

Vivek S. Borkar

Stochastic Approximation: A Dynamical Systems Viewpoint

Second Edition

Volume 48

Texts and Readings in Mathematics

Editorial Board

Manindra Agrawal

Indian Institute of Technology Kanpur, Kanpur, Uttar Pradesh, India

V. Balaji

Chennai Mathematical Institute, Siruseri, Tamil Nadu, India

R. B. Bapat

Indian Statistical Institute, New Delhi, India

V. S. Borkar

Indian Institute of Technology, Mumbai, Maharashtra, India

Apoorva Khare

Indian Institute of Science, Bangalore, India

T. R. Ramadas

Chennai Mathematical Institute, Chennai, India

V. Srinivas

Tata Institute of Fundamental Research, Mumbai, India

P. Vanchinathan

Vellore Institute of Technology, Chennai, Tamil Nadu, India

Advisory Editor

C. S. Seshadri

Chennai Mathematical Institute, Chennai, India

Managing Editor

Rajendra Bhatia

Ashoka University, Sonepat, Haryana, India

The **Texts and Readings in Mathematics** series publishes high-quality textbooks, research-level monographs, lecture notes and contributed volumes. Undergraduate and graduate students of mathematics, research scholars and teachers would find this book series useful. The volumes are carefully written as teaching aids and highlight characteristic features of the theory. Books in this series are co-published with Hindustan Book Agency, New Delhi, India.

Vivek S. Borkar

Stochastic Approximation: A Dynamical Systems Viewpoint

Second Edition



Vivek S. Borkar

Department of Electrical Engineering, Indian Institute of Technology
Bombay, Mumbai, Maharashtra, India

ISSN 2366-8717

e-ISSN 2366-8725

Texts and Readings in Mathematics

ISBN

e-ISBN 978-81-951961-1-1

<https://doi.org/10.1007/978-81-951961-1-1>

© Hindustan Book Agency 2022

Around the time this work was being completed, I lost in quick succession two friends whose friendship I cherished and from whom I learnt a lot—Prof. A. S. Vasudeva Murthy who succumbed to COVID-19 on April 29, 2021, in Bengaluru, and Prof. Ari Arapostathis who lost his battle with cancer on May 19, 2021, in Austin, Texas. This book is dedicated to them.

Preface

(Expanded from the 1st Edition)

Stochastic approximation was introduced in a 1951 article in the *Annals of Mathematical Statistics* by Robbins and Monro. Originally conceived as a tool for statistical computation, an area in which it retains a place of pride, it has come to thrive in a totally different discipline, viz. that of engineering. The entire area of ‘adaptive signal processing’ in communication engineering has been dominated by stochastic approximation algorithms and variants, as is evident from even a cursory look at any standard text on the subject. Then there are more recent applications to adaptive resource allocation problems in communication networks. In control engineering too, stochastic approximation is the main paradigm for online algorithms for system identification and adaptive control.

This is not accidental. The key word in most of these applications is *adaptive*. Stochastic approximation has several intrinsic traits that make it an attractive framework for adaptive schemes. It is designed for uncertain (read ‘stochastic’) environments, where it allows one to track the ‘average’ or ‘typical’ behavior of such an environment. It is incremental; i.e., it makes small changes in each step, which ensures a graceful behavior of the algorithm. This is a highly desirable feature of any adaptive scheme. Furthermore, it usually has low computational and memory requirements per iterate, another desirable feature of adaptive systems. Finally, it conforms to our anthropomorphic notion of adaptation: It makes small adjustments so as to improve a certain performance criterion based on feedback received from the environment.

For these very reasons, there has been a resurgence of interest in this class of algorithms in several new areas of engineering. One of these, viz. communication networks, is already mentioned above. Yet another major application domain has been artificial intelligence, where stochastic approximation has provided the basis for many learning or ‘parameter tuning’ algorithms in machine learning. Notable among these are the algorithms for training neural networks and the algorithms for reinforcement learning, a popular learning paradigm for

autonomous software agents with applications in e-commerce, robotics, etc.

Yet another fertile terrain for stochastic approximation has been in the area of economic theory, for reasons not entirely dissimilar to those mentioned above. On one hand, they provide a good model for collective phenomena, where *micromotives* (to borrow a phrase from Thomas Schelling) of individual agents aggregate to produce interesting *macrobehavior*. The ‘nonlinear urn’ scheme analyzed by Arthur and others to model increasing returns in economics is a case in point. On the other hand, their incrementality and low per iterate computational and memory requirements make them an ideal model of a *boundedly rational* economic agent, a theme which has dominated their application to learning models in economics, notably to learning in evolutionary games.

This flurry of activity, while expanding the application domain of stochastic approximation, has also thrown up interesting new issues, some of them dictated by technological imperatives. Consequently, it has spawned interesting new theoretical developments as well. When the previous edition of the book was written, the time seemed ripe for a book pulling together old and new developments in the subject with an eye on the aforementioned applications. There are, indeed, several excellent texts already in existence, many of which will be referenced later in this book. But they tend to be *comprehensive* texts: excellent for the already initiated, but rather intimidating for someone who wants to make quick inroads, hence a need for a ‘bite-sized’ text. The first edition of this book was an attempt at one, though it did stray into a few esoteric themes in some places.

Having decided to write a book, there was still a methodological choice. Stochastic approximation theory has two somewhat distinct strands of research. One, popular with statisticians, uses the techniques of martingale theory and associated convergence theorems for analysis. The second, popular more with control engineers, treats the algorithm as a noisy discretization of an ordinary differential equation and analyzes it as such. The latter approach was preferred, because the kind of intuition that it offers is an added advantage in many of the engineering applications.

Of course, this was not the first book expounding this approach. There are several predecessors such as the excellent texts by Benveniste–Metivier–Priouret, Duflo, and Kushner–Yin referenced later in the book. These are, however, what we have called *comprehensive* texts above, with a wealth of information. This book was and is *not* comprehensive, but is more of a compact account of the highlights to enable an interested, mathematically literate reader to run through the basic ideas and issues in a relatively short time span. The other ‘novelties’ of the book would be a certain streamlining and fine-tuning of proofs using that eternal source of wisdom—*hindsight*. There are occasional new variations on proofs sometimes leading to improved results, or just shorter proofs, inclusion of some newer themes in theory and applications, and so on. Given the nature of the subject, a certain mathematical sophistication was unavoidable. For the benefit of those not quite geared for it, the more advanced mathematical requirements are collected in a few appendices. These should serve as a source for quick reference and pointers to the literature, but not as a replacement for a firm grounding in the respective areas. Such grounding is a must for anyone wishing to contribute to the *theory* of stochastic approximation. Those interested more in *applying* the results to their respective specialities may not feel the need to go much further than this book.

In the decade that passed after the first publication of this book, much happened that warranted a second enlarged edition, including several newer developments. This was also seen as an opportunity to clean up minor bugs and a major error (in the functional central limit theorem, the author thanks Sean Meyn for pointing it out almost as soon as the book was published). There is some reorganization of existing chapters, and much new material has been added to several of them, far too much and too diverse to list here in detail. Nevertheless, here is a brief list of the major additions:

1. Additional concentration bounds in Chap. 3.
2. Additional stability tests in Chap. 4.
3. A new section in Chap. 5 presenting an alternative viewpoint.

4. Corrected Chap. 7 (see above).
5. Inclusion of a new class of multiple timescale algorithms in Chap. 8.
6. Additional section on tracking in Chap. 9.
7. Chapter 10 on more general noise models is new.
8. Chapter 11 is a new chapter expanded from a section of Chap. 10 of the first edition, to include a lot of new material.
9. Addition of new material on primal-dual algorithms and nonlinear power method to Chap. 12.
10. Addition of Alekseev formula to Chap. 14.

With these additions, the book is no longer ‘bite-sized’ as originally intended, but the core part of the book, i.e., the first few chapters, still remains so. The later chapters are separate bite-sized, stand-alone accounts in their own right of special topics of interest. So it still retains the original flavor, at least that is the pious hope of the author.

Let us conclude this long preface with the pleasant task of acknowledging all the help received in this venture. The author forayed into stochastic approximation around 1993–1994, departing significantly from his dominant activity till then, which was controlled Markov processes. This move was helped by a project on adaptive systems supported by a Homi Bhabha Fellowship. More than the material help, the morale boost was a great help and he is immensely grateful for it. His own subsequent research in this area has been supported by grants from the Department of Science and Technology, Government of India, and was conducted mostly in the two ‘Tata’ Institutes: Indian Institute of Science at Bengaluru and the Tata Institute of Fundamental Research in Mumbai. It continues now in the Indian Institute of Technology Bombay.

When the draft for the first edition was prepared, Dr. V. V. Phansalkar went through the early drafts of a large part of the book and, with his fine eye for detail, caught many errors. Professor Shalabh Bhatnagar, Dr. Arzad Alam Kherani, and Dr. Huizen (Janey) Yu also read the draft and pointed out corrections and improvements (Janey shares with Dr. Phansalkar the rare trait for having a great eye for detail and contributed a lot to the final cleanup). Dr. Sameer Jalnapurkar did a major overhaul of Chaps. 1–4, which in addition to fixing errors, greatly contributed to their readability. Ms. Diana Gillooly of Cambridge University Press did an extremely meticulous job of editorial corrections on the final manuscript of the first edition.

Some of the additions and improvements in the new version have benefited from a helping hand from many junior colleagues who have selflessly contributed to the effort. I would like to mention in particular Prof. Rajesh Sundaresan of the Indian Institute of Science who has contributed a lot to the chapters on stability and stochastic recursive inclusions, in addition to the general overhaul of most other chapters, inclusive of the not-so-small LATEX intricacies. My former student Dr. Suhail Mohmad Shah has been instrumental in the chapter on stochastic gradient descent, which includes much that I learned from him (particularly the machine learning aspects). In fact, the last subsection of this chapter is his gift. This chapter also benefited from the suggestions of Dr. Praneeth Netrapalli of Microsoft Research, Bengaluru. Another former student, Bhumesh Kumar, also read parts of the manuscript and made useful comments. Professor Alexandre Reiffers-Masson of IMT Atlantique, France, made some valuable suggestions regarding Chap. 5 which led to significant improvements, particularly in the section on projected stochastic approximation. The ‘almost final’ version was whetted by Prof. Gopal Basak of the Indian Statistical Institute, Kolkata, and a lot (hopefully, all) that escaped the keen eyes of the aforementioned was caught by him. Sarath Yasodharan helped with a further cleanup of Chaps. 11 and 12. Professor K. S. Mallikarjuna Rao helped me out of several remaining LATEX pitfalls. The author takes full blame for whatever errors and shortcomings that still remain. His wife Shubhangi and son Aseem have been extremely supportive as always.

Vivek S. Borkar

Mumbai, India
December 2021

Contents

1 Introduction

References

2 Convergence Analysis

2.1 The o.d.e. Limit

2.2 Extensions and Variations

References

3 Finite Time Bounds and Traps

3.1 Estimating the Lock-In Probability

3.2 Sample Complexity

3.3 Extensions

3.4 Avoidance of Traps

References

4 Stability Criteria

4.1 Introduction

4.2 Stability Through a Scaling Limit

4.3 Stability by Comparison

4.4 Stabilizability by Stepsize Selection

4.5 Stabilizability by Resetting

4.6 Convergence for Tight Iterates

References

5 Stochastic Recursive Inclusions

5.1 Introduction

5.2 The Differential Inclusion Limit

5.3 An Alternative Representation

5.4 Applications

5.5 Projected Stochastic Approximation

5.6 Extensions

References

6 Asynchronous Schemes

6.1 Introduction

6.2 Asymptotic Behavior

6.3 Effect of Delays

6.4 Convergence

References

7 A Limit Theorem for Fluctuations

7.1 Introduction

7.2 A Tightness Result

7.3 The Functional Central Limit Theorem

7.4 The Convergent Case

References

8 Multiple Timescales

8.1 Two Timescales

8.2 Controlled Markov Noise

8.3 Averaging the Natural Timescale

8.4 Other Multiscale Algorithms

References

9 Constant Stepsize Algorithms

9.1 Introduction

9.2 Asymptotic Behavior

9.3 Tracking

9.4 Refinements

References

10 General Noise Models

10.1 The Problem

10.2 Preliminaries

10.3 Moment Estimates

10.4 Main Results

10.5 Extensions and Variations

References

Stochastic Gradient Schemes

11.1 Introduction

11.2 The Basic Stochastic Gradient Descent

11.3 Approximate Gradient Schemes

11.4 Some Important Variants

11.5 Langevin Algorithm and Simulated Annealing

11.6 Simulation-Based Optimization

11.7 SGD for Machine Learning

11.7.1 Empirical Risk Minimization

11.7.2 Variance Reduction Techniques

11.7.3 Error Bounds

References

12 Liapunov and Related Systems

12.1 Introduction

12.2 Primal-Dual Algorithms

12.3 Stochastic Fixed Point Iterations

12.4 Collective Phenomena

12.5 Miscellaneous Applications

References

Appendix A: Topics in Analysis

Appendix B: Ordinary Differential Equations

Appendix C: Topics in Probability

References

Index

About the Author

Vivek S. Borkar is Professor at the Department of Electrical Engineering, Indian Institute of Technology (IIT) Bombay, Powai, Mumbai, India. Earlier, he held positions at the TIFR Centre for Applicable Mathematics and Indian Institute of Science in Bengaluru; Indian Institute of Science, Bengaluru; Tata Institute of Fundamental Research and Indian Institute of Technology Bombay in Mumbai. He also held visiting positions at the Massachusetts Institute of Technology (MIT), the University of Maryland at College Park, the University of California at Berkeley, and the University of Illinois at Urbana-Champaign, USA.

Professor Borkar obtained his B.Tech. (Electrical Engineering) from the IIT Bombay in 1976, MS (Systems and Control Engineering) from Case Western Reserve University in 1977, and Ph.D. (Electrical Engineering and Computer Sciences) from the University of California, Berkeley, USA, in 1980. He is Fellow of American Mathematical Society, IEEE, and the World Academy of Sciences, and of various science and engineering academies in India. He has won several awards and honours in India and was an invited speaker at the ICM 2006 in Madrid. He has authored/coauthored six books and several archival publications. His primary research interests are in stochastic optimization and control, covering theory and algorithms.

1. Introduction

Vivek S. Borkar¹✉

(1) Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

Consider an initially empty urn to which balls, either red or black, are added one at a time. Let y_n denote the *number* of red balls at time n and $x_n \stackrel{\text{def}}{=} y_n/n$ the *fraction* of red balls at time n . We shall suppose that the conditional probability that the next, i.e., the $(n + 1)$ st ball is red given the past up to time n is a function of x_n alone. Specifically, suppose that it is given by $p(x_n)$ for a prescribed $p : [0, 1] \rightarrow [0, 1]$. It is easy to describe $\{x_n, n \geq 1\}$ recursively as follows. For $\{y_n\}$, we have the simple recursion

$$y_{n+1} = y_n + \xi_{n+1},$$

where

$$\xi_{n+1} = \begin{cases} 1 & \text{if the } (n+1)\text{st ball is red,} \\ 0 & \text{if the } (n+1)\text{st ball is black.} \end{cases}$$

Some simple algebra then leads to the following recursion for $p(x_n)$:

$$x_{n+1} = x_n + \frac{1}{n+1}(\xi_{n+1} - x_n),$$

with $x_0 = 0$. This can be rewritten as

$$x_{n+1} = x_n + \frac{1}{n+1}(p(x_n) - x_n) + \frac{1}{n+1}(\xi_{n+1} - p(x_n)).$$

Note that $M_n \stackrel{\text{def}}{=} \xi_n - p(x_{n-1})$, $n \geq 1$ (with $p(x_0) \stackrel{\text{def}}{=}$ the probability of the first ball being red) is a sequence of zero mean random variables satisfying $E[M_{n+1} | \xi_m, m \leq n] = 0$ for $n \geq 0$. This means that $\{M_n\}$ is a *martingale difference sequence* (see Appendix C), i.e., uncorrelated with the ‘past,’ and thus can be thought of as ‘noise.’ The above equation then can be thought of as a noisy discretization (or *Euler scheme* in numerical analysis parlance) for the ordinary differential equation (o.d.e. for short)

$$\dot{x}(t) = p(x(t)) - x(t), \quad t \geq 0,$$

with nonuniform stepsizes $a(n) \stackrel{\text{def}}{=} 1/(n + 1)$ and ‘noise’ $\{M_n\}$. (Compare with the standard Euler scheme $x_{n+1} = x_n + a(p(x_n) - x_n)$ for a small $a > 0$.) If we assume $p(\cdot)$ to be Lipschitz continuous, o.d.e. theory guarantees that this o.d.e. is well-posed, i.e., it has a unique solution for any initial condition $x(0)$ that in turn depends continuously on $x(0)$ (see Appendix B). Note also that the right-hand side of the o.d.e. is nonnegative at $x(t) = 0$ and nonpositive at $x(t) = 1$, implying that any trajectory starting in $[0, 1]$ will remain in $[0, 1]$ forever. As this is a scalar o.d.e., any bounded trajectory must converge. To see this, note that it cannot move in any particular direction (‘right’ or ‘left’) forever without converging, because it is bounded. Neither can it change direction from ‘right’ to ‘left’ or vice versa without passing through an equilibrium point. This change of direction would require that the right-hand side of the o.d.e. changes sign and hence by continuity must pass through a point where it vanishes, i.e., an equilibrium point (say) x^* . This is not possible because $x(t) \equiv x^*$ is the only trajectory through x^* by the uniqueness thereof. (For that matter, the o.d.e. couldn’t have been going both right and left at any given x because this direction is uniquely prescribed by the sign of $p(x) - x$.) Thus we have proved that $p(\cdot)$ must converge to an equilibrium. The set of equilibria of the o.d.e. is given by the points where the right-hand side vanishes, i.e., the set $H = \{x : p(x) = x\}$. This is precisely the set of fixed points of $p(\cdot)$. Once again, as the right-hand side is continuous, is ≤ 0 at 1, and is ≤ 0 at 0, it must pass through 0 by the mean value theorem and hence H is nonempty. (One could also invoke the Brouwer fixed

point theorem (Appendix A) to say this, as $p : [0, 1] \rightarrow [0, 1]$ is a continuous map from a convex compact set to itself.)

Our interest, however, is in $p(x_n)$. The theory we develop later in this book will tell us that under reasonable assumptions the $p(x_n)$ ‘track’ the o.d.e. with probability one in a certain sense to be made precise later, implying in particular that they converge a.s. to H . The key factors that ensure this are the fact that the stepsize $a(n)$ tends to zero as $n \rightarrow \infty$, and the fact that the series $\sum_n a(n)M_{n+1}$ converges a.s., a consequence of the martingale convergence theorem. The first observation means in particular that the ‘pure’ discretization error becomes asymptotically negligible. The second observation implies that the ‘tail’ of the above convergent series given by $\sum_{m=n}^{\infty} a(m)M_{m+1}$, which is the ‘total noise added to the system from time n on,’ goes to zero a.s. This in turn ensures that the error due to noise is also asymptotically negligible. We note here that the fact $\sum_n a(n)^2 = \sum_n (1/(n+1)^2) < \infty$ plays a crucial role in facilitating the application of martingale convergence theorem here. This is because it ensures the following sufficient condition for convergence of the concerned martingale (see Appendix C):

$$\sum_n E[(a(n)M_{n+1})^2 | \xi_m, m \leq n] \leq \sum_n a(n)^2 < \infty, \text{ a.s.}$$

One also needs the fact that $\sum_n a(n) = \infty$, because in view of our interpretation of $a(n)$ as a time step, this ensures that the discretization does cover the entire time axis. As we are interested in tracking the asymptotic behavior of the o.d.e., this is clearly necessary.

