Stochastic Approximation: A Survey
Harold Kushner, Brown University, November, 2008[1]

Abstract. Stochastic recursive algorithms, also known as stochastic approximation, take many forms and have numerous applications. It is the asymptotic properties that are of interest. The early history, starting with the work of Robbins and Monro, is discussed. An approach to proofs of convergence with probability one is illustrated by a stability-type argument. For general noise processes and algorithms the most powerful current approach is what is called the ODE (ordinary differential equations) method. The algorithm is interpolated into a continuous-time process, which is shown to converge to the solution of an ODE, whose asymptotic properties are those of the algorithm. There are probability one and weak convergence methods, the latter being the easiest to use and the most powerful. After discussing the basic ideas and giving some standard proofs, extensions are outlined. These include multiple time scales, tracking of time changing systems, state-dependent noise, rate of convergence, and random direction methods for high-dimensional problems.

Introductory comments. Stochastic Approximation (SA) deals with asymptotic properties of recursive stochastic algorithms of the types

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n, \tag{1.1}$$

$$\theta^\epsilon_{n+1} = \theta^\epsilon_n + \epsilon Y^\epsilon_n, \tag{1.2}$$

$$\theta_{n+1} = \Pi_H[\theta_n + \epsilon_n Y_n] = \theta_n + \epsilon_n(Y_n + z_n), \tag{1.3}$$

$$\theta^\epsilon_{n+1} = \Pi_H[\theta^\epsilon_n + \epsilon Y^\epsilon_n] = \theta^\epsilon_n + \epsilon(Y^\epsilon_n + z^\epsilon_n), \tag{1.4}$$

where $\theta_n \in I\!R^r$, Euclidean r-space, $\epsilon_n > 0, \epsilon > 0$, Y_n and Y^ϵ_n are functions of θ_n and random noise, and $E|Y_n| < \infty$ for each n. The Π_H denotes projection onto some compact non-empty constraint set H and z_n, z^ϵ_n are the projection terms. The case of random ϵ_n is important, but unless otherwise noted, ϵ_n is real-valued and tends to zero. It is always assumed that

$$\sum_n \epsilon_n = \infty. \tag{1.5}$$

The references to be given illustrate the large number of applications. Only a few key methods and some brief comments on proofs will be given to give the flavor of the subject. The next section briefly surveys some of the original work and gives a "stability" argument for the convergence of a basic algorithm. The so-called ODE (ordinary differential equations method) is perhaps the most powerful in use now, and is introduced in Section 2. Section 3 deals with the weak convergence form of this method, the easiest to use. Various important extensions are described in Section 4.

[1] The work was partially supported by NSF grant DMS 0804822

1. Brief comments on early work. The original work was that of Robbins and Monro [57] on (1.1) that sought the unique zero $\bar{\theta}$ of a continuous real-valued and unknown function $\bar{g}(\cdot)$ on the real line. Only noise corrupted observations of the form $Y_n = \bar{g}(\theta_n) + \delta M_n$ were available. The observation noise δM_n is a martingale difference sequence (with respect to the filtration induced by $\{\theta_0, Y_n, n < \infty\}$) with uniformly bounded variances, $\bar{g}(\theta)(\theta - \bar{\theta}) < 0, \theta \neq \bar{\theta}$, $\bar{g}(\cdot)$ has at most a linear growth, and outside of each neighborhood of $\bar{\theta}$, $\bar{g}(\theta)$ is bounded away from zero. Also

$$\sum_n \epsilon_n^2 < \infty. \tag{1.6}$$

(For the complete set of assumptions for the early work see the references or [71].) It was shown that $\theta_n \to \bar{\theta}$ in mean square. This seminal paper was the start of an enormous literature that remains vibrant to this day. Condition (1.6) assures that the martingale sequence $\sum \epsilon_n \delta M_n$ converges w.p.1, and (1.5) is necessary if θ_n is not to eventually get stuck at some point that is not $\bar{\theta}$. Generalizations to w.p.1 convergence for more general and vector-valued processes were given by Blum [8] , Dvoretsky [23], and Schmetterer [63]. Other extensions are in Dupač [18, 19], Gladyshev [27], Robbins and Siegmund [58] and in the references in these articles and surveys cited below. Sakrison [62] (see also its references to the Prague Conferences) considered a modification with continuous observations. The observations had multiplicative ergodic noise instead of additive martingale difference noise.

Sacks [60] (see also [48, Section 10.2]) gave the rate of convergence for the one-dimensional model of Robbins-Monro. Let $\epsilon_n = K/n, K > 0$. Suppose that $\bar{g}(\cdot)$ is continuously differentiable in a neighborhood of $\bar{\theta}$ with $g_\theta(\bar{\theta}) < 0, K|g_\theta(\bar{\theta})| > 1/2$, and that for each $\rho > 0$, $E(\delta M_n)^2 I_{\{|\theta_n - \bar{\theta}| \leq \rho\}} \to \sigma^2$ in mean and $\{(dM_n)^2 I_{\{|\theta_n - \bar{\theta}| \leq \rho\}}\}$ is uniformly integrable. Then $\sqrt{n}(\theta - \bar{\theta})$ converges in distribution to a normally distributed random variable with mean zero and variance $K^2\sigma^2/(2K|g_\theta(\bar{\theta})| - 1)$. In this sense, the optimal value of K is $1/|g_\theta(\bar{\theta})|$. The rate $\epsilon_n = O(1/n)$ is often too fast for applications in that, in a practical sense, θ_n can "get stuck" far from $\bar{\theta}$.

Kiefer and Wolfowitz [35] used a finite-difference form to locate the minimum of an unknown differentiable real-valued function $\bar{g}(\cdot)$, with a uniformly continuous derivative, with a unique stationary point that is a minimum, where the only available data are noise corrupted observations $Y_n^{\pm}$. Their algorithm is

$$\theta_{n+1} = \theta_n + \epsilon_n [Y_n^+ - Y_n^-]/2c_n, \tag{1.7}$$

where $Y_n^{\pm} = \bar{g}(\theta_n \pm c_n) + \delta M_n^{\pm}$, $\delta M_n^{\pm}$ are martingale differences, and $0 < c_n \to 0$ is a finite difference interval. Also $\sum_n \epsilon_n c_n < \infty, \sum_n \epsilon_n^2/c_n^2 < \infty$ and $\sup_n E(\delta M_n^{\pm})^2 < \infty$. The first condition assures that the biases in the finite-difference approximation to the derivative are asymptotically negligible. The last two conditions assure the convergence of the martingale $\sum (\epsilon_n/c_n)(\delta M_n^+ + \delta M_n^-)$. The method is usually used with a constant difference interval since it reduces the noise effects and yields a more robust algorithm, even at the expense of a small bias. Further references to the early work as well as precise conditions and extensions can be found in the surveys by Wasan [71], Nevelson and Khasminskii [54], Tsypkin [69], Lai [49], Schmetterer [64], Ruppert [59], Fabian [25] and Dupač [19].

