_n(\omega)\}$ *converges to a unique* S_i.

Remark. In many applications where $-\bar{g}(\cdot)$ is a gradient and the truncation bounds are large enough, there is only one stationary point of (2.1), and that is globally asymptotically stable. Then $\{\theta_n\}$ converges w.p.1 to that point. For simplicity, we use *equicontinuity* to mean "equicontinuity in the extended sense," as defined in the definition preceding Theorem 4.2.2.

Proof: Part 1. Convergence of the martingale and equicontinuity. Recall that $\delta M_n = Y_n - \bar{g}(\theta_n) - \beta_n$, and decompose the algorithm (1.3) as

$$\theta_{n+1} = \theta_n + \epsilon_n \bar{g}(\theta_n) + \epsilon_n Z_n + \epsilon_n \delta M_n + \epsilon_n \beta_n. \tag{2.4}$$

Then we can write

$$\begin{aligned} \theta^n(t) = \theta_n &+ \sum_{i=n}^{m(t+t_n)-1} \epsilon_i g(\theta_i) + \sum_{i=n}^{m(t+t_n)-1} \epsilon_i Z_i \\ &+ \sum_{i=n}^{m(t+t_n)-1} \epsilon_i \delta M_i + \sum_{i=n}^{m(t+t_n)-1} \epsilon_i \beta_i. \end{aligned} \tag{2.5}$$

Define $M_n = \sum_{i=0}^{n-1} \epsilon_i \delta M_i$. This is a martingale sequence (with associated σ-algebras $\mathcal{F}_n$), since we have centered the summands about their conditional expectations, given the "past." By (4.1.4), for each $\mu > 0$,

$$P\left\{\sup_{n \geq m} |M_n| \geq \mu\right\} \leq \frac{E\left|\sum_{i=m}^{n} \epsilon_i \delta M_i\right|^2}{\mu^2}.$$

By (A2.1), (A2.4), and the fact that $E\delta M_i \delta M_j' = 0$ for $i \neq j$, the right side is bounded above by $K \sum_{i=m}^{\infty} \epsilon_i^2$, for some constant K. Thus, for each $\mu > 0$,

$$\lim_m P\left\{\sup_{n \geq m} |M_n| \geq \mu\right\} = 0. \tag{2.6}$$

Since $\theta^n(\cdot)$ is piecewise constant, we can rewrite (2.5) as

$$\theta^n(t) = \theta_n + \int_0^t g(\theta^n(s))ds + Z^n(t) + M^n(t) + B^n(t) + \rho^n(t), \tag{2.7}$$

where $\rho^n(t)$ is due to the replacement of the first sum in (2.5) by an integral. $\rho^n(t) = 0$ at the times $t = t_k - t_n$ at which the interpolated processes have jumps, and it goes to zero uniformly in t as $n \to \infty$. By (2.6) and (A2.5), there is a null set N such that for $\omega \notin N$, $M^n(\omega, \cdot)$ and $B^n(\omega, \cdot)$ go to zero uniformly on each bounded interval in $(-\infty, \infty)$ as $n \to \infty$.

Let $\omega \notin N$. Then, $M^n(\omega, \cdot)$ and $B^n(\omega, \cdot)$ go to zero on any finite interval as $n \to \infty$. By the definition of N, for $\omega \notin N$ the functions of t on

the right side of (2.7) (except for $Z^n(\cdot)$) are equicontinuous in n and the limits of $M^n(\cdot)$, $B^n(\cdot)$, and $\rho^n(\cdot)$ are zero. It will next be shown that the equicontinuity of $\{Z^n(\omega,\cdot), n<\infty\}$ is a consequence of the fact that

$$Z_n(\omega) \in -C(\theta_{n+1}(\omega)). \tag{2.8}$$

For $\omega \notin N$, $\theta_{n+1}(\omega) - \theta_n(\omega) \to 0$. If $Z^n(\omega,\cdot)$ is not equicontinuous, then there is a subsequence that has a jump asymptotically; that is, there are integers $\mu_k \to \infty$, uniformly bounded times s_k, $0 < \delta_k \to 0$ and $\rho > 0$ (all depending on ω) such that

$$|Z^{\mu_k}(\omega, s_k+\delta_k) - Z^{\mu_k}(\omega, s_k)| \geq \rho.$$

The changes of the terms other than $Z^n(\omega,t)$ on the right side of (2.7) go to zero on the intervals $[s_k, s_k+\delta_k]$. Furthermore $\epsilon_n Y_n(\omega) = \epsilon_n \bar{g}(\theta_n(\omega)) + \epsilon_n \delta M_n(\omega) + \epsilon_n \beta_n \to 0$ and $Z_n(\omega) = 0$ if $\theta_{n+1}(\omega) \in H^0$, the interior of H. Thus, this jump cannot force the iterate to the interior of the hyperrectangle H, and it cannot force a jump of the $\theta^n(\omega,\cdot)$ along the boundary either. Consequently, $\{Z^n(\omega,\cdot)\}$ is equicontinuous.

Part 2. Characterizing the limit of a convergent subsequence: Applying the Arzelà–Ascoli Theorem. Let $\omega \notin N$, and let n_k denote a subsequence such that

$$\{\theta^{n_k}(\omega,\cdot), Z^{n_k}(\omega,\cdot)\}$$

converges, and denote the limit by $(\theta(\omega,\cdot), Z(\omega,\cdot))$. Then

$$\theta(\omega,t) = \theta(\omega,0) + \int_0^t \bar{g}(\theta(\omega,s))ds + Z(\omega,t). \tag{2.9}$$

Note that $Z(\omega,0) = 0$ and $\theta(\omega,t) \in H$ for all t. To characterize $Z(\omega,t)$, use (2.8) and the fact that $\theta_{n+1}(\omega) - \theta_n(\omega) \to 0$. These facts, together with the upper semicontinuity property (4.3.2) and the continuity of $\theta(\omega,\cdot)$, imply that (4.3.4) and (4.3.5) hold. In fact, it follows from the method of construction of $Z(\omega,\cdot)$ that the function simply serves to keep the dynamics $\bar{g}(\cdot)$ from forcing $\theta(\omega,\cdot)$ out of H. Thus, for $s>0$, $|Z(\omega,t+s) - Z(\omega,t)| \leq \int_t^{t+s} |\bar{g}(\theta(\omega,u)|du$. Hence $Z(\omega,\cdot)$ is Lipschitz continuous, and $Z(\omega,t) = \int_0^t z(\omega,s)ds$, where $z(\omega,t) \in -C(\theta(\omega,t))$ for almost all t.

Recall the definition of the set A_H in (A2.6). Suppose that $\{\theta_n(\omega)\}$ has a limit point $x_0 \notin L_H^1 \cup A_H$. Then there is a subsequence m_k such that $\theta^{m_k}(\omega,\cdot)$ converges to a solution of (2.1) with initial condition x_0. Let $\delta > \delta_1 > 0$ be arbitrarily small. Since the trajectory of (2.1) starting at x_0 ultimately enters $N_{\delta_1}(A_H)$ by (A2.6), $\theta_n(\omega)$ must be in $N_{\delta_1}(A_H)$ infinitely often. It will be seen that escape from $N_\delta(A_H)$ infinitely often is impossible. Suppose that escape from $N_{\delta_1}(A_H)$ occurs infinitely often. Then, since

Suppose that (A2.7) *holds. Then, for almost all* ω, $\{\theta_n(\omega)\}$ *converges to a unique* S_i.

Remark. In many applications where $-g(\cdot)$ is a gradient and the truncation bounds are large enough, there is only one stationary point of (2.1), and that is globally asymptotically stable. Then $\{\theta_n\}$ converges w.p.1 to that point. For simplicity, we use *equicontinuity* to mean "equicontinuity in the extended sense," as defined in the definition preceding Theorem 4.2.2.

Proof: Part 1. Convergence of the martingale and equicontinuity. Recall that $\delta M_n = Y_n - \bar{g}(\theta_n) - \beta_n$, and decompose the algorithm (1.3) as

$$\theta_{n+1} = \theta_n + \epsilon_n \bar{g}(\theta_n) + \epsilon_n Z_n + \epsilon_n \delta M_n + \epsilon_n \beta_n. \tag{2.4}$$

Then we can write

$$\begin{aligned} \theta^n(t) = \theta_n &+ \sum_{i=n}^{m(t+t_n)-1} \epsilon_i \bar{g}(\theta_i) + \sum_{i=n}^{m(t+t_n)-1} \epsilon_i Z_i \\ &+ \sum_{i=n}^{m(t+t_n)-1} \epsilon_i \delta M_i + \sum_{i=n}^{m(t+t_n)-1} \epsilon_i \beta_i. \end{aligned} \tag{2.5}$$

Define $M_n = \sum_{i=0}^{n-1} \epsilon_i \delta M_i$. This is a martingale sequence (with associated σ-algebras $\mathcal{F}_n$), since we have centered the summands about their conditional expectations, given the "past." By (4.1.4), for each $\mu > 0$,

$$P\left\{\sup_{n \geq m} |M_n| \geq \mu\right\} \leq \frac{E\left|\sum_{i=m}^{n} \epsilon_i \delta M_i\right|^2}{\mu^2}.$$

By (A2.1), (A2.4), and the fact that $E\delta M_i \delta M_j' = 0$ for $i \neq j$, the right side is bounded above by $K \sum_{i=m}^{\infty} \epsilon_i^2$, for some constant K. Thus, for each $\mu > 0$,

$$\lim_m P\left\{\sup_{n \geq m} |M_n| \geq \mu\right\} = 0. \tag{2.6}$$

Since $\theta^n(\cdot)$ is piecewise constant, we can rewrite (2.5) as

$$\theta^n(t) = \theta_n + \int_0^t \bar{g}(\theta^n(s))ds + Z^n(t) + M^n(t) + B^n(t) + \rho^n(t), \tag{2.7}$$

where $\rho^n(t)$ is due to the replacement of the first sum in (2.5) by an integral. $\rho^n(t) = 0$ at the times $t = t_k - t_n$ at which the interpolated processes have jumps, and it goes to zero uniformly in t as $n \to \infty$. By (2.6) and (A2.5), there is a null set N such that for $\omega \notin N$, $M^n(\omega, \cdot)$ and $B^n(\omega, \cdot)$ go to zero uniformly on each bounded interval in $(-\infty, \infty)$ as $n \to \infty$.

Let $\omega \notin N$. Then, $M^n(\omega, \cdot)$ and $B^n(\omega, \cdot)$ go to zero on any finite interval as $n \to \infty$. By the definition of N, for $\omega \notin N$ the functions of t on

the right side of (2.7) (except for $Z^n(\cdot)$) are equicontinuous in n and the limits of $M^n(\cdot)$, $B^n(\cdot)$, and $\rho^n(\cdot)$ are zero. It will next be shown that the equicontinuity of $\{Z^n(\omega,\cdot), n<\infty\}$ is a consequence of the fact that

$$Z_n(\omega) \in -C(\theta_{n+1}(\omega)). \tag{2.8}$$

For $\omega \notin N$, $\theta_{n+1}(\omega) - \theta_n(\omega) \to 0$. If $Z^n(\omega,\cdot)$ is not equicontinuous, then there is a subsequence that has a jump asymptotically; that is, there are integers $\mu_k \to \infty$, uniformly bounded times s_k, $0 < \delta_k \to 0$ and $\rho > 0$ (all depending on ω) such that

$$|Z^{\mu_k}(\omega, s_k+\delta_k) - Z^{\mu_k}(\omega, s_k)| \geq \rho.$$

The changes of the terms other than $Z^n(\omega,t)$ on the right side of (2.7) go to zero on the intervals $[s_k, s_k+\delta_k]$. Furthermore $\epsilon_n Y_n(\omega) = \epsilon_n \bar{g}(\theta_n(\omega)) + \epsilon_n \delta M_n(\omega) + \epsilon_n \beta_n \to 0$ and $Z_n(\omega) = 0$ if $\theta_{n+1}(\omega) \in H^0$, the interior of H. Thus, this jump cannot force the iterate to the interior of the hyperrectangle H, and it cannot force a jump of the $\theta^n(\omega,\cdot)$ along the boundary either. Consequently, $\{Z^n(\omega,\cdot)\}$ is equicontinuous.

Part 2. Characterizing the limit of a convergent subsequence: Applying the Arzelà–Ascoli Theorem. Let $\omega \notin N$, and let n_k denote a subsequence such that

$$\{\theta^{n_k}(\omega,\cdot), Z^{n_k}(\omega,\cdot)\}$$

converges, and denote the limit by $(\theta(\omega,\cdot), Z(\omega,\cdot))$. Then

$$\theta(\omega,t) = \theta(\omega,0) + \int_0^t \bar{g}(\theta(\omega,s))ds + Z(\omega,t). \tag{2.9}$$

Note that $Z(\omega,0) = 0$ and $\theta(\omega,t) \in H$ for all t. To characterize $Z(\omega,t)$, use (2.8) and the fact that $\theta_{n+1}(\omega) - \theta_n(\omega) \to 0$. These facts, together with the upper semicontinuity property (4.3.2) and the continuity of $\theta(\omega,\cdot)$, imply that (4.3.4) and (4.3.5) hold. In fact, it follows from the method of construction of $Z(\omega,\cdot)$ that the function simply serves to keep the dynamics $\bar{g}(\cdot)$ from forcing $\theta(\omega,\cdot)$ out of H. Thus, for $s>0$, $|Z(\omega,t+s) - Z(\omega,t)| \leq \int_t^{t+s} |\bar{g}(\theta(\omega,u)|du$. Hence $Z(\omega,\cdot)$ is Lipschitz continuous, and $Z(\omega,t) = \int_0^t z(\omega,s)ds$, where $z(\omega,t) \in -C(\theta(\omega,t))$ for almost all t.

Recall the definition of the set A_H in (A2.6). Suppose that $\{\theta_n(\omega)\}$ has a limit point $x_0 \notin L_H^1 \cup A_H$. Then there is a subsequence m_k such that $\theta^{m_k}(\omega,\cdot)$ converges to a solution of (2.1) with initial condition x_0. Let $\delta > \delta_1 > 0$ be arbitrarily small. Since the trajectory of (2.1) starting at x_0 ultimately enters $N_{\delta_1}(A_H)$ by (A2.6), $\theta_n(\omega)$ must be in $N_{\delta_1}(A_H)$ infinitely often. It will be seen that escape from $N_\delta(A_H)$ infinitely often is impossible. Suppose that escape from $N_{\delta_1}(A_H)$ occurs infinitely often. Then, since

$\theta_{n+1}(\omega) - \theta_n(\omega) \to 0$, there are integers n_k such that $\theta_{n_k}(\omega)$ converges to some point x_1 on $\partial N_{\delta_1}(A_H)$, and the path $\theta^{n_k}(\omega, \cdot)$ converges to a solution of (2.1) starting at x_1. The path of (2.1) starting at x_1 (whatever the chosen convergent subsequence) never leaves $N_\delta(A_H)$ and ultimately stays in $N_{\delta_1}(A_H)$. This implies that $\theta_n(\omega)$ cannot exit $N_\delta(A_H)$ infinitely often.

