Wiley Encyclopedia of
Operations Research and
Management Science

Introduction to Stochastic Approximation

Journal:	*Wiley Encyclopedia of Operations Research and Management Science*
Manuscript ID:	EORMS-09-0052
Wiley - Manuscript type:	Introductory
Date Submitted by the Author:	27-Feb-2009
Complete List of Authors:	Yin, G. George
Keywords:	Stochastic approximation, stochastic optimization, ordinary differential equation method

Introduction to Stochastic Approximation

G. Yin*

February 27, 2009

Abstract

This chapter is concerned with stochastic approximation method. It provides a survey to the subject together with a brief account on the literature and recent development. The connection of the dynamic system aspect of the stochastic approximation algorithm is illustrated through the discrete iterations and ordinary differential equations.

Key Words. Stochastic approximation, stochastic optimization, ordinary differential equation method.

Stochastic approximation (SA) methods were introduced by Robbins and Monro in their pioneering work [10] in 1951. The original purpose was for locating zeros of a nonlinear continuous function. That is, find the roots $f(x) = 0$. When either the precise form of this function is unknown, or it is too complicated to compute. However, one is able to take "noisy" measurements or observations at desired values. The procedure is known as Robbins-Monro (RM) algorithm. The suggested solution method is a stochastic recursive algorithm. A classical example is to find appropriate dosage level of a drug, provided only $f(x)$+noise is available, where x is the level of dosage and $f(x)$ is the probability of success leading to the recovery of the patient at dosage level x. In 1952, Kiefer and Wolfowitz proposed another algorithm that is concerned with the minimization of a real-valued function using only noisy functional measurements [4]. That algorithm is commonly referred to as Kiefer-Wolfowitz (KW) algorithm.

Owing to its importance, stochastic approximation has had a long history and has drawn much attention in the past five decades. Significant progress has been made in the study of such stochastic recursive algorithms. The interesting theoretical issues in the analysis of iteratively defined stochastic processes focus on the basic paradigm of stochastic difference equations. Much of the development has been originated from a wide range of applications in optimization, control theory, economic systems, signal processing, communication theory, learning, pattern classification, neural network, and many other related fields. A number of monographs have been written; each of them has its own distinct features. To mention just a few, we cite the books of Albert and Gardner [1], Wasan [13], Nevel'son and Khasminskii [8], Kushner and Clark [5], Benveniste, Métivier, and Priouret [2], Chen [3], and Kushner and Yin [6] among others.

1 Historical Development

To put things in the historical perspective, we may divide the development of the stochastic approximation methods into several periods. The early development around 1950's and 1960's used

*Department of Mathematics, Wayne State University, Detroit, Michigan 48202, gyin@math.wayne.edu. Research of this author was supported in part by the National Science Foundation under DMS-0603287, in part by the National Security Agency under grant MSPF-068-029.

mainly basic probabilistic tools and traditional statistical assumptions such as independent and identically distributed noise together with certain restrictions on the functions such as assuming $f(x)$ to be increasing for instance. Since stochastic approximation is a recursive scheme, convergence and rates of convergence are crucial. In the original work [10], convergence in probability was obtained. Summarizing much of the early historical development, the book [13] included the with probability one (w.p.1) convergence proof for multidimensional problems of Blum, the asymptotic normality study of Sacks, and the work of Fabian among others. Soon researchers recognized that the stochastic approximation algorithms are closely related to stochastic dynamic systems. As a representative, Nevelson and Khasminskii's book [8] treated stochastic approximation as stochastic processes and dealt with martingale difference type noise processes. The use of Liapunov function was also brought in together with recursive estimation. As time went on, many applications arising in control and optimization forced researchers to examine the algorithms in which the noise encountered was correlated. In the middle 1970's, Ljung studied the problem from a dynamic system point of view [7]. The idea is: In lieu of the discrete recursion, one treats a continuous-time dynamic system given by an ordinary differential equation. For example, for the root finding problems, once the associated ordinary differential equation is obtained, the roots being sought becomes the stable points of the ordinary differential equations. Such an idea was further developed in [5]. By combining analysis and probabilistic argument, Kushner and Clark set up a framework by considering asymptotic properties of suitably scaled sequences. The work of Benveniste, Métivier, and Priouret [2] emphasized the close connection of stochastic approximation and adaptive system. One of the distinct features is the use of Markovian setting and the treatment of the Poisson equations. In addition, it exploits many identification, estimation, and tracking problems. The book of Chen [3] summarizes the work of using random varying truncation bounds developed by Chen and Zhu; it also provides trajectory-subsequence based convergence proof. The work of Kushner and Yin [6] presents a comprehensive development of the modern theory of stochastic approximation, or recursive stochastic algorithms, for both constrained and unconstrained algorithms, with step sizes that either go to zero or are constant and small and possibly random. In addition, stochastic approximation techniques have been employed in the analysis of global stochastic optimization [14], and emerging applications have been found in wireless communications, discrete optimization, and financial engineering; see the most recent work [16, 17, 18] and the references in.