A Liapunov function proof of convergence of the RM algorithm. A proof of convergence of the vector-valued form of (1.1) will now be given by a perturbed Liapunov function argument. W.l.o.g., let $\bar{\theta} = 0$. Let E_n denote the expectation conditioned on $\mathcal{F}_n = \{\theta_0, Y_i, i < n\}$. The $K_i > 0$ are constants.

Theorem 1.1. *Let $0 \leq V(\cdot)$ be real-valued, continuous, twice continuously differentiable with bounded mixed second partial derivatives and $V(0) = 0, EV(\theta_0) < \infty$. For each $\epsilon > 0$, let there be $\delta > 0 : V(\theta) \geq \delta$ for $|\theta| \geq \epsilon$, and δ does not decrease as ϵ increases. Write $E_n Y_n = \bar{g}(\theta_n)$. For each $\epsilon > 0$ let there be $\delta_1 > 0 :$ $V_\theta'(\theta)\bar{g}(\theta) \equiv -k(\theta) \leq -\delta_1$ for $|\theta| \geq \epsilon$. Suppose that $E_n|Y_n|^2 \leq K_2 k(\theta_n)$ when $|\theta_n| \geq K_0$. Let*

$$E\sum_{i=1}^{\infty} \epsilon_i^2 |Y_i|^2 I_{\{|\theta_i| \leq K_0\}} < \infty. \tag{1.8}$$

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

Proof. A truncated Taylor series expansion and the hypotheses yields

$$\begin{aligned} E_n V(\theta_{n+1}) - V(\theta_n) &\leq \epsilon_n V_\theta'(\theta_n) E_n Y_n + \epsilon_n^2 K_1 E_n |Y_n|^2 \\ &= -\epsilon_n k(\theta_n) + \epsilon_n^2 K_1 E_n |Y_n|^2. \end{aligned} \tag{1.9}$$

The hypotheses imply that $EV(\theta_n) < \infty$ for each n. By shifting the time origin we can suppose that $K_1 K_2 \epsilon_n^2 < \epsilon_n/2$. Define

$$\delta V_n = K_1 E_n \sum_{i=n}^{\infty} \epsilon_i^2 |Y_i|^2 I_{\{|\theta_i| \leq K_0\}},$$

and the perturbed Liapunov function $V_n(\theta_n) = V(\theta_n) + \delta V_n \geq 0$. Note that

$$E_n \delta V_{n+1} - \delta V_n = -K_1 \epsilon_n^2 E_n |Y_n|^2 I_{\{|\theta_n| \leq K_0\}}.$$

This, together with the hypotheses and (1.9), yields

$$E_n V_{n+1}(\theta_{n+1}) - V_n(\theta_n) \leq -\epsilon_n k(\theta_n)/2,$$

which implies that $\{V_n(\theta_n)\}$ is an $\mathcal{F}_n$-supermartingale. By the supermartingale convergence theorem, there is a $\tilde{V} \geq 0 : V_n(\theta_n) \to \tilde{V}$ w.p.1. Since $\delta V_n \to 0$ w.p.1, $V(\theta_n) \to \tilde{V}$ w.p.1.

For integers N and m,

$$-V_N(\theta_N) \leq E_N V_{N+m}(\theta_{N+m}) - V_N(\theta_N) \leq -\sum_{i=N}^{N+m-1} E_N \epsilon_i k(\theta_i)/2. \tag{1.10}$$

If $P\{\tilde{V} > 0\} > 0$, then by the properties of $V(\cdot)$, θ_n is asymptotically outside of some neighborhood of the origin, with a positive probability. This, the fact that $\sum \epsilon_i = \infty$, and the properties of $k(\cdot)$, imply that the sum on the right side of (1.10) goes to infinity as $m \to \infty$ with a positive probability, leading to a contradiction. Thus $\tilde{V} = 0$ w.p.1. ■

Extensions of the perturbed Liapunov function method used in the above proof are a powerful tool for more general problems [5, 38, 48, 66]. Combinations of the Stability and ODE methods that do not require constraints are in [9],[48, Sections 5.4, 6,7, 8,5].

2. Introduction to the (ordinary differential equation) ODE approach. In recent decades the complexity of potential applications of recursive algorithms has increased considerably, and the classical methods of proof are not adequate. There are more complicated stochastic dependencies of the Y_n observation processes, including state-dependence of the noise, more flexible requirements on the step sizes, constraints on the iterates, multiple time-scales, decentralized algorithms, regression functions that are an average cost over an infinite time interval, tracking of time varying systems, etc.

The algorithms (1.1)–(1.4) can be viewed as a finite-difference equations with step sizes ϵ_n or ϵ. Hence there is a natural connection with differential equations. The basic idea of what is now called the ODE method was introduced by Ljung [51] and was extensively developed by Kushner and coworkers (see [39, 42, 48] and references therein and [5]), and will be the subject of the rest of this article. One shows by a "local" analysis that the noise effects average out so that the asymptotic behavior is determined by that of a "mean" ODE. Essentially, the dynamical term at a value θ is obtained by averaging the Y_n as though the parameter were fixed at θ. The ODE might be replaced by a constrained form or a differential inclusion.The ODE method is perhaps the most versatile and powerful approach for the analysis of stochastic algorithms.

To get the limit mean ODE one works with continuous-time interpolations. Define $t_0 = 0$ and $t_n = \sum_{i=0}^{n-1} \epsilon_i$. Define $\theta^0(\cdot)$ on $(-\infty, \infty)$ by $\theta^0(t) = \theta_0$ for $t \leq 0$. For $0 \leq t_n \leq t < t_{n+1}$, set $\theta^0(t) = \theta_n$. Define the sequence of shifted processes $\theta^n(\cdot)$ by $\theta^n(t) = \theta^0(t_n + t)$. 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 $m(t) = 0$.

Constrained algorithms. In practice, in one fashion or another, the iterates are usually constrained to lie in some compact set. The constraint might be due to physical limits on the θ_n. In general, it is unreasonable to suppose that the user would allow the θ_n to go to infinity, without some appropriate intervention. Let H be a compact, connected and non-empty constraint set, and suppose that if the iterate leaves H, it is immediately returned to the closest point in H. If the constraint is added artificially for practical reasons, then it should be large enough so that the limits are inside. In an application, this might require some experimentation. For originally unconstrained problems, Chen and Zhu [14, 15] show convergence using gradually increasing bounds.

