Preface and Introduction

The basic stochastic approximation algorithms introduced by Robbins and Monro and by Kiefer and Wolfowitz in the early 1950s have been the subject of an enormous literature, both theoretical and applied. This is due to the large number of applications and the interesting theoretical issues in the analysis of "dynamically defined" stochastic processes. The basic paradigm is a stochastic difference equation such as $\theta_{n+1} = \theta_n + \epsilon_n Y_n$, where θ_n takes its values in some Euclidean space, Y_n is a random variable, and the "step size" $\epsilon_n > 0$ is small and might go to zero as $n \to \infty$. In its simplest form, θ is a parameter of a system, and the random vector Y_n is a function of "noise-corrupted" observations taken on the system when the parameter is set to θ_n. One recursively adjusts the parameter so that some goal is met asymptotically. This book is concerned with the qualitative and asymptotic properties of such recursive algorithms in the diverse forms in which they arise in applications. There are analogous continuous time algorithms, but the conditions and proofs are generally very close to those for the discrete time case.

The original work was motivated by the problem of finding a root of a continuous function $\bar{g}(\theta)$, where the function is not known but the experimenter is able to take "noisy" measurements at any desired value of θ. Recursive methods for root finding are common in classical numerical analysis, and it is reasonable to expect that appropriate stochastic analogs would also perform well.

In one classical example, θ is the level of dosage of a drug, and the function $\bar{g}(\theta)$, assumed to be increasing with θ, is the probability of success at dosage level θ. The level at which $\bar{g}(\theta)$ takes a given value v is sought.

The probability of success is known only by experiment at whatever values of θ are selected by the experimenter, with the experimental outcome being either success or failure. Thus, the problem cannot be solved analytically. One possible approach is to take a sufficient number of observations at some fixed value of θ, so that a good estimate of the function value is available, and then to move on. Since most such observations will be taken at parameter values that are not close to the optimum, much effort might be wasted in comparison with the stochastic approximation algorithm $\theta_{n+1} = \theta_n + \epsilon_n[v - \text{observation at } \theta_n]$, where the parameter value moves (on the average) in the correct direction after each observation. In another example, we wish to minimize a real-valued continuously differentiable function $f(\cdot)$ of θ. Here, θ_n is the nth estimate of the minimum, and Y_n is a noisy estimate of the negative of the derivative of $f(\cdot)$ at θ_n, perhaps obtained by a Monte Carlo procedure. The algorithms are frequently constrained in that the iterates θ_n are projected back to some set H if they ever leave it. The mathematical paradigms have posed substantial challenges in the asymptotic analysis of recursively defined stochastic processes.

A major insight of Robbins and Monro was that, if the step sizes in the parameter updates are allowed to go to zero in an appropriate way as $n \to \infty$, then there is an implicit averaging that eliminates the effects of the noise in the long run. An excellent survey of developments up to about the mid 1960s can be found in the book by Wasan [250]. More recent material can be found in [16, 48, 57, 67, 135, 225]. The book [192] deals with many of the issues involved in stochastic optimization in general.

In recent years, algorithms of the stochastic approximation type have found applications in new and diverse areas, and new techniques have been developed for proofs of convergence and rate of convergence. The actual and potential applications in signal processing and communications have exploded. Indeed, whether or not they are called stochastic approximations, such algorithms occur frequently in practical systems for the purposes of noise or interference cancellation, the optimization of "post processing" or "equalization" filters in time varying communication channels, adaptive antenna systems, adaptive power control in wireless communications, and many related applications. In these applications, the step size is often a small constant $\epsilon_n = \epsilon$, or it might be random. The underlying processes are often nonstationary and the optimal value of θ can change with time. Then one keeps ϵ_n strictly away from zero in order to allow "tracking." Such tracking applications lead to new problems in the asymptotic analysis (e.g., when ϵ_n are adjusted adaptively); one wishes to estimate the tracking errors and their dependence on the structure of the algorithm.

New challenges have arisen in applications to adaptive control. There has been a resurgence of interest in general "learning" algorithms, motivated by the training problem in artificial neural networks [7, 51, 97], the on-line learning of optimal strategies in very high-dimensional Markov decision processes [113, 174, 221, 252] with unknown transition probabilities,

in learning automata [155], recursive games [11], convergence in sequential decision problems in economics [175], and related areas. The actual recursive forms of the algorithms in many such applications are of the stochastic approximation type. Owing to the types of simulation methods used, the "noise" might be "pseudorandom" [184], rather than random.