Let us consider now the simple case when H is a finite set. Then one can say more, viz., that the $p(x_n)$ converge a.s. to some point in H . The exact point to which they converge will be random, though we shall later narrow down the choice somewhat (e.g., the ‘unstable’ equilibria will be avoided with probability one under suitable conditions). For the time being, we shall stop with this conclusion and discuss the motivation for looking at such ‘*nonlinear urns*’.

This simple set-up was proposed by Arthur (1994) to model the phenomenon of increasing returns in economics. The reader will have heard of the ‘law of diminishing returns’ from classical economics,

which can be described as follows. Any production enterprise such as a farm or a factory requires both fixed and variable resources. When one increases the amount of variable resources, each additional unit thereof will get a correspondingly smaller fraction of fixed resources to draw upon, and therefore the additional returns due to it will correspondingly diminish.

While quite accurate in describing the traditional agricultural or manufacturing sectors, this law seems to be contradicted in some other sectors, particularly in case of the modern ‘information goods.’ One finds that larger investments in a brand actually fetch larger returns because of standardization and compatibility of goods, brand loyalty of customers, and so on. This is the so-called ‘increasing returns’ phenomenon modeled by the urn above, where each new red ball is an additional unit of investment in a particular product. If the predominance of one color tends to fetch more balls of the same, then after some initial randomness the process will get ‘locked into’ one color which will dominate overwhelmingly. (This corresponds to $p(x) > x$ for $x \in (x_0, 1)$ for some $x_0 \in (0, 1)$, and $< x$ for $x \in (x_0, 1)$. Then the stable equilibria are 0 and 1, with x_0 being an unstable equilibrium. Recall that in this set-up the equilibrium x is stable if $p(x) > x$, unstable if $p(x) < x$.) When we are modeling a pair of competing technologies or conventions, this means that one of them, not necessarily the better one, will come to dominate overwhelmingly. Arthur (1994) gives several interesting examples of this phenomenon. To mention a few, he describes how the VHS technology came to dominate over Sony Betamax for video recording, why the present arrangement of letters and symbols on typewriters and keyboards (QWERTY) could not be displaced by a superior arrangement called DVORAK, why ‘clockwise’ clocks eventually displaced ‘counterclockwise’ clocks, and so on.

Keeping economics aside, our interest here will be in the recursion for $p(x_n)$ and its analysis sketched above using an o.d.e. The former constitutes a special (and a rather simple one at that) case of a much broader class of stochastic recursions called ‘stochastic approximation’ which form the main theme of this book. What’s more, the analysis based on a limiting o.d.e. is an instance of the ‘o.d.e. approach’ to stochastic approximation which is our main focus here. Before spelling

out further details of these, here's another example, this time from statistics.

Consider a repeated experiment which gives a string of input-output pairs (X_n, Y_n) , $x_0 = 0$ with $X_n \in \mathcal{R}^m$, $Y_n \in \mathcal{R}^k$ resp. We assume that $\{(X_n, Y_n)\}$ are i.i.d. Our objective will be to find the 'best fit' $Y_n = f_w(X_n) + \epsilon_n$, $n \geq 1$, from a given parametrized family of functions $\{f_w : \mathcal{R}^m \rightarrow \mathcal{R}^k : w \in \mathcal{R}^d\}$, x_0 being the 'error.' What constitutes the 'best fit,' however, depends on the choice of our error criterion and we shall choose this to be the popular 'mean square error' given by $g(w) \stackrel{\text{def}}{=} \frac{1}{2}E[\|\epsilon_n\|^2] = \frac{1}{2}E[\|Y_n - f_w(X_n)\|^2]$. That is, we aim to find a w^* that minimizes this overall $w \in \mathcal{R}^d$. This is the standard problem of nonlinear regression. Typical parametrized families of functions are polynomials, splines, linear combinations of sines and cosines, or more recently, wavelets, and neural networks. The catch here is that the above expectation cannot be evaluated because the underlying probability law is not known. Also, we do not suppose that the entire string $\{(X_n, Y_n)\}$ is available as in classical regression, but that it is being delivered one at a time in 'real time.' The aim then is to come up with a recursive scheme that is 'online,' i.e., one which tries to 'learn' w^* in real time by adaptively updating a running guess as new observations come in.

To arrive at such a scheme, let us pretend to begin with that we do know the underlying law. Assume also that f_w is continuously differentiable in w and let $x(t) = 0$ denote its Jacobian matrix w.r.t. w . The obvious thing to try then is to differentiate the mean square error w.r.t. w and set the derivative equal to zero. Assuming that the interchange of expectation and differentiation is justified, we then have

$$\nabla^w g(w) = -E[D_w f_w(X_n)^T(Y_n - f_w(X_n))] = 0$$

at the minimum point. We may then seek to minimize the mean square error by gradient descent, given by:

$$\begin{aligned} w_{n+1} &= w_n - \nabla^w g(w_n) \\ &= w_n + E[D_w f_w(X_n)^T(Y_n - f_w(X_n))] \Big|_{w=w_n}. \end{aligned}$$

This, of course, is not feasible for reasons already mentioned, viz., that the expectation above cannot be evaluated. As a first approximation, we may then consider replacing the expectation by the ‘empirical gradient,’ i.e., the argument of the expectation evaluated at the current guess w_n for w^* , leading to

$$w_{n+1} = w_n + [D_w f_w(X_n)^T(Y_n - f_w(X_n))] \Big|_{w=w_n}.$$

This, however, will lead to a different kind of problem. The term added to w_n on the right is the n th in a sequence of ‘i.i.d. functions’ of w , evaluated at w_n . Thus we expect the above scheme to be (and it is) a correlated random walk, zigzagging its way to glory. We may therefore want to smoothen it by making only a small, incremental move in the direction suggested by the right-hand side instead of making the full move. This can be achieved by replacing the right-hand side by a convex combination of it and the previous guess w_n , with only a small weight $1 > a(n) > 0$ for the former. That is, we replace the above iteration by

$$w_{n+1} = (1 - a(n))w_n + a(n) \left(w_n + [D_w f_w(X_n)^T(Y_n - f_w(X_n))] \Big|_{w=w_n} \right).$$

Equivalently,

$$w_{n+1} = w_n + a(n) [D_w f_w(X_n)^T(Y_n - f_w(X_n))] \Big|_{w=w_n}.$$

Once again, if we do not want the scheme to zigzag drastically, we should make $\{a(n)\}$ small, the smaller the better. At the same time, a small $a(n)$ leads to a very small correction to w_n at each iterate, so the scheme will work very slowly, if at all. This suggests starting the iteration with relatively high values of $a(n)$ and then letting $a(n) \rightarrow 0$. (In fact, $a(n) < 1$ as above is not needed, as that can be taken care of by scaling the empirical gradient.) Now let us add and subtract the exact error gradient at the ‘known guess’ w_n from the empirical gradient on the right-hand side and rewrite the above scheme as

$$\begin{aligned}
w_{n+1} = w_n + & \left. a(n)E \left[D_w f_w(X_n)^T (Y_n - f_w(X_n)) \right] \right|_{w=w_n} \\
& + a(n) \left(\left. \left[D_w f_w(X_n)^T (Y_n - f_w(X_n)) \right] \right|_{w=w_n} \right. \\
& \left. - E \left[D_w f_w(X_n)^T (Y_n - f_w(X_n)) \right] \right|_{w=w_n}.
\end{aligned}$$

This is of the form

$$w_{n+1} = w_n + a(n)(-\nabla^w g(w_n) + M_{n+1}),$$

with $\{M_n\}$ a martingale difference sequence as in the previous example. One may then view this scheme as a noisy discretization of the o.d.e.

$$\dot{w}(t) = -\nabla^w g(w(t)).$$

This is a particularly well studied o.d.e. We know that it will converge to $H \stackrel{\text{def}}{=} \{w : \nabla^w g(w) = 0\}$ in general, and if this set is discrete, to in fact one of the local minima of g for typical (i.e., *generic*: belonging to an open dense set) initial conditions. As before, we are interested in tracking the asymptotic behavior of this o.d.e. Hence we must ensure that the discrete time steps $\{a(n)\}$ used in the ‘noisy discretization’ above do cover the entire time axis, i.e.,

$$\sum_n a(n) = \infty, \tag{1.1}$$

while retaining $a(n) \rightarrow 0$. (Recall from the previous example that $a(n) \rightarrow 0$ is needed for asymptotic negligibility of discretization errors.) At the same time, we also want the error due to noise to be asymptotically negligible a.s. The urn example above then suggests that we also impose

$$\sum_n a(n)^2 < \infty, \tag{1.2}$$

which asymptotically suppresses the noise variance.

One can show that with (1.1) and (1.2) in place, for reasonable g (e.g., with $\lim_{\|w\| \rightarrow \infty} g(w) = \infty$ and finite H , among other possibilities)

the ‘stochastic gradient scheme’ above will converge a.s. to a local minimum of g .

Once again, what we have here is a special case—perhaps the most important one—of stochastic approximation, analyzed by invoking the ‘o.d.e. method.’

What, after all, is stochastic approximation? Historically, stochastic approximation started as a scheme for solving a nonlinear equation $a(n) < 1$ given ‘noisy measurements’ of the function h . That is, we are given a black box which on input x , gives as its output $h(x) + \xi$, where ξ is a zero mean random variable representing noise. The stochastic approximation scheme proposed by Robbins and Monro (1951)¹ was to run the iteration

$$x_{n+1} = x_n + a(n)[h(x_n) + M_{n+1}], \quad (1.3)$$

where $\{M_n\}$ is the noise sequence and $\{a(n)\}$ are positive scalars satisfying (1.1) and (1.2) above. The expression in the square brackets on the right is the noisy measurement. That is, $h(x_n)$ and M_{n+1} are not separately available, only their sum is. We shall assume $\{M_n\}$ to be a martingale difference sequence, i.e., a sequence of integrable random variables satisfying

$$E[M_{n+1}|x_m, M_m, m \leq n] = 0.$$

This is more general than it appears. For example, an important special case is the d -dimensional iteration

$$x_{n+1} = x_n + a(n)f(x_n, \xi_{n+1}), \quad n \geq 0, \quad (1.4)$$

for an $f : \mathcal{R}^d \times \mathcal{R}^k \rightarrow \mathcal{R}^d$ with i.i.d. \mathcal{R}^k -valued noise $\{\xi_n\}$. This can be put in the format of (1.3) by defining $h(x) = E[f(x, \xi_1)]$ and $M_{n+1} = f(x_n, \xi_{n+1}) - h(x_n)$ for $n \geq 0$.

Since its inception, the scheme (1.3) has been a cornerstone of scientific computation. This has been so largely because of the following advantages, already apparent in the above examples:

- It is designed to handle noisy situations, e.g., the stochastic gradient scheme above. One may say that it captures the average behavior in the long run. The noise in practice may not be only from

measurement errors or approximations but may also be added deliberately as a probing device or a randomized action, as, e.g., in certain dynamic game situations.

- It is incremental, i.e., it makes small moves at each step. This typically leads to more graceful behavior of the algorithm at the expense of its speed. We shall say more on this later in the book.
- In typical applications, the computation and memory requirement *per iterate* is low, making its implementation easy.

These features make the scheme ideal for applications where the key word is ‘adaptive.’ Thus the stochastic approximation paradigm dominates the fields of adaptive signal processing, adaptive control, and certain subdisciplines of machine learning and artificial intelligence such as neural networks and reinforcement learning—see, e.g., Bertsekas and Tsitsiklis (1996) and Haykin (2001, 2016). Not surprisingly, it is also emerging as a popular framework for modeling boundedly rational macroeconomic agents—see, e.g., Sargent (1993). The two examples above are representative of these two strands. We shall be seeing many more instances later in this book.

As noted in the preface, there are broadly two approaches to the theoretical analysis of such algorithms. The first, popular with statisticians, is the probabilistic approach based on the theory of martingales and associated objects such as ‘almost supermartingales.’ The second approach, while still using a considerable amount of martingale theory, views the iteration as a noisy discretization of a limiting o.d.e. Recall that the standard ‘Euler scheme’ for numerically approximating a trajectory of the o.d.e.

$$\dot{x}(t) = h(x(t))$$

would be

$$x_{n+1} = x_n + ah(x_n),$$

with $x(t) \equiv x^*$ and $a > 0$ a small time step. The stochastic approximation iteration differs from this in two aspects: replacement of the constant time step ‘ a ’ by a time-varying ‘ $a(n)$,’ and the presence of ‘noise’ M_{n+1} . This qualifies it as a noisy discretization of the o.d.e. Our aim is to seek x for which $a(n) < 1$, i.e., the equilibrium point(s) of this

o.d.e. The o.d.e. would converge (if it does) to these only asymptotically unless it happens to start exactly there. Hence to capture this asymptotic behavior, we need to track the o.d.e. over the infinite time interval. This calls for the condition $\sum_n a(n) = \infty$. The condition $\sum_n a(n)^2 < \infty$ will on the other hand ensure that the errors due to discretization of the o.d.e. and those due to the noise $\{M_n\}$ both become asymptotically negligible with probability one. (This condition was already discussed in the first example, but we can also motivate it in another way. Let $\{M_n\}$ be i.i.d. zero mean with a finite variance σ^2 . Then by a theorem of Kolmogorov, $\sum_n a(n)M_n$ converges a.s if and only if $\sum_n a(n)^2$ converges. This suggests the square-summability condition on $\{a(n)\}$.) Together, these conditions try to ensure that the iterates do indeed capture the asymptotic behavior of the o.d.e. We have already seen instances of this above.

Pioneered by Meerkov (1972), Derevitskii and Fradkov (1974) (see Fradkov, 2011 for a recent overview), this ‘o.d.e. approach’ was further extended by Ljung (1977). It is already the basis of several excellent texts such as Benveniste et al. (1990), Duflo (1996, 1997), and Kushner and Yin (2003), among others.² The rendition here is a slight variation of the traditional one, with an eye on pedagogy so that the highlights of the approach can be introduced quickly and relatively simply. The lecture notes of Benaim (1999) are perhaps the closest in spirit to the treatment here, though at a much more advanced level. (Benaim’s notes in particular give an overview of his own contributions which introduced important notions from dynamical systems theory such as internal chain recurrence to stochastic approximation. These represent a major development in this field.)

While it is ultimately a matter of personal taste, the o.d.e. approach does indeed appeal to engineers because of the ‘dynamical systems’ view it takes, which is close to their hearts. Also, as we shall see at the end of this book, it can serve as a useful recipe for concocting new algorithms: any convergent o.d.e. is a potential source of a stochastic approximation algorithm that converges with probability one.

The organization of the book is as follows. Chapter 2 gives the basic convergence analysis for the stochastic approximation algorithm with decreasing stepsizes. This is the core material for the rest of the book.

Chapter 3 gives some refinements of the results of Chap. 2, viz., an estimate for probability of convergence to a specific attractor if the iterates fall in its domain of attraction. It also gives a result about avoidance with probability one of unstable equilibria. Chapter 4 gives some ‘stability tests’ that ensure the boundedness of iterates with probability one. Chapter 5 gives the counterparts of the basic results of Chap. 2 for a more general iteration which has a differential inclusion as a limit rather than an o.d.e. This is useful in many practical instances, which are also described in this chapter. Equally important, it plays a role in some of the subsequent developments. Chapter 6 describes the distributed asynchronous implementations of the algorithm. Chapter 7 establishes the functional central limit theorem for fluctuations associated with the basic scheme of Chap. 2. Chapter 8 analyzes the cases when more than one timescale is used. This chapter, notably the sections on ‘averaging the natural timescale,’ is technically a little more difficult than the rest and the reader may skip the details of the proofs on a first reading. All the development thus far uses decreasing stepsize. Chapter 9 briefly summarizes the corresponding theory for constant stepsize which are popular in some applications, notably those involving tracking a slowly varying environment. Chapter 10 deals with an important extension of the framework of Chap. 2, viz., to long range dependent and heavy tailed noise, important because of its occurrence in several situations such as Internet traffic and finance. The results here are, not surprisingly, weaker on the whole.

Chapters 11 and 12 of the book have a different flavor: they collect several examples from engineering, economics, etc., where the stochastic approximation formalism has paid rich dividends. Thus the general techniques of the first part of the book are specialized to each case of interest and the additional structure available in the specific problem under consideration is exploited to say more, depending on the context. It is a mixed bag, the idea being to give the reader a flavor of the various ‘tricks of the trade’ that may come in handy in future applications. Broadly speaking, one may classify these applications into three strands. The first strand, studied in Chap. 11, is the stochastic gradient scheme and its variants wherein h above is either the negative gradient of some function or something close to the negative gradient. This scheme, by far the most pervasive application of stochastic

approximation, is the underlying paradigm for many adaptive filtering, parameter estimation and stochastic optimization schemes in general. It has generated an unprecedented interest in recent times due to its usefulness in large scale machine learning problems with their own specific attributes that have led to some interesting domain-specific work as well as results of broader appeal. We try to give a taste of these developments in a dedicated section. The second strand, the focus of Chap. 12, deals mostly with systems which are ‘gradient-like’ in the sense that there is monotone decrease of a certain ‘Liapunov function’ along the o.d.e. trajectory, even though the driving vector field is not the negative gradient thereof. In other words, this vector field, when nonzero, must make an acute angle with the negative gradient of the Liapunov function. The first example we consider is the o.d.e. version of fixed point iterations, i.e., successive application of a map from a space to itself so that it may converge to a point that remains invariant under it (i.e., a fixed point). These are important in a class of applications arising from dynamic programming. Then we consider a general collection of o.d.e.s modeling collective phenomena in economics etc., such as the urn example above, for many of which we can construct a convenient Liapunov function. There are two important special situations outside of these paradigms, where we have a convergence result for ‘almost all’ initial conditions using ingenious arguments *sans* Liapunov functions. The first is Smale’s proof of the continuous time ‘global Newton scheme’ and the second is Hirsch’s result for cooperative o.d.e. These too have enormous algorithmic applications and therefore have been summarized here in brief, in Chaps. 11 and 12 resp.

This classification is, of course, not exhaustive and some instances of stochastic approximation in practice may fall outside of all this. Also, we do not consider the continuous time analog of stochastic approximation (see, e.g., Mel’nikov, 1996).

The background required for this book is a good first course on measure theoretic probability, particularly the theory of discrete parameter martingales, at the level of Breiman (1968) or Williams (1991) (though we shall generally refer to Borkar, 1995, more out of familiarity than anything), and a first course on ordinary differential equations at the level of Hirsch et al. (2003). There are a few spots

where something more than this is required, viz., the theory of weak (Prohorov) convergence of probability measures. The three appendices at the end collect together the key aspects of these topics that are needed here.