For simplicity we will suppose that H is a hyperrectangle. Alternatives are:

$H = \{x : q_i(x) \leq 0, i \leq p\}$, where the $q_i(\cdot)$ are continuously differentiable real-valued functions, with gradients $q_{i,x}(\cdot)$ and $q_{i,x}(x) \neq 0$ if $q_i(x) = 0$; or H is an $\mathbb{R}^{r-1}$-dimensional connected compact surface with a continuously differentiable outer normal. These latter sets are dealt with in [48]. Define the set $C(x)$ as follows. For $x \in H^0$, the interior of H, $C(x) = \{0\}$; for $x \in \partial H$, the boundary of H, $C(x)$ is the infinite convex cone generated by the outer normals at x of the faces on which x lies.

If the iterates are not constrained, then a stability method such as [48, Sections 5.4, 10.5] can be used to assure that the they do not explode.

Compactness and continuous-time interpolations. W.p.1 convergence for the constrained algorithm. The following extension of the Arzela-Ascoli Lemma will be used. Suppose that for each n, $f_n(\cdot)$ is an $\mathbb{R}^r$-valued measurable function on $(-\infty, \infty)$ and $\{f_n(0)\}$ is bounded. Also suppose that for each T and $\epsilon > 0$, there is a $\delta > 0$ such that

$$\limsup_n \sup_{0 \leq t-s \leq \delta,\ |t| \leq T} |f_n(t) - f_n(s)| \leq \epsilon.$$

Then we say that $\{f_n(\cdot)\}$ is equicontinuous in the extended sense, and then there is a subsequence that converges to some continuous limit, uniformly on each bounded interval.

The Kushner-Clark approach [39, 48]. To illustrate an approach to w.p.1. convergence via the ODE method, consider a simple case of (1.3). Let $Y_n = g_n(\theta_n) + \psi_n + \beta_n$, where $\beta_n \to 0$ w.p.1. For $n \geq 0$, define $\Psi^n(t) = \sum_{i=n}^{m(t_n+t)-1} \epsilon_i \psi_i$, $B^n(t) = \sum_{i=n}^{m(t_n+t)-1} \epsilon_i \beta_i$, $Z^n(t) = \sum_{i=n}^{mt_n+t)-1} \epsilon_i z_i$. Suppose that for some $T > 0$,

$$\limsup_n \sup_{j \geq n} \max_{0 \leq t \leq T} \left| \Psi^0(jT + t) - \Psi^0(jT) \right| = 0 \text{ w.p.1.} \tag{2.1}$$

(2.1) is also a necessary condition for convergence [70]. It is assured by

$$\lim_n P \left\{ \sup_{j \geq n} \max_{0 \leq t \leq T} \left| \sum_{i=m(jT)}^{m(jT+t)-1} \epsilon_i \psi_i \right| \geq \mu \right\} = 0, \quad \textit{each } \mu > 0. \tag{2.2}$$

Suppose that the $g_n(\cdot)$ are continuous uniformly in n, and that there is a continuous function $\bar{g}(\cdot)$: for each $\theta \in H$ and $t > 0$,

$$\lim_n \left| \sum_{i=n}^{m(t_n+t)} \epsilon_i \left[g_i(\theta) - \bar{g}(\theta) \right] \right| \to 0. \tag{2.3}$$

Theorem 2.1. *Under the above conditions $\{\theta^n(\cdot)\}$ is equicontinuous in the extended sense. W.p.1, the pathwise limits $\theta(\cdot)$ satisfy the constrained ODE $\dot{\theta} = g(\theta) + z, z(t) \in -C(\theta(t))$, and θ_n converges to a limit set of the ODE.*

Proof.

$$\theta^n(t) = \theta_n + \sum_{i=n}^{m(t+t_n)-1} \epsilon_i g_i(\theta_i) + \sum_{i=n}^{m(t+t_n)-1} \epsilon_i(\psi_i + \beta_i) + \sum_{i=n}^{m(t+t_n)-1} \epsilon_i z_i. \quad (2.5)$$

The functions defined by the middle sum go to zero w.p.1, and the set defined by the first sum is equicontinuous in the extended sense. Suppose that $\{Z^n(\cdot)\}$ is not equicontinuous in the extended sense. Then there is a subsequence that has an asymptotic jump in the sense that 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 $|Z^{\mu_k}(s_k + \delta_k) - Z^{\mu_k}(s_k)| \geq \rho$. The facts that $z_n = 0$ if $\theta_{n+1} \in H^0$, the interior of H, and that $z_n \in -C(\theta_{n+1})$ can be used to complete the proof of equicontinuity and that any limit must satisfy the constrained ODE $\dot{\theta} = g(\theta) + z, \quad z(t) \in -C(\theta(t))$. The limits of θ_n must lie in a limit set of the ODE.

Assuring (2.1) for the martingale-difference noise case [48, Section 5.3]. Let $\psi_n = \delta M_n$, a martingale difference. Define $M_n = \sum_{i=0}^{n-1} \delta M_i$. For some even integer p, suppose that

$$\sum_n \epsilon_n^{p/2+1} < \infty, \quad \sup_n E|\delta M_n|^p < \infty. \quad (2.3)$$

Burkholder's inequality [48, Section 4.1] and the martingale bound

$$P_{\mathcal{F}_n}\left\{\sup_{n \leq m \leq N} |M_m - M_n| \geq \lambda\right\} \leq \frac{E_{\mathcal{F}_n} q(M_N - M_n)}{q(\lambda)}, \quad \text{each } \lambda > 0,$$

where $0 \leq q(\cdot)$ is a nondecreasing convex function, yield

$$\sum_j P\left\{\max_{0 \leq t \leq T} \left|\sum_{i=m(jT)}^{m(jT+t)-1} \epsilon_i \delta M_i\right| \geq \mu\right\} < \infty, \quad \text{each } \mu > 0,$$

which, via the Borel-Cantelli Lemma, implies (2.2).

Alternatively, suppose that $\epsilon_n \leq \gamma_n / \log n, \gamma_n \to 0$ and that the moments of δM_n grow no faster than Gaussian in that there is $K < \infty$ such that for each component $\delta M_{n,j}$ of δM_n, $E_n e^{\gamma(\delta M_{n,j})} \leq e^{\gamma^2 K/2}$. Suppose also the innocuous condition that for $T < \infty$, there is $c_1(T) < \infty$: for all n, $\sup_{n \leq i \leq m(t_n+T)} \epsilon_i/\epsilon_n \leq c_1(T)$. Then (2.2) holds [48, Section 5.3]. Criteria assuring (2.1) for more general noise processes are in [39]. If $g(\theta_n, \psi_n)$ replaces $g_n(\theta_n) + \psi_n$, then an analysis can still be done [5, 39], but becomes more difficult. The weak convergence approach below will be simpler.