Methods such as infinitesimal perturbation analysis [101] for the estimation of the pathwise derivatives of complex discrete event systems enlarge the possibilities for the recursive on-line optimization of many systems that arise in communications or manufacturing. The appropriate algorithms are often of the stochastic approximation type and the criterion to be minimized is often the average cost per unit time over the infinite time interval.

Iterate and observation averaging methods [6, 149, 216, 195, 267, 268, 273], which yield nearly optimal algorithms under broad conditions, have been developed. The iterate averaging effectively adds an additional time scale to the algorithm. Decentralized or asynchronous algorithms introduce new difficulties for analysis. Consider, for example, a problem where computation is split among several processors, operating and transmitting data to one another asynchronously. Such algorithms are only beginning to come into prominence, due to both the developments of decentralized processing and applications where each of several locations might control or adjust "local variables," but where the criterion of concern is global.

Despite their successes, the classical methods are not adequate for many of the algorithms that arise in such applications. Some of the reasons concern the greater flexibility desired for the step sizes, more complicated dependence properties of the noise and iterate processes, the types of constraints that might occur, ergodic cost functions, possibly additional time scales, nonstationarity and issues of tracking for time-varying systems, data-flow problems in the decentralized algorithm, iterate-averaging algorithms, desired stronger rate of convergence results, and so forth.

Much modern analysis of the algorithms uses the so-called ODE (ordinary differential equation) method introduced by Ljung [164] and extensively developed by Kushner and coworkers [123, 135, 142] to cover quite general noise processes and constraints by the use of weak ergodic or averaging conditions. The main idea is to show that, asymptotically, the noise effects average out so that the asymptotic behavior is determined effectively by that of a "mean" ODE. The usefulness of the technique stems from the fact that the ODE is obtained by a "local analysis," where the dynamical term of the ODE at parameter value θ is obtained by averaging the Y_n as though the parameter were fixed at θ. Constraints, complicated state dependent noise processes, discontinuities, and many other difficulties can be handled. Depending on the application, the ODE might be replaced by a constrained (projected) ODE or a differential inclusion. Owing to its versatility and naturalness, the ODE method has become a fundamental technique in the current toolbox, and its full power will be apparent from the results in this book.

The first three chapters describe applications and serve to motivate the algorithmic forms, assumptions, and theorems to follow. Chapter 1 provides the general motivation underlying stochastic approximation and describes various classical examples. Modifications of the algorithms due to robustness concerns, improvements based on iterate or observation averaging methods, variance reduction, and other modeling issues are also introduced. A Lagrangian algorithm for constrained optimization with noise corrupted observations on both the value function and the constraints is outlined. Chapter 2 contains more advanced examples, each of which is typical of a large class of current interest: animal adaptation models, parametric optimization of Markov chain control problems, the so-called Q-learning, artificial neural networks, and learning in repeated games. The concept of state-dependent noise, which plays a large role in applications, is introduced. The optimization of discrete event systems is introduced by the application of infinitesimal perturbation analysis to the optimization of the performance of a queue with an ergodic cost criterion. The mathematical and modeling issues raised in this example are typical of many of the optimization problems in discrete event systems or where ergodic cost criteria are involved. Chapter 3 describes some applications arising in adaptive control, signal processing, and communication theory, areas that are major users of stochastic approximation algorithms. An algorithm for tracking time varying parameters is described, as well as applications to problems arising in wireless communications with randomly time varying channels. Some of the mathematical results that will be needed in the book are collected in Chapter 4.