Whether or not there is a constraint set H, if boundedness with probability one of the sequence $\{\theta_n\}$ is assumed, then the preceding arguments show that (with probability one) the limits of $\{\theta^n(\omega, \cdot)\}$ are bounded solutions to (2.1) (which is (2.3) if there is no constraint) on the doubly infinite time interval $(-\infty, \infty)$. Thus the entire trajectory of a limit $\theta(\omega, \cdot)$ must lie in a bounded invariant set of (2.1) by the definition of an invariant set. The fact that $\{\theta_n(\omega)\}$ converges to some invariant set of (2.1) then follows; otherwise there would be a limit of a convergent subsequence satisfying (2.1) but not lying entirely in an invariant set.

These arguments do not depend on how the "sections" of $\theta^0(\omega, \cdot)$ are chosen. Any set of "sections" other than $\theta^n(\omega, \cdot)$ could have been used, as long as the initial times went to infinity. The statement of the theorem concerning $\{p_n\}$ then follows from what has been done.

Part 3. The case when $-\bar{g}(\cdot)$ is a gradient. Now assume (A2.7) and suppose that $\bar{g}(\cdot) = -f_\theta(\theta)$ for some continuously differentiable function $f(\cdot)$. As will be shown, the conclusion concerning the limits actually follows from what has been done.

We continue to work with $\omega \notin N$. Suppose for simplicity that there are only a finite number of S_i, namely, $S_0, \ldots, S_M$. In (2.1), $|z(t)| \leq |\bar{g}(\theta(t))|$. Thus, if $\bar{g}(\cdot) = -f_\theta(\cdot)$, the derivative along the solution of (2.1) at $\theta \in H$ is $f'_\theta(\theta)[-f_\theta(\theta) + z] \leq 0$, and we see that all solutions of (2.1) tend to the set of stationary points defined by (2.2). For each c, the set $\{\theta : f(\theta) \leq c\}$ is locally asymptotically stable in the sense of Liapunov, assuming that it is not empty. Then the previous part of the proof implies that $f(\theta_n(\omega))$ converges to some constant (perhaps depending on ω), and $\theta_n(\omega)$ converges to the set of stationary points.

It remains to be shown that $\{\theta_n(\omega)\}$ converges to a unique S_i. If the claimed convergence does not occur, the path will eventually oscillate back and forth between arbitrarily small neighborhoods of distinct S_i. This implies that there is a limit point outside the set of stationary points. □

An elaboration of the proof for the gradient descent case. For future use, and as an additional illustration of the ODE method, we will elaborate the proof for the case where $\bar{g}(\cdot)$ is a gradient. The ideas are just those used in the previous proof. The details to be given are of more general applicability and will be used in Theorems 4.2 and 4.3 in combination with a Liapunov function technique.

We start by supposing that the path $\theta^0(\omega, t)$ oscillates back and forth between arbitrarily small neighborhoods of distinct S_i. This will be seen

contradict the "gradient descent" property of the ODE (2.1). The proof simply sets up the notation required to formalize this idea.

Since $\{\theta_n(\omega)\}$ converges to $S_H = \cup_i S_i$, there is a subsequence m_k such that $\theta_{m_k}(\omega)$ tends to some point $x_0 \in S_H$. Suppose that $x_0 \in S_0$. We will show that $\theta_n(\omega) \to S_0$. Suppose that this last convergence hypothesis is false. Then there is an x_1 in some $S_i, i \neq 0$ (call it S_1 for specificity), and a subsequence $\{q_k\}$ such that $\theta_{q_k}(\omega) \to x_1$.

Continuing this process, let $S_0, \ldots, S_R, R > 0$, be all the sets that contain limit points of the sequence $\{\theta_n(\omega)\}$. Order the sets such that $f(S_R) = \liminf_n f(\theta_n(\omega))$, and suppose that S_R is the (assumed for simplicity) unique set on which the liminf is attained. The general (nonunique) case requires only a slight modification. Let $\delta > 0$ be such that $f(S_R) < f(S_i) - 2\delta, i \neq R$. For $\rho > 0$, define the ρ-neighborhood $N_\rho^f(S_R)$ of S_R by $N_\rho^f(S_R) = \{x : f(x) - f(S_R) < \rho\}$. By the definition of δ, $N_{2\delta}^f(S_R)$ contains no point in any $S_i, i \neq R$. By the hypothesis that more than one S_i contains limit points of $\{\theta_n(\omega)\}$, the neighborhood $N_\delta^f(S_R)$ of S_R is visited infinitely often and $N_{2\delta}^f(S_R)$ is exited infinitely often by $\{\theta_n(\omega)\}$. Thus there are $\nu_k \to \infty$ (depending on ω) such that $\theta_{\nu_k - 1}(\omega) \in N_\delta^f(S_R)$, $\theta_{\nu_k}(\omega) \notin N_\delta^f(S_R)$, and after time ν_k the path does not return to $N_\delta^f(S_R)$ until after it leaves $N_{2\delta}^f(S_R)$.

By the equicontinuity of $\{\theta^n(\omega, \cdot)\}$, there is a $T > 0$ such that $\theta_{\nu_k}(\omega) \to \partial N_\delta^f(S_R)$, the boundary of $N_\delta^f(S_R)$, and for large k, $\theta^{\nu_k}(\omega, t) \notin N_\delta^f(S_R)$ for $t \in [0, T]$.

There is a subsequence $\{\mu_m\}$ of $\{\nu_k\}$ such that $(\theta^{\mu_m}(\omega, \cdot), Z^{\mu_m}(\omega, \cdot))$ converges to some limit $(\bar{\theta}(\omega, \cdot), \bar{Z}(\omega, \cdot))$ that satisfies (2.1) with $\bar{\theta}(\omega, 0) \in \partial N_\delta^f(S_R)$, the boundary of $N_\delta^f(S_R)$, and $\bar{\theta}(\omega, t) \notin N_\delta^f(S_R)$ for $t \leq T$. This is a contradiction because, by the gradient descent property of (2.1) (with $g(\cdot) = -f_\theta(\cdot)$) and the definitions of δ and $N_\delta^f(S_R)$, any solution to (2.1) starting on $\partial N_\delta^f(S_R)$ must stay in $N_\delta^f(S_R)$ for all $t > 0$. □

The argument in Part 3 of the preceding proof is of more general use and leads to the following result.

Theorem 2.2. *Let $\{\theta^n(\omega, \cdot), Z^n(\omega, \cdot)\}$ be equicontinuous, with all limits satisfying (2.1). Suppose that $\{\theta_n(\omega)\}$ visits a set A_0 infinitely often, where A_0 is locally asymptotically stable in the sense of Liapunov. Then $\theta_n(\omega) \to A_0$, the closure of A_0.*

Remark on the proof of Theorem 2.1. Let us review the structure of the proof. First, the increment was partitioned to get the convenient representation (2.7), with which we could work on one part at a time. Then it was shown that with probability one the martingale term M_n converged, and this implied that $M^n(\cdot)$ converged with probability one to the "zero" process. Then the probability one convergence to zero of the bias $\{B^n(\cdot)\}$

was shown. The asymptotic continuity of $Z^n(\cdot)$ was obtained by a direct use of the properties of the Z_n as reflection terms. Then, by fixing ω not in some "bad" null set, and taking convergent subsequences of $\{\theta^n(\omega,\cdot), Z^n(\omega,\cdot)\}$, we were able to characterize the limit as a solution to the mean limit ODE. It then followed that the sequence $\{\theta_n(\omega)\}$ converged to some invariant set of the ODE.

A more general constraint set. Using the same basic structure of the proof, Theorem 2.1 can be readily generalized in several useful directions with little extra work. (A2.1) and (A2.4) will be weakened in the next section. Appropriate dependence on n of $\bar{g}(\cdot)$ can be allowed, and the hyperrectangle H can be replaced by a more general constraint set. The techniques involved in the required minor alterations in the proofs will be important in the analysis of the "dependent" noise case in Chapter 6.

In Theorem 2.3, the constraint form (A4.3.2) or (A4.3.3) will be used, where (A4.3.2) includes (A4.3.1). For $\theta \in I\!R^r$, let $\Pi_H(\theta)$ denote the closest point in H to θ. If the closest point is not unique, select a closest point such that the function $\Pi_H(\cdot)$ is measurable. We will work with the algorithm

$$\theta_{n+1} = \Pi_H\left[\theta_n + \epsilon_n Y_n\right], \tag{2.10a}$$

that will be written as

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n + \epsilon_n Z_n, \tag{2.10b}$$

where Z_n is the projection term. Recall the definition of $C(\theta)$ from Section 4.3. It follows from the calculus that $Z_n \in -C(\theta_{n+1})$, under (A4.3.3). Under (A4.3.2), this is proved by applying the Kuhn-Tucker Theorem of nonlinear programming to the problem of minimizing $|x - (\theta_n + \epsilon_n Y_n)|^2$ subject to the constraints $q_i(x) \le 0, i \le p$, where $x = \theta_{n+1}$. That theorem says that there are $\lambda_i \ge 0$ with $\lambda_i = 0$ if $q_i(x) < 0$ such that

$$(x - (\theta_n + \epsilon_n Y_n)) + \sum_i \lambda_i q_{i,x}(x) = 0,$$

which implies that $Z_n \in -C(\theta_{n+1})$.

The following assumption generalizes (A2.2) and will in turn be relaxed in the next section.

(A2.8) There are functions $g_n(\cdot)$ of θ, which are continuous uniformly in n, a continuous function $\bar{g}(\cdot)$ and random variables β_n such that

$$E_n Y_n = g_n(\theta_n) + \beta_n, \tag{2.11}$$

and for each $\theta \in H$

$$\lim_n \left| \sum_{i=n}^{m(t_n+t)} \epsilon_i \left[g_i(\theta) - \bar{g}(\theta)\right] \right| \to 0 \tag{2.12}$$

for each $t > 0$. (In other words, $\bar{g}(\cdot)$ is a "local average" of the functions $g_n(\cdot)$.)

Dependence of $\bar{g}(\cdot)$ on the past. Note that in all the algorithmic forms, $g_n(\theta_n)$ can be replaced by dependence on the past of the form $g_n(\theta_n, \ldots, \theta_{n-K})$ provided that the continuity of $g_n(\cdot)$ is replaced by the continuity in x of $g_n(x_0, \ldots, x_K)$ on the "diagonal" set $x = x_0 = \cdots = x_K$, uniformly in n.

Theorem 2.3. *Assume the conditions of Theorem* 2.1 *but use the algorithm* $\theta_{n+1} = \Pi_H[\theta_n + \epsilon_n Y_n]$ *with any of the constraint set conditions* (A4.3.1), (A4.3.2), *or* (A4.3.3) *holding, and* (A2.8) *with* $\beta_n \to 0$ *with probability one replacing* (A2.2) *and* (A2.5). *Then the conclusions of Theorem* 2.1 *continue to hold.*

Remark on the proof. The proof is essentially the same as that of Theorem 2.1, and we concentrate on the use of (A2.8) and the equicontinuity of $\{Z^n(\cdot)\}$ with probability one. The equicontinuity proof exploits the basic character of Z_n as *projection terms* to get the desired result, and the proof can readily be used for the general cases of Section 4 and Chapter 6.

Proof. Define

$$\bar{G}^n(t) = \sum_{i=n}^{m(t_n+t)-1} \epsilon_i \bar{g}(\theta_i), \quad \tilde{G}^n(t) = \sum_{i=n}^{m(t_n+t)-1} \epsilon_i \left[g_i(\theta_i) - \bar{g}(\theta_i)\right].$$

For simplicity, we only work with $t \geq 0$. With these definitions,

$$\theta^n(t) = \theta_n + \bar{G}^n(t) + \tilde{G}^n(t) + B^n(t) + M^n(t) + Z^n(t). \tag{2.13}$$

As in Theorem 2.1, (A2.1) and (A2.4) imply that $M^n(\cdot)$ converges to the "zero" process with probability one as $n \to \infty$. Since $\beta_n \to 0$ with probability one, the process $B^n(\cdot)$ also converges to zero with probability one. Since $g_n(\cdot)$ and $\bar{g}(\cdot)$ are uniformly bounded on H, the set $\{\bar{G}^n(\omega, \cdot), \tilde{G}^n(\omega, \cdot)\}$ is equicontinuous for each ω. These bounds and convergences imply that the jumps in $\bar{G}^n(\cdot) + \tilde{G}^n(\cdot) + M^n(\cdot) + B^n(\cdot)$ on any finite interval go to zero with probability one as $n \to \infty$. Consequently, with probability one the distance of $\theta_n + \epsilon_n Y_n$ to H goes to zero as $n \to \infty$.

Now fix attention on the case where H satisfies (A4.3.2). [The details under (A4.3.3) are left to the reader.] Then, if we were to ignore the effects of the terms $Z^n(\cdot)$, $\{\theta^n(\omega, \cdot)\}$ would be equicontinuous on $(-\infty, \infty)$ for ω not in a null set N, the set of nonconvergence to zero of $B^n(\omega, \cdot)$ or of $M^n(\omega, \cdot)$. Thus the only possible problem with equicontinuity would originate from Z_n. That this is not possible follows from the following argument.

Suppose that for some $\omega \notin N$, there are $\delta_1 > 0$, $m_k \to \infty$ and $0 < \Delta_k \to 0$ such that

$$|Z^{m_k}(\omega, \Delta_k)| \geq \delta_1.$$

Then the paths $Z^{m_k}(\omega, \cdot)$ will "asymptotically" have a jump of at least δ_1 at $t = 0$. This jump cannot take the path $\theta^{m_k}(\omega, \cdot)$ into the interior of H, since $Z_n(\omega) = 0$ if $\theta_{n+1}(\omega)$ is in the interior of H. Thus, the effect of the assumed "asymptotic" jump in $Z^{m_k}(\omega, \cdot)$ is an "'asymptotic" jump of $\theta^{m_k}(\omega, \cdot)$ from one point on the boundary of H to another point on the boundary of H. But this contradicts the fact that $\epsilon_n Z_n(\omega)$ goes to zero and acts as either an inward normal at $\theta_{n+1}(\omega)$, if $\theta_{n+1}(\omega)$ is not on an edge or corner of H, or as a non-negative linear combination of the linearly independent set of the inward normals of the adjacent faces if $\theta_{n+1}(\omega)$ is on an edge or corner. The Lipschitz continuity follows from the same argument that was used in the proof of Theorem 2.1. A similar reasoning for the equicontinuity of $\{Z^n(\cdot)\}$ can be used when the constraint set is defined by (A4.3.3).

The rest of the reasoning is as in Theorem 2.1; we need only identify the limits of $\tilde{G}^n(\omega, \cdot)$ and $G^n(\omega, \cdot)$ along convergent subsequences and for ω not in N. Fix $\omega \notin N$, and let n_k index a convergent subsequence of $\{\theta^n(\omega, \cdot)\}$ with limit $\theta(\omega, \cdot)$. Then the equicontinuity of $\{\theta^n(\omega, \cdot)\}$, the uniform (in n) continuity of $g_n(\cdot)$, and (2.12) imply that $\tilde{G}^{n_k}(\omega, t) \to 0$ and $G^{n_k}(t) \to \int_0^t \bar{g}(\theta(\omega, s))ds$ for each t. □

Theorem 2.4. (Random ϵ_n.) *Let $\epsilon_n \geq 0$ be $\mathcal{F}_n$-measurable and satisfy*

$$\sum_n \epsilon_n^2 < \infty \text{ w.p. 1.} \tag{2.14}$$

Then, under the other conditions of Theorem 2.3, the conclusions of the theorem remain valid.