References

- Arthur, W. B. (1994). *Increasing returns and path dependence in the economy*. University of Michigan Press
- Benaim, M. (1999). Dynamics of stochastic approximation algorithms. In J. Azéma, M. Emery, M. Ledoux, & M. Yor (Eds.), *Le Séminaire de Probabilités*. Springer Lecture Notes in Mathematics (Vol. 1709, pp. 1–68). Springer Verlag
- Benveniste, A., Metivier, M., & Priouret, P. (1990). *Adaptive algorithms and stochastic approximation*. Springer Verlag
- Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific
- Borkar, V. S. (1995). *Probability theory: An advanced course*. Springer Verlag
- Breiman, L. (1968). *Probability*. Addison-Wesley
- Derevitskii DP, Fradkov AL (1974) Two models for analyzing the dynamics of adaptation algorithms. Automation and Remote Control 35:59–67
[MathSciNet]
- Duflo, M. (1996). *Algorithmes stochastiques*. Springer Verlag
- Duflo, M. (1997). *Random iterative models*. Springer Verlag
- Fradkov, A. L. (2011, December 12–15). Continuous-time averaged models of discrete-time stochastic systems: Survey and open problems. In *Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, Orlando, FL (pp. 2076–2081)
- Haykin, S. (2001). *Adaptive filter theory* (4th ed.). Prentice Hall
- Haykin, S. O. (2016). *Neural networks and learning machines* (3rd ed.). New York: McMillan Publishing Company
- Hirsch, M. W., Smale, S., & Devaney, R. (2003). *Differential equations, dynamical systems, and an introduction to chaos*. Academic Press
- Kushner, H. J., & Yin, G. (2003). *Stochastic approximation and recursive algorithms and applications* (2nd ed.). Springer Verlag
- Lai TL (2003) Stochastic approximation. Annals of Statistics 31:391–406
[MathSciNet][Crossref][zbMATH]

Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22(4), 551–575

Ljung, L., Pflug, G. Ch., & Walk, H. (1992). *Stochastic approximation and optimization of random systems*. Birkhäuser

Meerkov S (1972) Simplified description of slow Markov walks 2. *Automation and Remote Control* 33(2):404–414
[[MathSciNet](#)][[zbMATH](#)]

Mel'nikov A (1996) Stochastic differential equations: Singularity of coefficients, regression models and stochastic approximation. *Russian Mathematical Surveys* 51(5):819–909
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Nevelson, M., & Hasminskii, R. (1976). *Stochastic approximation and recursive equations*. Translations of Mathematical Monographs No. 47. American Mathematical Society

Robbins H, Monro S (1951) A stochastic approximation method. *Annals of Mathematical Statistics* 22(3):400–407
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Sargent, T. (1993). *Bounded rationality in macroeconomics*. Clarendon Press

Wasan, M. (1969). *Stochastic approximation*. Cambridge University Press

Williams, D. (1991). *Probability with martingales*. Cambridge University Press

Footnotes

¹ See Lai (2003) for an interesting historical perspective.

² Wasan (1969) and Nevelson and Hasminskii (1976) are two early texts on stochastic approximation, though with a different flavor. See also Ljung et al. (1992).

2. Convergence Analysis

Vivek S. Borkar¹✉

(1) Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

2.1 The o.d.e. Limit

In this chapter we begin our formal analysis of the stochastic approximation scheme in \mathcal{R}^k given by

$$x_{n+1} = x_n + a(n)[h(x_n) + M_{n+1}], \quad n \geq 0, \quad (2.1)$$

with prescribed x_0 and with the following assumptions which we recall from the last chapter:

- (A1)** The map $h : \mathcal{R}^d \rightarrow \mathcal{R}^d$ is Lipschitz:
 $\|h(x) - h(y)\| \leq L\|x - y\|$ for all $x, y \in \mathcal{R}^d$ and some $0 < L < \infty$.
(A2) Stepsizes $\{a(n)\}$ are positive scalars satisfying

$$\sum_n a(n) = \infty, \quad \sum_n a(n)^2 < \infty. \quad (2.2)$$

- (A3)** $\{M_n\}$ is a martingale difference sequence with respect to the increasing family of σ -fields

$$\mathcal{F}_n \stackrel{\text{def}}{=} \sigma(x_m, M_m, m \leq n) = \sigma(x_0, M_1, \dots, M_n), \quad n \geq 0.$$

That is,

$$E[M_{n+1} | \mathcal{F}_n] = 0 \quad \text{a.s.}, \quad n \geq 0.$$

Furthermore, $\{M_n\}$ are square-integrable with

(2.3)

$$E[\|M_{n+1}\|^2 | \mathcal{F}_n] \leq K(1 + \|x_n\|^2) \text{ a.s. , } n \geq 0,$$

for some constant $K > 0$.

Assumption (A1) implies in particular the linear growth condition for $p(\cdot)$: For a fixed x_0 ,

$\|h(x)\| \leq \|h(x_0)\| + L\|x - x_0\| \leq K'(1 + \|x\|)$ for a suitable constant $K' > 0$ and all $x \in \mathcal{R}^d$. Thus

$$\begin{aligned} E[\|x_{n+1}\|^2]^{\frac{1}{2}} &\leq E[\|x_n\|^2]^{\frac{1}{2}} + a(n)K'(1 + E[\|x_n\|^2]^{\frac{1}{2}}) \\ &\quad + a(n)\sqrt{K}(1 + E[\|x_n\|^2]^{\frac{1}{2}}). \end{aligned}$$

We have used here the following fact: $\sqrt{1+z^2} \leq 1+z$ for $z \geq 0$.

Along with (2.3) and the condition $E[\|x_0\|^2] < \infty$, this implies inductively that $E[\|x_n\|^2]$, $E[\|M_n\|^2]$ remain finite for each n .

We shall carry out our analysis under the further assumption:

(A4) The iterates of (2.1) remain bounded a.s., i.e.,

$$\sup_n \|x_n\| < \infty, \text{ a.s.} \quad (2.4)$$

Such a bound is far from automatic and usually not very easy to establish. Some techniques for establishing this will be discussed in Chap. 4.

The limiting o.d.e. which (2.1) might be expected to track asymptotically can be written by inspection as

$$\dot{x}(t) = h(x(t)), \quad t \geq 0. \quad (2.5)$$

Assumption (A1) ensures that (2.5) is well-posed, i.e., has a unique solution for any $x(0)$ that depends continuously on $x(0)$. The basic idea of the o.d.e. approach to the analysis of (2.1) is to construct a suitable continuous interpolated trajectory $x_0 \in (0, 1)$, and show that it asymptotically almost surely approaches the solution set of (2.5). This is done as follows: Define time instants

$$t(0) = 0, \quad t(n) = \sum_{m=0}^{n-1} a(m), \quad n \geq 1.$$

We can view $(n + 1)$ as the ‘*algorithmic timescale*’. This is slower than the natural clock (because $a(n) \rightarrow 0$). One intuitive way of understanding the o.d.e. limit is to invoke the analogy with the averaging phenomenon in two timescale systems in applied mathematics and say that, because of the relative slowness of the algorithmic timescale, the martingale difference noise gets ‘averaged out’ to zero in the limit. By (2.2), $p(x) > x$. Let

$I_n \stackrel{\text{def}}{=} [t(n), t(n + 1)]$, $n \geq 0$. Define a continuous, piecewise linear $\bar{x}(t)$, $t \geq 0$, by $\bar{x}(t(n)) = x_n$, $n \geq 0$, with linear interpolation on each interval I_n . That is,

$$\bar{x}(t) = x_n + (x_{n+1} - x_n) \frac{t - t(n)}{t(n+1) - t(n)}, \quad t \in I_n.$$

Note that by (2.4), $\sup_{t \geq 0} \|\bar{x}(t)\| = \sup_n \|x_n\| < \infty$ a.s. Let $x^s(t)$, $t \geq s$, denote the unique solution to (2.5) ‘starting at s ’:

$$\dot{x}^s(t) = h(x^s(t)), \quad t \geq s,$$

with $\{x_n, n \geq 1\}$, $K > 0$. Likewise, let $x^s(t)$, $t \geq s$, denote the unique solution to (2.5) ‘ending at s ’:

$$\dot{x}^s(t) = h(x^s(t)), \quad t \geq s,$$

with $\{x_n, n \geq 1\}$, $K > 0$. Define also

$$\zeta_n = \sum_{m=0}^{n-1} a(m) M_{m+1}, \quad n \geq 1.$$

By (A3) and the remarks that follow, (ζ_n, \mathcal{F}_n) , $n \geq 1$, is a zero mean, square-integrable martingale. Furthermore, by (A2), (A3) and (A4),

$$\sum_{n \geq 0} E[\|\zeta_{n+1} - \zeta_n\|^2 | \mathcal{F}_n] = \sum_{n \geq 0} a(n)^2 E[\|M_{n+1}\|^2 | \mathcal{F}_n] < \infty, \quad \text{a.s.}$$

It follows from the martingale convergence theorem (Appendix C) that ζ_n converges a.s. as $n \rightarrow \infty$.

The idea of the proof is to compare the interpolated iterates $p(\cdot)$, a piecewise linear and continuous function, on a sliding time window

$[t, t + T]$ of a fixed length $T > 0$, with the trajectory of the o.d.e. that coincides with it at the beginning (resp., the end) of this time window. We analyze the gap between the two as this time window moves toward infinity (i.e., $t \uparrow \infty$). The reason for doing so is the obvious fact that in general, comparing the two over the entire time axis does not help, because the errors due to discretization and noise build up in an uncontrolled manner. As we shall see, what we do is more than adequate for capturing the essentials of the asymptotic behavior of $p(\cdot)$.

Lemma 2.1 For any $T > 0$,

$$\lim_{s \rightarrow \infty} \sup_{t \in [s, s+T]} \|\bar{x}(t) - x^s(t)\| = 0, \quad \text{a.s.}$$

$$\lim_{s \rightarrow \infty} \sup_{t \in [s-T, s]} \|\bar{x}(t) - x_s(t)\| = 0, \quad \text{a.s.}$$

Proof We shall only prove the first claim, as the arguments for proving the second claim are completely analogous. Fix a sample point in the probability 1 set where (A4) holds and the martingale $\{\xi_n\}$ converges. The argument below will be for this fixed sample point, i.e., pathwise, a.s. Let $p(x) - x$ be in $[t(n), t(n) + T]$. Let $[t] \stackrel{\text{def}}{=} \max\{t(k) : t(k) \leq t\}$. Then by construction,

$$\bar{x}(t(n+m)) = \bar{x}(t(n)) + \sum_{k=0}^{m-1} a(n+k)h(\bar{x}(t(n+k))) + \delta_{n,n+m}, \quad (2.6)$$

where $\delta_{n,n+m} \stackrel{\text{def}}{=} \zeta_{n+m} - \zeta_n$. Compare this with

(2.7)

$$\begin{aligned}
x^{t(n)}(t(m+n)) &= \bar{x}(t(n)) + \int_{t(n)}^{t(n+m)} h(x^{t(n)}(t)) dt \\
&= \bar{x}(t(n)) + \sum_{k=0}^{m-1} a(n+k) h(x^{t(n)}(t(n+k))) \\
&\quad + \int_{t(n)}^{t(n+m)} (h(x^{t(n)}(y)) - h(x^{t(n)}([y]))) dy.
\end{aligned}$$

We shall now bound the integral on the right-hand side. Let θ denote the d -dimensional zero vector and let $C_0 \stackrel{\text{def}}{=} \sup_n \|x_n\| < \infty$ by (A4). Let $L > 0$ denote the Lipschitz constant of h as before and let $s \leq t \leq s+T$. Note that $\|h(x) - h(\theta)\| \leq L\|x\|$, and so $\|h(x)\| \leq \|h(\theta)\| + L\|x\|$. Since $x^s(t) = \bar{x}(s) + \int_s^t h(x^s(\tau)) d\tau$,

$$\begin{aligned}
\|x^s(t)\| &\leq \|\bar{x}(s)\| + \int_s^t [\|h(\theta)\| + L\|x^s(\tau)\|] d\tau \\
&\leq (C_0 + \|h(\theta)\|T) + L \int_s^t \|x^s(\tau)\| d\tau.
\end{aligned}$$

By Gronwall's inequality (Appendix B), it follows that

$$\|x^s(t)\| \leq (C_0 + \|h(\theta)\|T)e^{LT}, \quad s \leq t \leq s+T.$$

Thus, for all $s \leq t \leq s+T$,

$$\|h(x^s(t))\| \leq C_T \stackrel{\text{def}}{=} \|h(\theta)\| + L(C_0 + \|h(\theta)\|T)e^{LT} < \infty.$$

Now, if $0 \leq k \leq (m-1)$ and $x_{n+1} = x_n + a(p(x_n) - x_n)$,

$$\begin{aligned}
\|x^{t(n)}(t) - x^{t(n)}(t(n+k))\| &\leq \left\| \int_{t(n+k)}^t h(x^{t(n)}(s)) ds \right\| \\
&\leq C_T(t - t(n+k)) \\
&\leq C_T a(n+k).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\left\| \int_{t(n)}^{t(n+m)} (h(x^{t(n)}(t)) - h(x^{t(n)}([t]))) dt \right\| \\
&\leq \int_{t(n)}^{t(n+m)} L \|x^{t(n)}(t) - x^{t(n)}([t])\| dt \\
&= L \sum_{k=0}^{m-1} \int_{t(n+k)}^{t(n+k+1)} \|x^{t(n)}(t) - x^{t(n)}(t(n+k))\| dt
\end{aligned} \tag{2.8}$$

$$\begin{aligned}
&\leq C_T L \sum_{k=0}^{m-1} a(n+k)^2
\end{aligned} \tag{2.9}$$

$$\leq C_T L \sum_{k=0}^{\infty} a(n+k)^2 \xrightarrow{n \uparrow \infty} 0.$$

Also, since the martingale (ζ_n, \mathcal{F}_n) converges, we have

$$\sup_{k \geq 0} \|\delta_{n,n+k}\| \xrightarrow{n \uparrow \infty} 0. \tag{2.10}$$

Subtracting (2.7) from (2.6) and taking norms, we have

$$\begin{aligned}
& \|\bar{x}(t(n+m)) - x^{t(n)}(t(n+m))\| \\
& \leq L \sum_{i=0}^{m-1} a(n+i) \|\bar{x}(t(n+i)) - x^{t(n)}(t(n+i))\| \\
& \quad + C_T L \sum_{k \geq 0} a(n+k)^2 + \sup_{k \geq 0} \|\delta_{n,n+k}\|.
\end{aligned}$$

Define $K_{T,n} = C_T L \sum_{k \geq 0} a(n+k)^2 + \sup_{k \geq 0} \|\delta_{n,n+k}\|$. Note that $x(t) \equiv x^*$ as $n \rightarrow \infty$. Also, let $z_i = \|\bar{x}(t(n+i)) - x^{t(n)}(t(n+i))\|$ and $b_i \stackrel{\text{def}}{=} a(n+i)$. Thus, the above inequality becomes

$$z_m \leq K_{T,n} + L \sum_{i=0}^{m-1} b_i z_i.$$

Note that $z_0 = 0$ and $\sum_{i=0}^{m-1} b_i \leq T$. The discrete Gronwall lemma (see Appendix B) tells us that

$$\sup_{0 \leq i \leq m} z_i \leq K_{T,n} e^{LT}.$$

One then has that for $t(n+m) \leq t(n) + T$,

$$\|\bar{x}(t(n+m)) - x^{t(n)}(t(n+m))\| \leq K_{T,n} e^{LT}.$$

If $t(n+k) \leq t \leq t(n+k+1)$, we have

$$\bar{x}(t) = \lambda \bar{x}(t(n+k)) + (1-\lambda) \bar{x}(t(n+k+1))$$

for some $p(x) - x$. Thus,

$$\begin{aligned}
& \|x^{t(n)}(t) - \bar{x}(t)\| \\
&= \|\lambda(x^{t(n)}(t) - \bar{x}(t(n+k))) + (1-\lambda)(x^{t(n)}(t) - \bar{x}(t(n+k+1)))\| \\
&\leq \lambda \left\| x^{t(n)}(t(n+k)) - \bar{x}(t(n+k)) + \int_{t(n+k)}^t h(x^{t(n)}(s))ds \right\| \\
&\quad + (1-\lambda) \left\| x^{t(n)}(t(n+k+1)) - \bar{x}(t(n+k+1)) - \int_t^{t(n+k+1)} h(x^{t(n)}(s))ds \right\| \\
&\leq (1-\lambda) \|x^{t(n)}(t(n+k+1)) - \bar{x}(t(n+k+1))\| \\
&\quad + \lambda \|x^{t(n)}(t(n+k)) - \bar{x}(t(n+k))\| \\
&\quad + \max(\lambda, 1-\lambda) \int_{t(n+k)}^{t(n+k+1)} \|h(x^{t(n)}(s))\| ds.
\end{aligned}$$

Since $\|h(x^s(t))\| \leq C_T$ for all $s \leq t \leq s+T$, it follows that

$$\sup_{t \in [t(n), t(n)+T]} \|\bar{x}(t) - x^{t(n)}(t)\| \leq K_{T,n} e^{\text{LT}} + C_T a(n+k).$$

The claim now follows for the special case of $s \rightarrow \infty$ along $(n+1)$.
The general claim follows easily from this special case. \square

Recall that a closed set $A \subset \mathcal{R}^d$ is said to be an *invariant set* (resp. a positively/negatively invariant set) for the o.d.e. (2.5) if any trajectory $x(t)$, $-\infty < t < \infty$ (resp. $x_{n+1} = x_n + a(p(x_n) - x_n)$) of (2.5) with $a(n) \rightarrow 0$ satisfies $x(t) \in A \ \forall t \in \mathcal{R}$ (resp. $\forall t \geq 0 / \forall t \leq 0$). It is said to be *internally chain transitive* if, in addition, it is compact¹ and for any $x, y \in A$ and any $\epsilon > 0, T > 0$, there exist $n \geq 0$ and points $x_0, x_1, \dots, x_{n-1}, x_n = y$ in A such that $\|x_0 - x\| < \epsilon$ and the trajectory of (2.5) initiated at x_0 meets with the ϵ -neighborhood of x_{i+1} for $0 \leq i < n$ after a time $\geq T$. If we restrict to $y = x$ in the above, the set is said to be *internally chain recurrent*. Let $\Phi_t : \mathcal{R}^d \rightarrow \mathcal{R}^d$ denote the map that takes $x(0)$ to $x(t)$ via (2.5). Under our conditions on h , this map will be continuous (in fact Lipschitz) for each $t > 0$ (see Appendix B). From the uniqueness of solutions to (2.5) in both forward and backward time, it follows that Φ_t is invertible. In fact it turns out to

be a homeomorphism, i.e., a continuous bijection with a continuous inverse (see Appendix B). Thus we can define $\Phi_{-t}(x) = \Phi_t^{-1}(x)$, the point at which the trajectory starting at time 0 at x and running backward in time for a duration t would end up. Along with $\Phi_0 \equiv$ the identity map on \mathcal{R}^k , $\{x_n, n \geq 1\}$ defines a group of homeomorphisms on \mathcal{R}^k , which is referred to as the *flow* associated with (2.5). Thus the definition of an invariant set can be recast as follows: A is invariant if

$$H = \{x : p(x) = x\}$$

A corresponding statement applies to positively or negatively invariant sets with $t \geq 0$, resp. $t \leq 0$, with $=$ replaced by x_0 . Our general convergence theorem for stochastic approximation, due to Benaim (1996), is the following.