Chain recurrence. The above theorem state that the limits of θ_n are in a limit or invariant set of the ODE. It was shown by Benaïm [1, 2, 3], [48], (assuming uniqueness of the solution for each initial condition) that the limits are in the possibly smaller set of chain recurrent points within the limit set. The chain recurrent points are defined as follows. Let $x(t|y)$ denote the solution to either $\dot{x} = \bar{g}(x)$ or the constrained form $\dot{x} = \bar{g}(x) + z, z \in -C(x)$, at time t, given $x(0) = y$. A point x is said to be *chain recurrent* [1, 2, 28] 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

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

where $y_i = x(T_{i-1}|u_{i-1})$ for $i = 1, \ldots, k+1$. Not all points in the limit or invariant set are chain recurrent.

3. Weak convergence. The classical theory of SA was concerned with w.p.1 convergence to a unique limit. Henceforth we work with weak convergence. It is much easier to use in that convergence can be proved under weaker and more easily verifiable conditions and generally with less effort. The approach yields almost the same information on the asymptotic behavior. The weak convergence methods have advantages when dealing with problems involving correlated noise and state dependent noise, decentralized or asynchronous algorithms, and discontinuities in the algorithm.

In applications, there is always a rule that tells us when to stop and accept some function of the recent iterates as the “final value.” Then the information that we have about the final value is distributional and there is no difference in the conclusions provided by the probability one and the weak convergence methods. If the application is of concern over long time interval, then the actual physical model might drift, in which case the step size is not allowed to go to zero, and there is no general alternative to the weak convergence methods. In practical on-line applications, the step size ϵ_n is not usually allowed to decrease to zero, due to considerations concerning robustness and to allow some tracking of the desired parameter as the system changes slowly over time. Then probability one convergence does not apply. In signal processing applications it is usual to keep the step size bounded away from zero. The proofs of probability one results are more technical. They can be hard for multiscale, state-dependent-noise cases or decentralized/asynchronous algorithms.

There is always value in knowing convergence w.p.1, when appropriate. Although no iteration continues indefinitely, it is still encouraging to know that if it were, then it will assuredly converge. However, the concerns that have been raised suggest that methods with slightly more limited convergence goals can be as useful, if they are technically easier, require weaker conditions, and are no less informative when dealing with the practical conditions in applications

With the ODE approach, where the ODE is obtained by locally averaging the dynamics, the conditions required for the averaging with weak convergence are weaker than those needed for w.p.1 convergence and the proofs are simpler. Furthermore, the $\{\theta_n\}$ spends “nearly all” of its time in an arbitrarily small neighborhood of the limit point or set. 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 near this points can often be used to get convergence w.p.1 See, for example [40, Sections 6.9 and 6.10] and [22]. Even when only weak convergence can be proved, if θ_n is close to a stable limit point at iterate n, then under broad conditions the mean escape time from a small neighborhood of that limit point is at least of the order of e^{c/ϵ_n} for some $c > 0$.

Definitions. W.p.1 convergence required a type of (w.p.1) compactness of the paths of the SA sequence. Given this compactness, one then showed that the limits satisfy the ODE. With weak convergence we can no longer work with the individual paths, and the compactness is of the set of probability measures. This can be translated into information on the asymptotic properties of the paths. Let $D^r[0,\infty)$ denote the space of $I\!\!R^r$-valued functions on $[0,\infty)$ that are right-continuous with left-hand limits. The Skorokhod topology is used. The exact definition is unimportant here [7, 24], except to note that the limits of a sequence that is compact in this topology and whose discontinuities go to zero, has continuous limits. This topology is common in weak convergence analysis, and is convenient since we are working with piecewise-constant interpolations $\theta^n(\cdot)$. One could use piecewise-linear interpolations, but the criteria for compactness and the details of proof are simpler with our choice.

A set of of probability measures $\{P_n(\cdot)\}$ on $D^r[0,\infty)$ is said to be tight or weakly sequentially compact if for each $\delta > 0$ there is a compact set S_δ such that $P_n(S_\delta) \geq 1 - \delta$ for all n. P_n is said to converge to P weakly if $\int f(x)P_n(dx) \to \int f(x)P(dx)$ for all bounded and continuous real-valued functions $f(\cdot)$. A tight sequence always contains a weakly convergent subsequence. The P_n, P are the measures of processes with paths in $D^r[0,\infty)$, and it is these processes that are of concern. If the set of measures is tight or converges weakly, then abusing notation, we use the same terminology for the processes. More detail on the use in SA theory is in [48].

Consider (1.3) and, henceforth, let $\{Y_n\}$ be uniformly integrable. Then $\{\theta^n(\cdot)\}$ is tight. Also, henceforth, suppose that $\{Y_n^\epsilon\}$ in (1.4) is uniformly integrable. Then $\{\theta^\epsilon(T+\cdot), T < \infty, \epsilon > 0\}$ is tight. In both cases, the limit processes, as $n \to \infty$, or $\epsilon \to 0, T \to \infty$ have Lipschitz continuous paths w.p.1. For the analog of Theorem 2.1 with martingale difference noise, the sequence $\{\Psi^n(\cdot)\}$ is tight with zero weak-sense limits if only $\epsilon_n \to 0$. Then all weak-sense limits satisfy the constrained ODE. The analog of (2.2) is the much simpler requirement: For any $T > 0$ and all $\delta > 0$,

$$\lim_n P\left\{\max_{0\leq t\leq T} |\Psi^n(t)| \geq \delta\right\} = 0.$$

Correlated noise: A convergence theorem. Constant ϵ will be used to simplify the notation. The results are analogous if $\epsilon_n \to 0$. Rewrite (1.4) as

$$\theta_{n+1}^\epsilon = \theta_n^\epsilon + \epsilon E_n Y_n^\epsilon + \epsilon \delta M_n^\epsilon + \epsilon z_n^\epsilon, .$$

where $\delta M_n^\epsilon = Y_n^\epsilon - E_n Y_n^\epsilon$ and $E_n Y_n^\epsilon = g(\theta_n^\epsilon, \xi_{n+1}^\epsilon)$ where ξ_n^ϵ is a random sequence. Define the martingale $M_n^\epsilon = \sum_{i=0}^n \epsilon \delta M_i^\epsilon$, with $M^\epsilon(\cdot)$ being the continuous time interpolation. Analogously, define $Z^\epsilon(\cdot)$ from the $\{z_n^\epsilon\}$.