Proof. The proof is essentially the same as that of Theorem 2.3. Modify the process on a set of arbitrarily small measure such that $E\sum \epsilon_n^2 < \infty$ and then prove convergence of the modified process. In particular, for $K > 0$ define $\epsilon_{n,K} = \epsilon_n$ until the first n that $\sum_{i=0}^{n-1} \epsilon_i^2 \geq K$, and set $\epsilon_{n,K} = 1/n$ at and after that time. The proof of Theorem 2.3 holds if the $\epsilon_{n,K}$ are used. Since $\lim_{K\to\infty} P\{\epsilon_n \neq \epsilon_{n,K}, \text{ any } n\} = 0$, the theorem follows. □

5.2.2 Chain Recurrence

In the previous parts of this chapter and in the following sections, it is shown that θ_n and $\theta^n(\cdot)$ converge with probability one to an invariant or limit set of the ODE, indeed to some bounded invariant or limit set contained within some other set, say the constraint set. In the absence of other information, we may have to assume that the invariant set is the

largest one. But, sometimes the largest invariant or limit sets contain points to which convergence clearly cannot occur. Consider the following example.

Example. Let x be real-valued, with $\bar{g}(x) = x(1-x), H = [0,1]$. Then the entire interval $[0,1]$ is an invariant set for the ODE. It is clear that if any arbitrarily small neighborhood of $x = 1$ is entered infinitely often with probability $\mu > 0$, then (with probability one, relative to that set), the only limit point is $x = 1$. Furthermore, if each small neighborhood of $x = 0$ is exited infinitely often with probability $\mu > 0$, then it is obvious that the limit point will be $x = 1$ (with probability one, relative to that set). One does not need a sophisticated analysis to characterize the limit in such a case, and an analogous simple analysis can often be done in applications. But the idea of *chain recurrence* as introduced by Benaim [6] can simplify the analysis in general, since it can be shown that the convergence must be to the subset of the invariant set that is chain recurrent, as defined below. In this example, the only chain recurrent points are $\{0,1\}$.

Definition. Let $\Phi_t(x)$ denote the solution to the ODE at time t given that the initial condition is x. A point x is said to be *chain recurrent* [6, 62] if for each $\delta > 0$ and $T > 0$ there is an integer k and points $u_i, T_i, 0 \le i \le k$, with $T_i \ge T$, such that

$$|x - u_0| \le \delta,\ |y_1 - u_1| \le \delta,\ \ldots,\ |y_k - u_k| \le \delta,\ |y_{k+1} - x| \le \delta, \qquad (2.15)$$

where $y_i = \Phi_{T_{i-1}}(u_{i-1})$ for $i = 1, \ldots, k+1$. We also say that two points x and $\bar{x}$ are *chain connected* if, with the above terminology,

$$|x - u_0| \le \delta,\ |y_1 - u_1| \le \delta,\ \ldots,\ |y_k - u_k| \le \delta,\ |y_{k+1} - \bar{x}| \le \delta \qquad (2.16)$$

and with a similar perturbed path taking $\bar{x}$ to x.

See Figure 2.1 for an illustration of a chain connectedness. All points in L_H (the limit set for paths of the ODE starting in H) are chain recurrent. But not all chain recurrent points are in L_H.

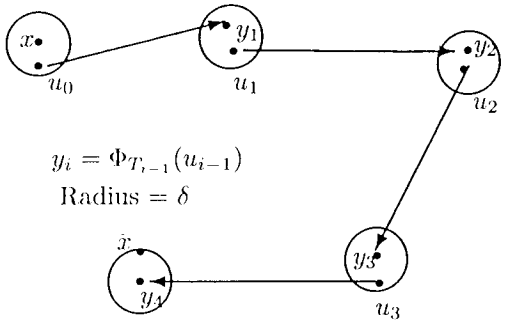

Figure 2.1. An example of chain connectedness.

Example. Figure 2.2 illustrates an example of chain recurrence. The flow lines are drawn. We suppose that $\bar{g}(\cdot)$ is Lipschitz continuous and $\bar{g}(x) = 0$ at the points $\{a, b, c, d, e\}$. Hence these points are stationary. Suppose that the point e is attracting for points interior to the rectangle. The square with corners $\{a, b, c, d\}$ is an invariant set. But the only chain recurrent points (in some neighborhood of the box) are $\{e\}$ and the lines connecting $\{a, b, c, d\}$. The limit points for the ODE (with initial conditions in some neighborhood of the box) are only $\{a, b, c, d, e\}$.

The paths of the stochastic approximation and chain recurrent points. As seen in Theorem 2.1, for almost all ω the path $\theta^n(\omega, \cdot)$ follows the solution to the ODE closely for a time that increases to infinity as $n \to \infty$. Let ω not be in the null set N of Theorem 2.1. Given $\delta > 0$, there are $T_0^n \to \infty$ such that $|\theta^n(\omega, t) - \Phi_t(\theta^n(\omega, 0))| \leq \delta$ on $[0, T_0^n]$. Now repeat the procedure. There are times T_k^n such that $T_{k+1}^n - T_k^n \to \infty$ for each large n and such that

$$|\theta^n(\omega, T_k^n + t) - \Phi_t(\theta^n(\omega, T_k^n))| \leq \delta, \ t \leq T_{k+1}^n - T_k^n.$$

Thus, asymptotically, the path of the stochastic approximation will follow a path of the ODE which is restarted periodically at a value close to the value at the end of the previous section.

Consider a sequence of paths of the ODE on $[0, s_n]$, with initial condition x_n and where $s_n \to \infty$. Then, for any $\mu > 0$, the fraction of time that the paths spend in $N_\mu(L_H)$ goes to infinity as $n \to \infty$.

It follows from these arguments that the path of the stochastic approximation spends an increasing amount of time (the fraction going to one) close to L_H as $n \to \infty$.

It also follows from these arguments and the definition of chain recurrence that if any small neighborhood of a point x is returned to infinitely often by $\theta_n(\omega)$, then that point will be chain recurrent. The point can be illustrated by the example of Figure 2.2. The paths of the stochastic approximation process will first be drawn to the center or to the boundary of the box. A path $\theta^n(\omega, \cdot)$ that is near the line $[a, b]$ will be drawn toward the point b, but if it is very slightly inside the box, as it gets close to b, it will be drawn towards c or to the center. The noise might eventually force it slightly outside the box, so that it will not necessarily end up at the point e. But if it does not go to e, it will stay close to the boundary of the box.

Although the process will eventually spend most of its time in an arbitrarily small neighborhood of the limit points, it might visit any small neighborhood of some chain recurrent point again and again, but the time intervals between such visits to a small neighborhood of a nonlimit point will go to infinity.

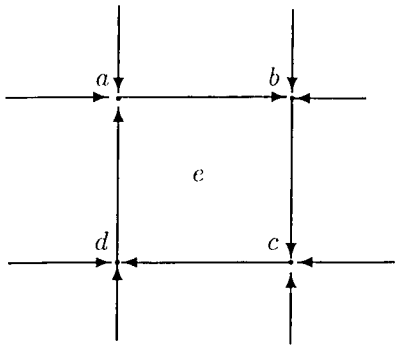

Figure 2.2. An example of chain recurrence vs. invariance.

Comment. The proof of Theorem 2.5 implies that if any neighborhood of a point x is visited infinitely often for ω in some set Ω_x, and x and $\bar{x}$ are not chain connected, then any small enough neighborhood of $\bar{x}$ can be visited only finitely often with probability one relative to Ω_x. Furthermore, there is a set that the path must enter infinitely often, which is disjoint from some neighborhood of $\bar{x}$, and which is locally asymptotically stable in the sense of Liapunov. In particular, there are arbitrarily small neighborhoods of this set such that the "flow is strictly inward" on the boundaries. Thus, there is a Liapunov function argument that can be used to show a contradiction to (2.18) below. In applications where the limit set of the mean limit ODE might be complicated, one needs to do further analysis.

Theorem 2.5. *Let $\bar{g}(\cdot)$ be Lipschitz continuous and let the ODE be*

$$\dot{\theta} = \bar{g}(\theta) + z, \tag{2.17}$$

where z is the reflection term. Assume the other conditions (except for (A2.6)) of any of the Theorems 2.1 to 2.4. For points x and $\bar{x}$, let there be $\mu > 0$ such that

$$P\{\theta_n \in N_\delta(x),\ \theta_n \in N_\delta(\bar{x}),\ \textit{infinitely often}\} \geq \mu \tag{2.18}$$

for all $\delta > 0$. Then x and $\bar{x}$ are chain connected. [We can have $x = \bar{x}$.] Thus the assertions concerning convergence to an invariant or limit set of the mean limit ODE can be replaced by convergence to a set of chain recurrent points within that invariant or limit set.

Proof. For the constrained problem the neighborhoods are relative to the constraint set. The Lipschitz condition is used to assure that the time required to reach a point where $\bar{g}(x) = 0$ is infinite, whether the path moves forward or backward in time. Let Ω_x denote the set whose probability is taken in (2.18). Let $R(A; T, \infty)$ denote the closure of the range of the

solution of the ODE on the interval $[T, \infty)$ when the initial conditions are in the set A. Define $R(A) = \lim_{T\to\infty} R(A; T, \infty)$. Let $N_\delta(A)$ denote the δ-neighborhood of the set A. For $\delta > 0$, set $R_1^\delta(x) = R(N_\delta(x))$. For $n > 1$ define, recursively, $R_n^\delta(x) = R(N_\delta(R_{n-1}^\delta(x)))$, and let $R_\infty^\delta(x)$ be the closure of $\lim_n R_n^\delta(x)$. Note that $R_n^\delta(x) \subset R_{n+1}^\delta(x)$. For purposes of the proof, even without a constraint, we can suppose without loss of generality that all of the above sets are bounded. If either x or $\bar{x}$ is not chain recurrent, then for small enough $\delta > 0$, either $\bar{x} \notin R_\infty^\delta(x)$ or $x \notin R_\infty^\delta(x)$. Suppose the former option, without loss of generality.

By the ODE method and the definition of $R_\infty^\delta(x)$, for almost all ω in Ω_x each small neighborhood of $R_\infty^\delta(x)$ must be entered infinitely often. We need to show that if $\bar{x} \notin R_\infty^\delta(x)$, then there is some set which excludes a small neighborhood of $\bar{x}$ and which cannot be exited infinitely often (with probability one relative to Ω_x). Since this will contradict (2.18), the theorem will be proved.

Let $\delta_i > 0$ be small enough such that for $\delta \leq \delta_1$, $N_{2\delta_2}(\bar{x}) \cap R_\infty^\delta(x) = \phi$, the empty set. The ν_i used below will be smaller than $\min\{\delta_1, \delta_2\}$, and δ will be less than δ_1. The δ and δ_i are fixed. Given $\nu_i > 0$, define the sets

$$\begin{aligned} S_1(\nu_1) &= N_{\nu_1}(R_\infty^\delta(x)), \\ S_2(\nu_2) &= \overline{N_{\nu_2}(R_\infty^\delta(x)) - R_\infty^\delta(x)}, \\ S_3(\nu_3) &= N_{\nu_3}(R_\infty^\delta(x)). \end{aligned}$$

We will show that, for any (small enough) $\nu_3 > 0$, there are $0 < \nu_1 < \nu_2 < \nu_3$ such that any solution to the ODE which starts in $S_2(\nu_2) - S_1(\nu_1)$ cannot exit $S_3(\nu_3)$ and must return to $S_1(\nu_1)$ by a time $T(\nu_3)$ which is bounded in the initial condition in $S_2(\nu_2)$. This assertion can be used with the ODE method to prove the theorem, since the ODE method would then imply that exit infinitely often from any arbitrarily small neighborhood of $R_\infty^\delta(x)$ can occur only with probability zero.

We now prove the assertion. Let $\nu_2^n \to 0$, and suppose that for small $\nu_3 > 0$ there are $x_n \in S_2(\nu_2^n)$ and $T_n < \infty$ such that the trajectory $\Phi_t(x_n)$ first exits $S_3(\nu_3)$ at time T_n. Suppose that (take a subsequence, if necessary) $T_n \to T < \infty$. Then there is a point $y \in R_\infty^\delta(x)$ with $g(y) \neq 0$ and a path $\Phi_t(y)$ which exits $S_3(\nu_3)$ by time T. By the continuity of the path of the ODE in the initial condition, the "δ-perturbation" method used for constructing the $R_n^\delta(x)$, the definition of $R_\infty^\delta(x)$, and the fact that any path starting at a point $y \in R_\infty^\delta(x)$ with $\bar{g}(y) \neq 0$ is in $R_\infty^\delta(x)$ for all t, this path would be contained in $R_\infty^\delta(x)$, a contradiction.

Now suppose that $T_n \to \infty$. Then by the "δ-perturbation" method which was used to construct the sets $R_n^\delta(x)$ and the definition of $R_\infty^\delta(x)$, we see that some point on the boundary of $S_3(\nu_3)$ would be in $R_\infty^\delta(x)$, a contradiction. We can conclude from this argument and from the definition of $R_\infty^\delta(x)$ that for small enough ν_3 there is a $\nu_2 > 0$ such that any solution

which starts in $S_2(\nu_2)$ cannot exit $S_3(\nu_3)$ and also that it must eventually return to $S_1(\nu_1)$ for any $\nu_1 > 0$.

We need only show that the time required to return to $S_1(\nu_1)$, $\nu_1 < \nu_2$, is bounded uniformly in the initial condition in $S_2(\nu_2)$ for small enough ν_2 and ν_3. We have shown that the paths starting in $S_2(\nu_2) - S_1(\nu_1)$ cannot exit $S_3(\nu_3)$ and must eventually converge to $S_1(\nu_1)$. Suppose that for each small $\nu_i, i \leq 3$, there is a sequence of initial conditions x_n in $S_2(\nu_2) - S_1(\nu_1)$ such that the time required for the path to reach $S_1(\nu_1)$ goes to infinity as $n \to \infty$. Then there is a path starting in $S_2(\nu_2) - S_1(\nu_1)$ and which stays in $S_3(\nu_3) - S_1(\nu_1)$ for an infinite amount of time, for small enough $\nu_i, i \leq 3$. But then the "δ-perturbation" definition of the $R_n^\delta(x)$ and the definition of $R_\infty^\delta(x)$ imply that such paths would be in $R_\infty^\delta(x)$ for small enough $\nu_i, i \leq 3$, a contradiction. □

Corollary. *Drop the Lipschitz condition, but assume continuity of $\bar{g}(\cdot)$, uniqueness of the solution of the ODE for each initial condition and that the path takes an infinite amount of time (going either forward or backward) to reach any point where $\bar{g}(\theta) = 0$. Then the conclusions of the theorem hold.*

Comment on differential inclusions. There is an extension of the result to the case where the mean limit ODE is replaced by a differential inclusion $\dot{\theta} \in G(\theta) + z$, where $G(\cdot)$ is upper semicontinuous, and some other mild conditions are imposed. The details are omitted to avoid complicating things further.