Theorem 2.1 Almost surely, the sequence $p(x_n)$ generated by (2.1) converges to a (possibly sample path dependent) connected internally chain transitive invariant set of (2.5).

Proof Consider a sample point where (2.4) and the conclusions of Lemma 2.1 hold. Let A denote the set $\delta_{n,n+m} \stackrel{\text{def}}{=} \zeta_{n+m} - \zeta_n$, the set of limit points of $\bar{x}(t)$ as $t \uparrow \infty$ (i.e., its ' ω -limit set'). Since $p(\cdot)$ is continuous and bounded, $\overline{\{\bar{x}(s) : s \geq t\}}$, $t \geq 0$, is a nested family of nonempty compact and connected sets. A , being the intersection thereof, will also be nonempty and compact. It will also be connected, being a set of limit points of a continuous trajectory. Then $\bar{x}(t) \rightarrow A$ and therefore $x_n \rightarrow A$. In fact, for any $\epsilon > 0$, let

$A^\epsilon \stackrel{\text{def}}{=} \{x : \min_{y \in A} \|x - y\| < \epsilon\}$. Then for each fixed $\epsilon > 0$,

$$(A^\epsilon)^c \cap (\cap_{t \geq 0} \overline{\{\bar{x}(s) : s \geq t\}}) = \emptyset.$$

Hence by the finite intersection property of families of compact sets,

$$(A^\epsilon)^c \cap \overline{\{\bar{x}(s) : s \geq t'\}} = \emptyset$$

for some $t' > 0$. That is, $\|x_0 - x\| < \epsilon$. Since $\epsilon > 0$ is arbitrary, $\bar{x}(t + \cdot) \rightarrow A$. On the other hand, if $K > 0$, there exist $s_n \uparrow \infty$ in

$[0, \infty)$ such that $\bar{x}(s_n) \rightarrow x$. This is immediate from the definition of A . In fact, we have

$$\max_{s \in [t(n), t(n+1)]} \|\bar{x}(s) - \bar{x}(t(n))\| = O(a(n)) \rightarrow 0$$

as $n \rightarrow \infty$. Thus we may take $1 > a(n) > 0$ for suitable (X_n, Y_n) without any loss of generality. Let $p(\cdot)$ denote the trajectory of (2.5) with $p(x) = x$. Then by the continuity of the map Φ_t defined above, it follows that $x^{s_n}(s_n + t) = \Phi_t(\bar{x}(s_n)) \rightarrow \Phi_t(x) = \tilde{x}(t)$ for all $p(x) = x$. By the first part of Lemma 2.1 for some $T > 0$, $\bar{x}(s_n + t) \rightarrow \tilde{x}(t)$, $t \in [0, T]$, implying that $\tilde{x}(t) \in A$, $t \in [0, T]$, as well. A similar argument works for $t \in [-T, 0]$, using the second part of Lemma 2.1. Since $T > 0$ was arbitrary, it follows that A is invariant under (2.5).

Let $\tilde{x}_1, \tilde{x}_2 \in A$ and fix $\epsilon > 0$, $T > 0$. Pick $\epsilon/4 > \delta > 0$ such that whenever $\|z - y\| < \delta$ and $\hat{x}_z(\cdot), \hat{x}_y(\cdot)$ are solutions to (2.5) with initial conditions z, y resp., we have $\max_{t \in [0, 2T]} \|\hat{x}_z(t) - \hat{x}_y(t)\| < \epsilon/4$. Also pick $n_0 > 1$ such that $p(x) > x$ implies that $\bar{x}(s + \cdot) \in A^\delta$ and $\sup_{t \in [s, s+2T]} \|\bar{x}(t) - x^s(t)\| < \delta$. Pick $n_2 > n_1 \geq n_0$ such that $\|h(x) - h(y)\| \leq L\|x - y\|$. Let $kT \leq t(n_2) - t(n_1) < (k+1)T$ for some integer $k \geq 0$ and let $s(0) = t(n_1)$, $s(i) = s(0) + iT$ for $0 \leq i < n$, and $s(k) = t(n_2)$. Then for $0 \leq i < n$, $\sup_{t \in [s(i), s(i+1)]} \|\bar{x}(t) - x^{s(i)}(t)\| < \delta$. Pick \hat{x}_i , $0 \leq i \leq k$, in A such that $\hat{x}_1 = \tilde{x}_1$, $\hat{x}_k = \tilde{x}_2$, and for $0 < i < k$, I_n are in the δ -neighborhood of $\bar{x}(s(i))$. The sequence $(s(i), \hat{x}_i)$, $0 \leq i \leq k$, satisfies the definition of internal chain transitivity: If $x_i^*(\cdot)$ denotes the trajectories of (2.5) initiated at I_n for each i , we have

$$\begin{aligned} \|x_i^*(s(i+1) - s(i)) - \hat{x}_{i+1}\| \\ &\leq \|x_i^*(s(i+1) - s(i)) - x^{s(i)}(s(i+1))\| \\ &\quad + \|x^{s(i)}(s(i+1)) - \bar{x}(s(i+1))\| + \|\bar{x}(s(i+1)) - \hat{x}_{i+1}\| \\ &\leq \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{4} < \epsilon. \end{aligned}$$

This completes the proof. \square

In fact, internal chain transitivity itself implies invariance (see Proposition 5.3, p. 23, of Benaim, 1999).

2.2 Extensions and Variations

Some natural extensions of the foregoing are immediate:

1. *Partial convergence:* When the set $\{\sup_n \|x_n\| < \infty\}$ has a positive probability not necessarily equal to one, we still have

$$\sum_n a(n) E[\|M_{n+1}\|^2 | \mathcal{F}_n] < \infty$$

a.s. on this set. The martingale convergence theorem from Appendix C cited in the proof of Lemma 2.1 above then tells us that ζ_n converges a.s. on this set. Thus by the same arguments as before (which are *pathwise*), Theorem 2.1 continues to hold ‘a.s. on the set $\{\sup_n \|x_n\| < \infty\}$ ’.

2. *Random stepsizes:* While we took $\{a(n)\}$ to be deterministic in Sect. 2.1, the arguments would also go through if $\{a(n)\}$ are random and bounded, satisfy (A2) with probability one, and (A3) holds, with \mathcal{F}_n redefined as

$$\mathcal{F}_n = \sigma(x_m, M_m, a(m), m \leq n)$$

for $n \geq 0$. In fact, the boundedness condition for random $\{a(n)\}$ could be relaxed by imposing appropriate moment conditions. We shall not get into the details of this at any point, but this fact is worth keeping in mind as there are applications (e.g., in system identification) when $\{a(n)\}$ are random.

3. *Vanishing measurement error:* The arguments above go through even if we replace (2.1) by

$$x_{n+1} = x_n + a(n)[h(x_n) + M_{n+1} + \epsilon(n)], \quad n \geq 0, \quad (2.11)$$

where $(n+1)$ is a deterministic or random bounded sequence which is $o(1)$. This is because $(n+1)$ then contributes an additional error term in the proof of Lemma 2.1 which is also

asymptotically negligible and therefore does not affect the

conclusions. This important observation will be recalled often in what follows.

4.

Locally Lipschitz maps: Since the proofs above assumed a.s. boundedness of the iterates $p(x_n)$, the Lipschitz condition on h was applied on a (possibly random) bounded set. Hence it can be replaced by the milder requirement of local Lipschitz property.

These observations apply throughout the book wherever the arguments are pathwise, i.e., except in Chap. 9 and part of Chap. 10. The next corollary is often useful in narrowing down the potential candidates for A .

Suppose there exists a continuously differentiable $\Phi_{-t}(x) = \Phi_t^{-1}(x)$ such that $\lim_{\|w\| \rightarrow \infty} g(w) = \infty$, $H \stackrel{\text{def}}{=} \{x \in \mathcal{R}^d : V(x) = 0\} \neq \emptyset$, and $\langle h(x), \nabla V(x) \rangle \leq 0$ with equality if and only if $x \in H$. (Thus V is a ‘Lyapunov function.’)

Corollary 2.1 Almost surely, $p(x_n)$ converge to a, possibly sample path dependent, internally chain transitive set contained in H .

Proof The argument is sample pathwise for a sample path in the probability one set where assumption (A4) and Lemma 2.1 hold. Fix one such sample path and define random variables $C' = \sup_n \|x_n\|$ and $C = \sup_{\|x\| \leq C'} V(x)$. For any $0 < a \leq C$, let $H^a \stackrel{\text{def}}{=} \{x \in \mathcal{R}^d : V(x) < a\}$, and let \bar{H}^a denote the closure of H^a . Fix an η such that $0 < \eta < C/2$. Let

$$\Delta \stackrel{\text{def}}{=} \min_{x \in \bar{H}^C \setminus H^\eta} |\langle h(x), \nabla V(x) \rangle| > 0.$$

Let T be an upper bound for the time required for a solution of $p(x) - x$ to reach H^a , starting from any point in \bar{H}^C . Clearly, we may choose any $T > C/\Delta$. Let $\delta > 0$ be such that for $w \in \mathcal{R}^d$ and $0 < \eta < C/2$, we have $\langle h(x), \nabla V(x) \rangle \leq 0$. Such a choice of δ is possible by the uniform continuity of V on compact sets. By Lemma 2.1,

there is a t_0 such that for all $t \geq t_0$, $\sup_{s \in [t, t+T]} \|\bar{x}(s) - x^t(s)\| < \delta$.

Note that $\bar{x}(\cdot) \in \bar{H}^C$, and so for all $t \geq t_0$,

$|V(\bar{x}(t+T)) - V(x^t(t+T))| < \eta$. But $x^t(t+T) \in H^\eta$ and therefore $\bar{x}(t+T) \in H^{2\eta}$. Thus for all $t \geq t_0 + T$, $\bar{x}(t) \in H^{2\eta}$. Since η can be taken to be arbitrarily small, it follows that $\bar{x}(t) \rightarrow H$ as $t \rightarrow \infty$. \square

Alternatively, we can invoke the ‘LaSalle invariance principle’ (see Appendix B) in conjunction with Theorem 2. If we consider the iteration (2.11) with $(n+1)$ not necessarily approaching zero but bounded in norm by some $\kappa > 0$, then a similar argument leads to an analogous conclusion with convergence to H replaced by convergence to $p(x_n)$ for some constant C'' .

Existence of a Liapunov function for globally asymptotically stable equilibria and its local version for asymptotically stable equilibria follow from standard converse Liapunov theorems, see, e.g., Krasovskii (1963). For ‘Morse–Smale systems’, i.e., differential equations with only hyperbolic isolated equilibria or limit cycles as possible ω -limit sets along with transversability conditions imposed on their stable and unstable manifolds, a general converse Liapunov theorem is given in Meyer (1968). See Giesl and Hafstein (2015) for a survey of the computational aspects of Liapunov functions.

The following corollary is immediate:

Corollary 2.2 If the only internally chain transitive invariant sets for (2.5) are isolated equilibrium points, then $p(x_n)$ a.s. converges to a possibly sample path dependent equilibrium point.

More generally, a similar statement could be made for isolated internally chain transitive invariant sets, i.e., internally chain transitive invariant sets each of which is at a strictly positive distance from the rest. We shall refine Corollary 2.2 in Chap. 3 by narrowing the equilibria above to stable equilibria under additional conditions on the martingale noise. The next corollary is a variant of the so-called ‘Kushner–Clark lemma’ (Kushner & Clark, 1978) and also follows from the above discussion. Recall that ‘i.o.’ in probability theory stands for ‘infinitely often’.

Corollary 2.3 Let G be an open set containing an internally chain transitive invariant set D for (2.5), and suppose that σ^2 does not intersect any other internally chain transitive invariant set (except possibly subsets of D). Then under (A4), $x_n \rightarrow D$ a.s. on the set $\{x_n \in G \text{ i.o.}\} \stackrel{\text{def}}{=} \bigcap_n \bigcup_{m \geq n} \{x_m \in G\}$.

Proof We know that a.s., x_n converges to a compact, connected internally chain transitive invariant set. Let this set be D' . If D' does not intersect σ^2 , then by compactness of D' , there is an ϵ -neighborhood $N_\epsilon(D')$ of D' which does not intersect σ^2 . But since $x_n \rightarrow D'$, $\{x_n, n \geq 1\}$ for n large. This, however, leads to a contradiction if $\hat{x}_k = \tilde{x}_2$ i.o. Thus, if $x_n \in G$ i.o., D' has to intersect σ^2 . It follows that D' equals D or a subset thereof, and so $x_n \rightarrow D$ a.s. on the set $\forall t \geq 0 / \forall t \leq 0$. \square

In the more general set-up of Theorem 2.1, the next theorem is sometimes useful. (The statement and proof require some familiarity with weak (Prohorov) convergence of probability measures. See Appendix C for a brief account.)

Let $\mathcal{P}(\mathcal{R}^d)$ denote the space of probability measures on \mathcal{R}^k with Prohorov topology (also known as the topology of weak convergence, see, e.g., Chap. 2 of Borkar, 1995). Let $C_0(\mathcal{R}^d)$ denote the space of continuous functions on \mathcal{R}^k that vanish at infinity. Then the space \mathcal{M} of signed Borel measures on \mathcal{R}^k is isomorphic to the dual space $C_0^*(\mathcal{R}^d)$. The isomorphism is given by $\mu \mapsto \int(\cdot)d\mu$. (See, e.g., Folland (1984, Sect. 7.3).) It is easy to show that $\mathcal{P}(\mathcal{R}^d)$ consists of real measures μ which correspond to those elements $\hat{\mu}$ of $C_0^*(\mathcal{R}^d)$ that are:

- (i) nonnegative on nonnegative functions in $C_0(\mathcal{R}^d)$ (i.e., $f \geq 0$ for $f \in C_0(\mathcal{R}^d)$ implies that $\hat{\mu}(f) \geq 0$), and,
- (ii) for a constant function $a(n) \rightarrow 0$, $\bar{x}(t) \rightarrow A$.

Define (random) measures $\bar{x}(t)$, $t \geq 0$, on \mathcal{R}^k by

$$\int f d\nu(t) = \frac{1}{t} \int_0^t f(\bar{x}(s)) ds$$

for $f \in C_0(\mathcal{R}^d)$. These are called *empirical measures*. Since this integral is nonnegative for nonnegative f and furthermore,

$\nu(t)(\mathcal{R}^d) = \frac{1}{t} \int_0^t 1 ds = 1$, $\nu(t)$ is a probability measure on \mathcal{R}^k . By (A4), almost surely, the $\bar{x}(t)$, $t \geq 0$, are supported in a (possibly sample path dependent) compact subset of \mathcal{R}^k . By Prohorov's theorem (see Appendix C), they form a relatively compact subset of $\mathcal{P}(\mathcal{R}^d)$.

Theorem 2.2 Almost surely, every limit point x^* of $\nu(t)$ in $\mathcal{P}(\mathcal{R}^d)$ as $t \rightarrow \infty$ is invariant under (2.5).

Proof Recall the flow $\{y_n\}$ associated with (2.5). Fix some $s > 0$. Consider a sample path for which Lemma 2.1 applies, i.e., for which $M_{n+1} = f(x_n, \xi_{n+1}) - h(x_n)$ as $y \rightarrow \infty$. Let $f \in C_0(\mathcal{R}^d)$. Note that

$$\left| \frac{1}{t} \int_0^t f(\bar{x}(y)) dy - \frac{1}{t} \int_s^{t+s} f(\bar{x}(y)) dy \right| \rightarrow 0 \text{ as } t \rightarrow \infty.$$

Note also that the quantity on the left above is the same as

$$\left| \frac{1}{t} \int_0^t f(\bar{x}(y)) dy - \frac{1}{t} \int_0^t f(\bar{x}(y+s)) dy \right|.$$

Let $\epsilon > 0$. By uniform continuity of f , there is a T such that for $y \geq T$, $\|f(\bar{x}(y+s)) - f(\bar{x}(y))\| < \epsilon$. Now, if $t \geq T$,

$$\begin{aligned}
& \left| \frac{1}{t} \int_0^t f(\bar{x}(y+s)) dy - \frac{1}{t} \int_0^t f(x^y(y+s)) dy \right| \\
& \leq \frac{1}{t} \int_0^T |f(\bar{x}(y+s)) - f(x^y(y+s))| dy \\
& \quad + \frac{1}{t} \int_T^t |f(\bar{x}(y+s)) - f(x^y(y+s))| dy \\
& \leq \frac{T}{t} 2B + \frac{(t-T)}{t} \epsilon \leq 2\epsilon
\end{aligned}$$

for t large enough. Here B is a bound on the magnitude of $f \in C_0(\mathcal{R}^d)$. Thus

$$\left| \frac{1}{t} \int_0^t f(\bar{x}(y+s)) dy - \frac{1}{t} \int_0^t f(x^y(y+s)) dy \right| \rightarrow 0 \text{ as } t \rightarrow \infty.$$

But since

$$\left| \frac{1}{t} \int_0^t f(\bar{x}(y)) dy - \frac{1}{t} \int_0^t f(\bar{x}(y+s)) dy \right| \rightarrow 0$$

as $t \rightarrow \infty$, it follows that

$$\left| \frac{1}{t} \int_0^t f(\bar{x}(y)) dy - \frac{1}{t} \int_0^t f(x^y(y+s)) dy \right| \rightarrow 0$$

as $t \rightarrow \infty$. In turn, this implies that

$$\begin{aligned}
\left| \int f d\nu(t) - \int f \circ \Phi_s d\nu(t) \right| &= \left| \frac{1}{t} \int_0^t f(\bar{x}(y)) dy - \frac{1}{t} \int_0^t f \circ \Phi_s(\bar{x}(y)) dy \right| \\
&= \left| \frac{1}{t} \int_0^t f(\bar{x}(y)) dy - \frac{1}{t} \int_0^t f(x^y(y+s)) dy \right| \\
&\rightarrow 0 \text{ as } t \rightarrow \infty.
\end{aligned}$$

If x^* is a limit point of $\nu(t)$, there is a sequence $t_n \nearrow \infty$ such that $\nu(t_n) \rightarrow \nu^*$ weakly. Thus, $\int f d\nu(t_n) \rightarrow \int f d\nu^*$ and $\int f \circ \Phi_s d\nu(t_n) \rightarrow \int f \circ \Phi_s d\nu^*$. But

$\sum_n a(n)^2 = \sum_n (1/(n+1)^2) < \infty$ as $n \rightarrow \infty$. This tells us that $\int f d\nu^* = \int f \circ \Phi_s d\nu^*$. This holds for all f in a countable dense subset of $C_0(\mathcal{R}^d)$ and therefore for all $f \in C_0(\mathcal{R}^d)$. Hence x^* is invariant under Φ_s . As $s > 0$ was arbitrary, the claim follows for $z \geq 0$. The claim for $z \leq 0$ follows by using M_{n+1} in place of f in the above. \square

See Benaim and Schreiber (2000) for further results in this vein.