Let $g(\theta,\xi^\epsilon_{n+1})$ be continuous in θ, in the mean, uniformly in n. Suppose that there is a function $\bar{g}(\cdot)$ such that for each θ

$$\lim_{n_\epsilon\to\infty,\,\epsilon n_\epsilon\to 0}\frac{1}{n_\epsilon}\sum_{i=m}^{m+n_\epsilon-1}E_m g(\theta,\xi^\epsilon_{i+1})=\bar{g}(\theta) \tag{3.1}$$

in mean, and uniformly in m. Due to the use of the conditional expectation, this is a weak form of the law of large numbers. (Discontinuous and θ-dependent noise can also be handled, but is more complicated.)

Theorem 3.1. *Under the above conditions, $\{\theta^\epsilon(T+\cdot)\}$ is tight and any weak-sense limit as $\epsilon\to 0, T\to\infty$, is in the limit or invariant set of the constrained ODE. If the solution is unique for each initial condition, then the limit points are chain recurrent.*

Proof. Write

$$\begin{aligned}\theta^\epsilon(T+t)-\theta^\epsilon(T)=&\sum_{i=T/\epsilon}^{(T+t)/\epsilon-1}\epsilon E_i g(\theta^\epsilon_i,\xi^\epsilon_{i+1})\\&+[M^\epsilon(T+t)-M^\epsilon(T)]+[Z^\epsilon(T+t)-Z^\epsilon(T)].\end{aligned}\tag{3.2}$$

If $\sup_{\epsilon,n}E|Y^\epsilon_n|^2<\infty$, then the martingale property implies that

$$E\sup_{s\le t}|M^\epsilon(T+s)-M^\epsilon(T)|^2=E\sum_{i=T/\epsilon}^{(T+t)/\epsilon-1}[\epsilon\delta M^\epsilon_i]^2=O(t\epsilon),$$

uniformly in T, implying that $M^\epsilon(T_\epsilon+\cdot)-M^\epsilon(T_\epsilon)$ converges weakly to zero, no matter what the sequence T_ϵ. This latter property also holds under only uniform integrability. Let $T_\epsilon\to\infty$. The sequence $\{\theta^\epsilon(T_\epsilon+\cdot)\}$ is tight and any weak-sense limit has Lipschitz continuous paths w.p.1 (by the uniform integrability assumption). Since the weak-sense limit of the martingale terms is zero and those of the sum in (3.2) have Lipschitz continuous paths, the weak-sense limits of $Z^\epsilon(T_\epsilon+\cdot)-Z^\epsilon(T_\epsilon)$ are Lipshitz continuous. Fix a weakly convergent subsequence, index it by ϵ for notational simplicity, and denote the limit processes by $\theta(\cdot)$ and $Z(\cdot)$, with $z(t)=\dot{Z}(t)$.

To get the limit ODE we need to consider the first sum in (3.2). It will be seen that for continuous $f(\cdot)$, any integer q, any $t>0$, and any $s_i\le t, i\le q$, $\tau>0$,

$$Ef(\theta(s_i),i\le q)\left[\theta(t+\tau)-\theta(t)-\int_t^{t+\tau}\bar{g}(\theta(u))du+Z(t+\tau)-Z(t)\right]=0, \tag{3.3}$$

where $\dot{Z}(t)\in -C(\theta(t))$. This implies that $\theta(t)-\theta(0)-\int_0^t\bar{g}(\theta(u))du-Z(t)$ is a martingale. The (absolute) mean value is zero, hence it is zero. Thus $\theta(\cdot)$ satisfies the constrained ODE.

The proof proceeds by using local averaging, using the fact that θ_n^ϵ varies slowly. We can write

$$Ef(\theta^\epsilon(s_i), i \le q) \\ \times \left[\theta^\epsilon(t+\tau) - \theta^\epsilon(t) - \epsilon \sum_{j=t/\epsilon}^{(t+\tau)/\epsilon - 1} g(\theta_j^\epsilon, \xi_{j+1}^\epsilon) - [Z^\epsilon(t+\tau) - Z^\epsilon(t)]\right] = 0.$$

Without loss of generality we can let $T_\epsilon = 0$. Consider the term with the sum and collect terms in groups of size $n_\epsilon \to \infty$, with $\delta_\epsilon = \epsilon n_\epsilon \to 0$, and rewrite:

$$Ef(\theta^\epsilon(s_i), i \le q) \left\{\theta^\epsilon(t+\tau) - \theta^\epsilon(t) - \sum_{l:l\delta_\epsilon = t}^{t+\tau-\delta_\epsilon} \delta_\epsilon \left[\frac{1}{n_\epsilon} \sum_{j=ln_\epsilon}^{ln_\epsilon + n_\epsilon - 1} E_{ln_\epsilon} g(\theta_j^\epsilon, \xi_{j+1}^\epsilon)\right]\right\}.$$

By the continuity and uniform integrability, modulo negligible terms,

$$Ef(\theta^\epsilon(s_i), i \le q) \left\{\theta^\epsilon(t+\tau) - \theta^\epsilon(t) - \sum_{l:ln_\epsilon = t}^{t+\tau-\delta_\epsilon} \delta_\epsilon \left[\frac{1}{n_\epsilon} \sum_{j=ln_\epsilon}^{ln_\epsilon + n_\epsilon - 1} E_{ln_\epsilon} g(\theta_{ln_\epsilon}^\epsilon, \xi_{j+1}^\epsilon)\right]\right\}.$$

As $\epsilon \to 0$, this expression is approximated by

$$Ef(\theta^\epsilon(s_i), i \le q) \left[\theta^\epsilon(t+\tau) - \theta^\epsilon(t) - \sum_{l:ln_\epsilon = t}^{t+\tau-\delta_\epsilon} \delta_\epsilon \bar{g}(\theta_{ln_\epsilon}^\epsilon)\right],$$

where the sum is asymptotically replaceable by an integral. The reflection term is treated similarly to what was done in Theorem 2.1. ■

4. Extensions. 4.1. Multiple time-scales. In multiple-time-scale systems, the components of the iterate are divided into groups with each group having its own step-size sequence, say for two groups, ϵ_n, μ_n, and they are of different orders in that $\epsilon_n/\mu_n \to 0$. The methods follow the same lines as above, with the usual modifications used in singular perturbations of essentially dealing with the fast system first, getting its limits as a function of the fixed slow variable, and then showing that the slow system can be dealt with the limit fast variables used [6], [37, 48, Section 8.6]. The rate of convergence can be very slow.