5.3 A General Compactness Method

5.3.1 The Basic Convergence Theorem

The proofs of Theorems 2.1 to 2.3 used the square summability condition (A2.4) to guarantee that the martingale M_n converged with probability one as $n \to \infty$. It was also supposed that $\beta_n \to 0$ with probability one. These were key points in the proofs that the sequence $\{\theta^n(\omega, \cdot)\}$ was equicontinuous with probability one, which allowed us to show that the limit points of $\theta^n(\cdot)$ were determined by the asymptotic behavior of the ODE determined by the "mean dynamics." An alternative approach, which was initiated in [99], starts with general conditions that guarantee the equicontinuity, and hence the limit theorem, and then proceeds to find specific and more verifiable conditions that guarantee the general conditions. Many such verifiable sets of conditions were given in [99]. The general conditions and the approach are very natural. They are of wide applicability and will be extended further in the next chapter. It has been shown that for certain classes of problems, the general conditions used are both necessary and sufficient [180].

We will continue to suppose (A2.8), namely,

$$E_n Y_n = g_n(\theta_n) + \beta_n, \tag{3.1}$$

and will work with algorithm (2.10).

Definition: Asymptotic "rate of change" conditions. Recall the definition

$$M^0(t) = \sum_{i=0}^{m(t)-1} \epsilon_i \delta M_i, \quad \delta M_n = Y_n - E_n Y_n,$$

and the analogous definition for $B^0(\cdot)$. Instead of using (A2.1) and (A2.4) (which implied the desired convergence of $\{M^n(\cdot)\}$), and the assumption (A2.5) to deal with the β_n effects, we will suppose that the *rates of change* of $M^0(\cdot)$ and $B^0(\cdot)$ go to zero with probability one as $t \to \infty$. By this, it is meant that for some positive number T,

$$\limsup_n \max_{j \geq n} \max_{0 \leq t \leq T} |M^0(jT + t) - M^0(jT)| = 0 \text{ w.p.1} \tag{3.2}$$

and

$$\limsup_n \max_{j \geq n} \max_{0 \leq t \leq T} |B^0(jT + t) - B^0(jT)| = 0 \text{ w.p.1.} \tag{3.3}$$

If (3.2) and (3.3) hold for some positive T, then they hold for all positive T. Note that (3.2) does not imply convergence of $\{M_n\}$. For example, the function $\log(t+1)$ for $t > 0$ satisfies (3.2) but does not converge. Condition (3.2) is guaranteed by (A2.1) and (A2.4), but we will show that it is much weaker. The conditions (3.2) and/or (3.3) will be referred to either by saying that the *asymptotic rate of change is zero with probability one* or that *the asymptotic rate of change goes to zero with probability one.*

Note that Theorem 3.1 does not require that Y_n be random, provided that there is some decomposition of the form $Y_n = g_n(\theta_n) + \delta M_n + \beta_n$ and (3.2) and (3.3) holds for whatever sequence $\{\delta M_n, \beta_n\}$ is used. Conditions (3.2) and (3.3) are equivalent to

$$\limsup_n \sup_{|t| \leq T} |M^n(t)| = 0, \quad \limsup_n \sup_{|t| \leq T} |B^n(t)| = 0. \tag{3.4}$$

With assumptions (3.2) and (3.3) used to eliminate many of the details, the proofs of Theorems 2.1 to 2.3 give us their conclusions, without the necessity of (A2.1), (A2.4), and (A2.5). We thus have the following theorem.

Theorem 3.1. *Suppose* (1.1) *and that* $E|Y_n| < \infty$ *for each* n. *Assume* (3.2), (3.3), (A2.8), *and any of the constraint set conditions* (A4.3.1), (A4.3.2), *or* (A4.3.3). *If* $\bar{g}(\cdot)$ *is a gradient, assume* (A2.7). *Then the conclusions of Theorems* 2.1 *to* 2.3 (*which do not require* (A2.6)) *continue to hold.*

Under (A2.6), *the limit points are contained in* $L_H^1 \cup A_H$.

Under the additional conditions of Theorem 2.5 *or its corollary* (*but not using* (A2.6)), *for almost all* ω, $\theta_n(\omega)$ *converges to the set of chain recurrent points.*

A sufficient condition for the asymptotic rate of change assumption (3.2). The main problem is the verification of (3.2) when (A2.4) fails to hold. The next theorem sets up a general framework for obtaining perhaps the weakest possible replacement for (A2.4), and this is illustrated by the examples in the next section. The general approach is reminiscent of the "large deviations" upper bounds.

Theorem 3.2. *Let* $E|Y_n| < \infty$ *for each* n. *For each* $\mu > 0$ *and some* $T > 0$, *suppose either that*

$$\lim_n P\left\{\sup_{j\geq n}\max_{0\leq t\leq T}\left|\sum_{i=m(jT)}^{m(jT+t)-1}\epsilon_i\delta M_i\right| \geq \mu\right\} = 0 \tag{3.5}$$

or

$$\sum_j q_j'(\mu) < \infty, \tag{3.6}$$

where $q_j'(\mu)$ *is defined by*

$$q_j'(\mu) = P\left\{\max_{0\leq t\leq T}\left|\sum_{i=m(jT)}^{m(jT+t)-1}\epsilon_i\delta M_i\right| \geq \mu\right\}. \tag{3.7}$$

Then (3.2) *holds for each* T.

Proof. If the conditions hold for some positive T, then they hold for all positive T. Equation (3.5) implies (3.2) for each T. Under (3.6), the Borel Cantelli Lemma says that the event $\sup_{|t|\leq T}|M^n(t)| \geq \mu$ occurs only finitely often with probability one for each $\mu > 0$ and $T < \infty$. This implies (3.2) for each T. □

5.3.2 *Sufficient Conditions for the Rate of Change Condition*

We will use the following conditions. Modifications for the Kiefer-Wolfowitz scheme will be given in the next subsection.

(A3.1) For each $\mu > 0$,

$$\sum_n e^{-\mu/\epsilon_n} < \infty. \tag{3.8}$$

(A3.2) For *some* $T < \infty$, there is a $c_1(T) < \infty$ such that for all n,

$$\sup_{n \le i \le m(t_n+T)} \frac{\epsilon_i}{\epsilon_n} \le c_1(T). \tag{3.9}$$

(A3.3) There is a real $K < \infty$ such that for small real γ, all n, and each component $\delta M_{n,j}$ of δM_n,

$$E_n e^{\gamma(\delta M_{n,j})} \le e^{\gamma^2 K/2}. \tag{3.10}$$

(A3.2) is unrestrictive in applications. (A3.1) holds if $\epsilon_n \le \gamma_n / \log n$, for any sequence $\gamma_n \to 0$. If γ_n does not tend to zero, then (excluding degenerate cases) it is not possible to get probability one convergence, since $\{\epsilon_n\}$ is in the "simulated annealing" range, where convergence is at best in the sense of convergence in probability (or in the sense of weak convergence; see Chapter 7). The sets of conditions (A2.1), (A2.4) and (A3.1), (A3.3) represent the extremes of the possibilities. In the intermediate cases, the speed at which the step sizes must go to zero for probability one convergence depends on the rate of growth of the moments $E|\delta M_n|^k$ of the noise as $k \to \infty$.

Theorem 3.3. (A3.1) *to* (A3.3) *imply* (3.6) *for real* ϵ_n. [*If the* ϵ_n *are random, suppose that* ϵ_n *is* $\mathcal{F}_n$*-measurable and that there are real* $\tilde{\epsilon}_n$ *satisfying* (A3.1) *and* (A3.2) *and that* $\epsilon_n \le \tilde{\epsilon}_n$ *for all but a finite number of* n *with probability one.*]

Proof. It is sufficient to work with one component of δM_i at a time, so we suppose that δM_i are real-valued henceforth. The case of random ϵ_n is a straightforward extension of the nonrandom case, and we suppose that ϵ_n are nonrandom. To prove (3.6), it is enough to show that for some positive T there is a real $\alpha > 0$ (that can depend on j) such that for $\mu > 0$, $q_j(\mu)$ defined by

$$P\left\{\max_{0 \le t \le T} \exp\left[\alpha \sum_{i=m(jT)}^{m(jT+t)-1} \epsilon_i \delta M_i\right] \ge e^{\alpha\mu}\right\} \equiv q_j(\mu) \tag{3.11}$$

is summable. [To deal with negative excursions, just replace δM_i with $-\delta M_i$.] Since M_n is a martingale and the exponential is a convex function, (4.1.4) implies that

$$q_j(\mu) \le e^{-\alpha\mu} E\left\{\exp\left[\alpha \sum_{i=m(jT)}^{m(jT+T)-1} \epsilon_i \delta M_i\right]\right\}. \tag{3.12}$$

The summability of $q_j(\mu)$ will follow from (A3.1) to (A3.3) by evaluating (3.12) with an appropriate choice of α. By (3.10) and using conditional

expectations and $m > n$,

$$\begin{aligned} & E\left\{\exp\left[\alpha\sum_{i=n}^{m}\epsilon_i\delta M_i\right]\right\} \\ & \quad = E\left\{\exp\left[\alpha\sum_{i=n}^{m-1}\epsilon_i\delta M_i\right]E_m e^{\alpha\epsilon_m\delta M_m}\right\} \\ & \quad \le E\exp\left[\alpha\sum_{i=n}^{m-1}\epsilon_i\delta M_i\right]e^{\alpha^2\epsilon_m^2K/2}. \end{aligned} \tag{3.13}$$

Repeating this procedure on the right side of (3.13) yields

$$q_j(\mu) \le \exp\left[K\alpha^2\sum_{i=m(jT)}^{m(jT+T)-1}\epsilon_i^2/2\right]e^{-\alpha\mu}. \tag{3.14}$$

Minimizing the exponent in (3.14) with respect to α yields that

$$\alpha_{\min} = \mu\Big/\left[K\sum_{i=m(jT)}^{m(jT+T)-1}\epsilon_i^2\right].$$

Thus

$$q_j(\mu) \le \exp\left[\frac{-\mu^2}{2K\sum_{i=m(jT)}^{m(jT+T)-1}\epsilon_i^2}\right].$$

By (A3.2),

$$\alpha_{\min} \ge \frac{\mu}{K\epsilon_{m(jT)}c_1(T)T} \equiv \alpha_0.$$

Using $\alpha = \alpha_0$ in (3.14) yields [using (A3.2) again]

$$\begin{aligned} q_j(\mu) & \le \exp\left[\frac{K\mu^2\sum_{i=m(jT)}^{m(jT+T)-1}\epsilon_i^2}{2K^2\epsilon_{m(jT)}^2c_1^2(T)T^2} - \frac{\mu^2}{K\epsilon_{m(jT)}c_1(T)T}\right] \\ & \le \exp\left(\frac{-\mu^2}{2Kc_1(T)T\epsilon_{m(jT)}}\right) \equiv q_j''(\mu). \end{aligned} \tag{3.15}$$

The terms $q_j''(\mu)$ are summable for each $\mu > 0$ by (A3.1). Note that it is enough for (3.10) to hold only for small γ because in (3.13) $\alpha_0\epsilon_i$ effectively replaces the γ in (3.10), and $\epsilon_i\alpha_0 = O(\mu)$, which is arbitrarily small. □

Examples of condition (A3.3). It is sufficient to work with real-valued random variables.

Example 1. Suppose that ξ_n are Gaussian, mutually independent, with mean zero and uniformly bounded variances σ_n^2. Let $\delta M_n = \nu_n(\theta_n)\xi_n$, where $\{\nu_n(\theta), \theta \in H\}$ is bounded. Then

$$E_n \{\exp[\gamma \nu_n(\theta_n)\xi_n]\} = \exp[\gamma^2 \sigma_n^2 \nu_n^2(\theta_n)/2],$$

and (3.10) holds.

Example 2. Let there be a $K_1 < \infty$ such that for all n, $k < \infty$, δM_n satisfies

$$E_n|\delta M_n|^{2k} \leq K_1^k k!. \tag{3.16}$$

Then (3.10) holds. Without loss of generality, let $K_1 > 1$. Inequality (3.16) holds for the Gaussian distribution of Example 1, since $E|\xi_n|^{2k} = (2k-1)(2k-3)\cdots 3 \cdot 1 \cdot \sigma_n^{2k}$. Hence $K_1 = 2 \sup_n \sigma_n^2$. Also, (3.16) holds if $\{\delta M_n\}$ is bounded. One canonical model takes the form $Y_n = g_n(\theta_n, \xi_n)$ with $\{\xi_n\}$ being mutually independent and independent of θ_0. Then (3.16) is essentially a condition on the moments of $g_n(\theta, \xi_n)$ for $\theta \in H$.

Comment on (3.16). Recall the discussion of robustness in Section 1.3.4. When truncation procedures are used to cull or truncate high values of Y_n to "robustify" the performance of the algorithm, (3.16) would generally be satisfied. It is not desirable to have the performance of the stochastic approximation procedure be too sensitive to the structure of the tails of the distribution functions of the noise terms.

To prove the sufficiency of (3.16), use $X = \delta M_n$ and

$$E_n \{\exp[\gamma X]\} \leq 1 + \gamma E_n X + \frac{\gamma^2}{2} E_n X^2 + \sum_{k=3}^{\infty} \frac{\gamma^k}{k!} E_n |X|^k. \tag{3.17}$$

Since X is a martingale difference with respect to $\mathcal{F}_n$, $E_n X = 0$. By (3.16), there is a real K_2 such that

$$\frac{\gamma^{2k} E_n |X|^{2k}}{(2k)!} \leq \frac{\gamma^{2k} K_1^k k!}{(2k)!} \leq \frac{\gamma^{2k} K_2^k}{k!}, \quad k \geq 1.$$

For odd exponents ($k \geq 2$), Hölder's inequality yields

$$E_n |X|^{2k-1} \leq E_n^{(2k-1)/2k} |X|^{2k},$$

which implies that there is a $K_3 < \infty$ such that

$$\frac{\gamma^{2k-1} E_n |X|^{2k-1}}{(2k-1)!} \leq \frac{\gamma^{2k-1} K_3^k}{k!}, \quad k \geq 2.$$

Now writing $\gamma^{2k-1} \leq \gamma^{2k-2} + \gamma^{2k}$ and using upper bounds where needed yields that there is a $K < \infty$ such that

$$E_n \{\exp[\gamma X]\} \leq 1 + \sum_{k=1}^{\infty} \frac{\gamma^{2k} K^k}{2^k k!} = \exp\left[\gamma^2 K/2\right],$$

which yields (3.10).

5.3.3 *The Kiefer-Wolfowitz Algorithm*

Theorems 2.1 2.3 and 3.1-3.3 can be modified to hold for the Kiefer-Wolfowitz form of the projected algorithm (2.1). The few modifications required for Theorem 3.3 will be discussed next. The precise form of Y_n will depend on how one iterates among the coordinates. We will work with a form that includes the various cases discussed in Section 1.2, so that quite general choices of the coordinate(s) to be iterated on at each step are covered; see, for example, (1.2.4), (1.2.14), or the forms where one cycles among the coordinates. By modifying (A3.1)-(A3.3) appropriately, we will see that Theorem 3.1 continues to hold under general conditions.

Suppose that the observation can be written as

$$Y_n = g_n(\theta_n) + \beta_n + \frac{\delta M_n}{2c_n}, \tag{3.18}$$

where δM_n is a martingale difference. The term $\delta M_n/(2c_n)$ arises from the observation noise divided by the finite difference interval c_n. Redefine $M^0(\cdot)$ as

$$M^0(t) = \sum_{i=0}^{m(t)-1} \frac{\epsilon_i}{2c_i} \delta M_i. \tag{3.19}$$

Theorem 3.4. *Assume the conditions of Theorem* 3.1 *for the case where* $g(\cdot)$ *is a gradient, but with observations of the form* (3.18) *and the new definition* (3.19) *of* $M^0(\cdot)$ *used. Then the conclusions of Theorem* 3.1 *hold.*

Sufficient conditions for (3.2) with $M^0(\cdot)$ defined by (3.19). We next obtain a sufficient condition for (3.2) under the new definition (3.19). Assume the following.