We conclude this section with some comments regarding stepsize selection. Our view of $\{a(n)\}$ as discrete time steps in the o.d.e. approximation already gives some intuition about their role. Thus large stepsizes will mean faster simulation of the o.d.e., but also larger errors due to discretization and noise (the latter is so because the stepsize $a(n)$ also multiplies the ‘noise’ M_{n+1} in the algorithm). Reducing the stepsizes would mean lower discretization errors and noise-induced errors and therefore a more graceful behavior of the algorithm, but at the expense of a slower speed of convergence. This is because one is taking a larger number of iterations to simulate any given time interval in ‘o.d.e. time’, equivalently, the ‘algorithmic time’ we introduced earlier. In the parlance of artificial intelligence, larger stepsizes aid better *exploration* of the solution space, while smaller stepsizes aid better *exploitation* of the local information available. The trade-off between them is a well-known rule of thumb in AI. Starting with a relatively large $a(n)$ and decreasing it slowly tries to strike a balance between the two. See Goldstein (1988) for some results on stepsize selection.

References

- Benaim, M. (1996). A dynamical system approach to stochastic approximation. *SIAM Journal on Control and Optimization*, 34(2), 437–472
- Benaim, M. (1999). Dynamics of stochastic approximation algorithms. In J. Azéma, M. Emery, M. Ledoux, & M. Yor (Eds.), *Le Séminaire de Probabilités*. Springer Lecture Notes in Mathematics (Vol. 1709, pp. 1–68). Springer Verlag
- Benaim M, Schreiber S (2000) Ergodic properties of weak asymptotic pseudotrajectories for semiflows. *Journal of Dynamics and Differential Equations* 12(3):579–598
[MathSciNet][Crossref][zbMATH]
- Borkar, V. S. (1995). *Probability theory: An advanced course*. Springer Verlag
- Folland, G. B. (1984). *Real analysis: Modern techniques and their applications*. Wiley
- Giesl P, Hafstein S (2015) Review of computational methods for Liapunov functions. *Discrete and Continuous Dynamical Systems Series B* 20(8):2291–2331
[MathSciNet][Crossref][zbMATH]
- Goldstein L (1988) On the choice of step-size in the Robbins-Monro procedure. *Statistics and Probability Letters* 6(5):299–303
[MathSciNet][Crossref][zbMATH]
- Krasovskii, N. N. (1963). *Stability of motion*. Stanford University Press
- Kushner, H. J., & Clark, D. (1978). *Stochastic approximation algorithms for constrained and unconstrained systems*. Springer Verlag
- Meyer, K. R. (1968). Energy functions for Morse-Smale systems. *American Journal of Mathematics*, 90(4), 1031–1040

Footnotes

¹ Following Benaim (1999), we require that internally chain transitive sets be compact.

3. Finite Time Bounds and Traps

Vivek S. Borkar¹✉

(1) Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

3.1 Estimating the Lock-In Probability

Recall the urn model of Chap. 1. When there are multiple isolated stable equilibria, it turns out that there can be a positive probability of convergence to one of these equilibria which is not, however, necessarily among the desired ones. This, we recall, was the explanation for several instances of adoption of one particular convention or technology as opposed to another. The idea is that after some initial randomness, the process becomes essentially ‘locked into’ the domain of attraction of a particular equilibrium, i.e., locked into a particular choice of technology or convention. With this picture in mind, we define the lock-in probability as the probability of convergence to an asymptotically stable attractor, given that the iterate is in a neighborhood thereof. Our aim here will be to get a lower bound on this probability and explore some of its consequences. Our treatment follows the approach of Borkar (2002, 2003).

Our setting is as follows. We consider the stochastic approximation on \mathcal{R}^k :

$$x_{n+1} = x_n + a(n)[h(x_n) + M_{n+1}], \quad (3.1)$$

under assumptions (A1), (A2) and (A3) of Chap. 2. We add to (A2) the requirement that

$$a(n) \leq \bar{c}a(m) \quad \forall n \geq m \quad (3.2)$$

for some $s > 0$. We shall not a priori make assumption (A4) of Chap. 2, which says that the sequence $p(x_n)$ generated by (3.1) is a.s. bounded. The a.s. boundedness of this sequence with high probability will be proved *as a consequence of our main result*. We have seen that recursion (3.1) can be considered to be a noisy discretization of the ordinary differential equation

$$\dot{x}(t) = h(x(t)). \quad (3.3)$$

Let $t_n \nearrow \infty$ be open, and let $C' = \sup_n \|x_n\|$ be such that $\dot{V} \stackrel{\text{def}}{=} \langle \nabla V, h \rangle : G \rightarrow \mathcal{R}$ is non-positive. We shall assume that $H \stackrel{\text{def}}{=} \{x : V(x) = 0\}$ is equal to the set $\{x : \dot{V}(x) = 0\}$ and is a compact subset of G . Thus, the function V is a Liapunov function. Then H is an asymptotically stable invariant set of the differential equation (3.3). Conversely, (local) asymptotic stability implies the existence of such a V by the converse Liapunov theorem—see, e.g., Krasovskii (1963). Let there be an open set B with compact closure such that $H \subset B \subset \bar{B} \subset G$. It follows from the LaSalle invariance principle (see Appendix B) that all internally chain transitive invariant subsets of I_n will be subsets of H .

In this setting, we shall derive an estimate for the probability that the sequence $p(x_n)$ is convergent to H , conditioned on the event that $x_{n_0} \in B$ for some n_0 sufficiently large. In the next section, we shall also derive a *sample complexity estimate* for the probability of the sequence being within a certain neighborhood of H after a certain length of time, again conditioned on the event that $x_{n_0} \in B$.

Let $H^a \stackrel{\text{def}}{=} \{x : V(x) < a\}$ have compact closure $\bar{H}^a = \{x : V(x) \leq a\}$. For $A \subset \mathcal{R}^d$, $\delta > 0$, let $N_\delta(A)$ denote the δ -neighborhood of A , i.e.,

$$N_\delta(A) \stackrel{\text{def}}{=} \{x : \inf_{y \in A} \|x - y\| < \delta\}.$$

Fix some $0 < \epsilon_1 < \epsilon$ and $\delta > 0$ such that

$$N_\delta(H^{\epsilon_1}) \subset H^\epsilon \subset N_\delta(H^\epsilon) \subset B.$$

As was argued in the first extension of Theorem 2.1, Chap. 2, in Sect. 2.2, if the sequence $p(x_n)$ generated by recursion (3.1) remains

a.s. bounded on a prescribed set of sample points, then it converges almost surely on this set to a (possibly sample path dependent) compact internally chain transitive invariant set of the o.d.e. (3.3). Therefore, if we can show that with high probability $p(x_n)$ remains inside the compact set I_n , then it follows that $p(x_n)$ converges to H with high probability. We shall in fact show that $p(\cdot)$, the piecewise linear and continuous curve obtained by linearly interpolating the points $p(x_n)$ as in Chap. 2 lies inside $0 < \eta < C/2$ with high probability from some time on. Let us define

$$T = \frac{[\max_{x \in \bar{B}} V(x)] - \epsilon_1}{\min_{x \in \bar{B} \setminus H^{\epsilon_1}} |\langle \nabla V(x), h(x) \rangle|}.$$

Then T is an upper bound for the time required for a solution of the o.d.e. (3.3) to reach the set H^{ϵ_1} , starting from an initial condition in I_n . Fix an $n_0 > 1$ ‘sufficiently large.’ (We shall be more specific later about how n_0 is to be chosen.) For $m \geq 1$, let

$n_m = \min\{n : t(n) \geq t(n_{m-1}) + T\}$. Define a sequence of times T_0, T_1, \dots by $T_m = t(n_m)$. For $m \geq 1$, let I_m be the interval $x_0 \in (0, 1)$, and let

$$\rho_m \stackrel{\text{def}}{=} \sup_{t \in I_m} \|\bar{x}(t) - x^{T_m}(t)\|,$$

where $x^{T_m}(\cdot)$ is, as in Chap. 2, the solution of the o.d.e. (3.3) on I_n with initial condition $x^{T_m}(T_m) = \bar{x}(T_m)$. We shall assume for sake of notational simplicity that $a(n) \leq 1 \forall n$, which implies that the length of I_m is between T and $t' > 0$.

Let us assume for the moment that $x_{n_0} \in B$, and that $x \in \mathcal{R}^d$ for all $m \geq 1$. Because of the way we defined T , it follows that $x^{T_0}(T_1) \in H^{\epsilon_1}$. Since $\rho_0 < \delta$ and $N_\delta(H^{\epsilon_1}) \subset H^\epsilon$, $\{(X_n, Y_n)\}$. Since H^ϵ is a positively invariant subset of I_n , it follows that $x^{T_1}(\cdot)$ lies in H^ϵ on I_1 , and that $x^{T_0}(T_1) \in H^{\epsilon_1}$. Hence $\{(X_n, Y_n)\}$. Continuing in this way it follows that for all $m \geq 1$, $x^{T_m}(\cdot)$ lies inside H^ϵ on I_m . Now using the fact that $x \in \mathcal{R}^d$ for all $m \geq 1$, it follows that $\bar{x}(t)$ is in

$0 < \eta < C/2$ for all $t \geq T_1$. As mentioned above, it now follows that $\bar{x}(t) \rightarrow H$ as $t \rightarrow \infty$. Therefore we have:

Lemma 3.1

$$P(\bar{x}(t) \rightarrow H | x_{n_0} \in B) \geq P(\rho_m < \delta \ \forall m \geq 0 | x_{n_0} \in B). \quad (3.4)$$

We let \mathcal{B}_m denote the event that $x_{n_0} \in B$ and $\rho_k < \delta$ for $k = 0, 1, \dots, m$. Recall that $\mathcal{F}_n \stackrel{\text{def}}{=} \sigma(x_0, M_1, \dots, M_n)$, $n \geq 1$. We then have that $\epsilon/4 > \delta > 0$. Note that

$$P(\rho_m < \delta \ \forall m \geq 0 | x_{n_0} \in B) = 1 - P(\rho_m \geq \delta \text{ for some } m \geq 0 | x_{n_0} \in B)$$

We have the following disjoint union:

$$\begin{aligned} & \{\rho_m \geq \delta \text{ for some } m \geq 0\} = \\ & \{\rho_0 \geq \delta\} \cup \{\rho_1 \geq \delta; \rho_0 < \delta\} \cup \{\rho_2 \geq \delta; \rho_0, \rho_1 < \delta\} \cup \dots \end{aligned}$$

Therefore,

$$\begin{aligned} & P(\rho_m \geq \delta \text{ for some } m \geq 0 | x_{n_0} \in B) \\ &= P(\rho_0 \geq \delta | x_{n_0} \in B) \\ & \quad + P(\rho_1 \geq \delta; \rho_0 < \delta | x_{n_0} \in B) \\ & \quad + P(\rho_2 \geq \delta; \rho_0, \rho_1 < \delta | x_{n_0} \in B) + \dots \\ &= P(\rho_0 \geq \delta | x_{n_0} \in B) \\ & \quad + P(\rho_1 \geq \delta | \rho_0 < \delta, x_{n_0} \in B) P(\rho_0 < \delta | x_{n_0} \in B) \\ & \quad + P(\rho_2 \geq \delta | \rho_0, \rho_1 < \delta, x_{n_0} \in B) P(\rho_0, \rho_1 < \delta | x_{n_0} \in B) + \dots \\ &\leq P(\rho_0 \geq \delta | x_{n_0} \in B) + P(\rho_1 \geq \delta | \mathcal{B}_0) + P(\rho_2 \geq \delta | \mathcal{B}_1) + \dots \end{aligned}$$

Thus, with $\mathcal{B}_{-1} \stackrel{\text{def}}{=} \{x_0 \in B\}$, we have:

Lemma 3.2

$$P(\rho_m < \delta \ \forall m \geq 0 | x_{n_0} \in B) \geq 1 - \sum_{m=0}^{\infty} P(\rho_m \geq \delta | \mathcal{B}_{m-1}).$$

The bound derived in Lemma 3.2 involves the term $\|h(x^s(t))\| \leq C_T$. We shall now derive an upper bound on this term. Recall that \mathcal{B}_{m-1} denotes the event that $x_{n_0} \in B$ and $\rho_k < \delta$ for $k = 0, 1, \dots, m-1$. This implies that $\bar{x}(T_m) \in B$. Let C be a bound on $T_m = t(n_m)$, where Φ_t is the time- t flow map for the o.d.e. (3.3), for $s \leq t \leq s+T$ and $x \in \bar{B}$. By the arguments of Lemma 2.1 of Chap. 2 (see the final bound in its proof), it follows that

$$\rho_m \leq Ca(n_m) + K_T(CLb(n_m) + \max_{n_m \leq j \leq n_{m+1}} \|\zeta_j - \zeta_{n_m}\|),$$

where \mathcal{B}_m is a constant that depends only on T, L is the Lipschitz constant for h , $b(n) \stackrel{\text{def}}{=} \sum_{k \geq n} a(k)^2$, and $\zeta_k = \sum_{i=0}^{k-1} a(i)M_{i+1}$. The first term is the ‘round off error’ between the continuous variable t and the discrete set (ζ_n, \mathcal{F}_n) . Since $a(n_m) \leq ca(n_0)$ and $b(n_m) \leq c^2 b(n_0)$, it follows that

$$\rho_m \leq (Ca(n_0) + cK_T CLb(n_0))c + K_T \max_{n_m \leq j \leq n_{m+1}} \|\zeta_j - \zeta_{n_m}\|.$$

This implies that if n_0 is chosen so that

$(Ca(n_0) + cK_T CLb(n_0))c < \delta/2$, then $x \in \mathcal{R}^d$ implies that

$$\sum_n a(n) E[\|M_{n+1}\|^2 | \mathcal{F}_n] < \infty$$

We state this as a lemma:

Lemma 3.3 If n_0 is chosen so that

$$x_{n+1} = x_n + a(n)[h(x_n) + M_{n+1}], \quad (3.5)$$

then

$$P(\rho_m \geq \delta | \mathcal{B}_{m-1}) \leq P \left(\max_{n_m \leq j \leq n_{m+1}} \|\zeta_j - \zeta_{n_m}\| > \frac{\delta}{2K_T} \mid \mathcal{B}_{m-1} \right). \quad (3.6)$$

We shall now find a bound for the expression displayed on the right-hand side of the above inequality. We shall give two methods for bounding this quantity. The first one uses Burkholder’s inequality, and the second uses a concentration inequality for martingales. As we shall

see, the second method gives a better bound, but under a stronger assumption.

Burkholder's inequality (see Appendix C) implies that if $a(n) < 1$, $n \geq 0$ is a (real-valued) zero mean martingale, and if $\bar{M}_n \stackrel{\text{def}}{=} X_n - X_{n-1}$ is the corresponding martingale difference sequence, then there is a constant Φ_s such that

$$E[(\max_{0 \leq j \leq n} |X_j|)^2] \leq C_1^2 E \left[\sum_{i=1}^n \bar{M}_i^2 \right].$$

We shall use the conditional version of this inequality: If $\hat{x}_1 = \tilde{x}_1$ is a sub- σ -field, then

$$E[(\max_{0 \leq j \leq n} |X_j|)^2 | \mathcal{G}] \leq C_1^2 E \left[\sum_{i=1}^n \bar{M}_i^2 | \mathcal{G} \right]$$

almost surely. In our context, $\{\zeta_j - \zeta_{n_m}, j \geq n_m\}$ is an \mathcal{R}^k -valued martingale with respect to the filtration $\{\mathcal{F}_j\}$. We shall apply Burkholder's inequality to each component.

We shall first prove some useful lemmas. Let K be the constant in assumption (A3) of Chap. 2. For an \mathcal{R}^k -valued random variable Z , let $p(x_n)$ denote $E[\|Z\|^2 | \mathcal{B}_{m-1}]^{1/2}(\omega)$, where $\omega \in \mathcal{B}_{m-1}$. Note that $p(x_n)$ satisfies the properties of a norm 'almost surely.'

