Stochastic Approximation and Its Applications

# Nonconvex Optimization and Its Applications

Volume 64



*The titles published in this series are listed at the end of this volume.*

# Stochastic Approximation and Its Applications

by

Han-Fu Chen

*Institute of Systems Science,*
*Academy of Mathematics and System Science,*
*Chinese Academy of Sciences,*
*Beijing, P.R. China*

**KLUWER ACADEMIC PUBLISHERS**

NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-306-48166-9
Print ISBN: 1-4020-0806-6



Created in the United States of America

Visit Kluwer Online at: http://kluweronline.com
and Kluwer's eBookstore at: http://ebooks.kluweronline.com

# Contents

# Preface

Estimating unknown parameters based on observation data containing information about the parameters is ubiquitous in diverse areas of both theory and application. For example, in system identification the unknown system coefficients are estimated on the basis of input-output data of the control system; in adaptive control systems the adaptive control gain should be defined based on observation data in such a way that the gain asymptotically tends to the optimal one; in blind channel identification the channel coefficients are estimated using the output data obtained at the receiver; in signal processing the optimal weighting matrix is estimated on the basis of observations; in pattern classification the parameters specifying the partition hyperplane are searched by learning, and more examples may be added to this list.

All these parameter estimation problems can be transformed to a root-seeking problem for an unknown function. To see this, let $y_k$ denote the observation at time $k$, i.e., the information available about the unknown parameters at time $k$. It can be assumed that the parameter under estimation denoted by $x^0$ is a root of some unknown function $f(\cdot)$: $f(x^0) = 0$. This is not a restriction, because, for example, $\|x - x^0\|^2$ may serve as such a function. Let $x_k$ be the estimate for $x^0$ at time $k$. Then the available information $y_{k+1}$ at time $k+1$ can formally be written as

$$y_{k+1} = f(x_k) + \epsilon_{k+1},$$

where

$$\epsilon_{k+1} = y_{k+1} - f(x_k).$$

Therefore, by considering $y_{k+1}$ as an observation on $f(\cdot)$ at $x_k$ with observation error $\epsilon_{k+1}$, the problem has been reduced to seeking the root $x^0$ of $f(\cdot)$ based on $\{y_k\}$.

It is clear that for each problem to specify $f(\cdot)$ is of crucial importance. The parameter estimation problem is possible to be solved only if $f(\cdot)$

is appropriately selected so that the observation error $\{\epsilon_k\}$ meets the requirements figured in convergence theorems.

If $f(\cdot)$ and its gradient can be observed without error at any desired values, then numerical methods such as Newton-Raphson method among others can be applied to solving the problem. However, this kind of methods cannot be used here, because in addition to the obvious problem concerning the existence and availability of the gradient, the observations are corrupted by errors which may contain not only the purely random component but also the structural error caused by inadequacy of the selected $f(\cdot)$.

Aiming at solving the stated problem, Robbins and Monro proposed the following recursive algorithm

$$x_{k+1} = x_k + a_k y_{k+1}, \qquad a_k > 0,$$

to approximate the sought-for root $x^0$, where $a_k$ is the step size. This algorithm is now called the Robbins-Monro (RM) algorithm. Following this pioneer work of stochastic approximation, there have been a large amount of applications to practical problems and research works on theoretical issues.

At beginning, the probabilistic method was the main tool in convergence analysis for stochastic approximation algorithms, and rather restrictive conditions were imposed on both $f(\cdot)$ and $\{\epsilon_k\}$. For example, it is required that the growth rate of $f(x)$ is not faster than linear as $\|x\|$ tends to infinity and $\{\epsilon_k\}$ is a martingale difference sequence [78]. Though the linear growth rate condition is restrictive, as shown by simulation it can hardly be simply removed without violating convergence for RM algorithms.

To weaken the noise conditions guaranteeing convergence of the algorithm, the ODE (ordinary differential equation) method was introduced in [72, 73] and further developed in [65]. Since the conditions on noise required by the ODE method may be satisfied by a large class of $\{\epsilon_k\}$ including both random and structural errors, the ODE method has been widely applied for convergence analysis in different areas. However, in this approach one has to *a priori* assume that the sequence of estimates $\{x_k\}$ is bounded. It is hard to say that the boundedness assumption is more desirable than a growth rate restriction on $f(\cdot)$.