2 Algorithms

RM Algorithm. We begin with the RM algorithms for finding zeros of a nonlinear function. Let $f : \mathbb{R}^r \mapsto \mathbb{R}^r$ be a continuous function. We wish to find $f(x) = 0$, but only noisy measurements

$$y_n = f(x_n) + \xi_n$$

are available, where $\{\xi_n\}$ denotes a sequence of random noise. Note that n is a positive integer referred to as "discrete time," representing the iteration number. The algorithm proposed by Robbins and Monro takes the form

$$x_{n+1} = x_n + a_n y_n, \tag{2.1}$$

where $\{a_n\}$ is a sequence of nonnegative real numbers satisfying $\sum_n a_n = \infty$ and $a_n \to 0$ as $n \to \infty$. The sequence $\{a_n\}$ is usually referred to as a sequence of step sizes or gains. The conditions on the step sizes indicate that they cannot be too small. If they are too small (i.e., $\sum_n a_n < \infty$), then the iterates may not ever converge to the desired value. To see this, take the noise free case $\xi_n = 0$ and

suppose that $f(\cdot)$ is a bounded function. Then

$$\sum_{j=0}^{\infty} |x_{j+1} - x_j| \leq \sum_{j=0}^{\infty} a_j |f(x_j)| \leq \kappa.$$

[Here and throughout the paper, $\kappa > 0$ is used as a generic constant; its value may change for different usage. Thus by our convention, $\kappa + \kappa = \kappa$ and $\kappa\kappa = \kappa$.] The above argument indicates that $\sum_j (x_{j+1} - x_j)$ converges absolutely. Nevertheless, by telescoping

$$\sum_{j=0}^{n} (x_{j+1} - x_j) = x_{n+1} - x_0.$$

Thus, $x_n \not\to x^*$, the true parameter we are approximating unless x_0 is sufficiently close to x^*.

KW Algorithm. Here, we wish to minimize a real-valued function $f(x) : \mathbb{R}^r \mapsto \mathbb{R}$ based on noise corrupted observations $F(x, \zeta_n)$ with $EF(x, \zeta_n) = f(x)$, where $\{\zeta_n\}$ represents the noise. The difficulty is that we know neither $F(\cdot)$ nor $f(\cdot)$. To approximate the minimizer, we use the finite difference approximation to the gradient of $f(x)$. Denote the finite difference interval by $\{c_n\}$ (with $c_n \to 0$ as $n \to \infty$). Use x_n to denote the nth estimate of the minimum. Suppose that for each i and each n, we observe

$$y_{n,i} = -\frac{F(x_n + c_n e_i, \zeta_n^+) - F(x_n - c_n e_i, \zeta_n^-)}{2c_n},$$

where $\zeta_n^{\pm}$ are random noise sequences. Denote $y_n = (y_{n,1}, \ldots, y_{n,r})$. Then the approximation algorithm is again given by $x_{n+1} = x_n + a_n y_n$, which is the same form as that of (2.1). By introducing

$$\xi_n = (\xi_{n,1}, \ldots, \xi_{n,r}), \ \beta_n = (\beta_{n,1}, \ldots, \beta_{n,r}),$$

with

$$\xi_{n,i} = [f(x_n + c_n e_i) - F(x_n + c_n e_i, \zeta_{n,i}^+)] - [f(x_n - c_n e_i) - F(x_n - c_n e_i, \zeta_{n,i}^-)]$$
$$\beta_{n,i} = f_{x_i}(x_n) - \frac{f(x_n + c_n e_i) - f(x_n - c_n e_i)}{2c_n},$$

where $f_x(\cdot)$ denotes the gradient of $f(\cdot)$ w.r.t. x. Now the above algorithm can be rewritten as

$$x_{n+1} = x_n - a_n f_x(x_n) + a_n \frac{\xi_n}{2c_n} + a_n \beta_n. \tag{2.2}$$