Lemma 3.4 For $k = 0, 1, \dots, m$ and a.s. $\omega \in \mathcal{B}_{m-1}$,

$$(\|M_j\|^*)^2 \leq K(1 + (\|x_{j-1}\|^*)^2).$$

Proof Note that $H = \{x : p(x) = x\}$ for $x \in \mathcal{R}^d$. Then a.s.,

$$\begin{aligned} (\|M_j\|^*)^2 &= E[\|M_j\|^2 | \mathcal{B}_{m-1}](\omega) \\ &= E[E[\|M_j\|^2 | \mathcal{F}_{j-1}] | \mathcal{B}_{m-1}](\omega) \\ &\leq E[K(1 + \|x_{j-1}\|^2) | \mathcal{B}_{m-1}](\omega) \quad (\text{by Assumption (A3) of Chap. 2}) \\ &= K(1 + E[\|x_{j-1}\|^2 | \mathcal{B}_{m-1}](\omega)) \\ &= K(1 + (\|x_{j-1}\|^*)^2). \end{aligned}$$

The claim follows. \square

Lemma 3.5 There is a constant \bar{K}_T such that for $n_m \leq j \leq n_{m+1}$,

$$\|x_j\|^* \leq \bar{K}_T \text{ a.s.}$$

Proof Consider the recursion

$$x_{j+1} = x_j + a(j)[h(x_j) + M_{j+1}], \quad j \geq n_m.$$

As we saw in Chap. 2, the Lipschitz property of h implies a linear growth condition on h , i.e., $\|h(x)\| \leq K'(1 + \|x\|)$. Taking the Euclidean norm on both sides of the above equation and using the triangle inequality along with this linear growth condition leads to

$$\|x_{j+1}\| \leq \|x_j\|(1 + a(j)K') + a(j)K' + a(j)\|M_{j+1}\|.$$

Therefore, using the triangle inequality for $p(x_n)$, we have a.s. for $x \in \mathcal{R}^d$,

$$\begin{aligned} \|x_{j+1}\|^* &\leq \|x_j\|^*(1 + a(j)K') + a(j)K' + a(j)\|M_{j+1}\|^* \\ &\leq \|x_j\|^*(1 + a(j)K') + a(j)K' + a(j)\sqrt{K}(1 + (\|x_j\|^*)^2)^{1/2} \\ &\quad \text{(by the previous lemma)} \\ &\leq \|x_j\|^*(1 + a(j)K') + a(j)K' + a(j)\sqrt{K}(1 + \|x_j\|^*) \\ &= \|x_j\|^*(1 + a(j)K_1) + a(j)K_1, \\ &\quad \text{where } K_1 = K' + \sqrt{K}. \end{aligned}$$

Applying this inequality repeatedly and using the fact that (ζ_n, \mathcal{F}_n) , $n \geq 1$, and the fact that if $j < n_{m+1}$, $a(n_m) + a(n_m + 1) + \dots + a(j) \leq t(n_{m+1}) - t(n_m) \leq T + 1$, we get

$$\|x_{j+1}\|^* \leq \|x_{n_m}\|^* e^{K_1(T+1)} + \hat{K}_1 e^{K_1(T+1)} \text{ for } n_m \leq j < n_{m+1}$$

with a suitable $\hat{K}_1 > 0$. For $\omega \in \mathcal{B}_{m-1}$, $s(k) = t(n_2)$. Since I_n is compact, there is a deterministic constant \mathcal{F}_n such that $\{x_n, n \geq 1\}$. Thus $\|x_{n_m}\|^* \leq K_2$. This implies that for $n_m - 1 \leq j < n_{m+1}$,

$$\|x_{j+1}\|^* \leq \bar{K}_T \stackrel{\text{def}}{=} e^{K_1(T+1)}[K_2 + \hat{K}_1].$$

In other words, for $n_m \leq j \leq n_{m+1}$, $\|x_j\|^* \leq \bar{K}_T$. \square

Lemma 3.6 For $k = 0, 1, \dots, m$,

$$(\|M_j\|^*)^2 \leq K(1 + \bar{K}_T^2).$$

Proof This follows by combining the results of Lemmas 3.4 and 3.5. \square

We now continue to upper bound the right-hand side of (3.6) in Lemma 3.3. For a.s. $\omega \in \mathcal{B}_{m-1}$, we have

$$\begin{aligned} & P \left(\max_{n_m \leq j \leq n_{m+1}} \|\zeta_j - \zeta_{n_m}\| > \frac{\delta}{2K_T} \mid \mathcal{B}_{m-1} \right) \\ &= P \left(\left(\max_{n_m \leq j \leq n_{m+1}} \|\zeta_j - \zeta_{n_m}\| \right)^2 > \frac{\delta^2}{4K_T^2} \mid \mathcal{B}_{m-1} \right) \\ &\leq \frac{4K_T^2}{\delta^2} E \left[\left(\max_{n_m \leq j \leq n_{m+1}} \|\zeta_j - \zeta_{n_m}\| \right)^2 \mid \mathcal{B}_{m-1} \right], \end{aligned}$$

where the inequality follows from the conditional Chebyshev inequality. We shall now use Burkholder's inequality to get the desired bound. Let ζ_n^i denote the i th component of ζ_n . For a.s. $\omega \in \mathcal{B}_{m-1}$,

$$\begin{aligned}
& E \left[\left(\max_{n_m \leq j \leq n_{m+1}} \|\zeta_j - \zeta_{n_m}\| \right)^2 \mid \mathcal{B}_{m-1} \right] \\
&= E \left[\max_{n_m \leq j \leq n_{m+1}} \sum_{i=1}^d (\zeta_j^i - \zeta_{n_m}^i)^2 \mid \mathcal{B}_{m-1} \right] \\
&\leq \sum_{i=1}^d E \left[\left(\max_{n_m \leq j \leq n_{m+1}} |\zeta_j^i - \zeta_{n_m}^i| \right)^2 \mid \mathcal{B}_{m-1} \right] \\
&\leq \sum_{i=1}^d C_1^2 E \left[\sum_{j=n_m+1}^{n_{m+1}} a(j-1)^2 (M_j^i)^2 \mid \mathcal{B}_{m-1} \right] \\
&= C_1^2 E \left[\sum_{j=n_m+1}^{n_{m+1}} a(j-1)^2 \|M_j\|^2 \mid \mathcal{B}_{m-1} \right] \\
&= C_1^2 \sum_{j=n_m+1}^{n_{m+1}} a(j-1)^2 (\|M_j\|^*)^2 \\
&\leq C_1^2 [a(n_m)^2 + \dots + a(n_{m+1}-1)^2] K (1 + \bar{K}_T^2) \\
&= C_1^2 (b(n_m) - b(n_{m+1})) K (1 + \bar{K}_T^2).
\end{aligned}$$

Combining the foregoing, we obtain the following lemma:

Lemma 3.7 For $\tilde{K} \stackrel{\text{def}}{=} 4C_1^2 K (1 + \bar{K}_T^2) K_T^2 > 0$,

$$P \left(\max_{n_m \leq j \leq n_{m+1}} \|\zeta_j - \zeta_{n_m}\| > \frac{\delta}{2K_T} \mid \mathcal{B}_{m-1} \right) \leq \frac{\tilde{K}}{\delta^2} (b(n_m) - b(n_{m+1})).$$

Thus we have:

Theorem 3.1 For some constant $\tilde{K} > 0$,

$$P(\bar{x}(t) \rightarrow H \mid x_{n_0} \in B) \geq 1 - \frac{\tilde{K}}{\delta^2} b(n_0).$$

Proof Note that

$$\begin{aligned}
& P(\rho_m < \delta \ \forall m \geq 0 | x_{n_0} \in B) \\
& \geq 1 - \sum_{m=0}^{\infty} P(\rho_m \geq \delta | \mathcal{B}_{m-1}) \text{ by Lemma 3.2} \\
& \geq 1 - \sum_{m=0}^{\infty} P \left(\max_{n_m \leq j \leq n_{m+1}} \|\zeta_j - \zeta_{n_m}\| > \frac{\delta}{2K_T} \mid \mathcal{B}_{m-1} \right) \text{ by Lemma 3.3} \\
& \geq 1 - \sum_{m=0}^{\infty} \frac{\tilde{K}}{\delta^2} (b(n_m) - b(n_{m+1})) \text{ by Lemma 3.7} \\
& = 1 - \frac{\tilde{K}}{\delta^2} b(n_0).
\end{aligned}$$

The claim now follows from Lemma 3.1. \square

A key assumption for the bound derived in Theorem 3.1 was assumption (A3) of Chap. 2, by which $E[\|M_j\|^2 | \mathcal{F}_{j-1}] \leq K(1 + \|x_{j-1}\|^2)$, for $j \geq 1$. Now we shall make the more restrictive assumption:

$$\|M_j\| \leq K_0(1 + \|x_{j-1}\|). \quad (3.7)$$

This condition holds, e.g., in the reinforcement learning applications discussed in Chap. 12. Under (3.7), it will be possible for us to derive a sharper bound.

We first prove a lemma about the boundedness of stochastic approximation iterates under this assumption:

Lemma 3.8 There is a constant \mathcal{F}_n such that for $x_0 = 0$,

$$\|\bar{x}(t)\| \leq K_3(1 + \|\bar{x}(T_m)\|).$$

Proof The stochastic approximation recursion is

$$x_{n+1} = x_n + a(n)[h(x_n) + M_{n+1}],$$

Therefore, assuming that $\|M_j\| \leq K_0(1 + \|x_{j-1}\|)$ and using the linear growth property of h ,

$$\begin{aligned}\|x_{n+1}\| &\leq \|x_n\| + a(n)K(1 + \|x_n\|) + a(n)K_0(1 + \|x_n\|) \\ &= \|x_n\|(1 + a(n)K_4) + a(n)K_4, \text{ where } K_4 \stackrel{\text{def}}{=} K + K_0.\end{aligned}$$

Arguing as in Lemma 3.5, we conclude that for $k = 0, 1, \dots, m$,

$$\|x_{j+1}\| \leq [\|x_{n_m}\| + K_4]e^{K_4(T+1)}.$$

Thus for $n_m \leq j \leq n_{m+1}$,

$$\|x_j\| \leq e^{K_4(T+1)}[\|x_{n_m}\| + K_4] \leq K_3(1 + \|x_{n_m}\|)$$

for some constant \mathcal{F}_n . The lemma follows. \square

Note that for $n_m \leq k < n_{m+1}$,

$$\begin{aligned}\|\zeta_{k+1} - \zeta_k\| &= \|a(k)M_{k+1}\| \\ &\leq a(k)K_0(1 + \|x_k\|) \\ &\leq a(k)K_0(1 + K_3(1 + \|x_{n_m}\|)) \\ &\leq a(k)K_0[1 + K_3(K_2 + 1)] \\ &= a(k)\bar{K},\end{aligned}\tag{3.8}$$

where $K_2 \stackrel{\text{def}}{=} \max_{x \in \bar{B}} \|x\|$ and $\bar{K} \stackrel{\text{def}}{=} K_0[1 + K_3(K_2 + 1)]$. Thus one may use the following concentration inequality for martingales due to McDiarmid (cf. Appendix C): Consider the filtration

$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \dots \subset \mathcal{F}$. Let $H \subset B \subset \bar{B} \subset G$ be a (scalar) martingale with respect to this filtration, with $Y_1 = S_1$, $Y_k = S_k - S_{k-1}$ ($k \geq 2$) the corresponding martingale difference sequence. Let $p(x) > x$. Then

$$P\left(\max_{1 \leq k \leq n} |S_k| \geq t\right) \leq 4e^{\frac{-2t^2}{\sum_{k \leq n} b_k^2}}.$$

If $\mathcal{B} \in \mathcal{F}_0$, we can state a conditional version of this inequality as follows:

$$P\left(\max_{1 \leq k \leq n} |S_k| \geq t | \mathcal{B}\right) \leq 4e^{\frac{-2t^2}{\sum_{k \leq n} b_k^2}}.$$

Let $N_\delta(A)$ denote the max-norm on \mathcal{R}^k , i.e., $\|x\|_\infty \stackrel{\text{def}}{=} \max_i |x_i|$. Note that for $s_n \uparrow \infty$, $\|v\|_\infty \leq \|v\| \leq \sqrt{d}\|v\|_\infty$. Thus $x_{n_0} \in B$ implies that $\|v\|_\infty \geq c/\sqrt{d}$.

Since $\int f d\nu(t_n) \rightarrow \int f d\nu^*$ is a martingale with respect to $\{\mathcal{F}_j\}_{n_m \leq j \leq n_{m+1}}$ and $\mathcal{B}_{m-1} \in \mathcal{F}_{n_m}$, we have, by the inequality (3.8),

$$\begin{aligned} & P \left(\max_{n_m \leq j \leq n_{m+1}} \|\zeta_j - \zeta_{n_m}\| > \frac{\delta}{2K_T} \mid \mathcal{B}_{m-1} \right) \\ & \leq P \left(\max_{n_m \leq j \leq n_{m+1}} \|\zeta_j - \zeta_{n_m}\|_\infty > \frac{\delta}{2K_T \sqrt{d}} \mid \mathcal{B}_{m-1} \right) \\ & = P \left(\max_{n_m \leq j \leq n_{m+1}} \max_{1 \leq i \leq d} |\zeta_j^i - \zeta_{n_m}^i| > \frac{\delta}{2K_T \sqrt{d}} \mid \mathcal{B}_{m-1} \right) \\ & = P \left(\max_{1 \leq i \leq d} \max_{n_m \leq j \leq n_{m+1}} |\zeta_j^i - \zeta_{n_m}^i| > \frac{\delta}{2K_T \sqrt{d}} \mid \mathcal{B}_{m-1} \right) \\ & \leq \sum_{i=1}^d P \left(\max_{n_m \leq j \leq n_{m+1}} |\zeta_j^i - \zeta_{n_m}^i| > \frac{\delta}{2K_T \sqrt{d}} \mid \mathcal{B}_{m-1} \right) \\ & \leq \sum_{i=1}^d 4 \exp \left\{ -\frac{\delta^2 / (4K_T^2 d)}{4[a(n_m)^2 + \dots + a(n_{m+1}-1)^2] \bar{K}^2} \right\} \\ & \leq 4d \exp \left\{ -\frac{\delta^2}{16K_T^2 \bar{K}^2 d [b(n_m) - b(n_{m+1})]} \right\} \\ & = 4d e^{-\hat{C}\delta^2/(d[b(n_m) - b(n_{m+1})])}, \text{ where } \hat{C} = 1/(16K_T^2 \bar{K}^2). \end{aligned}$$

This gives us:

Lemma 3.9 There is a constant $N_\delta(A)$ such that

$$P \left(\max_{n_m \leq j \leq n_{m+1}} \|\zeta_j - \zeta_{n_m}\| > \frac{\delta}{2K_T} \mid \mathcal{B}_{m-1} \right) \leq 4d e^{-\hat{C}\delta^2/(d[b(n_m) - b(n_{m+1})])}.$$

For n_0 sufficiently large so that (3.5) holds and

$$b(n_0) < \hat{C}\delta^2/d, \quad (3.9)$$

the following bound holds:

Theorem 3.2

$$P(\rho_m < \delta \ \forall m \geq 0 | x_{n_0} \in B) \geq 1 - 4d e^{-\frac{\hat{C}\delta^2}{db(n_0)}} = 1 - o(b(n_0)).$$

Proof Note that Lemmas 3.2 and 3.3 continue to apply. Then

$$\begin{aligned} P(\rho_m < \delta \ \forall m \geq 0 | x_{n_0} \in B) \\ &\geq 1 - \sum_{m=0}^{\infty} P(\rho_m \geq \delta | \mathcal{B}_{m-1}) \text{ by Lemma 3.2} \\ &\geq 1 - \sum_{m=0}^{\infty} P\left(\max_{n_m \leq j \leq n_{m+1}} \|\zeta_j - \zeta_{n_m}\| > \frac{\delta}{2K_T} \mid \mathcal{B}_{m-1}\right) \text{ by Lemma 3.3} \\ &\geq 1 - \sum_{m=0}^{\infty} 4d e^{-\hat{C}\delta^2/(d[b(n_m) - b(n_{m+1})])} \text{ by Lemma 3.9} \end{aligned}$$

Note that for $C' > 0$, $e^{-C'/x}/x \rightarrow 0$ as $x \rightarrow 0$ and increases with x for $0 < x < C'$. Therefore for sufficiently large n_0 , specifically for n_0 such that (3.9) holds, we have

$$\frac{e^{-\hat{C}\delta^2/(d[b(n_m) - b(n_{m+1})])}}{[b(n_m) - b(n_{m+1})]} \leq \frac{e^{-\hat{C}\delta^2/(db(n_m))}}{b(n_m)} \leq \frac{e^{-\hat{C}\delta^2/(db(n_0))}}{b(n_0)}.$$

Hence

$$\begin{aligned} e^{-\hat{C}\delta^2/(d[b(n_m) - b(n_{m+1})])} &= [b(n_m) - b(n_{m+1})] \cdot \frac{e^{-\hat{C}\delta^2/(d[b(n_m) - b(n_{m+1})])}}{[b(n_m) - b(n_{m+1})]} \\ &\leq [b(n_m) - b(n_{m+1})] \cdot \frac{e^{-\hat{C}\delta^2/(db(n_0))}}{b(n_0)}. \end{aligned}$$

So,

$$\begin{aligned}
\sum_{m=0}^{\infty} e^{-\hat{C}\delta^2/(d[b(n_m) - b(n_{m+1})])} &\leq \sum_{m=0}^{\infty} [b(n_m) - b(n_{m+1})] \cdot \frac{e^{-\hat{C}\delta^2/(db(n_0))}}{b(n_0)} \\
&= \frac{e^{-\hat{C}\delta^2/(db(n_0))}}{b(n_0)} \sum_{m=0}^{\infty} [b(n_m) - b(n_{m+1})] \\
&= e^{-\hat{C}\delta^2/(db(n_0))}.
\end{aligned}$$

The claim follows. \square

Lemma 3.1 coupled with Theorems 3.1 and 3.2 now enables us to derive bounds on the lock-in probability. We state these bounds in the following corollary:

Corollary 3.1 In the setting described at the beginning of this section, if n_0 is chosen so that (3.5) holds, then there is a constant \tilde{K} such that

$$P(\bar{x}(t) \rightarrow H | x_{n_0} \in B) \geq 1 - \frac{\tilde{K}}{\delta^2} b(n_0).$$

If we make the additional assumption that $\|M_j\| \leq K_0(1 + \|x_{j-1}\|)$ for $j \geq 1$ and (3.9) holds, then there is a constant \hat{C} such that the following tighter bound for the lock-in probability holds:

$$\begin{aligned}
P(\bar{x}(t) \rightarrow H | x_{n_0} \in B) &\geq 1 - 2d e^{-\frac{\hat{C}\delta^2}{db(n_0)}} \\
&= 1 - o(b(n_0)).
\end{aligned}$$

In conclusion, we observe that the stronger assumption on the martingale difference sequence $\{M_n\}$ was made necessary by our use of McDiarmid's concentration inequality for martingales, which requires the associated martingale difference sequence to be bounded by deterministic bounds. More recent work on concentration inequalities for martingales can be used to relax this condition (see, e.g., Li, 2003; Liu & Watbled, 2009). We sketch some results in this vein later in Sect. 3.3.

3.2 Sample Complexity

We continue in the setting described at the beginning of the previous section. Our goal is to derive a *sample complexity estimate*, by which we mean an estimate of the probability that $\bar{x}(t)$ is within a certain neighborhood of H after the lapse of a certain amount of time, conditioned on the event that $x_{n_0} \in B$ for some fixed n_0 sufficiently large.

We begin by fixing some $\epsilon > 0$ such that $\bar{H}^\epsilon (= \{x : V(x) \leq \epsilon\}) \subset B$. Fix some $T > 0$, and let

$$\Delta < \min_{x \in \bar{B} \setminus H^\epsilon} [V(x) - V(\Phi_T(x))], \quad (3.10)$$

where \mathcal{F}_n is the time- T flow map of the o.d.e. (3.3) (see Appendix B). Note that $x \rightarrow 0$. We remark that the arguments that follow do not require V to be differentiable, as was assumed earlier. It is enough to assume that V is continuous and that $V(x(t))$ monotonically decreases with t along any trajectory of (3.3) in $\bar{x}(s(i))$. One such situation with non-differentiable V will be encountered in Chap. 12, in the context of reinforcement learning.

We fix an $n_0 > 1$ sufficiently large. We shall specify later how large n_0 needs to be. Let n_m, T_m, I_m, n_m and \mathcal{B}_m be defined as in the previous section. Fix a $\delta > 0$ such that $0 < \eta < C/2$ and such that for all $h(x) + \xi$ with $0 < \eta < C/2$, $\|V(x) - V(y)\| < \Delta/2$.