The book also develops "stability" and combined "stability–ODE" methods for unconstrained problems. Nevertheless, a large part of the work concerns constrained algorithms, because constraints are generally present either explicitly or implicitly. For example, in the queue optimization problem of Chapter 2, the parameter to be selected controls the service rate. What is to be done if the service rate at some iteration is considerably larger than any possible practical value? Either there is a problem with the model or the chosen step sizes, or some bizarre random numbers appeared. Furthermore, in practice the "physics" of models at large parameter values are often poorly known or inconvenient to model, so that whatever "convenient mathematical assumptions" are made, they might be meaningless at large state values. No matter what the cause is, one would normally alter the unconstrained algorithm if the parameter θ took on excessive values. The simplest alteration is truncation. Of course, in addition to truncation, a practical algorithm would have other safeguards to ensure robustness against "bad" noise or inappropriate step sizes, etc. It has been somewhat traditional to allow the iterates to be unbounded and to use stability methods to prove that they do, in fact, converge. This approach still has its place and is dealt with here. Indeed, one might even alter the dynamics by introducing "soft" constraints, which have the desired stabilizing effect.

However, allowing unbounded iterates seems to be of greater mathematical than practical interest. Owing to the interest in the constrained algorithm, the "constrained ODE" is also discussed in Chapter 4. The chapter contains a brief discussion of stochastic stability and the perturbed stochastic Liapunov function, which play an essential role in the asymptotic analysis.

The first convergence results appear in Chapter 5, which deals with the classical case where the Y_n can be written as the sum of a conditional mean $g_n(\theta_n)$ and a noise term, which is a "martingale difference." The basic techniques of the ODE method are introduced, both with and without constraints. It is shown that, under reasonable conditions on the noise, there will be convergence with probability one to a "stationary point" or "limit trajectory" of the mean ODE for step-size sequences that decrease at least as fast as $\alpha_n / \log n$, where $\alpha_n \to 0$. If the limit trajectory of the ODE is not concentrated at a single point, then the asymptotic path of the stochastic approximation is concentrated on a limit or invariant set of the ODE that is also "chain recurrent" [9, 89]. Equality constrained problems are included in the basic setup.

Much of the analysis is based on interpolated processes. The iterates $\{\theta_n\}$ are interpolated into a continuous time process with interpolation intervals $\{\epsilon_n\}$. The asymptotics (large n) of the iterate sequence are also the asymptotics (large t) of this interpolated sequence. It is the paths of the interpolated process that are approximated by the paths of the ODE.

If there are no constraints, then a stability method is used to show that the iterate sequence is recurrent. From this point on, the proofs are a special case of those for the constrained problem. As an illustration of the methods, convergence is proved for an animal learning example (where the step sizes are random, depending on the actual history) and a pattern classification problem. In the minimization of convex functions, the subdifferential replaces the derivative, and the ODE becomes a differential inclusion, but the convergence proofs carry over.

Chapter 6 treats probability one convergence with correlated noise sequences. The development is based on the general "compactness methods" of [135]. The assumptions on the noise sequence are intuitively reasonable and are implied by (but weaker than) strong laws of large numbers. In some cases, they are both necessary and sufficient for convergence. The way the conditions are formulated allows us to use simple and classical compactness methods to derive the mean ODE and to show that its asymptotics characterize that of the algorithm. Stability methods for the unconstrained problem and the generalization of the ODE to a differential inclusion are discussed. The methods of large deviations theory provide an alternative approach to proving convergence under weak conditions, and some simple results are presented.

In Chapters 7 and 8, we work with another type of convergence, called *weak convergence*, since it is based on the theory of weak convergence of a sequence of probability measures and is weaker than convergence with

probability one. It is actually much easier to use in that convergence can be proved under weaker and more easily verifiable conditions and generally with substantially less effort. The approach yields virtually the same information on the asymptotic behavior. The weak convergence methods have considerable theoretical and modeling advantages when dealing with complex problems involving correlated noise, state dependent noise, decentralized or asynchronous algorithms, and discontinuities in the algorithm. It will be seen that the conditions are often close to minimal. Only a very elementary part of the theory of weak convergence of probability measures will be needed; this is covered in the second part of Chapter 7. The techniques introduced are of considerable importance beyond the needs of the book, since they are a foundation of the theory of approximation of random processes and limit theorems for sequences of random processes.

When one considers how stochastic approximation algorithms are used in applications, the fact of ultimate convergence with probability one can be misleading. Algorithms do not continue on to infinity, particularly when $\epsilon_n \to 0$. There is always a stopping rule that tells us when to stop the algorithm and to accept some function of the recent iterates as the "final value." The stopping rule can take many forms, but whichever it takes, all that we know about the "final value" at the stopping time is information of a distributional type. There is no difference in the conclusions provided by the probability one and the weak convergence methods. In applications that are of concern over long time intervals, the actual physical model might "drift." Indeed, it is often the case that the step size is not allowed to go to zero, and then there is no general alternative to the weak convergence methods at this time.