In the above, ξ_n represents the noise, and β_n denotes the bias. We have used two-sided finite difference. One-sided finite difference can also be used. However, in practice, the two-sided finite difference method appears to be more preferable since it has smaller bias. This is easily seen by taking a Taylor expansion of the finite difference quotient in β_n.

More General Algorithms and Variants. In various applications such as those arising in signal processing and adaptive controls, one often needs to treat stochastic approximation algorithms with nonadditive noise of the form

$$x_{n+1} = x_n + a_n f(x_n, \xi_n). \tag{2.3}$$

Including both (2.1) and (2.2) as special cases, such an algorithms arise, for example, in "equalization" filters in communication channels, adaptive antenna array processing etc. In addition, $f(\cdot)$ can also depend on n:

$$x_{n+1} = x_n + a_n f_n(x_n, \xi_n). \tag{2.4}$$

To be able to track slight parameter variations, one often uses an algorithm with constant step size of the form

$$x_{n+1} = x_n + \varepsilon f_n(x_n, \xi_n), \tag{2.5}$$

where $\varepsilon > 0$ is a small parameter. Suppose that H is a constraint set. We require the iterates be in the set. Write the algorithm as

$$x_{n+1} = \Pi_H \left(x_n + a_n f(x_n, \xi_n) \right), \tag{2.6}$$

where Π_H denotes the projection onto the constraint set H. If the iterate is within the projection region, we simply keep the recursion running; if it is outside the region, we project it back. Define a "reflection" or correction term z_n as

$$a_n z_n = x_{n+1} - x_n - a_n f(x_n, \xi_n),$$

i.e., it is the vector of shortest Euclidean length needed to take $x_n + a_n f(x_n, \xi_n)$ back to the constraint set H if it is not in H. Using this notation, (2.6) can be rewritten as

$$x_{n+1} = x_n + a_n f(x_n, \xi_n) + a_n z_n. \tag{2.7}$$

The constraint set can be as simple as a rectangular region, or as complex as a complicated set bounded by nonlinear functions; see [6, Section 4.3] for various possible projection regions.

In addition to the algorithms mentioned above, many other variants have been studied. For example, if one can get the observation y_n, but the sequence $\{x_n\}$ emerges in a random manner and is not at one's disposal. Then one is in the so-called passive stochastic approximation framework. Sometimes, we may wish to use constraints but wish to relax the "hard constraints," and allow them to be violated slightly from time to time. Roughly, such constraints are "soft constraints." In addition, as was mentioned, one may also define a sequence of randomly generated truncation bounds. This will enable one to relax the conditions on the growth rates of the function under consideration or the priori knowledge of the constraint set.

3 Convergence and Rates of Convergence

Convergence. To study the convergence of the algorithm, we use the ordinary differential equation method. Instead of working with the discrete iteration directly, take a continuous-time interpolation. To get some insight, we first give some heuristic argument.

Consider (2.1). Suppose that the function $f(\cdot)$ is continuous. Although rather complex noise processes can be dealt with (see [6]), for simplicity, in this survey, we assume the measurement or observation error is a sequence of independent and identically distributed random variables. Also, for simplicity, assume the iterates $\{x_n\}$ is bounded. The boundedness assumption is not crucial but can simplify the presentation here; see [6] for more general treatment. Choose a small $\Delta > 0$ such that

$$\sum_{j=n}^{n+m_n^\Delta - 1} a_j \approx \Delta \quad \text{or equivalently}$$

$$m_n^\Delta = \max \left\{ m; \sum_{j=n}^{n+m-1} a_j < \Delta \right\}.$$

Iterating on (2.1) yields

$$x_{n+m_n^\Delta} = x_n + \sum_{j=n}^{n+m_n^\Delta-1} a_j f(x_j) + \sum_{j=n}^{n+m_n^\Delta-1} a_j \xi_j.$$