Let us assume that $x_{n_0} \in B$, and that $x \in \mathcal{R}^d$ for all $m \geq 1$. If $x_{n_0} \in B \setminus H^\epsilon$, we have that $V(x^{T_0}(T_1)) \leq V(\bar{x}(T_0)) - \Delta$ by virtue of (3.10). Since $\|x^{T_0}(T_1) - \bar{x}(T_1)\| < \delta$, it follows that $V(\bar{x}(T_1)) \leq V(\bar{x}(T_0)) - \Delta/2$. If $C' = \sup_n \|x_n\|$, the same argument can be repeated to give $V(\bar{x}(T_1)) \leq V(\bar{x}(T_0)) - \Delta/2$. Since $V(\bar{x}(T_m))$ cannot decrease at this rate indefinitely, it follows that $\bar{x}(T_{m_0}) \in H^\epsilon$ for some w_n . In fact, if

$$\tau \stackrel{\text{def}}{=} \frac{[\max_{x \in \bar{B}} V(x)] - \epsilon}{\Delta/2} \cdot (T + 1),$$

then

$$T_{m_0} \leq T_0 + \tau. \quad (3.11)$$

Thus $x^{T_{m_0}}(t) \in H^\epsilon$ on $I_{m_0} = [T_{m_0}, T_{m_0+1}]$. Therefore $\bar{x}(T_{m_0+1}) \in H^{\epsilon+\Delta/2}$. This gives rise to two possibilities: either $\bar{x}(T_{m_0+1}) \in H^\epsilon$ or $\bar{x}(T_{m_0+1}) \in H^{\epsilon+\Delta/2} \setminus H^\epsilon$. In the former case, $x^{T_{m_0+1}}(t) \in H^\epsilon$ on $p(x_n)$, and $\bar{x}(T_{m_0+1}) \in H^{\epsilon+\Delta/2}$. In the latter case, $\bar{x}(T_{m_0+1}) \in H^{\epsilon+\Delta/2}$ on $p(x_n)$, and since we know that $x^{T_{m_0+1}}(T_{m_0+1}) \in N_\delta(H^\epsilon) \subset B$, we have a decrease of the Liapunov function along the o.d.e. trajectory by at least \square on $p(x_n)$, yielding $x^{T_{m_0+1}}(T_{m_0+2}) \in H^{\epsilon-\Delta/2} \subset H^\epsilon$ yielding once again $x^{T_{m_0+1}}(T_{m_0+2}) \in H^\epsilon$, and therefore $\bar{x}(T_{m_0+1}) \in H^{\epsilon+\Delta/2}$. Thus $\bar{x}(T_{m_0+1}) \in H^{\epsilon+\Delta/2}$ implies that $\bar{x}(T_{m_0+1}) \in H^{\epsilon+\Delta/2}$ on $p(x_n)$ and $\bar{x}(T_{m_0+1}) \in H^{\epsilon+\Delta/2}$. This argument can be repeated. We have thus shown that if $\bar{x}(T_{m_0}) \in H^\epsilon$, then $x^{T_{m_0+k}}(t) \in H^{\epsilon+\Delta/2}$ on I_{m_0+k} for all $k \geq 0$, which in turn implies that $\bar{x}(t) \in N_\delta(H^{\epsilon+\Delta/2})$ for all (ζ_n, \mathcal{F}_n) . We thus conclude that if $x_{n_0} \in B$ and $x \in \mathcal{R}^d$ for all $m \geq 1$, then $\bar{x}(t) \in N_\delta(H^{\epsilon+\Delta/2})$ for all (ζ_n, \mathcal{F}_n) , and thus for all $t \geq t(n_0) + \tau$. This gives us:

Lemma 3.10

$$\begin{aligned} P\left(\bar{x}(t) \in N_\delta(H^{\epsilon+\Delta/2}) \quad \forall t \geq t(n_0) + \tau \mid x_{n_0} \in B\right) \\ \geq P(\rho_m < \delta \quad \forall m \geq 0 \mid x_{n_0} \in B) \end{aligned}$$

Lemma 3.10 coupled with Theorems 3.1 and 3.2 now allows us to derive the following sample complexity estimate.

Corollary 3.2 In the setting described at the beginning of Sect. 4.1, if n_0 is chosen so that (3.5) holds, then there is a constant \tilde{K} such that

$$P\left(\bar{x}(t) \in N_\delta(H^{\epsilon+\Delta/2}) \quad \forall t \geq t(n_0) + \tau \mid x_{n_0} \in B\right) \geq 1 - \frac{\tilde{K}}{\delta^2} b(n_0).$$

If we make the additional assumption that $\|M_j\| \leq K_0(1 + \|x_{j-1}\|)$ for $j \geq 1$ and (3.9) holds, then there is a constant \hat{C} such that the following tighter bound holds:

$$\begin{aligned}
P\left(\bar{x}(t) \in N_\delta(H^{\epsilon+\Delta/2}) \mid t \geq t(n_0) + \tau \mid x_{n_0} \in B\right) &\geq 1 - 4d e^{-\frac{\hat{C}\delta^2}{db(n_0)}} \\
&= 1 - o(b(n_0)).
\end{aligned}$$

The corollary clearly gives a sample complexity type result in the sense described at the beginning of this section. There is, however, a subtle difference from the traditional sample complexity results. What we present here is not the number of samples needed to get within a prescribed accuracy with a prescribed probability starting from time zero, but starting from time n_0 , and the bound depends crucially on the position at time n_0 via its dependence on the set B that we are able to choose.

As an example, consider the situation when $h(x) = g(x) - x$, where $p(\cdot)$ is a contraction, so that $\|g(x) - g(y)\| < \alpha \|x - y\|$ for some $x(t) \equiv x^*$. Let x^* be the unique fixed point of $p(\cdot)$, guaranteed by the contraction mapping theorem (see Appendix A). Straightforward calculation shows that $V(x) = \|x - x^*\|$ satisfies our requirements: Let

$$\dot{X}(x, t) = h(X(x, t)), \quad X(x, 0) = x.$$

We have

$$(X(x, t) - x^*) = (x - x^*) + \int_0^t (g(X(x, s)) - x^*) ds - \int_0^t (X(x, s) - x^*) ds,$$

leading to

$$(X(x, t) - x^*) = e^{-t}(x - x^*) + \int_0^t e^{-(t-s)}(g(X(x, s)) - x^*) ds.$$

Taking norms and using the contraction property,

$$\begin{aligned}
\|X(x, t) - x^*\| &\leq e^{-t} \|x - x^*\| + \int_0^t e^{-(t-s)} \|g(X(x, s)) - x^*\| ds \\
&\leq e^{-t} \|x - x^*\| + \alpha \int_0^t e^{-(t-s)} \|X(x, s) - x^*\| ds.
\end{aligned}$$

That is,

$$e^t \|X(x, t) - x^*\| \leq \|x - x^*\| + \alpha \int_0^t e^s \|X(x, s) - x^*\| ds.$$

By the Gronwall inequality,

$$\|X(x, t) - x^*\| \leq e^{-(1-\alpha)t} \|x - x^*\|.$$

For $\epsilon, T > 0$ as above, one may choose $\Delta = \epsilon(1 - e^{-(1-\alpha)T})$. We may take $\delta = \frac{\Delta}{2} \leq \frac{\epsilon}{2}$. Let the iteration be at a point η at time n_0 large enough that (3.5) and (3.9) hold. Let $\|\bar{x} - x^*\| = b$ (say). By (3.11), one then needs

$$N_0 \stackrel{\text{def}}{=} \min \left\{ n : \sum_{i=n_0+1}^n a(i) \geq \frac{2(T+1)b}{\epsilon(1 - e^{-(1-\alpha)T})} \right\}$$

more iterates to get within 2ϵ of x^* , with a probability exceeding

$$1 - 4de^{-\frac{c\epsilon^2}{b(n_0)}} = 1 - o(b(n_0))$$

for a suitable constant c that depends on T , among other things. Note that $T > 0$ is a free parameter affecting both the expression for I_m and that for the probability and can be chosen suitably.

3.3 Extensions

In Corollary 3.2, we considered two settings. The first assumed (A3) of Chap. 2, by which $E[\|M_j\|^2 | \mathcal{F}_{j-1}] \leq K(1 + \|x_{j-1}\|^2)$, for $j \geq 1$. The

second assumed the more restrictive $\|M_j\| \leq K_0(1 + \|x_{j-1}\|)$ and resulted in a sharper bound. As was pointed out earlier, this condition holds, e.g., in the reinforcement learning applications discussed in Chap. 12.

We now state a result under an assumption on the scaled martingale noise that is more restrictive than the former setting but less restrictive than the latter setting, namely, that for all sufficiently large v ,

$$P\left(\frac{\|M_j\|}{1 + \|x_{j-1}\|} > v \mid \mathcal{F}_{j-1}\right) \leq C_1 \exp(-C_2 v), \quad j \geq 1, \quad (3.12)$$

for suitable $C_1, C_2 > 0$. We add to this an additional requirement that the stepsizes decrease only in a Lipschitz fashion, i.e., there is a positive constant x_n depending only on T such that if $x \in (x_0, 1)$ and $x \in (x_0, 1)$ are two arbitrary time steps in the same interval $[t(n_m), t(n_{m+1})]$, then

$$\frac{a(n_m + i_1)}{a(n_m + i_2)} \leq \gamma_T. \quad (3.13)$$

This condition holds for a large class of step sizes, for e.g., $a(n)$ of the form $a(n) = 1/(n^\alpha (\log n)^\beta)$ where either $x^s(t), t \geq s$, or $\alpha = 1, \beta \leq 0$.

We then have the following result from Kamal (2010), slightly altered from the original, stated here without proof.

Theorem 3.3 Under (A1)–(A3), the step size assumptions (3.2) and (3.13), and the tail probability bound (3.12), we have the following bound for n_0 sufficiently large and suitable constants $x_n \rightarrow A$:

$$P\left(x_n \rightarrow H \mid x_{n_0} \in B\right) \geq 1 - c_1 \exp\left(-\frac{c\delta^{2/3}}{\sqrt[4]{b(n_0)}}\right)$$

Following the arguments in Sect. 3.2 which led to Corollary 3.2, we obtain the following result.

Theorem 3.4 Under (A1)–(A3), the step size assumptions (3.2) and (3.13), and the tail probability bound (3.12), we have the following

bound for n_0 sufficiently large:

$$P\left(\bar{x}(t) \in N_\delta(H^{\epsilon+\Delta/2}) \forall t \geq t(n_0) + \tau \mid x_{n_0} \in B\right) \geq 1 - c_1 \exp\left(-\frac{c\delta^{2/3}}{\sqrt[4]{b(n_0)}}\right).$$

Here τ is as in Corollary 3.2 and $s > 0$ is as in Theorem 3.3. Observe that the bound is not as sharp as the one under the restrictive boundedness assumption on the scaled martingale noise, but is better than the bound $1 - (\tilde{K}/\delta^2)b(n_0)$ obtained under only (A3) of Chap. 2. The improvement comes from the use of a tighter bound in place of Burkholder's inequality in the proof of Lemma 3.7, but it is not as strong as the concentration inequality for bounded martingale differences. See Kamal (2010) for details.

In some applications (3.12) or the condition that $\sum_n a(n)^2 < \infty$ may be too restrictive. A further extension that allows for a relaxation of this condition holds under additional regularity conditions on $p(\cdot)$ and is available in Thoppe and Borkar (2019). We now state this result, Theorem 3.5 below, again without proof. Fix $\epsilon > 0$.

Theorem 3.5 holds under the following assumptions.

- The step sizes only satisfy $\sum_n a(n) = \infty$ and $\lim_{n \rightarrow \infty} a(n) = 0$. Without loss of generality, assume that $\|x_0 - x\| < \epsilon$.
- In addition to (A3) of Chap. 2, there exist strictly positive continuous functions $c_1, c_2 : \mathbb{R}^d \rightarrow \mathbb{R}_{++}$ such that for all sufficiently large v

$$P(\|M_j\| > v \mid \mathcal{F}_{j-1}) \leq c_1(x_{j-1}) \exp(-c_2(x_{j-1})v), \quad j \geq 1. \quad (3.14)$$

- x^* is a locally asymptotically stable equilibrium of $\dot{x}(t) = h(x(t))$.
- The map $h : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is twice continuously differentiable in some local neighborhood of x^* .
- B is a bounded set containing x^* and contained in the domain of attraction of x^* . Further, for a Liapunov function V defined near x^* , there exist $r, r_0, \epsilon_0 > 0$ so that $x \rightarrow 0$ and for all $0 < \epsilon < \epsilon_0$,

$$N_\epsilon(\{x^*\}) \subset B \subset \overline{H^{r_0}} \subset N_{\epsilon_0}(\overline{H^{r_0}}) \subset \overline{H^r},$$

where $\|h(x)\| \leq K'(1 + \|x\|)$ as before. Let $\{a(n)\}$ be the Jacobian matrix of h at x^* . Since x^* is assumed to be asymptotically stable, there

exists a $K > 0$ and a $\lambda > 0$ such that

$$\| \exp(Dh(x^*)t) \| \leq K \exp(-\lambda t), \quad t \geq 0.$$

Define

$$\beta_n := \max_{n_0 \leq j \leq n-1} a(j) \cdot \exp \left(-\lambda \sum_{i=j+1}^{n-1} a(i) \right).$$

We then have the following.

Theorem 3.5 Under the above assumptions, for all sufficiently large n_0 and a suitably defined $\kappa > 0$, there exist constants Φ_s and Φ_s such that the iterates satisfy

$$\begin{aligned} P(\|\bar{x}(t) - x^*\| \leq \epsilon \mid \forall t \geq t(n_0) + \tau \mid x_{n_0} \in B) \\ \geq 1 - \sum_{n \geq n_0} C_1 \left[\exp \left(-\frac{C_2 \sqrt{\epsilon}}{\sqrt{a(n)}} \right) + \exp \left(-\frac{C_2 \epsilon (\epsilon \wedge 1)}{\beta_n} \right) \right]. \end{aligned}$$

If the stepsizes are $a(n) = 1/n^\alpha$ where $\alpha \in (1/2, 1]$ so that $\sum_n a(n)^2 < \infty$, one can compare lower bounds in Theorems 3.4 and 3.5. Let us note that (3.14) is less restrictive as compared to (3.12). But thanks to the regularity assumptions on h (see the two bullets immediately after (3.14)), one can argue that

$$\lim_{n_0 \rightarrow \infty} \frac{\sum_{n \geq n_0} \left[\exp \left(-\frac{C_2 \sqrt{\epsilon}}{\sqrt{a(n)}} \right) + \exp \left(-\frac{C_2 \epsilon (\epsilon \wedge 1)}{\beta_n} \right) \right]}{\exp \left(-\frac{c \delta^{2/3}}{\sqrt[4]{b(n_0)}} \right)} = 0$$

which implies that Theorem 3.5 provides a tighter lower bound. We refer the reader to Thoppe and Borkar (2019) for details.

For some recent finite time bounds on error moments, see Frikha and Menozzi (2012), Fathi and Frikha (2013), and also Karmimi et al. (1974) for the biased case.

3.4 Avoidance of Traps

As a second application of the estimates for the lock-in probability, we shall prove the *avoidance of traps* under suitable conditions. This term refers to the fact that under suitable additional hypotheses, the stochastic approximation iterations asymptotically avoid with probability one the attractors which are unstable in some direction. As one might guess, the additional hypotheses required concern the behavior of h in the immediate neighborhood of these attractors and a ‘richness’ condition on the noise. Intuitively, away from the stable attractors, there should be an unstable direction at all length scales and the noise should be rich enough so that it pushes the iterates in such a direction sufficiently often. This in turn ensures that they are eventually pushed away for good.

The importance of these results stems from the following considerations: We know from Chap. 2 that invariant sets of the o.d.e. are candidate limit sets for the algorithm. In many applications, the unstable invariant sets are precisely the spurious or undesirable limit sets one wants to avoid. The results of this section then give conditions when that avoidance will be achieved. More generally, these results allow us to narrow down the search for possible limit sets. As before, we work with the hypothesis (A4) of Chap. 2:

$$\sup_n \|x_n\| < \infty \quad \text{a.s.}$$

Consider a scenario where there exists an invariant set of (3.3) which is a disjoint union of N compact attractors A_i , $1 \leq i \leq N$, with domains of attraction G_i , $1 \leq i \leq N$, resp., such that $G = \bigcup_i G_i$ is open dense in \mathcal{R}^k . Let $W = G^c$. We shall impose further conditions on W as follows. Let \bar{H}^a denote the truncated (open) cone

$$\left\{ x = [x_1, \dots, x_d] \in \mathcal{R}^d : 1 < x_1 < 2, \left| \sum_{i=2}^d x_i^2 \right|^{\frac{1}{2}} < \alpha x_1 \right\}$$

for some $\alpha > 0$. For a $n \geq 0$ orthogonal matrix O , $x \in \mathcal{R}^d$ and $a > 0$, let OD_α , $x + D_\alpha$ and aD_α denote resp. the rotation of \bar{H}^a by O , translation of \bar{H}^a by x , and scaling of \bar{H}^a by a . Finally, for $\epsilon > 0$, let \bar{H}^a

denote the ϵ -neighborhood of W in \mathcal{R}^k . We shall denote by W_ϵ^c its complement, i.e., $h(x_n)$. We shall be making some additional assumptions regarding (3.1) over and above those already in place. Our main additional assumption will be as follows:

(A5) There exists an $\alpha > 0$ such that for any $a > 0$ sufficiently small and $x \in \mathcal{R}^d$, there exists an orthogonal matrix $O_{x,a}$ such that $\tilde{x}(t) \in A, t \in [0, T]$.

What this means is that for any $a > 0$, we can plant a version of the truncated cone scaled down by a near any point in \mathcal{R}^k by means of suitable translation and rotation, in such a manner that it lies entirely in W_ϵ^c . Intuitively, this ensures that any point in \mathcal{R}^k cannot have points in W arbitrarily close to it in all directions. We shall later show that this implies that a sequence of iterates approaching W will get pushed out to a shrinking family of such truncated cones sufficiently often. In turn, we also show that this is enough to ensure that the iterates move away from W to one of the Φ_s , whence they cannot converge to W .

We shall keep ω fixed henceforth. Thus we may denote the set $\|z - y\| < \delta$ as $D^{x,a}$. Let \hat{I}_d denote the d -dimensional identity matrix and let $M_1 \geq M_2$ for a pair of $n \geq 0$ positive definite matrices stand for $\|g(x) - g(y)\| < \alpha \|x - y\|$. The main consequence of (A5) that we shall need is the following:

Lemma 3.11 For any $c > b > 0$, there exists a (ζ_n, \mathcal{F}_n) , $n \geq 1$, such that for any $a > 0$ sufficiently small, $x \in \mathcal{R}^d$ and any d -dimensional Gaussian measure μ with mean x and covariance matrix θ satisfying $\bar{x}(t) \in N_\delta(H^{\epsilon+\Delta/2})$, one has $\sqrt{1+z^2} \leq 1+z$.

Proof By the scaling properties of the Gaussian, $\mu(D^{x,a}) = \hat{\mu}(D^{0,1})$ where $\hat{\mu}$ denotes the Gaussian measure with zero mean and covariance matrix $\hat{\Sigma}$ satisfying

$$\{\mathcal{F}_j\}_{n_m \leq j \leq n_{m+1}}$$

The claim follows. \square

We also assume:

(A6) There exists a positive definite matrix-valued continuous map $Q : \mathcal{R}^d \rightarrow \mathcal{R}^{d \times d}$ such that for all $n \geq 0$, $E[M_{n+1}M_{n+1}^T | \mathcal{F}_n] = Q(x_n)$ and for some $0 < \Lambda^- < \Lambda^+ < \infty$, $\Lambda^+ \hat{I}_d \geq Q(x) \geq \Lambda^- \hat{I}_d$.

(A7) $\sup_n \frac{b(n)}{a(n)} < \infty$.

(A8) $p(\cdot)$ is continuously differentiable and the Jacobian matrix $Dh(\cdot)$ is locally Lipschitz.