(A3.4) For each $\mu > 0$,

$$\sum_n e^{-\mu c_n^2/\epsilon_n} < \infty. \tag{3.20}$$

(A3.5) For *some* $T < \infty$, there is a $c_1(T) < \infty$ such that for all n,

$$\sup_{n \leq i \leq m(t_n+T)} \frac{\epsilon_i/c_i^2}{\epsilon_n/c_n^2} \leq c_1(T). \tag{3.21}$$

(A3.6) There is a real $K < \infty$ such that for small real γ, all n, and each component $\delta M_{n,j}$ of δM_n,

$$E_n e^{\gamma(\delta M_{n,j})} \leq e^{\gamma^2 K/2}. \tag{3.22}$$

If ϵ_n and c_n are random, suppose that they are $\mathcal{F}_n$-measurable and that there are $\tilde{\epsilon}_n, \tilde{c}_n$ satisfying (3.20) and (3.21) such that $\epsilon_n/c_n^2 \leq \tilde{\epsilon}_n/\tilde{c}_n^2$ for all but a finite number of n with probability one. Then (3.2) holds for the new definition of $M^0(\cdot)$.

The proof is a repetition of the argument of Theorem 3.3, where ϵ_i/c_i replaces ϵ_i in (3.14). Thus, we only need to show that

$$\sum_j \exp\left[\frac{-\mu^2}{2K\sum_{i=m(jT)}^{m(jT+T)-1} \epsilon_i^2/c_i^2}\right] < \infty \quad \text{for some } T < \infty \text{ and each } \mu > 0. \tag{3.23}$$

Using (3.21) yields the upper bound

$$\sum_{i=m(jT)}^{m(jT+T)-1} \frac{\epsilon_i^2}{c_i^2} \leq \frac{\epsilon_n}{c_n^2} \sum_{i=m(jT)}^{m(jT+T)-1} \epsilon_i c_1(T) \approx \frac{\epsilon_n}{c_n^2} c_1(T) T.$$

This and (3.20) yield (3.23).

Of course, M_n converges under the classical condition

$$\sum \epsilon_n^2/c_n^2 < \infty, \quad \sup_n E|\delta M_n|^2 < \infty. \tag{3.24}$$

The result, as stated, is quite general. Keep in mind that the algorithm should be designed so that $-\bar{g}(\cdot)$ is the gradient of the function we wish to minimize.

5.4 Stability and Stability–ODE Methods

Stability methods provide an alternative approach to proofs of convergence. They are most useful when the iterates are allowed to vary over an unbounded set and not confined to a compact set H. They can be used to prove convergence with probability one directly, but the conditions are generally weaker when they are used in combination with an ODE-type method. A stability method would be used to prove that the process is recurrent, that is, there is some compact set to which the stochastic approximation iterates return infinitely often with probability one. Then the

ODE method takes over, starting at the recurrence times, and is used to show that (asymptotically) the iterates follow the path of the mean limit ODE, as in Sections 2 and 3. Indeed, if the paths are not constrained, there might be no alternative to starting the analysis with some sort of stability method.

Stability methods are generally based on a Liapunov function, and this Liapunov function is generally a small perturbation of one for the underlying ODE. The "combined" stability and ODE method is a powerful tool when the constraint set is unbounded or when there are no constraints. It can also be used for the state dependent noise problem or where the algorithm is decentralized. In this section, we are still concerned with the martingale difference noise case. Extensions will be given in subsequent chapters. In this regard, a discussion of stochastic stability for processes driven by nonwhite noise is in [93].

Recurrence and probability one convergence. Two types of theorems and proofs are presented. The first is a more or less classical approach via the use of a perturbed Liapunov function, as in Section 4.4.3. To construct the perturbation and assure that it is finite with probability one, a "local" square summability condition on ϵ_n is used. The step sizes ϵ_n are allowed to be random. The theorem will be used in Section 7 to prove probability one convergence for the lizard learning problem of Section 2.1, where ϵ_n are random. The perturbed Liapunov function-type argument is quite flexible as seen in [93]. A combination of a stability and a "local" ODE method yields convergence under weaker conditions. In particular, the square summability will be dropped. Such an approach is quite natural since the limit mean ODE characterizes the "flow" for large n. This "combined" method will be presented in Theorem 4.2. For notational simplicity, both Theorems 4.1 and 4.2 suppose that $\theta = 0$ is globally stable in the sense of Liapunov for the mean limit ODE $\dot{\theta} = g(\theta)$ and establish the convergence of $\{\theta_n\}$ to zero.

Theorem 4.3 gives a sufficient condition for recurrence, and then a "local" ODE is again used to get convergence. The theorems represent a few of the many possible variations. Starting with some canonical model, stability-type proofs are commonly tailored to the special application at hand. The statement of Theorem 4.2 is complicated because it is intended to apply to cases where the mean limit ODE is an ordinary ODE, a differential inclusion, or where γ_n is an approximation to an element of a set of subgradients.

Theorem 4.1. *Assume* (1.1). *Let $V(\cdot)$ be a real-valued non-negative and continuous function on $\mathbb{R}^r$ with $V(0) = 0$, which is twice continuously differentiable with bounded mixed second partial derivatives. Suppose that for each $\epsilon > 0$, there is a $\delta > 0$ such that $V(\theta) \geq \delta$ for $|\theta| \geq \epsilon$, and δ does not decrease as ϵ increases. Let $\{\mathcal{F}_n\}$ be a sequence of nondecreasing σ-algebras, where $\mathcal{F}_n$ measures at least $\{\theta_0, Y_i, i < n\}$. Let $EV(\theta_0) < \infty$.*

Suppose that there is a function $\bar{g}(\cdot)$ *such that* $E_n Y_n = \bar{g}(\theta_n)$. *For each* $\epsilon > 0$ *let there be a* $\delta_1 > 0$ *such that*

$$V_\theta'(\theta)\bar{g}(\theta) = -k(\theta) \leq -\delta_1$$

for $|\theta| \geq \epsilon$. *Suppose that there are* $K_2 < \infty$ *and* $K < \infty$ *such that*

$$E_n|Y_n|^2 \leq K_2 k(\theta_n), \quad \textit{when } |\theta_n| \geq K. \tag{4.1}$$

Let

$$E\sum_{i=1}^{\infty} \epsilon_i^2 |Y_i|^2 I_{\{|\theta_i| \leq K\}} < \infty. \tag{4.2}$$

Then $\theta_n \to 0$ *with probability one.*

Now, suppose that the step sizes ϵ_n *are random. Let* $\epsilon_n \to 0$ *with probability one and be* $\mathcal{F}_n$*-measurable, with* $\sum \epsilon_i = \infty$ *with probability one. Let there be real positive* $\tilde{\epsilon}_n$ *such that* $\epsilon_n \leq \tilde{\epsilon}_n$ *for all but a finite number of* n *with probability one. Suppose that* (4.2) *is replaced by*

$$E\sum_{i=1}^{\infty} \tilde{\epsilon}_i^2 |Y_i|^2 I_{\{|\theta_i| \leq K\}} < \infty. \tag{4.2'}$$

Then the conclusion continues to hold.

Comment on the proof. A truncated Taylor series expansion and the boundedness of the second partial derivatives yield that there is a constant K_1 such that

$$E_n V(\theta_{n+1}) - V(\theta_n) \leq \epsilon_n V_\theta'(\theta_n) E_n Y_n + \epsilon_n^2 K_1 E_n |Y_n|^2.$$

By the hypotheses,

$$E_n V(\theta_{n+1}) - V(\theta_n) \leq -\epsilon_n k(\theta_n) + \epsilon_n^2 K_1 E_n |Y_n|^2.$$

The statement of the part of the theorem that uses (4.2) is now the same as that of Theorem 4.4.3. The proof under (4.2′) is a simple modification; the details are left to the reader.

The next result extends Theorem 4.1 in several directions. It allows the conditional mean of Y_n to depend on n, and it covers the case where the γ_n are obtained as subgradients. It also uses either a "local" square summability condition as in Theorem 4.1, or a localized form of (3.2) if square summability cannot be assumed.

One cannot generally assume that $E|Y_n|^2$ are bounded *a priori*. It is a common practice to assume a bound for $E_n|Y_n|^2$ in terms of the Liapunov function itself. This accounts for condition (4.6). Note that $\mathcal{F}_n$ and E_n are defined as in Theorem 4.1. The result will be presented in two parts; we first prove a lemma that will be needed in the theorem.

Lemma 4.1. *Consider the algorithm*

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n. \tag{4.3}$$

Let ϵ_n be $\mathcal{F}_n$-measurable such that $\epsilon_n \to 0$ as $n \to \infty$ and $\sum_i \epsilon_i = \infty$, both with probability one. Let the real-valued continuous function $V(\cdot)$ have continuous first and bounded second mixed partial derivatives. Suppose that $V(0) = 0$, $V(\theta) > 0$ for $\theta \neq 0$, and $V(\theta) \to \infty$ as $|\theta| \to \infty$.

If $E|Y_k| < \infty$ for any $k \geq 0$, write

$$E_k Y_k = \gamma_k. \tag{4.4}$$

Suppose that there are positive numbers K_1 and K_2 such that

$$E_k|Y_k|^2 \leq K_1|V'_\theta(\theta_k)\gamma_k| + K_1 \leq K_2 V(\theta_k) + K_2. \tag{4.5}$$

Assume that $EV(\theta_0) < \infty$. Then $EV(\theta_n) < \infty$ and $E|Y_n|^2 < \infty$ for each n.

Remark. A common model has $V(\theta)$ growing at most as $O(|\theta|^2)$ and $|V_\theta(\theta)| + |\gamma_n(\theta)|$ growing at most as $O(|\theta|)$ for large $|\theta|$.

Proof. The proof uses induction on n. By (4.5) and $EV(\theta_0) < \infty$, we have $E|Y_0|^2 < \infty$. Now suppose that $EV(\theta_n) < \infty$, for some n. Then $E|Y_n|^2 < \infty$ in view of (4.5). By a truncated Taylor series expansion and the boundedness of $V_{\theta\theta}(\cdot)$, we have

$$E_n V(\theta_{n+1}) - V(\theta_n) = \epsilon_n V'_\theta(\theta_n)\gamma_n + O(\epsilon_n^2) E_n|Y_n|^2. \tag{4.6}$$

The inequality (4.5) implies that there is a real K_3 such that the right side of (4.6) is bounded above by $\epsilon_n K_3[1 + V(\theta_n)]$. This, together with the induction hypothesis, implies that $EV(\theta_{n+1}) < \infty$. Hence (4.5) yields $E|Y_{n+1}|^2 < \infty$. Thus by induction, we have proven that $EV(\theta_n) < \infty$ and $E|Y_n|^2 < \infty$ for all n. □

Combined Stability–ODE methods.

Theorem 4.2. *Assume the conditions of Lemma 4.1 and suppose that γ_n has the following properties and dependence on θ_n. For each real K, $\gamma_n I_{\{|\theta_n| \leq K\}}$ is bounded uniformly in n. There are convex and upper semi-continuous (see the definition (4.3.2)) sets $G(\theta) \subset \mathbb{R}^r$ with $G(\theta)$ being uniformly bounded on each bounded θ-set and for each real K and all ω,*

$$\min\{|\gamma_n - y| : y \in G(\theta_n)\} I_{\{|\theta_n| \leq K\}} = \text{distance}(\gamma_n, G(\theta_n)) I_{\{|\theta_n| \leq K\}} \to 0,$$

as $n \to \infty$. Let $c(\delta)$ be a nondecreasing real-valued function with $c(0) = 0$ and $c(\delta) > 0$ for $\delta > 0$ such that for large n (that can depend on δ)

$$V'_\theta(\theta_n)\gamma_n < -c(\delta), \quad \text{if } V(\theta_n) \geq \delta. \tag{4.7}$$

Suppose that there is a nondecreasing real-valued function $c_0(\delta)$ *with* $c_0(0) = 0$ *and* $c_0(\delta) > 0$ *for* $\delta > 0$ *such that*

$$V_\theta'(\theta)\gamma \leq -c_0(\delta), \quad \textit{for all } \gamma \in G(\theta) \textit{ if } V(\theta) > \delta. \tag{4.8}$$

Assume that (3.2) *holds with* δM_i *replaced by* $\delta M_i I_{\{|\theta_i| \leq K\}}$ *for each positive* K. *Then* $\theta_n \to 0$ *with probability one.*

Proof. For large n, (4.5) and (4.7) imply that the right side of (4.6) is negative outside of a neighborhood of $\theta = 0$, which decreases to the origin as $n \to \infty$. Thus, outside of this "decreasing" neighborhood, $V(\theta_n)$ has the supermartingale property. The supermartingale convergence theorem then implies that each neighborhood of $\theta = 0$ is recurrent, that is, θ_n returns to it infinitely often with probability one.

Completion of the proof under (A2.4). Once recurrence of each small neighborhood of the origin is shown, the rest of the proof uses a "local analysis." To illustrate the general idea in a simpler context, the proof will first be completed under the stronger conditions that (A2.4) holds and that $\sup_n E|Y_n|^2 I_{\{|\theta_n| \leq K\}} < \infty$ for some positive K. Define $\delta M_n = Y_n - E_n Y_n = Y_n - \gamma_n$. Fix δ and $\Delta > 2\delta > 0$ small, recall the definition $Q_\lambda = \{\theta : V(\theta) \leq \lambda\}$, and let Q'_Δ denote the complement of the set Q_Δ. Let τ be a stopping time such that $\theta_\tau \in Q_\delta$. By (4.5b), (4.7) and (A2.4), for large n the terms γ_n cannot force θ_n out of Q_Δ. The only way that $\{\theta_k, k \geq \tau\}$ can leave Q_Δ is by the effects of $\{\delta M_k, k \geq \tau\}$. But the convergence of

$$\sum_{i=0}^{m(t)} \epsilon_i \delta M_i I_{\{V(\theta_i) \leq \Delta\}}, \tag{4.9}$$

which is implied by the hypotheses (A2.4) being added in this paragraph, assures that these martingale difference terms cannot force the path from Q_δ to Q'_Δ infinitely often. The convergence follows from this.

Completion of the proof under (3.2). Fix $\delta > 0$ and $\Delta > 2\delta$ as before. The proof is very similar to that of the proof of the gradient case assertions of Theorem 2.1. Let N denote the null set on which the asymptotic rate of change of (4.9) is not zero and let $\omega \notin N$. Suppose that there are infinitely many excursions of $\{\theta_n(\omega)\}$ from Q_δ to Q'_Δ. Then there are $n_k \to \infty$ (depending on ω) such that n_{k-1} is the last index at which the iterate is in $Q_{2\delta}$ before exiting Q_Δ.