For Δ small enough, and for $n \le j \le n + m_n$, $f(x_j)$ is "close" to $f(x_n)$ by the continuity. As a result

$$\sum_{j=n}^{n+m_n^\Delta-1} a_j f(x_j) \approx \sum_{j=n}^{n+m_n^\Delta-1} a_j f(x_n) \approx \Delta f(x_n),$$

and

$$x_{n+m_n^\Delta} - x_n \approx \Delta f(x_n) + \text{ error }.$$

To see the size of the error, we compute its variance:

$$\text{Var (error)} = \text{Var}\left(\sum_{j=n}^{n+m_n^\Delta-1} a_j \xi\right) = O\left(\sum_{j=n}^{n+m_n^\Delta-1} a_j^2\right) = O(\Delta a_n).$$

Therefore,

$$\frac{x_{n+m_n^\Delta} - x_n}{\Delta} \approx f(x_n) + \text{ error with diminishing variance.}$$

Over small interval, the mean change of the values of the parameter is much more important than that of the noise. The noise is averaged out in the limit, and the asymptotic behavior can be approximated by the differential equation (ODE)

$$\dot{x} = f(x). \tag{3.1}$$

If a projection algorithm is used, one obtains a projected ODE is obtained; see [6].

To prepare us for the study of the desired asymptotic properties, let us recall the notion of equicontinuity. Let $\{f_n\}$ be a sequence of $\mathbb{R}^r$-valued functions on $[0, \infty)$. The set is said to be *equicontinuous* in $C^r[0, \infty)$ (the set of $\mathbb{R}^r$-valued continuous functions defined on $[0, \infty)$) if $\{f_n(0)\}$ is bounded and for each T and $\varepsilon > 0$, there is a $\delta > 0$ such that for all n

$$\sup_{|t-s|\le\delta,\ |t|\le T} |f_n(t) - f_n(s)| \le \varepsilon. \tag{3.2}$$

The well-known *Arzelà–Ascoli* Theorem states: Let $\{f_n(\cdot)\}$ be a sequence of functions in $C^r[0, \infty)$, and let the sequence be equicontinuous. Then there is a subsequence that converges to some continuous limit, uniformly on each bounded interval.

For the stochastic approximation algorithm, take piecewise linear interpolations and define

$$t_n = \sum_{j=0}^{n-1} a_j, \quad m(t) = \max\{n : t_n \le t\},$$
$$x^0(t_n) = x_n, \quad x^0(t) = \frac{t_{n+1} - t}{a_n} x_n + \frac{t - t_n}{a_n} x_{n+1}, \quad t \in (t_n, t_{n+1}).$$

That is, the interpolation interval is $[t_n, t_{n+1})$. Next, to bring the asymptotic behavior of the process to the foreground, define a shifted sequence by $x^n(t) = x^0(t + t_n)$. Under suitable conditions, it can be shown that $\{x^n(\cdot)\}$ is uniformly bounded and equicontinuous. By Ascoli–Arzelá Theorem,

we can extract a convergent subsequence $x^{n_k}(\cdot)$ such that $x^{n_k}(\cdot) \to x(\cdot)$. Then we characterize the limit $x(\cdot)$ and prove that it is nothing but the solution of the ODE (3.1). Concerning the limit differential equations, we note that the stationary points of (3.1) are exactly the roots of $f(\cdot)$ that we are searching for. If x^* is the unique asymptotically stable point of (3.1), then we can show the iterates $x_n \to x^*$ w.p.1. If instead of a single point x^*, Z is a set of zeros of $f(\cdot)$ such that Z is asymptotically stable in the sense of Liapunov, then $x_n \to Z$ w.p.1.

In applications, it is often more convenient to work with piecewise constant interpolations. To use such an interpolation, we need to modify and extend the definition of the equicontinuity. In [6, p. 102], an extension, known as equicontinuity in the extended sense was given. Then a version of the Ascoli–Arzelá Theorem for sequences of functions that are equicontinuous in the extended sense can be used, which facilities the analysis for piecewise constant interpolated sequences.