4.2. Random directions. For high-dimensional Kiefer-Wolfowitz type problems, the original algorithms repeat (1.7) for each component. The rate of convergence can be very slow, and one can waste much time iterating on unimportant parameters. This suggests the possible value of iterating in random directions using an algorithm such as

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

where $Y_n^\pm$ are observations taken at parameter values $\theta_n \pm c_n d_n$, and the d_n are mutually independent and identically distributed random direction vectors. Such algorithms were discussed in [39] where the d_n were uniformly distributed on the unit sphere.. Spall [61, 67, 68] showed that there can be a considerable advantage if the directions $d_n = (d_{n,1}, \ldots, d_{n,r})$ were selected so that the set of components $\{d_{n,i}; n, i\}$ were mutually independent, with $d_{n,i}$ taking values ± 1 each with probability $1/2$; that is, the direction vectors are the corners of the unit cube in $I\!R^r$. In [48, Section 10.7]. it was shown that the result is the same if they are uniformly distributed on the unit sphere, with radius $\sqrt{r}$. An analysis of the rate of convergence suggests that the improvement is best if there are many components of θ that are relatively unimportant.

4.3. Linear system identification and tracking. The problem of estimating or tracking the value of a time-varying parameter in linear systems has applications in adaptive control theory, and in communications theory for adaptive equalizers and noise cancellation, etc., where the signal, noise, and channel properties change with time. Consider the observation model $y_n = \phi_n' \bar\theta + \nu_n$, where $\bar\theta$ is the value of the unknown physical parameter, ν_n is observation noise, and ϕ_n is a regression vector. The values of y_n, ϕ_n are observed. Owing to the importance of this problem there has been a vast literature [5, 17, 29, 30, 31, 32, 52, 53, 65, 72]. A basic algorithm, based on least squares approximation, has the form $\theta_{n+1}^\epsilon = \theta_n^\epsilon + \epsilon\phi_n[y_n - \phi_n'\theta_n^\epsilon]$, where θ_n^ϵ is the current estimate of $\bar\theta$. Now, replace $\bar\theta$ by $\bar\theta_n$, and let the distributions vary slowly. A lot of attention has been devoted to getting the optimal value of ϵ [5, 32]. A successful stochastic approximation scheme for tracking in the time-varying case was given in [10] and further developed in [45] and discussed in [48, Section 3.2]. Then ϵ_n replaces ϵ, but it does not converge to zero. The updating algorithm for ϵ_n is the actual SA algorithm. An application to an adaptive antenna system is in [12]. If the ϕ_n are not observable, then we have "blind optimization," and it is substantially harder [4]

4.4. Iterate averaging methods. Consider the multivariate RM algorithm. If ϵ_n goes to zero slower than $O(1/n)$ in that $\epsilon_n/\epsilon_{n+1} = 1 + o(\epsilon_n)$, then the rate of convergence of the average $\Theta_n = \sum_{n-N_n}^n \theta_i/N_n$, where $n - N_n \to \infty$, is nearly optimal under broad conditions, in that it is the rate that would be achieved if ϵ_n were replaced by $\epsilon_n K$ for the asymptotically optimal matrix K. This surprising result was shown by Polyak and Juditsky [34, 56] and Ruppert [59] , and analyzed under more general conditions in [13, 16, 44, 46, 47, 50, 73, 74] and [48, Section 11.1]. In particular, [44] showed that $\sqrt{n}(\Theta_n - \bar\theta)$ converges in distribution to a normally distributed random variable with mean zero and covariance $\bar V$, where $\bar V$ is the smallest possible [48, Section 11.1].

4.5. Rate of convergence. Classical rate of convergence results for the RM process concern the asymptotic distribution of $U_n = (\theta_n - \bar\theta)/\sqrt{\epsilon_n}$ and $U_n^\epsilon = (\theta_n^\epsilon - \bar\theta)/\sqrt{\epsilon}$, as $n \to \infty$ or $\epsilon \to 0, n\epsilon \to \infty$. Define the interpolations $U^n(\cdot)$ and $U^\epsilon(\cdot)$ with intervals ϵ_n and ϵ, resp. Suppose that $\theta^n(\cdot)$ (or $\theta^\epsilon(T+\cdot)$, as $\epsilon \to 0, T \to \infty$) converges weakly to $\bar\theta \in H^0$. The references [36, 39, 41] introduced the idea of studying the rate by showing that, under appropriate conditions, $U^n(\cdot)$ and $U^\epsilon(T+\cdot)$ converge to stationary mean-zero linear diffusions as $n \to \infty$ or $\epsilon \to 0, T \to \infty$, and the covariance of these limits serves as a measure of the rate of convergence. A more complete development is in [48, Chapter 10] and [5].

Consider the constant ϵ case where $E_nY_n^\epsilon = g(\theta_n, \xi_{n+1}^\epsilon)$, where $g(\cdot, \xi)$ is continuously diffentiable near $\bar{\theta}$. Let there be a Hurwitz matrix A such that

$$\frac{1}{m} \sum_{i=n}^{n+m-1} \left[E_n^\epsilon g_\theta'(\bar{\theta}, \xi_i^\epsilon) - A \right] \to 0$$

in probability as $\epsilon \to 0$ and $n, m \to \infty$. Suppose that the sequence defined by $\sqrt{\epsilon} \sum_{T/\epsilon}^{(t+T)/\epsilon} (g(\bar{\theta}, \xi_i^\epsilon) + \delta M_i^\epsilon)$ converges weakly to a Wiener process with covariance Σ. Then, under some other mild conditions, $U^\epsilon(T + \cdot)$ converges weakly to the stationary process $U(\cdot)$ defined by $dU = AUdt + dW$, where Σ is the covariance of the Wiener process W.

For $\epsilon_n = K/n$, where K is a positive definite matrix, the limit is $dU = (I/2 + KA)Udt + KdW$. In the sense of minimizing the trace of the stationary covariance matrix, the optimal value of K is $-A^{-1}$ and the corresponding covariance is $A^{-1}\Sigma(A^{-1})'$. If $\epsilon_n \to 0$ faster than $O(1/n)$ in that $(\epsilon_n/\epsilon_{n+1})^{1/2} = 1 + o(\epsilon_n)$, then drop the $I/2$ term. See [48, Chapter 10]. If the limit $\bar{\theta}$ is on the boundary of H, then there might be a a reflection term on the diffusion [11]. The rate of convergence of moments for the random directions algorithm is in [26].

An alternative point of view uses the so-called strong diffusion approximation and estimates the asymptotic difference between the normalized error process and a Wiener process in terms of an “iterated logarithm” [33, 55].