The stochastic approximation algorithm with expanding truncations was introduced in [27], and the analysis method has then been improved in [14]. In fact, this is an RM algorithm truncated at expanding bounds, and for its convergence the growth rate restriction on $f(\cdot)$ is not required. The convergence analysis method for the proposed algorithm is called the trajectory-subsequence (TS) method, because the analysis is carried out at trajectories where the noise condition is satisfied and in contrast to the ODE method the noise condition need not be verified on the whole sequence $\{x_k\}$ but is verified only along convergent subsequences $\{x_{n_k}\}$. This makes a great difference when dealing with the state-dependent noise $\{\epsilon_{k+1}(x_k)\}$, because a convergent subsequence $\{x_{n_k}\}$ is always bounded while the boundedness of the whole sequence $\{x_k\}$ is not guaranteed before establishing its convergence. As shown in Chapters 4, 5, and 6 for most of parameter estimation problems after transforming them to a root-seeking problem, the structural errors are unavoidable, and they are state-dependent.

The expanding truncation technique equipped with TS method appears a powerful tool in dealing with various parameter estimation problems: it not only has succeeded in essentially weakening conditions for convergence of the general stochastic approximation algorithm but also has made stochastic approximation possible to be successfully applied in diverse areas. However, there is a lack of a reference that systematically describes the theoretical part of the method and concretely shows the way how to apply the method to problems coming from different areas. To fill in the gap is the purpose of the book.

The book summarizes results on the topic mostly distributed over journal papers and partly contained in unpublished material. The book is written in a systematical way: it starts with a general introduction to stochastic approximation and then describes the basic method used in the book, proves the general convergence theorems and demonstrates various applications of the general theory.

In Chapter 1 the problem of stochastic approximation is stated and the basic methods for convergence analysis such as probabilistic method, ODE method, TS method, and the weak convergence method are introduced.

Chapter 2 presents the theoretical foundation of the algorithm with expanding truncations: the basic convergence theorems are proved by TS method; various types of noises are discussed; the necessity of the imposed noise condition is shown; the connection between stability of the equilibrium and convergence of the algorithm is discussed; the robustness of stochastic approximation algorithms is considered when the commonly used conditions deviate from the exact satisfaction, and the moving root tracking is also investigated. The basic convergence theorems are presented in Section 2.2, and their proof is elementary and purely deterministic.

Chapter 3 describes asymptotic properties of the algorithms: convergence rates for both cases whether or not the gradient of $f(\cdot)$ is degenerate; asymptotic normality of $\{x_k\}$ and asymptotic efficiency by averaging method.

Starting from Chapter 4 the general theory developed so far is applied to different fields. Chapter 4 deals with optimization by using stochastic approximation methods. Convergence and convergence rates of the Kiefer-Wolfowitz (KW) algorithm with expanding truncations and randomized differences are established. A global optimization method consisting in combination of the KW algorithms with search methods is defined, and its a.s. convergence as well as asymptotic behaviors are established. Finally, the global optimization method is applied to solving the model reduction problem.

In Chapter 5 the general theory is applied to the problems arising from signal processing. Applying the stochastic approximation method to blind channel identification leads to a recursive algorithm estimating the channel coefficients and continuously improving the estimates while receiving new signal in contrast to the existing "block" algorithms. Applying TS method to principal component analysis results in improving conditions for convergence. Stochastic approximation algorithms with expanding truncations with TS method are also applied to adaptive filters with and without constraints. As a result, conditions required for convergence have been considerably improved in comparison with the existing results. Finally, the expanding truncation technique and TS method are applied to the asynchronous stochastic approximation.

In the last chapter, the general theory is applied to problems arising from systems and control. The ideal parameter for operation is identified for stochastic systems by using the methods developed in this book. Then the obtained results are applied to the adaptive quadratic control problem. Adaptive regulation for a nonlinear nonparametric system and learning pole assignment are also solved by the stochastic approximation method.

The book is self-contained in the sense that there are only a few points using knowledge for which we refer to other sources, and these points can be ignored when reading the main body of the book. The basic mathematical tools used in the book are calculus and linear algebra based on which one will have no difficulty to read the fundamental convergence Theorems 2.2.1 and 2.2.2 and their applications described in the subsequent chapters. To understand other material, probability concept, especially the convergence theorems for martingale difference sequences are needed. Necessary concept of probability theory is given in Appendix A. Some facts from probability that are used at a few specific points are listed in Appendix A but without proof, because omitting the corresponding parts still makes the rest of the book readable. However, the proof of convergence theorems for martingales and martingale difference sequences is provided in detail in Appendix B.

The book is written for students, engineers and researchers working in the areas of systems and control, communication and signal processing, optimization and operation research, and mathematical statistics.

HAN-FU CHEN

# Acknowledgments

The support of the National Key Project of China and the National Natural Science Foundation of China is gratefully acknowledged. The author would like to express his gratitude to Dr. Haitao Fang for his helpful suggestions and useful discussions. The author would also like to thank Ms. Jinling Chang for her skilled typing and to thank my wife Shujun Wang for her constant support.