Rates of Convergence. Suppose that $x_n \to x^*$ in an appropriate sense (either with probability one or weakly). Then one is interested in the rate of convergence. For stochastic approximation, what do we mean by "rate of convergence?" Consider, for instance, Eq. (2.1). Suppose that $x_n \to x^*$ (the true parameter) w.p.1 as $n \to \infty$. To study the convergence rate, we take a suitably scaled sequence $u_n = (x_n - x^*)/a_n^\alpha$, for some $\alpha > 0$. In case of constant-step size algorithm, this is changed to $(x_n - x^*)/\varepsilon^\alpha$. The idea is to choose α such that u_n converges in distribution to a nontrivial limit. The scaling factor α together with the asymptotic covariance of the scaled sequence gives us the rate of convergence. That is, the scaling α tells us the dependence of the estimation error $x_n - x^*$ on the step size, and the asymptotic covariance is a mean of assessing "goodness" of the approximation.

The suitable scaling for (2.1) is $\sqrt{a_n}$. If a constant step size algorithm is used, the scaling is $\sqrt{\varepsilon}$. To illustrate, consider (2.1) with $a_n = 1/n$. Assume $f(x) = H(x - x^*) + O(|x - x^*|^2)$ such that $\widetilde{H} = H + I/2$ is a stable matrix (i.e., all of its eigenvalues have negative real parts). The rate of convergence is based on local analysis. Using the definition u_n together with (2.1), we obtain

$$u_{n+1} = u_n + \frac{1}{n}\widetilde{H}u_n + \frac{1}{\sqrt{n}}\xi_n + O\left(\frac{1}{n}|u_n|^2\right).$$

Define a piecewise constant interpolation $u^n(0) = u_n$ and $u^n(t) = u_k$ for $t \in [t_{k+n} - t_n, t_{k+n+1} - t_n)$. For i.i.d. sequence $\{\xi_n\}$ with zero mean and finite second moment, it is easily shown that for $\ell \geq 1$, $\sum_{k=\ell}^{\ell+m(t)-1} \frac{1}{\sqrt{k}}\xi_k$ converges weakly to a Brownian motion $\widetilde{w}(\cdot)$ with covariance Σt. We can further rewrite the $\widetilde{w}(\cdot)$ in terms of the standard Brownian motion $w(\cdot)$ as $\widetilde{w}(t) = \Sigma^{1/2}w(t)$. Furthermore, it can be shown that $u^n(\cdot)$ converges weakly to the solution of the stochastic differential equation

$$du = \widetilde{H}udt + \Sigma^{1/2}dw.$$

Note that the above stochastic differential equation is linear in u, and as a result, it has a unique solution for each initial condition u_0. The limit stochastic differential equation gives a better description of the dynamics of the suitably scaled estimation errors. We refer the reader to [6, Chpater 8] for further reading. Denote the asymptotic covariance by S. Then it can be shown that S satisfies the following Liapunov equation

$$S\widetilde{H} + \widetilde{H}'S = -\Sigma.$$

This further implies that under suitable conditions, one can show $\sqrt{n}(x_n - x^*)$ converges in distribution to a normal random variable $N(0, S)$. The scaling factor $\sqrt{n}$ together with the asymptotic covariance S then gives us a mean of evaluation of the rate of convergence.

4 Further Remarks

One of the most important matters is to improve the performance of stochastic approximation algorithms. Let us consider KW algorithms. Each step of the KW algorithm uses $2r$ observations (if two-sided finite difference is used). Thus $2r$ steps are needed to get a derivative estimate. An alternative method is to update only one direction at each iteration using a finite difference estimate and to choose the direction randomly at each step. This results in using only two observations at each step. Such an idea appeared in Kushner and Clark [5] together with the associated convergence and rate of convergence results. At that time, it was noted that if the random directions are chosen at the unit sphere, then there is a little advantage as compared to the KW method. The work of Spall [12] indicated that if the random directions are chosen on the unit cube, then the performance is better than the KW algorithms, which is a significant discovery and enables one to treat high dimensional problems effectively.