4.6. State-dependent noise. So far it has been assumed that $E_nY_n = g(\theta_n, \xi_{n+1})$ (or the analogous form for $\epsilon_n = \epsilon$) and it was implicitly supposed that that the evolution of the noise did not depend on the values of the iterates in that $P\left\{\xi_{n+1} \in \cdot \mid \xi_i, \theta_i, i \le n\right\} = P\left\{\xi_{n+1} \in \cdot \mid \xi_i, i \le n\right\}$. In numerous applications, there is a dependence that must be accounted for. The usual form is what is called Markov state dependence. It takes the following form. $P(\cdot, \cdot|\theta)$ is a Markov transition function parameterized by θ such that $P\left\{\xi_{n+1} \in \cdot \mid \xi_i, \theta_i, i \le n\right\} = P(\xi_n, \cdot|\theta_n)$. The analysis follows lines that are analogous to the simpler case, but is technically more complicated. The approach was initiated in [38, 42] and developed further in [43] for weak convergence. See [5] for a complete development of the w.p.1 convergence case. Some comments on the non Markov case are in [48, Section 8.4.].

4.7. Minimizing an ergodic cost function. The KW and related procedures are concerned with the minimization of a function whose values at the chosen parameters are observed subject to noise. In many applications, one wishes to minimize functions of the type $\lim_{T\to\infty} E\int_0^T c(\theta, x(t, \theta))dt/T$ where $x(\theta, \cdot)$ is a random process parameterized by θ, and only samples of the integrand are observed. Effective algorithms are discussed in [48, Section 9.1], [43].

4.8. Large deviations. In [48, Section 6.9], [20, 21, 22] large deviations methods are used to show that escape from a small neighborhood of a weak-sense limit point is impossible, under weak conditions.

References

[1] M. Benaïm. A dynamical systems approach to stochastic approximations. *SIAM J. Control Optim.*, 34:437–472, 1996.

[2] M. Benaïm. Dynamics of stochastic approximation algorithms. In *Lecture Notes in Mathematics: 1769: Séminaire de Probabilities*, pages 1–69. Sprtinger-Verlag, Berlin and New York, 1999.

[3] M. Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités, XXXIII, 1–68, Lecture Notes in Math., 1709.* Springer, Berlin and New York, 1999.

[4] A. Benveniste, M. Goursat, and G.Ruget. Robust identification of a non-minimum phase system: Blind adjustment of a linear equalizer in data communications. *IEEE Trans. Automatic Control*, AC-25:385–399, 1980.

[5] A. Benveniste, M. Metivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximation.* Springer-Verlag, Berlin and New York, 1990.

[6] S. Bhatnagar and V.S. Borkar. A two-time-scale stochastic approximation scheme for simulation based parametric optimization. *Probab. Eng. Inform. Sci.*, 12:519–531, 1998.

[7] P. Billingsley. *Convergence of Probability Measures; Second edition.* Wiley, New York, 1999.

[8] J.R. Blum. Multidimensional stochastic approximation. *Ann. Math. Statist.*, 9:737–744, 1954.

[9] V. S. Borkar and S. P.Meyn. The o.d.e. method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.*, 38:447–469, 2000.

[10] J.-M. Brossier. *Egalization Adaptive et Estimation de Phase: Application aux Communications Sous-Marines.* PhD thesis, Institut National Polytechnique de Grenoble, 1992.

[11] R. Buche and H.J. Kushner. Rate of convergence for constrained stochastic approximation algorithms. *SIAM. J. Control Optim.*, 40:1011–1041, 2001.

[12] R. Buche and H.J. Kushner. Adaptive optimization of least squares tracking algorithms: with applications to adaptive antennas arrays for randomly time-varying mobile communications systems. *IEEE Trans. on Autom. Cont.*, 50:1749–1760, 2005.

[13] H.-F. Chen. Asymptotically efficient stochastic approximation. *Stochastics Stochastics Rep.*, 45:1–16, 1993.

[14] H.-F. Chen. *Stochastic Approximation and Its Applications.* Kluwer Academic, Boston, 2002.

[15] H.-F. Chen and Y.M. Zhu. Stochastic approximation procedure with randomly varying truncations. *Sci. Sinica ser. A*, 29:914–926, 1986.

[16] B. Delyon and A. Juditsky. Stochastic optimization with averaging of trajectories. *Stochastics Stochastic Rep.*, 39:107–118, 1992.

[17] B. Delyon and A. Juditsky. Asymptotical study of parameter tracking algorithms. *SIAM J. Control Optim.*, 33:323–345, 1995.

[18] V. Dupač. On the Kiefer–Wolfowitz approximation method. *Casopis Pest. Mat.*, 82:47–75, 1957.

[19] V. Dupač. Stochastic approximation. *Kybernetica (Prague) Suppl.*, 17, 1981.

[20] P. Dupuis and H.J. Kushner. Stochastic approximation via large deviations: Asymptotic properties. *SIAM J. Control Optim.*, 23:675–696, 1985.

[21] P. Dupuis and H.J. Kushner. Asymptotic behavior of constrained stochastic approximations via the theory of large deviations. *Probab. Theory Related Fields*, 75:223–244, 1987.

[22] P. Dupuis and H.J. Kushner. Stochastic approximation and large deviations: Upper bounds and w.p.1 convergence. *SIAM J. Control Optim.*, 27:1108–1135, 1989.

[23] A. Dvoretzky. On stochastic approximation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 39–55, 1956.

[24] S.N. Ethier and T.G. Kurtz. *Markov Processes: Characterization and Convergence.* Wiley, New York, 1986.

[25] V. Fabian. Stochastic approximation. In J.S. Rustagi, editor, *Optimizing Methods in Statistics.* Academic Press, New York, 1971.

[26] L Gerencsér. Convergence rate of moments in stochastic approximation with simultaneous perturbation gradient approximation and resetting. *IEEE Trans. Automat. Control*, 44:894–905, 1999.

[27] E.G. Gladyshev. On stochastic approximation. *Theory Probab. Appl.*, 10:275–278, 1965.

[28] J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields.* Springer-Verlag, Berlin and New York, 1983.

[29] S. Gunnarsson and L. Ljung. Frequency domain tracking characteristics of adaptive algorithms. *IEEE Trans. Acoust. Speech Signal Process.*, ASSP-37:1072–1089, 1989.

[30] L. Guo. Stability of recursive tracking algorithms. *SIAM J. Control Optim.*, 32:1195–1225, 1994.

[31] L. Guo and L. Ljung. Performance analysis of general tracking algorithms. *IEEE Trans. Automatic Control*, AC-40:1388–1402, 1995.