The ODE approach to the limit theorems obtains the ODE by appropriately averaging the dynamics, and then by showing that some subset of the limit set of the ODE is just the set of asymptotic points of the $\{\theta_n\}$. The ODE is easier to characterize, and requires weaker conditions and simpler proofs when weak convergence methods are used. Furthermore, it can be shown that $\{\theta_n\}$ spends "nearly all" of its time in an arbitrarily small neighborhood of the limit point or set. The use of weak convergence methods can lead to better probability one proofs in that, once we know that $\{\theta_n\}$ spends "nearly all" of its time (asymptotically) in some small neighborhood of the limit point, then a *local analysis* can be used to get convergence with probability one. For example, the methods of Chapters 5 and 6 can be applied locally, or the local large deviations methods of [63] can be used. Even when we can only prove weak convergence, if θ_n is close to a stable limit point at iterate n, then under broad conditions the mean escape time (indeed, if it ever does escape) from a small neighborhood of that limit point is at least of the order of e^{c/ϵ_n} for some $c > 0$.

Section 7.2 is motivational in nature, aiming to relate some of the ideas of weak convergence to probability one convergence and convergence in distribution. It should be read only "lightly." The general theory is covered

in Chapter 8 for a broad variety of algorithms, using what might be called "weak local ergodic theorems." The essential conditions concern the rates of decrease of the conditional expectation of the future noise given the past noise, as the time difference increases. Chapter 9 illustrates the relative convenience and power of the methods of Chapter 8 by providing proofs of convergence for some of the examples in Chapters 2 and 3.

Chapter 10 concerns the rate of convergence. Loosely speaking, a standard point of view is to show that a sequence of suitably normalized iterates, say of the form $(\theta_n - \bar{\theta})/\sqrt{\epsilon_n}$ or $n^\beta(\theta_n - \bar{\theta})$ for an appropriate $\beta > 0$, converges in distribution to a normally distributed random variable with mean zero and finite covariance matrix $\bar{V}$. We will do a little better and prove that the continuous time process obtained from suitably interpolated normalized iterates converges "weakly" to a stationary Gauss–Markov process, whose covariance matrix (at any time t) is $\bar{V}$. The methods use only the techniques of weak convergence theory that are outlined in Chapter 7.

The use of stochastic approximation for the minimization of functions of a very high-dimensional argument has been of increasing interest. Owing to the high dimension, the classical Kiefer–Wolfowitz procedures can be very time consuming to use. As a result, there is much current interest in the so-called random-directions methods, where at each step n one chooses a direction d_n at random, obtains a noisy estimate $\widehat{Y}_n$ of the derivative in direction d_n, and moves an increment $-\epsilon_n \widehat{Y}_n$. Although such methods have been of interest and used in various ways for a long time [135], convincing arguments concerning their value and the appropriate choices of the direction vectors and scaling were lacking. The paper [226] proposed a different way of getting the directions and attracted needed attention to this problem. The proof of convergence of the random-directions methods that have been suggested to date are exactly the same as that for the classical Kiefer–Wolfowitz procedure (as in Chapter 5). The comparison of the rates of convergence under the different ways of choosing the random directions is given at the end of Chapter 10, and shows that the older and newer methods have essentially the same properties, when the norms of the direction vectors d_n are the same. It is seen that the random-directions methods can be quite advantageous, but care needs to be exercised in their use.

The performance of the stochastic approximation algorithms depends heavily on the choice of the step size sequence ϵ_n, and the lack of a general approach to getting good sequences has been a handicap in applications. In [195], Polyak and Juditsky showed that, if the coefficients ϵ_n go to zero "slower" than $O(1/n)$, then the averaged sequence $\sum_{i=1}^n \theta_i/n$ converges to its limit at an optimal rate. This implies that the use of relatively large step sizes, while letting the "off-line" averaging take care of the increased noise effects, will yield a substantial overall improvement. These results have since been corroborated by numerous simulations and extended mathematically. In Chapter 11, it is first shown that the averaging improves the

asymptotic properties whenever there is a "classical" rate of convergence theorem of the type derived in Chapter 10, including the constant $\epsilon_n = \epsilon$ case. This will give the minimal window over which the averaging will yield an improvement. The maximum window of averaging is then obtained by a direct computation of the asymptotic covariance of the averaged process. Intuitive insight is provided by relating the behavior of the original and the averaged process to that of a three-time-scale discrete-time algorithm where it is seen that the key property is the separation of the time scales.