We now repeat the argument of Theorem 2.1. By selecting a subsequence if necessary, we can suppose that $\{\theta_{n_k}(\omega)\}$ converges to a point on $\partial Q_{2\delta}$ and that $V(\theta_{n_k+i}(\omega)) \geq 2\delta$ until at least after the next time that $V(\theta_{n_k+i}(\omega)) \geq$

Δ. For $u \geq 0$,

$$\theta^{n_k}(u) - \theta_{n_k} - \sum_{i=n_k}^{m(t_{n_k}+u)-1} \epsilon_i \gamma_i = \sum_{i=n_k}^{m(t_{n_k}+u)-1} \epsilon_i \delta M_i. \tag{4.10}$$

For $\theta_i(\omega) \in Q_\Delta$, $\epsilon_i Y_i(\omega) \to 0$. This and the fact that the right side of (4.10) goes to zero if $\theta^{n_k}(\omega, s) \in Q_\Delta$ for $0 \leq s \leq u$ imply that there is a $T > 0$ such that for large k, $\theta^{n_k}(\omega, t) \in Q_\Delta$ for $t \leq T$, that $\{\theta^{n_k}(\omega, \cdot)\}$ is equicontinuous for $t \leq T$, and that $V(\theta^{n_k}(\omega, t)) \geq 2\delta$ for $t \in (0, T]$.

Let $\theta(\omega, \cdot)$ be the limit of a convergent subsequence of $\{\theta^{n_k}(\omega, \cdot), t \leq T\}$. Write the first sum in (4.10) at ω in an obvious way as

$$\int_0^u g^k(\omega, s) ds.$$

By the hypothesis, the distance between γ_i and the set $G(\theta_i)$ goes to zero for the indices i involved in the first sum in (4.10). Using this and the convexity and upper semicontinuity properties of $G(\theta)$, it follows that the limit (along the convergent subsequence) of the integral has the representation

$$\int_0^u \gamma(s) ds,$$

where $\gamma(s) \in G(\theta(\omega, s))$ for almost all s. Thus, the limit $\theta(\omega, \cdot)$ of any convergent subsequence satisfies the differential inclusion

$$\dot{\theta} \in G(\theta), \tag{4.11}$$

with $V(\theta(\omega, t)) \geq 2\delta$, for $t \leq T$ and $V(\theta(\omega, 0)) = 2\delta$. But (4.8) implies that $V(\theta(\omega, \cdot))$ is strictly decreasing until it reaches the value zero, which contradicts the assertion of the previous sentence and (hence) the assertion that there are an infinite number of escapes from Q_δ. □

A straightforward modification of the proof of Theorem 4.2 yields the following, the proof of which is left to the reader. The theorem gives a very useful combination of the ODE and the stability methods. It first assures that some compact set is recurrent with probability one. Then, looking at the sequence of processes starting at the recurrence time, the ODE method takes over and the asymptotic stability of the limit mean ODE when starting in the recurrence set guarantees the convergence of $\{\theta_n\}$ with probability one. The ODE-type argument is like that in the proof to gradient case assertions of Theorem 2.1.

Theorem 4.3. *Assume the conditions of Theorem 4.2 except let there be a λ_0 such that $c(\delta) > 0$ and $c_0(\delta) > 0$ for $\delta \geq \lambda_0$ and not necessarily otherwise. Then if $\lambda > \lambda_0$, Q_λ is a recurrence set for $\{\theta_n\}$. Let the asymptotic*

rate of change of (4.9) *be zero with probability one for* $\Delta = 2\lambda_0$. *For almost all* ω, *the limit trajectories of* $\{\theta^n(\omega,\cdot)\}$ *are trajectories of* (4.11) *in an invariant set of* (4.11) *in* Q_{λ_0}. *Under* (A2.6′), *with* L_H *and* A_H *replaced by equivalent sets in* Q_{λ_0}, *the invariant set is in* $L_H^1 \cup A_H$.

Let $G(\theta)$ *contain only the single point* $\bar{g}(\theta)$ *for each* θ. *If* $g(\cdot)$ *is a gradient, let* (A2.7) *hold. Then, for each* ω *not in some null set, convergence is to a single stationary set* S_i. *Otherwise, under the additional conditions of Theorem* 2.5 *or its corollary, the limit points for the algorithm are contained in the set of chain recurrent points, with probability one.*

5.5 Soft Constraints

In Sections 2 and 3, we used hard constraints of either the form $a_i \leq \theta_{n,i} \leq b_i$ or where θ_n was confined to a compact set H defined in terms of the constraint functions $q_i(\cdot), i \leq p$. In these cases, the iterate is required to be in the set H for all n. Sometimes, the given constraint functions are merely a guide in that they should not be violated by much, but they can be violated. Then we say that the constraint is *soft*. Soft constraints (equivalently, penalty functions) can be added to the algorithm directly, and stability methods such as those introduced in the last section can be used to prove convergence. The idea is more or less obvious and will now be illustrated by a very simple example. The discussion to follow is merely a guide to possible variations of the basic stochastic approximation algorithm. The soft constraint might be combined with hard constraints. The reader is invited to construct the general form.

In the example to follow, the soft constraint is the sphere S_0 in $\mathbb{R}^r$ with the origin as its center and with radius R_0. Define $q(\theta)$ to be the square of the distance of θ to S_0. Thus, $q(\theta) = (|\theta| - R_0)^2$ for $|\theta| \geq R_0$ and is zero for $|\theta| \leq R_0$. The gradient is $q_\theta(\theta) = 2\theta(1 - R_0/|\theta|)$ for $|\theta| \geq R_0$ and is zero otherwise.

Assume (1.1). The algorithm is

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n - \epsilon_n K_0 q_\theta(\theta_n) \tag{5.1}$$

for sufficiently large positive K_0. The purpose of the $K_0 q_\theta(\cdot)$ term is to assure that the iterates do not wander too far from the sphere. Suppose that $Eq(\theta_0) < \infty$ and that there is a $K_1 < \infty$ (not depending on n) such that if $Eq(\theta_n) < \infty$, then $E_n|Y_n|^2 \leq K_1(q(\theta_n)+1)$ and $E_n Y_n = \bar{g}(\theta_n)+\beta_n$, where $\beta_n \to 0$ with probability one and $\bar{g}(\cdot)$ is continuous. Suppose that for $K_0 > 0$ large enough there are $\alpha > 0$ and $C_0 \geq 0$ such that

$$q'_\theta(\theta)\,[\bar{g}(\theta) - K_0 q_\theta(\theta)] \leq -\alpha q(\theta) + C_0, \tag{5.2}$$

and it is assumed that our K_0 satisfies this condition.

Recall the definition $\delta M_n = Y_n - E_n Y_n$. Suppose that the asymptotic rate of change of

$$\sum_i \epsilon_i \delta M_i I_{\{|\theta_i| \le K\}}$$

is zero with probability one for each positive K. (See Section 3 for a useful criteria.) Then the conclusions of Theorem 2.1 continue to hold, with mean limit ODE

$$\dot{\theta} = \bar{g}(\theta) - K_0 q_\theta(\theta). \tag{5.3}$$

Note that if $\bar{g}(\theta) = -f_\theta(\theta)$ for a continuously differentiable real-valued function $f(\cdot)$, (5.1) is a gradient descent algorithm with the right side of (5.3) being $-[f_\theta(\theta) + K_0 q_\theta(\theta)]$.

The proof is essentially that of Theorems 4.2 and 4.3 and will now be outlined. The essential point is the demonstration of recurrence, that is, that some finite sphere is visited infinitely often by the sequence $\{\theta_n\}$ with probability one. As expected in such algorithms, the penalty function $q(\cdot)$ is used as the Liapunov function.

It will next be shown that

$$\sup_n E q(\theta_n) < \infty. \tag{5.4}$$

A truncated Taylor series expansion yields

$$\begin{aligned} q(\theta_{n+1}) - q(\theta_n) = \; & \epsilon_n q'_\theta(\theta_n)\left[\bar{g}(\theta_n) - K_0 q_\theta(\theta_n)\right] \\ & + O(\epsilon_n^2)\left[|Y_n|^2 + K_0^2 |q_\theta(\theta_n)|^2\right] + \epsilon_n q'_\theta(\theta_n)\delta M_n. \end{aligned}$$

Now, using (5.2), the bound on $E_n|Y_n|^2$ in terms of $q(\theta_n)$ and the fact that there is a real $\alpha_0 > 0$ such that $|q_\theta(\theta)|^2 \le \alpha_0[q(\theta) + 1]$, for large n (hence small ϵ_n),

$$\begin{aligned} q(\theta_{n+1}) \le (1 - \epsilon_n \alpha/2) q(\theta_n) + O(\epsilon_n^2)\left[|Y_n|^2 - E_n|Y_n|^2\right] \\ + \epsilon_n q'_\theta(\theta_n)\delta M_n + O(\epsilon_n^2) + \epsilon_n C_0. \end{aligned} \tag{5.5}$$

Thus,

$$E_n q(\theta_{n+1}) \le (1 - \epsilon_n \alpha/2) q(\theta_n) + O(\epsilon_n^2) + \epsilon_n C_0, \tag{5.6}$$

which implies (5.4).

The inequality (5.6) also implies that Q_λ is a recurrence set for $\{\theta_n\}$ for large enough λ (see Theorem 4.3 or Theorem 4.4.4). Now, Theorem 4.3 implies that the conclusions of Theorem 2.1 hold with the ODE being (5.3). In particular, for almost all ω the limit trajectories of $\theta^n(\omega, \cdot)$ satisfy the ODE (5.3) and are in a bounded invariant set for (5.3).

Finally, we remark that $\bar{g}(\cdot)$ can be replaced by $g_n(\cdot)$ in that $E_n Y_n = g_n(\theta_n) + \beta_n$ if (2.12) holds and (5.2) holds for $g_n(\cdot)$ replacing $\bar{g}(\cdot)$. The "soft constraint" can be used in all subsequent chapters as well.

5.6 Random Directions, Subgradients, and Differential Inclusions

The random directions algorithm. Refer to the random directions algorithm (1.2.11). Let $\mathcal{F}_n^d$ be the minimal σ-algebra that measures $\{\theta_0, Y_{i-1}, d_i, i \leq n\}$. Adding the constraint set H yields the projected algorithm

$$\theta_{n+1} = \Pi_H \left[\theta_n - \epsilon_n d_n \frac{Y_n^+ - Y_n^-}{2c_n} \right], \tag{6.1}$$

that we write in the expanded form as

$$\theta_{n+1} = \theta_n - \epsilon_n f_\theta(\theta_n) + \epsilon_n d_n \beta_n + \epsilon_n d_n \frac{\delta M_n}{2c_n} + \epsilon_n \tilde{\psi}_n + \epsilon_n Z_n, \tag{6.2}$$

where we redefine

$$\delta M_n = (Y_n^- - E_{\mathcal{F}_n^d} Y_n^-) - (Y_n^+ - E_{\mathcal{F}_n^d} Y_n^+). \tag{6.3}$$

β_n is the finite difference bias and $\tilde{\psi}_n = [I - d_n d_n'] f_\theta(\theta_n)$ is the "random directions noise."

More generally, the random directions algorithm takes the following form. Let there be measurable functions $\gamma_n(\cdot)$ such that

$$E_{\mathcal{F}_n^d} Y_n^{\pm} = \gamma_n(\theta_n \pm c_n d_n). \tag{6.4}$$

Suppose that there are functions $g_n(\cdot)$ that are continuous in θ, uniformly in n, and random variables β_n such that

$$\frac{\gamma_n(\theta_n - c_n d_n) - \gamma_n(\theta_n + c_n d_n)}{2c_n} = d_n' g_n(\theta_n) + \beta_n.$$

Write the algorithm in the expanded form

$$\begin{aligned} \theta_{n+1} &= \theta_n + \epsilon_n d_n d_n' g_n(\theta_n) + \epsilon_n d_n \beta_n + \epsilon_n d_n \frac{\delta M_n}{2c_n} + \epsilon_n Z_n \\ &= \theta_n + \epsilon_n g_n(\theta_n) + \epsilon_n d_n \beta_n + \epsilon_n d_n \frac{\delta M_n}{2c_n} \\ &\quad + \epsilon_n \left(d_n d_n' - I \right) g_n(\theta_n) + \epsilon_n Z_n. \end{aligned} \tag{6.5}$$

When the general form (6.5) is used, we will require that the asymptotic rate of change of

$$\sum_{n=0}^{m(t)-1} \epsilon_n \left(d_n d_n' - I \right) g_n(\theta) \tag{6.6}$$

be zero for each θ.

The following theorem follows directly from Theorem 3.1. See also the comments on the Kiefer–Wolfowitz procedure at the end of Section 3 concerning the condition (3.2) with $d_n \delta M_n/(2c_n)$ and $\tilde{\psi}_n$ replacing the δM_n there. If the constraint set H is dropped, then the stability theorems of Section 4 can be used, with $d_n \delta M_n/(2c_n)$ and $\tilde{\psi}_n$ replacing the δM_n and $d_n \beta_n$ replacing β_n there.

Theorem 6.1. *Assume algorithm* (6.5), *the conditions of Theorem* 3.1 *for the gradient case* $\bar{g}(\cdot) = -f_\theta(\cdot)$, *with* $d_n \delta M_n/(2c_n)$ *replacing* δM_n, *and* $d_n \beta_n$ *replacing* β_n. *Assume that the asymptotic rate of change of* (6.6) *is zero with probability one. Then for almost all* ω, $\theta_n(\omega)$ *converges to a unique stationary set* S_i. *In particular if* $f(\cdot)$ *has a unique constrained stationary point* $\bar{\theta}$ *in* H, θ_n *converges to* $\bar{\theta}$ *with probability one.*

Remark: Random directions and the Robbins–Monro procedure. The random directions idea can also be used with the Robbins–Monro procedure. This requires estimating the directional "increment." If the effort required to do this is commensurate with the effort required to get the components of Y_n and also much less than what is required to get Y_n for high dimensions, then it might be advantageous for high dimensions; see Chapter 10.

Example of algorithm (6.5). The use of $g_n(\cdot)$ rather than $-f_\theta(\cdot)$ in (6.5) arises if the direction of search is selected in a different subspace on successive iterations. For example, let $r = r_1 + r_2$, and suppose that we iterate in the subspace of the first r_1 (resp., the last r_2) components of θ on the even (resp., odd) iterations. Let $f_i(\cdot)$ denote the negative of the gradient of $f(\cdot)$ in the first r_1 (resp., second r_2) components. Then $g_{2n}(\cdot) = (f_1(\cdot), 0)$ and $g_{2n+1}(\cdot) = (0, f_2(\cdot))$ and $\bar{g}(\cdot) = (f_1(\cdot) + f_2(\cdot))/2$. On the even (resp., odd) numbered iterations, the last r_2 (resp., first r_1) components of the random direction vector are zero.

Convex function minimization and subgradients. Consider the constrained form of the algorithm (1.2.9), for the minimization of a convex function $f(\cdot)$ that is not necessarily continuously differentiable everywhere. Write the constrained form of (1.2.9) as

$$\theta_{n+1} = \theta_n - \epsilon_n \gamma_n + \epsilon_n \frac{\delta M_n}{2c_n} + \epsilon_n Z_n, \tag{6.7}$$

where Z_n is the reflection term, δM_n is the observation noise, and γ_n is a finite difference approximation to a subgradient of $f(\cdot)$ at θ_n. The required properties of γ_n were stated preceding (1.2.10). Again, the stability theorems of Section 4 can be applied if the constraint set H is dropped. Theorem 3.1 yields the following result.