[32] L. Guo, L. Ljung, and P. Priouret. Tracking performance analyses of the forgetting factor RLS algorithm. In *Proceedings of the* 31*st Conference on Decision and Control,* Tucson, Arizona, pages 688–693, New York, 1992. IEEE.

[33] J.A. Joslin and A.J.Heunis. Law of the iterated logarithm for constant-gain linear stochastic gradient algorithm. *SIAM J. Control Optim.*, 39:533–570, 2000.

[34] A. Juditsky. A stochastic estimation algorithm with observation averaging. *IEEE Trans. Automatic Control*, 38:794–798, 1993.

[35] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.*, 23:462–466, 1952.

[36] H.J. Kushner. Rates of convergence for sequential monte-carlo optimization methods. *SIAM J. Control Optim.*, 16:150–168, 1978.

[37] H.J. Kushner. Diffusion approximations to output processes of nonlinear systems with wide-band inputs, with applications. *IEEE Trans. on Inf. Theory*, 26:715–725, 1980.

[38] H.J. Kushner. *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory.* MIT Press, Cambridge, Mass., 1984.

[39] H.J. Kushner and D.S. Clark. *Stochastic Approximation for Constrained and Unconstrained Systems.* Springer-Verlag, Berlin and New York, 1978.

[40] H.J. Kushner and P. Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time.* Springer-Verlag, Berlin and New York, 1992. Second edition, 2001.

[41] H.J. Kushner and H. Huang. Rates of convergence for stochastic approximation type algorithms. *SIAM J. Control Optim.*, 17:607–617, 1979.

[42] H.J. Kushner and A. Shwartz. An invariant measure approach to the convergence of stochastic approximations with state dependent noise. *SIAM J. Control Optim.*, 22:13–27, 1984.

[43] H.J. Kushner and F.J. Vázquez-Abad. Stochastic approximation algorithms for systems over an infinite horizon. *SIAM J. Control Optim.*, 34:712–756, 1996.

[44] H.J. Kushner and J. Yang. Stochastic approximation with averaging of the iterates: Optimal asymptotic rates of convergence for general processes. *SIAM J. Control Optim.*, 31:1045–1062, 1993.

[45] H.J. Kushner and J. Yang. Analysis of adaptive step size SA algorithms for parameter tracking. *IEEE Trans. Automatic Control*, AC-40:1403–1410, 1995.

[46] H.J. Kushner and J. Yang. Stochastic approximation with averaging and feedback: faster convergence. In G.C. Goodwin K. Aström and P.R. Kumar, editors, *IMA Volumes in Mathematics and Applications, Volume 74, Adaptive Control, Filtering and Signal Processing*, pages 205–228. Springer-Verlag, Volume 74, the IMA Series, Berlin and New York, 1995.

[47] H.J. Kushner and J. Yang. Stochastic approximation with averaging and feedback: Rapidly convergent "on line" algorithms. *IEEE Trans. Automatic Control*, AC-40:24–34, 1995.

[48] H.J. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications.* Springer-Verlag, Berlin and New York, 2003.

[49] T. L. Lai. Stochastic approximation. *Ann. Statist*, 31:391–406, 2003.

[50] A. Le Breton. Averaging with feedback in Gaussian schemes in stochastic approximation. *Math. Methods Statist.*, 6:313–331, 1997.

[51] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Trans. Automatic Control*, AC-22:551–575, 1977.

[52] L. Ljung. *System Identification Theory for the User.* Prentice-Hall, Englewood Cliffs, NJ, 1986.

[53] L. Ljung and T. Söderström. *Theory and Practice of Recursive Identification.* MIT Press, Cambridge, Mass, 1983.

[54] M.B. Nevelson and R.Z. Khasminskii. *Stochastic Approximation and Recursive Estimation.* Amer. Math. Soc, Providence, RI, 1976. Translation of Math. Monographs, Vol. 47.

[55] H. Pezeshki-Esfanahani and A.J. Heunis. Strong diffusion approximations for recursive stochastic algorithms. *IEEE Trans. Inform. Theory*, 43:312–323, 1997.

[56] B.T. Polyak and A.B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30:838–855, 1992.

[57] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22:400–407, 1951.

[58] H. Robbins and D. Siegmund. A convergence theorem for some nonnegative almost supermartingales and and some applications. In J.S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, New York, 1971.

[59] D. Ruppert. Stochastic approximation. In B.K. Ghosh and P.K. Sen, editors, *Handbook in Sequential Analysis*, pages 503–529. Marcel Dekker, New York, 1991.

[60] J. Sacks. Asymptotic distribution of stochastic approximation processes. *Ann. Math. Statist.*, 29:373–405, 1958.

[61] P. Sadegh and J.C. Spall. Optimal random perturbations for stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automat. Control*, 43:1480–1484, 1998.

[62] D.J. Sakrison. A continuous Kiefer-Wolfowitz procedure for random processes. *Ann. Math. Statist.*, 35:590–599, 1964.

[63] L. Schmetterer. Stochastic approximation. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 587–609, Berkeley, 1960. University of California.

[64] L. Schmetterer. Multidimensional stochastic approximation. In P. R. Krishnaiah, editor, *Multivariate Analysis II*, pages 443–460. Academic Press, New York, 1969.

[65] V. Solo. The limit behavior of LMS. *IEEE Trans Acoust. Speech Signal Process.*, ASSP-37:1909–1922, 1989.

[66] V. Solo and X. Kong. *Adaptive Signal Processing Algorithms.* Prentice-Hall, Englewood Cliffs, NJ, 1995.

[67] J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automatic Control*, AC-37:331–341, 1992.

[68] J.C. Spall. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Trans. Automat. Control*, 45:1839–1853, 2000.

[69] Ya.Z. Tsypkin. *Adaptation and Learning in Automatic Systems.* Academic Press, New York, 1971.

[70] I.J. Wang, E.K.P. Chong, and S.R. Kulkarni. Equivalent necessary and sufficient conditions on noise sequences for stochastic approximation algorithms. *Adv. Appl. Probab.*, 28:784–801, 1996.

[71] M.T. Wasan. *Stochastic Approximation.* Cambridge University Press, Cambridge, UK, 1969.

[72] B. Widrow and S.D. Stearns. *Adaptive Signal Processing.* Prentice-Hall, Englewood Cliffs, NJ, 1985.

[73] G. Yin. On extensions of Polyak's averaging approach to stochastic approximation. *Stochastics Stochastics Rep.*, 36:245–264, 1991.

[74] G. Yin. Stochastic approximation via averaging: Polyak's approach revisited. In G. Pflug and U. Dieter, editors, *Lecture Notes in Economics and Mathematical Systems 374*, pages 119–134. Springer-Verlag, Berlin and New York, 1992.