Chapter 12 concerns decentralized and asynchronous algorithms, where the work is split between several processors, each of which has control over a different set of parameters. The processors work at different speeds, and there can be delays in passing information to each other. Owing to the asynchronous property, the analysis must be in "real" rather than "iterate" time. This complicates the notation, but all of the results of the previous chapters can be carried over. Typical applications are decentralized optimization of queueing networks and Q-learning.

Some topics are not covered. As noted, the algorithm in continuous time differs little from that in discrete time. The basic ideas can be extended to infinite-dimensional problems [17, 19, 66, 87, 144, 185, 201, 214, 219, 246, 247, 248, 277]. The function minimization problem where there are many local minima has attracted some attention [81, 130, 258], but little is known at this time concerning effective methods. Some effort [31] has been devoted to showing that suitable conditions on the noise guarantee that there cannot be convergence to an unstable or marginally stable point of the ODE. Such results are needed and do increase confidence in the algorithms. The conditions can be hard to verify, particularly in high-dimensional problems, and the results do not guarantee that the iterates would not actually spend a lot of time near such bad points, particularly when the step sizes are small and there is poor initial behavior. Additionally, one tries to design the procedure and use variance reduction methods to reduce the effects of the noise.

Penalty-multiplier and Lagrangian methods (other than the discussion in Chapter 1) for constrained problems are omitted and are discussed in [135]. They involve only minor variations on what is done here, but they are omitted for lack of space. We concentrate on algorithms defined on r-dimensional Euclidean space, except as modified by inequality or equality constraints. The treatment of the equality constrained problem shows that the theory also covers processes defined on smooth manifolds.

We express our deep gratitude to Paul Dupuis and Felisa Vazquèz-Abad, for their careful reading and critical remarks on various parts of the manuscript of the first edition. Sid Yakowitz also provided critical remarks for the first edition; his passing away is a great loss. The long-term support and encouragement of the National Science Foundation and the Army Research Office are also gratefully acknowledged.

Comment on the second edition. This second edition is a thorough revision, although the main features and the structure of the book remain unchanged. The book contains many additional results and more detailed discussion; for example, there is a fuller discussion of the asymptotic behavior of the algorithms, Markov and non-Markov state-dependent-noise, and two-time-scale problems. Additional material on applications, in particular, in communications and adaptive control, has been added. Proofs are simplified where possible.

Notation and numbering. Chapters are divided into sections, and sections into subsections. Within a chapter, (1.2) (resp., (A2.1)) denotes Equation 2 of Section 1 (resp., Assumption 2 of Section 1). Section 1 (Subsection 1.2, resp.) always means the first section (resp., the second subsection of the first section) in the chapter in which the statement is used. To refer to equations (resp., assumptions) in other chapters, we use, e.g., (1.2.3) (resp., (A1.2.3)) to denote the third equation (resp., the third assumption) in Section 2 of Chapter 1. When not in Chapter 1, Section 1.2 (resp., Subsection 1.2.3) means Section 2 (resp., Subsection 3 of Section 2) of Chapter 1.

Throughout the book, $|\cdot|$ denotes either a Euclidean norm or a norm on the appropriate function spaces, which will be clear from the context. A point x in a Euclidean space is a column vector, and the ith component of x is denoted by x_i. However, the ith component of θ is denoted by θ^i, since subscripts on θ are used to denote the value at a time n. The symbol $'$ denotes transpose. Moreover, both A' and $(A)'$ will be used interchangeably, e.g., both $g_n^{\epsilon,\prime}(\theta)$ and $(g_n^\epsilon(\theta))'$ denote the transpose of $g_n^\epsilon(\theta)$. Subscripts θ and x denote either a gradient or a derivative, depending on whether the variable is vector or real-valued. For convenience, we list many of the symbols at the end of the book.

Providence, Rhode Island, USA — Harold J. Kushner
Detroit, Michigan, USA — G. George Yin