Theorem 6.2. *Assume algorithm* (6.7), *where the* γ_n *are bounded and satisfy the condition preceding* (1.2.10). *Assume* (1.1), (3.2), (3.3), *and any of the constraint set conditions* (A4.3.1), (A4.3.2) *or* (A4.3.3), *with* $\delta M_n/(2c_n)$ *replacing* δM_n *in* (3.2). *Suppose that* $f(\cdot)$ *is not constant. Then the mean limit ODE is the differential inclusion*

$$\dot{\theta} \in -SG(\theta) + z, \ z(t) \in -C(\theta(t)). \tag{6.8}$$

With probability one, all limit points of θ_n *are stationary points* [*i.e., points where* $0 \in -SG(\theta) + z$]. *If there is a unique limit point* $\bar{\theta}$ *of the paths of* (6.8), *then* $\theta_n \to \bar{\theta}$ *with probability one.*

Differential inclusions. In some applications, (2.12) fails to hold since the statistics of the disturbing noise are not stationary, but where nevertheless the local averages of $g_n(\cdot)$ are in a suitable set, such that one can prove convergence. The differential inclusions form (6.8) might then be useful. Write the algorithm as

$$\theta_{n+1} = \Pi_H \left[\theta_n + \epsilon_n \left(g_n(\theta_n) + \delta M_n\right)\right]. \tag{6.9}$$

The proof of the following useful result is like that of Theorem 3.1; the details are left to the reader. If the constraint set H is dropped, then the stability theorems of Section 4 can be used.

Theorem 6.3. *Assume the conditions in the first paragraph of Theorem* 3.1 *for the algorithm* (6.9), *except for* (A2.8). *Suppose that*

$$\lim_{\Delta \to 0} \limsup_n \sup_{m(t_n+\Delta) \geq i \geq n} \frac{|\epsilon_i - \epsilon_n|}{\epsilon_n} = 0. \tag{6.10}$$

Let $g_n(\cdot)$ *be continuous on* H, *uniformly in* n, *and suppose that for each* θ, *there is a* $G(\theta)$ *that is upper semicontinuous in the sense of* (4.3.2) *such that*

$$\lim_{n,m} \text{distance} \left[\frac{1}{m} \sum_{i=n}^{n+m-1} g_i(\theta), G(\theta)\right] = 0. \tag{6.11}$$

Alternatively, replace (6.11) *by the following. The* $g_n(\theta_n)$ *are bounded and for all* α *and* α_i^n *in* H *such that*

$$\lim_{n, m \to \infty} \sup_{n \leq i \leq n+m} |\alpha_i^n - \alpha| = 0,$$

we have

$$\lim_{n, m \to \infty} \text{distance} \left[\frac{1}{m} \sum_{i=n}^{n+m-1} g_i(\alpha_i^n), G(\alpha)\right] = 0. \tag{6.11'}$$

Then, for almost all ω, *the limit points are contained in an invariant set of the differential inclusion*

$$\dot{\theta} \in G(\theta) + z, \quad z(t) \in -C(\theta(t)). \tag{6.12}$$

Under (A2.6'), *the invariant set is contained in the limit set* $L_H^1 \cup A_H$.

5.7 Convergence for the Lizard Learning and Pattern Classification Problems

Theorems 2.1, 3.1, and 4.1 will be illustrated by proving convergence for two examples from Chapters 1 and 2.

5.7.1 The Lizard Learning Problem

Using the stability Theorem 4.1, convergence with probability one will be proved for the lizard learning problem in Section 2.1. The algorithm is (2.1.3) [equivalently, (2.1.4)], and the assumptions stated in that section will be used. In Theorem 4.1, $\bar{\theta} = 0$ was used for notational simplicity. Here $\theta - \bar{\theta}$ replaces the θ in that theorem.

Define the Liapunov function $V(\theta) = (\theta - \bar{\theta})^2$. Note that because there is a constant C such that $E_n[\tau_n^2 + r_n^2] \leq C$ for all n and ω, we have

$$E_n\left[\theta_{n+1} - \theta_n\right]^2 = O(\epsilon_n^2)\left[1 + \bar{g}^2(\theta_n)\right].$$

Thus,

$$E_n V(\theta_{n+1}) - V(\theta_n) = 2\epsilon_n(\theta_n - \bar{\theta})\bar{g}(\theta_n) + O(\epsilon_n^2)\left[1 + \bar{g}^2(\theta_n)\right]. \tag{7.1}$$

By the properties of $\bar{g}(\cdot)$, the first term on the right side of (7.1) is bounded above by $-\epsilon_n\lambda|\theta_n - \bar{\theta}|^2$ for some positive number λ. Until further notice, suppose that $\epsilon_n \to 0$ with probability one. Then, using the fact that $|\bar{g}(\theta)| \leq c_0 + c_1|\theta - \bar{\theta}|$ for some positive c_i, for large n we have

$$E_n V(\theta_{n+1}) - V(\theta_n) \leq -\epsilon_n\lambda|\theta_n - \bar{\theta}|^2/2 + O(\epsilon_n^2).$$

It will next be shown that (4.2′) can be used with $\tilde{\epsilon}_n = (2K + 2\bar{\theta})/(nE\tau_1)$ and $\theta_n - \bar{\theta}$ replacing θ_n. Recall that $\epsilon_n = 1/W_n$ and $\theta_n = T_n/W_n$. If $|\theta_n - \bar{\theta}| \leq K$ for any real $K > 0$, then $1/W_n \leq (K + \bar{\theta})/T_n$. Recall that $T_n \geq \sum_{i=0}^{n-1} \tau_i$, where τ_n are identically distributed and mutually independent with a positive mean value. Hence, the strong law of large numbers implies that with probability one we eventually have $T_n \geq nE\tau_1/2$. Hence, eventually, with probability one,

$$\epsilon_n I_{\{|\theta_n - \theta| \leq K\}} \leq \frac{2K + 2\bar{\theta}}{nE\tau_1}.$$

Since the other conditions of Theorem 4.1 hold, it follows that $\theta_n \to \bar{\theta}$ with probability one.

It remains to be shown that $\epsilon_n \to 0$ with probability one. As was seen earlier, this will be true if $\theta_n \leq K_1$ infinitely often for some real K_1. Thus, we need to consider what happens if $\theta_n \to \infty$ with a positive probability. Note that as θ_n increases to infinity the probability of pursuit increases to

one. Thus, there are $\delta_0 > 0$ and $K_0 > 0$ sufficiently large such that for $\theta_n \geq K_0$, the probability of pursuit and capture (conditioned on the past) is greater than δ_0. Thus, considering these n such that $\theta_n \geq K_0$, we have a (time varying) Bernoulli sequence with probability of success $\geq \delta_0$, and the strong law of large numbers guarantees that (except on a null set) there is a $c_2 > 0$ such that we eventually have

$$W_{n+1} \geq \sum_{i=1}^{n} w_i J_i I_i I_{\{\theta_i \geq K_0\}} \geq c_2 \sum_{i=1}^{n} I_{\{\theta_i \geq K_0\}}$$

on the set where the right side goes to infinity, where I_i (resp., J_i) is the indicator function of pursuit (resp., capture). Thus, $\epsilon_n \to 0$ with probability one. Finally, we note that these arguments imply that $\epsilon_n = O(1/n)$ with probability one.

5.7.2 *The Pattern Classification Problem*

An application of Theorem 3.1 yields the probability one convergence of the projected form of the algorithm (1.1.18). (The stability Theorem 4.1 can be used if the algorithm is untruncated.) Recall that the sequence $\{y_n, \phi_n\}$ was assumed to be mutually independent. Let H denote the box $\prod_{i=1}^{r}[a_i, b_i]$, where $-\infty < a_i < b_i < \infty$. The algorithm is

$$\theta_{n+1} = \Pi_H \left[\theta_n + \epsilon_n \phi_n \left(y_n - \phi_n' \theta_n\right)\right], \tag{7.2}$$

and it is supposed that (1.1) holds. Suppose that

$$\sup_n E\left[|y_n \phi_n| + |\phi_n|^2\right] < \infty, \tag{7.3}$$

and define $\bar{S}_n = E y_n \phi_n$ and $Q_n = E\phi_n \phi_n'$. Suppose that there are matrices $\bar{S}$ and $Q > 0$ such that for each $t > 0$,

$$\lim_n \sum_{i=n}^{m(t_n+t)} \epsilon_i \left[\bar{S}_i - \bar{S}\right] = 0, \tag{7.4a}$$

$$\lim_n \sum_{i=n}^{m(t_n+t)} \epsilon_i \left[Q_i - Q\right] = 0. \tag{7.4b}$$

Define the martingale difference

$$\delta M_n = \left(\phi_n y_n - \bar{S}_n\right) - \left(\phi_n \phi_n' - Q_n\right)\theta_n,$$

and suppose that it satisfies (3.2) or the sufficient conditions in Theorem 3.2. The algorithm can be written as

$$\theta_{n+1} = \theta_n + \epsilon_n \left[S_n - Q_n \theta_n\right] + \epsilon_n \delta M_n + \epsilon_n Z_n. \tag{7.5}$$

Theorem 3.1 holds with ODE

$$\dot{\theta} = \bar{S} - Q\theta + z,\ z(t) \in -C(\theta(t)). \tag{7.6}$$

The limit points of $\{\theta_n\}$ are the stationary points of (7.6). If $\bar{\theta} = Q^{-1}\bar{S} \in H^0$, then this is the unique limit point. Otherwise it is the point θ on the boundary of H that satisfies $(\bar{S} - Q\theta) \in C(\theta)$. This will be the point in H closest to $\bar{\theta}$. As a practical matter, if the iterates cluster about some point on the boundary, one enlarges the box unless there is a reason not to. The averaging approach of Subsection 1.3.3 can be used to get an asymptotically optimal algorithm (see Chapters 9 and 11).

Now turn to the algorithm (1.1.16). Continue to assume the conditions listed above and that $n\Phi_n^{-1} \to Q^{-1}$ (equivalently, $\Phi_n/n \to Q$) with probability one. Define $\tilde{\epsilon}_n = \Phi_n^{-1}Q$, and write the algorithm as

$$\theta_{n+1} = \Pi_H\left[\theta_n + \tilde{\epsilon}_n Q^{-1}\phi_n\left(y_n - \phi_n'\theta_n\right)\right]. \tag{7.7}$$

By modifying the process on a set of arbitrarily small probability, it can be assumed (without loss of generality) that there are real $\delta_n \to 0$ such that $|n\tilde{\epsilon}_n - I| \leq \delta_n$. Then Theorem 3.1 continues to hold with the ODE

$$\dot{\theta} = Q^{-1}\bar{S} - \theta + z,\ z(t) \in -C(\theta(t)). \tag{7.8}$$

5.8 Convergence to a Local Minimum: A Perturbation Method

Let $g(\cdot) = -f_\theta(\cdot)$ for a real-valued continuously differentiable function $f(\cdot)$. Under appropriate conditions, the theorems in Sections 2 to 4 have shown that $\{\theta_n\}$ converges with probability one. If the algorithm is unconstrained, then the limit points are in the set S of stationary points, that is, the points satisfying $f_\theta(\theta) = 0$. If the algorithm is constrained, with constraint set H, the limit points are the points $\theta \in H$ that satisfy $-f_\theta(\theta) + z = 0$, $z \in -C(\theta)$.

The set S (and similarly for the constrained problem) contains local minima and other types of stationary points such as local maxima and saddles. In practice, the local maxima and saddles are often seen to be unstable points for the stochastic approximation algorithm. Yet this "practical" lack of convergence to the "bad" points does not follow from the given convergence theorems. The "bad" points tend to be unstable for the algorithm because of the interaction of the instability (or marginal stability) properties of the ODE near those points with the perturbing noise. In an intuitive sense, when the iterate is near a stationary point that is not a local minimum, the noise "shakes" the iterate sequence until it is "captured" by a path descending to a more stable point. This is similar in spirit to what happens with the simulated annealing procedure, except

that we are not looking for a global minimum here. Under appropriate "directional nondegeneracy" conditions on the noise, instability theorems based on Liapunov functions can be used to prove the repelling property of local maxima and local saddles. A one dimensional problem was treated in [132, Chapter 5, Theorem 3.1], in which it was shown that convergence to a local maximum point was not possible under certain "nondegeneracy" conditions on the noise and the dynamics; see also [18]. It is often hard to know whether the natural system noise is "sufficiently perturbing."

The main problem in actually proving such an instability result is that little is known about this perturbing noise in general. This is particularly true for high dimensional problems, so that depending on the system noise to "robustify" the algorithm can be risky. Practical "robustification" might require an explicit introduction of carefully selected perturbations. This would reduce the theoretical rate of convergence, as defined in Chapter 10, but we would expect that added robustness has some associated price.

In this section, it will be shown that a slight perturbation of the basic algorithm guarantees convergence with probability one to a local minimum. The proof is still "asymptotic." At this time, we are not suggesting the use of such a perturbed algorithm Indeed, the alteration is not a practical algorithm, although some sort of perturbation will no doubt be needed. In addition, even if an algorithm is constructed to guarantee asymptotic escape from whatever local maximum or saddle the iterates might wander near, there is not necessarily a guarantee of good behavior during any run of "practical length." Keep in mind that if the iterate sequence is near a "bad" point for some large n, then it might stay near that point for a long time afterwards, for either the original or the perturbed forms used here, particularly when the step sizes are small. This can be a problem in practice. However, the result shows the fundamental importance of the structure of the perturbing noise (or added perturbations) and the ODE and suggests other algorithms that will be essentially immune to the presence of "bad" points.

Generally, one might wish to treat the problem of stationary points that are not local minima in the context of the problem where there are many local minima, and convergence to a reasonably good local minimum is desired. The basic methods suggested to date for such a problem involve either appropriate multiple stochastic approximations, as in [187], or some sort of crude perturbation as in [54, 96]. The perturbation itself is equivalent to a "local" restarting, but without the step size sequence being reset.

The following demonstration is for illustrative purposes only. We are concerned with the main idea but not the generality of the results. For practical robustness, we might require that neither the step sizes nor the perturbations go to zero; this case is not covered by the following proof.

The perturbed algorithm. For the sake of simplicity of development, suppose that there is no constraint set and that $\{\theta_n\}$ is bounded with

probability one. Let S consist of a finite number of points, define S_1 to be the set of (finite) local minima, and set $S_2 = S - S_1$. Let ν_k be a sequence of integers going to infinity, define $T_k = t_{\nu_{k+1}} - t_{\nu_k}$, and suppose that $T_k \to \infty$. Let $\{\chi_k\}$ be a sequence of mutually independent random variables uniformly distributed on the unit sphere in $I\!R^r$, and let $\{b_k\}$ be a sequence of positive numbers tending to zero. Also suppose that, for each k, χ_k is independent of $\{Y_i, i \le \nu_k - 1\}$. Suppose that $\sup_n E|Y_n|^2 < \infty$. With the definition $E_n Y_n = \bar{g}(\theta_n) + \beta_n$, define $\delta M_n = Y_n - E_n Y_n$, where E_n denotes the expectation conditioned on $\{\theta_0, \theta_i, Y_{i-1}, i \le n\}$.