Along another line in the late 1980's Polyak [9] and Ruppert [11] independently discovered that averaging can help to improve the efficiency of stochastic approximation algorithms. The main idea is to first generate a sequence of rough estimates using large step sizes and then average the resulting estimate to smooth things out. The algorithm is given by

$$\begin{aligned} x_{n+1} &= x_n + \frac{1}{n^\gamma}\left(f(x_n) + \xi_n\right) \\ \overline{x}_{n+1} &= \overline{x}_n - \frac{1}{n+1}\overline{x}_n + \frac{1}{n+1}x_{n+1}, \end{aligned} \tag{4.1}$$

where $1/2 < \gamma < 1$. It can be shown that such algorithms are asymptotically optimal in that they have the best scaling α and the smallest variance. Moreover, the efficiency issue can also be studied in connection with the Cramér-Rao bounds.

The setting of stochastic approximation can be carried over to infinite dimensional spaces, e.g., Banach spaces and/or Hilbert spaces. In addition to the pure mathematical interest, the motivation of the study stems from the fact in various optimization problems, the solutions of the problems involve finding the root or the optimum for points not living in Euclidean spaces, but in function spaces.

To summarize, stochastic approximation methods have been the subject of an enormous literature, both theoretical and applied for five decades. Owing to the vast amount of literature accumulated, it is very difficult or virtually impossible to provide an exhaustive list of references on stochastic approximation. Our hope is that with the references provided at the end of this article, the reader will be able to pick out suitable references of his/her needs. Moreover, it is likewise very difficult to give extensive account on the technical development in a survey paper of this scale. As a result, we choose the road of discussing the main ideas and leaving most of the technical details aside. Appropriate references are provided, however.

References

[1] A.E. Albert and L.A. Gardner, *Stochastic Approximation and Nonlinear Regression*, MIT Press, Cambridge, MA, 1967.

[2] A. Benveniste, M. Métivier and P. Priouret, *Adaptive Algorithms and Stochastic Approximation*, Springer-Verlag, Berlin, 1990.

[3] H.-F. Chen. *Stochastic Approximation and Its Applications*. Kluwer Academic, Dordrecht, Netherlands, 2002.

[4] J. Kiefer and J. Wolfowitz, Stochastic estimation of the maximum of a regression function, *Ann. Math. Statist.* **23** (1952), 462-466.

[5] H.J. Kushner and D.S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, 1978.

[6] H.J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd Edition, Springer-Verlag, New York, 2003.

[7] L. Ljung, Analysis of recursive stochastic algorithms, *IEEE Trans. Automat. Control* **AC-22** (1977), 551-575.

[8] M.B. Nevel'son and R.Z. Khasminskii, *Stochastic Approximation and Recursive Estimation*, Translation of Math. Monographs, v47, AMS, Providence, 1976.

[9] B.T. Polyak, New method of stochastic approximation type, *Automat. Remote Control* **51** (1990), 937-946.

[10] H. Robbins and S. Monro, A stochastic approximation method, *Ann. Math. Statist.* **22** (1951), 400-407.

[11] D. Ruppert, Efficient estimations from a slowly convergent Robbins-Monro process, Technical Report, No. 781, School of Oper. Res. & Industrial Eng., Cornell Univ., 1988. [see also the chapter Stochastic approximation in *Handbook in Sequential Analysis*, B.K. Ghosh and P.K. Sen Eds., 503-529, Marcel Dekker, New York, 1991.]

[12] J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation, *IEEE Trans. Automat. Control* **AC-37** (1992), 331-341.

[13] M.T. Wasan, *Stochastic Approximation*, Cambridge Press, London, 1969.

[14] G. Yin, Rates of convergence for a class of global stochastic optimization algorithms, *SIAM J. Optim.*, **10** (1999), 99-120.

[15] G. Yin, Stochastic approximation: Theory and applications, in *Handbook of Stochastic Analysis and Applications*, D. Kannan and V. Lakshmikantham Eds., Marcel Dekker, New York, 2002, 577-624.

[16] G. Yin and V. Krishnamurthy, Least mean square algorithms with Markov regime switching limit, *IEEE Trans. Automatic Control*, **50** (2005), 577-593.

[17] G. Yin, V. Krishnamurthy, and C. Ion, Regime switching stochastic approximation algorithms with application to adaptive discrete stochastic optimization, *SIAM J. Optim.*, **14** (2004), 1187-1215.

[18] G. Yin, R.H. Liu, and Q. Zhang, Recursive algorithms for stock Liquidation: A stochastic optimization approach, *SIAM J. Optim.*, **13** (2002), 240-263.