The perturbed algorithm is

$$\theta_n = \theta_{n-1} + \epsilon_{n-1} Y_{n-1}, \quad n \ne \nu_k, \text{ any } k,$$

$$\theta_n = \theta_{n-1} + \epsilon_{n-1} Y_{n-1} + b_k \chi_k, \quad n = \nu_k$$

or in the expanded form:

$$\theta_n = \theta_{n-1} + \epsilon_{n-1}\bar{g}(\theta_{n-1}) + \epsilon_{n-1}\beta_{n-1} + \epsilon_{n-1}\delta M_{n-1}, \; n \ne \nu_k, \text{ any } k, \quad (8.1a)$$

$$\theta_n = \theta_{n-1} + \epsilon_{n-1}\bar{g}(\theta_{n-1}) + \epsilon_{n-1}\beta_{n-1} + \epsilon_{n-1}\delta M_{n-1} + b_k\chi_k, \; n = \nu_k, \quad (8.1b)$$

where β_n is a bias on which further conditions will be given later. Notice that ν_k are the perturbation times, and at each perturbation time ν_k, the iterate is changed by $b_k\chi_k$. Assume the conditions of Theorem 2.1 or 3.1, except for those on the constraint set. The asymptotic rate of change of the right continuous piecewise constant process

$$p(t) = \sum_{k:\nu_k \le m(t)} b_k \chi_k$$

is zero, since $T_k \to \infty$ and $b_k \to 0$. Thus, by Theorem 2.1 or 3.1, the limiting ODE is just

$$\dot{\theta} = \bar{g}(\theta) = -f_\theta(\theta), \quad (8.2)$$

as for the unperturbed algorithm. Then, by these theorems, we have probability one convergence to S.

The basic issue in the proof of convergence to a unique (perhaps ω-dependent) local minimum concerns the relations between $\{\nu_k\}$ and $\{b_k\}$, and appropriate assumptions will now be stated. The conditions will be discussed in detail after the proof. It will be seen that they are not very strong and are quite natural for the problem at hand. Let $\bar{g}(\cdot)$ be Lipschitz continuous, with Lipschitz constant K. Suppose that for some $\bar{K} > 2K$,

$$\sum_{i=\nu_k}^{\nu_{k+1}} \prod_{j=i+1}^{\nu_{k+1}} (1 + \bar{K}\epsilon_j)\epsilon_i^2 \to 0, \quad (8.3)$$

$$\sum_{i=\nu_k}^{\nu_{k+1}} \prod_{j=i+1}^{\nu_{k+1}} (1 + K\epsilon_j)\epsilon_i|\beta_i| \to 0 \text{ w.p.1}, \quad (8.4)$$

as $k \to \infty$.

Let $R_\rho(x)$ denote a sphere in $\mathbb{R}^r$ of radius ρ and center x. For a set $Q \subset \mathbb{R}^r$, define $R_\rho(Q) = \cup_{x \in Q} R_\delta(x)$, a ρ-neighborhood of the set Q. Let $\delta > 0$ denote the minimum distance between any pair of points in S. For $x \in S$, let $\mathcal{S}_x$ denote the set of stationary points y such that $f(y) < f(x)$.

Finally, there is the following critical "accessibility" condition [despite its arcane appearance, it is quite reasonable and will be discussed further after the proof]: For small enough $\delta_0 > \delta_1 > 0$, with $\delta > \delta_0$ there is a $\delta_2 > 0$ such that for each $x \in S_2$ and each $y \in R_{\delta_0}(x)$,

$$P\{\text{a path of (8.2) starting at random on } \partial R_{b_k}(y) \text{ reaches } R_{\delta_1}(\mathcal{S}_x) \text{ by } T_k\} \geq \delta_2. \tag{8.5}$$

Theorem 8.1. *Assume the conditions stated above. Then the limit points must be in S_1 with probability one.*

Proof. The proof starts by assuming that, with positive probability, θ_n converges to some point in S_2. The next step is to show that the stochastic approximation process stays very close to the path of the ODE over a sufficiently long time interval, when they both start from the same perturbed iterate value, and that on this time interval, with strictly positive probability, the ODE will reach a sufficiently small neighborhood of a point with lower value. Since asymptotically, the path cannot return from this neighborhood to any sufficiently small neighborhood of the point with the higher value, there will be a contradiction of convergence to a point in S_2 with positive probability. The main work of the proof involves estimates of the distance between the paths of the ODE (8.2) and the stochastic approximation process (8.1).

Define $\bar{X}_0^k = \theta_{\nu_k}$, and let $\bar{X}^k(\cdot)$ denote the solution to the ODE (8.2) with initial condition $\bar{X}_0^k$. Define the samples of the solution to the ODE: $\bar{X}_n^k = X^k(t_n - t_{\nu_k})$, for $n \geq \nu_k$. Define $\epsilon_n^k = \epsilon_{\nu_k + n}$, and define X_n^k by $X_0^k = \theta_{\nu_k}$ and

$$X_{n+1}^k = X_n^k + \epsilon_n^k \bar{g}(X_n^k), \quad \text{for } n < \nu_{k+1} - \nu_k, \tag{8.6}$$

a difference equation approximation to the ODE. Note that

$$\bar{X}_{n+1}^k = \bar{X}_n^k + \epsilon_n^k \bar{g}(\bar{X}_n^k) + O([\epsilon_n^k]^2), \quad n < \nu_{k+1} - \nu_k. \tag{8.7}$$

Define $\beta_n^k = \beta_{\nu_k + n}$. Define $\tilde{X}_n^k$ by $\tilde{X}_0^k = \theta_{\nu_k}$, and for $n < \nu_{k+1} - \nu_k$,

$$\tilde{X}_{n+1}^k = \tilde{X}_n^k + \epsilon_n^k \bar{g}(\tilde{X}_n^k) + \epsilon_n^k \beta_n^k, \quad n < \nu_{k+1} - \nu_k. \tag{8.8}$$

This is the stochastic approximation process without the δM_n terms.

The error between the solutions of (8.1) and (8.2) will be estimated via the inequality

$$\begin{aligned} |\bar{X}_n^k - \theta_{\nu_k + n}| &\leq |\bar{X}_n^k - X_n^k| + |X_k^n - \tilde{X}_n^k| + |\theta_{\nu_k + n} - \tilde{X}_n^k| \\ &\equiv \bar{\Delta}_n^k + \tilde{\Delta}_n^k + \hat{\Delta}_n^k. \end{aligned} \tag{8.9}$$

Subtracting (8.6) from (8.7) and taking absolute values yield

$$\bar{\Delta}_{n+1}^k \le \bar{\Delta}_n^k + \epsilon_n^k \left|\bar{g}(\bar{X}_n^k) - \bar{g}(X_n^k)\right| + O([\epsilon_n^k]^2).$$

Using the Lipschitz condition on $\bar{g}(\cdot)$ yields

$$\Delta_{n+1}^k \le \bar{\Delta}_n^k + K\epsilon_n^k \bar{\Delta}_n^k + O([\epsilon_n^k]^2).$$

Solving this inequality yields the upper bound

$$\sup_{0\le n<\nu_{k+1}-\nu_k} \bar{\Delta}_n^k \le K_1 \sum_{i=\nu_k}^{\nu_{k+1}} \prod_{j=i+1}^{\nu_{k+1}} (1 + K\epsilon_j)\epsilon_i^2$$

for some real K_1, which is bounded above by (K_1 times) the left side of the expression (8.3) since $K > K$. Analogously, for some real K_1, we obtain

$$\sup_{0\le n<\nu_{k+1}-\nu_k} \tilde{\Delta}_n^k \le K_1 \sum_{i=\nu_k}^{\nu_{k+1}} \prod_{j=i+1}^{\nu_{k+1}} (1 + K\epsilon_j)\epsilon_i|\beta_i|,$$

which is (K_1 times) the left side of the expression (8.4).

Define $\delta_n^k = \theta_{\nu_k+n} - \tilde{X}_n^k$ and $\delta M_n^k = \delta M_{\nu_k+n}$. Note that $\hat{\Delta}_n^k = |\delta_n^k|$. Then, for $n < \nu_{k+1} - \nu_k$,

$$\delta_{n+1}^k = \delta_n^k + \epsilon_n^k \left[\bar{g}(\theta_{\nu_k+n}) - \bar{g}(\tilde{X}_n^k)\right] + \epsilon_n^k \delta M_n^k. \tag{8.10}$$

Letting $\Delta_n^k = E|\delta_n^k|^2$ and using (8.10), the Lipschitz condition on $\bar{g}(\cdot)$ and the fact that δM_n^k is of mean zero, with uniformly bounded variance, and is orthogonal to the other random variables on the right side of (8.10) yield

$$\Delta_{n+1}^k \le \Delta_n^k + 2\epsilon_n^k K \Delta_n^k + [\epsilon_n^k]^2 K^2 \Delta_n^k + O([\epsilon_n^k]^2).$$

Since $\bar{K} > 2K$, solving this inequality for large k yields

$$\sup_{0\le n<\nu_{k+1}-\nu_k} \Delta_n^k = O(\text{lhs of (8.3)}).$$

Putting the pieces together yields

$$\begin{aligned} &E \sup_{0\le n<\nu_{k+1}-\nu_k} |X_n^k - \theta_{\nu_k+n}| \\ &\quad = O(\text{lhs of (8.3)}) + O(\text{lhs of (8.4)}) + O((\text{lhs of (8.3)})^{1/2}). \end{aligned} \tag{8.11}$$

Equation (8.11) and conditions (8.3) and (8.4) imply that the (interpolation of the) iterate path $\{\theta_{\nu_k+n}, n < \nu_{k+1} - \nu_k\}$ tracks the solution $\{\bar{X}^k(t), t < T_k\}$ arbitrarily closely with a probability arbitrarily close to one as $k \to \infty$. To complete the proof, we need to show a contradiction if it is supposed that the limit points of $\{\theta_n\}$ are contained in S_2 with a positive probability.

Recall the definitions of δ and δ_0 above (8.5). Without loss of generality, suppose that $\delta \gg \delta_0$. Let $\bar{\delta} > 0$ be such that for all $x \in S_2$, $\bar{\delta} \leq f(x) - f(\mathcal{S}_x)$.

Since $\{\theta_n\}$ converges with probability one, θ_n eventually lies in a $\delta_0/2$ sphere about its limit. Suppose that $x_0 \in S_2$ is a limit with a positive probability. Then (8.5) and (8.11), together with the conditions (8.3) and (8.4), imply that

$$\liminf_k P\left\{f(\theta_{\nu_{k+1}-1}) - f(\theta_{\nu_k}) \leq -\bar{\delta}/2 \mid \theta_{\nu_k} \in R_{\delta_0}(x_0)\right\} \geq \delta_2/2. \quad (8.12)$$

Inequality (8.12) and the Borel–Cantelli Lemma imply that (with probability one relative to the paths converging to x_0) an arbitrarily small neighborhood of a stationary point with a lower value of $f(\cdot)$ will eventually be reached and the iterate will be in that neighborhood infinitely often. As seen in Theorems 2.1 or 3.1, the path cannot return to a small neighborhood of x_0 infinitely often. Thus, no point in S_2 can be a limit point with positive probability. □

Comments on the conditions. Recall that k does not index the iterate number, but rather the number of perturbations to iterate ν_k. It will next be shown that the hypotheses (8.3)–(8.5) hold under some common conditions.

Let $\epsilon_n = 1/n$. Then for large k, (8.3) is asymptotically bounded by the order of

$$e^{\bar{K}T_k} T_k \epsilon_{\nu_k}.$$

A slight increase of $\bar{K}$ allows the use of simply

$$e^{KT_k} \epsilon_{\nu_k}. \quad (8.13)$$

The asymptotic relation between ν_k and T_k is given by

$$\int_{\nu_k}^{\nu_{k+1}} \frac{1}{s} ds \approx T_k.$$

Thus we have approximately $\nu_{k+1} = \nu_k e^{T_k}$ or

$$\nu_k \approx \exp\left[\sum_{i=0}^{k-1} T_i\right].$$

Let $T_0 = 1$ and (for $k \geq 1$) $T_k = q_k \log k$, where $q_k \to \infty$ slowly, a useful form that will be motivated in what follows. Then (8.13) is approximately

$$\frac{\exp \bar{K} T_k}{\exp\left[\sum_{i=0}^{k-1} T_i\right]} \approx \frac{k^{Kq_k}}{\prod_{i=1}^{k-1} [i^{q_i}]},$$

which goes to zero as $k \to \infty$ if $q_k \to \infty$ slowly enough. Thus (8.3) holds.

Next, consider (8.4), where β_n are due to the use of a finite difference method to estimate the derivative of $f(\theta)$. Let the difference intervals be $c_n = O(1/n^\alpha)$, where $\alpha \in (0,1)$ with $\beta_n = O(c_n)$. Then the approximation procedure yields that for K_1 slightly larger than K, (8.4) is asymptotically bounded by the order of

$$e^{K_1 T_k} |\beta_{\nu_k}|,$$

which is of the order of

$$\frac{k^{K_1 q_k}}{\nu_k^\alpha} \approx \frac{k^{K_1 q_k}}{\prod_{i=1}^{k-1} [i^{\alpha q_i}]},$$

which also goes to zero as $k \to \infty$ if $q_k \to \infty$ slowly enough. Thus (8.4) holds.

The sequences $\{T_k\}$ and $\{\beta_k\}$ are connected via (8.5). This condition assumes that a certain amount of time is required for the solution of the ODE (8.2) to escape from a small neighborhood of an unstable or marginally stable point x_0 when it starts very close to that point. This time will depend on how close to x_0 one starts. After the kth perturbation, one starts at a random point on the surface of a sphere with radius b_k and center θ_{ν_k}. If θ_{ν_k} is very close to the "bad" point x_0, we require that there be some "minimal part" of the surface of the sphere from which the solution of the ODE will get to a small neighborhood of a "better" point in time T_k. Consider the following simple example, from which the general motivation for and reasonableness of (8.5) should be clear. Let $f(\theta) = \cos\theta$. Then $\ddot{g}(\theta) = -f_\theta(\theta) = \sin\theta$. There is a local maximum at $\theta = 0$ and local minima at $\theta = \pm\pi$. Suppose we start the perturbed trajectory at a point $\alpha_k > 0$, where α_k is small. Then the solution of (8.2) will eventually reach any small neighborhood of the local minimum $\theta = \pi$. For small values of θ, the ODE is linearizable and is approximately

$$\dot{\theta} = \sin\theta \approx \theta, \quad \theta(0) = \alpha_k.$$

Thus, to reach a small neighborhood of $\theta = \pi$, one requires a time T_k such that $\alpha_k e^{T_k} = O(1)$. To be certain that the solution of the ODE will escape from a small neighborhood, we use $\alpha_k e^{T_k} \to \infty$. The more slowly α_k goes to zero, the smaller T_k can be. Now, suppose that $T_k = q_k \log k$, where $q_k \to \infty$, as used in the preceding discussion. Then, we need that $\alpha_k k^{q_k} \to \infty$. The argument for a negative perturbation is analogous. Now suppose that we start somewhere in a small neighborhood of the origin and not at the origin itself, and then perturb by $b_k \chi_k$, where the random χ_k takes the equally likely values ± 1, the $\{\chi_k\}$ are mutually independent and $b_k k^{q_k} \to \infty$. Then the perturbation by $b_k \chi_k$ can take the iterate closer either to the origin or to the nearest stable point (π or $-\pi$), each possibility occurring with probability 1/2. Thus, (8.5) holds.