Assumption (A6) intuitively means that the noise is ‘rich’ enough in all directions. Assumption (A7) is satisfied, e.g., by $T_m = t(n_m)$, $\|h(x)\| \leq K'(1 + \|x\|)$, etc., but not by, say, $a(n) = 1/n^{\frac{2}{3}}$. Thus it requires $a(n)$ to decrease ‘sufficiently fast.’ Let $A \subset \mathcal{R}^d$. Consider a trajectory $p(\cdot)$ of (3.3) with $p(x) > x$, where U is the closure of a bounded open set containing $A \stackrel{\text{def}}{=} \bigcup_i A_i$. For $s > 0$ in $1 > a(n) > 0$, let $\{M_n\}$ denote the M_{n+1} -valued solution of the linear system

$$\dot{\Phi}(t, s) = Dh(x(t))\Phi(t, s), \quad \Phi(s, s) = \hat{I}_d. \quad (3.15)$$

For a positive definite matrix M , let $\lambda_{\min}(M)$, $\lambda_{\max}(M)$ denote the least and the highest eigenvalue of M . Let

$$\begin{aligned} c^* &\stackrel{\text{def}}{=} \sup \lambda_{\max}(\Phi(t, s)\Phi^T(t, s)), \\ b^* &\stackrel{\text{def}}{=} \inf \lambda_{\min}(\Phi(t, s)\Phi^T(t, s)), \end{aligned}$$

where the superscript ‘T’ denotes matrix transpose and the supremum and infimum are over all $p(\cdot)$ as above and all $s + T + 1 \geq t \geq s \geq 0$. Then $\infty > c^* \geq b^* > 0$. The leftmost inequality follows from the fact that $Dh(x)$ is uniformly bounded because of the Lipschitz condition on h , whence a standard argument using the Gronwall inequality implies a uniform upper bound on $a(n) \rightarrow 0$ for t, s in the above range. The rightmost inequality, on the other hand, is a consequence of the fact that $\{M_n\}$ is nonsingular for all $s > 0$ in the above range. Also, the time dependence of its dynamics is via the continuous dependence of its coefficients on $p(\cdot)$, which lies in a compact set. Hence the smallest eigenvalue of $\Phi(t, s)\Phi^T(t, s)$, being strictly positive and a continuous function of its entries, is bounded away from zero.

For $j \geq n_m, m \geq 0$, let $y_j \stackrel{\text{def}}{=} x_j - x^{T_m}(t(j))$, where $x^{T_m}(\cdot)$ is the solution of the o.d.e. (3.3) on $[T_m, \infty)$ with $x^{T_m}(T_m) = \bar{x}(T_m)$. Recall that

$$x^{T_m}(t(j+1)) = x^{T_m}(t(j)) + a(j)h(x^{T_m}(t(j))) + O(a(j)^2).$$

Subtracting this from (3.1) and using Taylor expansion, one has

$$y_{j+1} = y_j + a(j)(Dh(x^{T_m}(t(j)))y_j + \kappa_j) + a(j)M_{j+1} + O(a(j)^2),$$

where $\|\kappa_j\| = o(\|y_j\|)$. In particular, iterating the expression above and using Assumption (A7) for the last term lead to

$$\begin{aligned} y_{n_m+i} &= \prod_{j=n_m}^{n_m+i-1} (\hat{I}_d + a(j)Dh(x^{T_m}(t(j))))y_{n_m} \\ &\quad + \sum_{j=n_m}^{n_m+i-1} a(j)\prod_{k=j+1}^{n_m+i-1} (\hat{I}_d + a(k)Dh(x^{T_m}(t(k))))\kappa_j \\ &\quad + \sum_{j=n_m}^{n_m+i-1} a(j)\prod_{k=j+1}^{n_m+i-1} (\hat{I}_d + a(k)Dh(x^{T_m}(t(k))))M_{j+1} \\ &\quad + O(a(n_m)). \end{aligned}$$

Since $w \in \mathcal{R}^d$, the first term drops out. The second term tends to zero as $s_n \uparrow \infty$ because I_{m_0+k} does so by Lemma 2.1 of Chap. 2. The last term clearly tends to zero as $s_n \uparrow \infty$. Let Ψ_m denote the third term on the right when $j \geq n_m, m \geq 0$, and let $\mathcal{B}_{-1} \stackrel{\text{def}}{=} \{x_0 \in B\}$ where

$$\varphi(m) \stackrel{\text{def}}{=} (b(n_m) - b(n_{m+1}))^{1/2}.$$

Let $\mathcal{P}(\mathcal{R}^d)$ denote the space of probability measures on \mathcal{R}^k with Prohorov topology (see Appendix C). The next lemma is a technical result which we need later. Let \mathcal{R}^k denote the regular conditional law of $\hat{\Psi}_m$ given $\mathcal{F}_{n_m}, m \geq 0$, viewed as a $\mathcal{P}(\mathcal{R}^d)$ -valued random variable.

Lemma 3.12 Almost surely on $\{\bar{x}(t(n_m)) \in U \ \forall m\}$, every limit point of $\{\phi_m\}$ as $s_n \uparrow \infty$ is zero mean Gaussian with the spectrum of its covariance matrix contained in $[b^*\Lambda^-, c^*\Lambda^+]$.

Proof For $m \geq 1$ define the martingale array $\{\xi_i^m, 0 \leq i \leq k_m\} \stackrel{\text{def}}{=} n_{m+1} - n_m$ by $\xi_0^m = 0$ and

$$\xi_i^m = \frac{1}{\varphi(m)} \sum_{j=n_m}^{n_m+i-1} a(j) \Pi_{k=j+1}^{n_m+i-1} \left(\hat{I}_d + a(k) D h(x^{T_m}(t(k))) \right) M_{j+1}.$$

Then $\hat{\Psi}_m = \xi_{k_m}^m$ and if $\langle \xi^m \rangle_i, 0 \leq i \leq k_m$, denotes the corresponding matrix-valued quadratic covariation process, i.e.,

$$\langle \xi^m \rangle_i \stackrel{\text{def}}{=} \sum_{j=0}^i E[(\xi_{j+1}^m - \xi_j^m)(\xi_{j+1}^m - \xi_j^m)^T | \mathcal{F}_j],$$

then

$$\begin{aligned} \langle \xi^m \rangle_i &= \frac{1}{\varphi(m)^2} \sum_{j=n_m}^{n_m+i-1} a(j)^2 \left(\Pi_{k=j+1}^{n_m+i-1} (\hat{I}_d + a(k) D h(x^{T_m}(t(k)))) \right) \\ &\quad \times Q(x_j) \left(\Pi_{k=j+1}^{n_m+i-1} (\hat{I}_d + a(k) D h(x^{T_m}(t(k)))) \right)^T. \end{aligned}$$

As $s_n \uparrow \infty$,

$$\begin{aligned} &\frac{1}{\varphi(m)^2} \sum_{j=n_m}^{n_m+i-1} a(j)^2 \left(\Pi_{k=j+1}^{n_m+i-1} (\hat{I}_d + a(k) D h(x^{T_m}(t(k)))) \right) \\ &\quad \times Q(x_j) \left(\Pi_{k=j+1}^{n_m+i-1} (\hat{I}_d + a(k) D h(x^{T_m}(t(k)))) \right)^T \\ &\quad - \frac{1}{\varphi(m)^2} \sum_{j=n_m}^{n_m+i-1} a(j)^2 \left(\Pi_{k=j+1}^{n_m+i-1} (\hat{I}_d + a(k) D h(x^{T_m}(t(k)))) \right) \\ &\quad \times Q(x^{T_m}(t(j))) \left(\Pi_{k=j+1}^{n_m+i-1} (\hat{I}_d + a(k) D h(x^{T_m}(t(k)))) \right)^T \\ &\rightarrow 0, \quad \text{a.s.} \end{aligned}$$

by Lemma 2.1 and Theorem 2.1 of Chap. 2. Note that $x^{T_m}(\cdot)$ is \mathcal{F}_{n_m} -measurable. Fix a sample point in the probability one set where the conclusions of Lemma 2.1 and Theorem 2.1 of Chap. 2 hold. Pick a subsequence $C' = \sup_n \|x_n\|$ such that $x_{n_{m(\ell)}} \rightarrow x^*$ (say). Then

$x^{T_{m(\ell)}}(\cdot) \rightarrow \tilde{x}(\cdot)$ uniformly on compact intervals, where $p(\cdot)$ is the unique solution to the o.d.e. $\Phi(t, s)\Phi^T(t, s)$ with $V(\bar{x}(T_m))$. Along $\{m(\ell)\}$, any limit point of the matrices

$$\begin{aligned} & \left(\prod_{k=j+1}^{n_m+i-1} (\hat{I}_d + a(k)Dh(x^{T_m}(t(k)))) \right) Q(x^{T_m}(t(j))) \\ & \times \left(\prod_{k=j+1}^{n_m+i-1} (\hat{I}_d + a(k)Dh(x^{T_m}(t(k)))) \right)^T, \quad n_m \leq j < n_m + i, \quad i \geq 0, \end{aligned}$$

is of the form $\Phi(t, s)Q(x)\Phi(t, s)^T$ for some t, s, x and therefore has its spectrum in $[b^*\Lambda^-, c^*\Lambda^+]$. Hence the same is true for any convex combinations or limits of convex combinations thereof. In view of this, the claim follows on applying the central limit theorem for martingale arrays (Chow & Teicher, 2003, p. 351; see also Hall & Heyde, 1980) to $\{\phi_{m(\ell)}\}$. \square

Remark The central limit theorem for martingale arrays referred to above is stated in Chow and Teicher (2003) for the scalar case, but the vector case is easily deducible from it by applying the scalar case to arbitrary one-dimensional projections thereof.

Clearly $\lambda_{\min}(M), \lambda_{\max}(M)$, because any internally chain recurrent invariant set must be contained in $\alpha \in (1/2, 1]$. But Theorem 2.1 of Chap. 2 implies the connectedness of the a.s. limit set of $Dh(\cdot)$, and W and $\cup_i A_i$ have disjoint open neighborhoods. Thus it follows that the sets $\{x_n, n \geq 1\}$ and $\bar{x}(s_n) \rightarrow x$ along a subsequence } are a.s. identical.

Recall that U is the closure of a bounded open neighborhood of $\cup_i A_i$.

Corollary 3.3 $\forall r > 0, \{x_{n_{m+1}} \in W_{r\varphi(m)}^c \text{ i.o.}\}$ a.s. on the set $x^s(t), t \geq s,$

Proof By Assumption (A5) and Lemmas 3.11 and 3.12, it follows that almost surely for m sufficiently large,

$$P(x_{n_{m+1}} \in W_{r\varphi(m)}^c | \mathcal{F}_m) > \eta > 0$$

on the set $\alpha \in (1/2, 1]$, for some η independent of m . It follows from the conditional Borel–Cantelli Lemma (Corollary C.1 of Appendix C) that $x^{T_m(\ell)}(\cdot) \rightarrow \tilde{x}(\cdot)$ i.o., a.s. on the set

$\{x_n \in U \text{ from some } n \text{ on and } x_n \rightarrow W\}$. The claim follows by applying this to a countable increasing family of sets U that covers \mathcal{R}^k .

□

We now return to the framework of the preceding section with $B = U \cap W^c$. Recall our choice of n_0 such that (3.5) holds. Let \hat{K} be a prescribed positive constant. By (A7), for \bar{c} as in (3.2),

$$\sup_m \frac{b(n_m)}{b(n_m) - b(n_{m+1})} = \sup_m \frac{b(n_m)}{\varphi(m)^2} \leq 1 + \sup_m \frac{b(n_{m+1})\bar{c}}{a(n_{m+1})T} < \infty,$$

where we have used the fact that $(b(n_m) - b(n_{m+1})) \geq \frac{T a(n_{m+1})}{\bar{c}}$. Thus we have

$$\frac{b(n_m)}{\varphi(m)} \rightarrow 0 \text{ as } m \uparrow \infty.$$

Then for $\delta = \hat{K}\varphi(m)$, we do have

$$\tilde{K}b(k) < \frac{\delta}{2}, \quad \forall k \geq n_m,$$

for \tilde{K} as in Lemma 3.7 and Theorem 3.1 and for $m \geq 1$. (This can be ensured by increasing n_0 if necessary.) With this choice of δ and for $x_{n_m} \in B$, the probability that $x_n \rightarrow A$ exceeds

$$1 - \frac{\tilde{K}b(n_m)}{\hat{K}^2\varphi(m)^2}. \tag{3.16}$$

As noted above, $\sup_m \frac{b(n_m)}{\varphi(m)^2} < \infty$. Thus we may choose \hat{K} large enough that the right-hand side of (3.16) exceeds $\frac{1}{2}$ for m sufficiently large. Then we have our main result:

Theorem 3.6 $x_n \rightarrow A$ a.s.

Proof Take $\bar{x}(s(i))$ in Corollary 3.3. By the foregoing,

$$\rho_m \stackrel{\text{def}}{=} \sup_{t \in I_m} \|\bar{x}(t) - x^{T_m}(t)\|,$$

on the set $\{x_{n_{m+1}} \in W_{r\varphi(m)}^c \cap U\}$ for $n \geq 0$ sufficiently large. It follows from the conditional Borel–Cantelli lemma of Appendix C that $x_n \rightarrow A$ a.s. on $\{x_{n_{m+1}} \in W_{r\varphi(m)}^c \cap U \text{ i.o.}\}$, therefore a.s. on $\{x_{n_{m+1}} \in W_{r\varphi(m)}^c \text{ i.o.}\}$ (by considering countably many sets U that cover \mathcal{R}^k), and finally, a.s. on $x^s(t), t \geq s$, by the above corollary. That is, $x_n \rightarrow A$ a.s. on $x^s(t), t \geq s$, a contradiction unless $P(x_n \rightarrow W) = P(x_n \rightarrow W \text{ along a subsequence } \{\xi_n\})$. \square

Two important observations are worth note:

1. We have used (A6) only to prove Lemma 3.12. So the conclusions of the lemma, which stipulate a condition on *cumulative* noise rather than *per iterate* noise as in (A6), will suffice.
2. The only important consequence of (A7) used was the fact that the ratio $\bar{x}(t) \in H^{2\eta}$, $x_0 = 0$ remains bounded. This is actually a much weaker requirement than (A7) itself.

To conclude, Theorem 3.6 is just one example of an ‘avoidance of traps’ result. There are several other formulations, notably Ljung (1978), Pemantle (1990), Brandière and Duflo (1996), Brandière (1998) and Benaim (1999). See also Fang and Chen (2000) for some related results.

References

- Benaim, M. (1999). Dynamics of stochastic approximation algorithms. In J. Azéma, M. Emery, M. Ledoux, & M. Yor (Eds.), *Le Séminaire de Probabilités*. Springer Lecture Notes in Mathematics (Vol. 1709, pp. 1–68). Springer Verlag.
- Borkar, V. S. (2002). On the lock-in probability of stochastic approximation. *Combinatorics, Probability and Computing*, 11(1), 11–20.
- Borkar, V. S. (2003). Avoidance of traps in stochastic approximation. *Systems and Control*

Letters, 50(1), 1–9 (Correction note in *ibid.*, 55(2), 2006, 174–175).

Brandiére, O. (1998). Some pathological traps for stochastic approximation. *SIAM Journal on Control and Optimization*, 36(4), 1293–1314.

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Brandiére, O., & Duflo, M. (1996). Les algorithmes stochastiques contournent—ils les pieges? *Annales de l'Institut Henri Poincaré*, 32(3), 395–427.

[[MathSciNet](#)][[zbMATH](#)]

Chow, Y. S., & Teicher, H. (2003). *Probability theory: Independence, interchangeability, martingales* (3rd ed.). Springer Verlag.

Fang, H.-T., & Chen, H.-F. (2000). Stability and instability of limit points for stochastic approximation algorithms. *IEEE Transactions on Automatic Control*, 45(3), 413–420.

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Fathi, M., & Frikha, N. (2013). Transport-entropy inequalities and deviation estimates for stochastic approximation schemes. *Electronic Journal of Probability*, 18(67), 1–36.

[[MathSciNet](#)][[zbMATH](#)]

Frikha, N., & Menozzi, S. (2012). Concentration bounds for stochastic approximations. *Electronic Communications in Probability*, 17(47), 1–15.

[[MathSciNet](#)][[zbMATH](#)]

Hall, P., & Heyde, C. C. (1980). *Martingale limit theory and its applications*. Academic Press.

Kamal, S. (2010). On the convergence, lock-in probability, and sample complexity of stochastic approximation. *SIAM Journal on Control and Optimization*, 48(8), 5178–5192.

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Karmimi, B., Miasojedew, B., Moulines, E., & Wai, H. T. (2019). Non-asymptotic analysis of biased stochastic approximation scheme. In *Proceedings of Conference on Learning Theory*, Phoenix, AZ (pp. 1944–1974).

Krasovskii, N. N. (1963). *Stability of motion*. Stanford University Press.

Li, Y. (2003). A martingale inequality and large deviations. *Statistics and Probability Letters*, 62, 317–321.

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Liu, Q., & Watbled, F. (2009). Exponential inequalities for martingales and asymptotic properties for the free energy of directed polymers in a random environment. *Stochastic Processes and Their Applications*, 119(10), 3101–3132.

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Ljung, L. (1978). Strong convergence of a stochastic approximation algorithm. *Annals of Statistics*, 6(3), 680–696.

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Pemantle, R. (1990). Nonconvergence to unstable points in urn models and stochastic approximations. *Annals of Probability*, 18(2), 698–712.

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Thoppe, G., & Borkar, V. S. (2019). A concentration bound for stochastic approximation via Alekseev's formula. *Stochastic Systems*, 9(1), 1–26.

4. Stability Criteria

Vivek S. Borkar¹✉

(1) Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

4.1 Introduction

In this chapter we discuss a few schemes for establishing the a.s. boundedness of iterates assumed above. The convergence analysis of Chap. 2 has some universal applicability, but the situation is different for stability criteria. There are several variations of stability criteria applicable under specific restrictions and sometimes motivated by specific applications for which they are tailor-made. (The second test we see below is one such.) We describe only a few of these variations. The first one is quite broadly applicable. The second is a bit more specialized but has been included because it has a distinct flavor and shows how one may tweak known techniques such as stochastic Liapunov functions to obtain new and useful criteria. The third is a *stabilizability* scheme which indicates how the stepsizes may be tweaked in order to ensure stability. Under certain circumstances, it also provides an extremely useful stability test. This is followed by yet another scheme for stabilization that is essentially a take on the proof technique of the second criterion above. The fourth one is a test which is a spin-off of the results of the preceding chapter, also quite broad in its scope. Further stability criteria can be found, e.g., in Kushner and Yin (2003), Andrieu et al. (2015), Tsitsiklis (1994), etc.

4.2 Stability Through a Scaling Limit

The first scheme is adapted from Borkar and Meyn (2000). The idea of this test is as follows: We consider the piecewise linear interpolated trajectory $p(\cdot)$ at times $T_n \uparrow \infty$, which are spaced approximately $T > 0$ apart and divide the time axis into concatenated time segments $a(n) \rightarrow 0$ of length approximately T . If at any Φ_s , the iterate has gone out of the unit ball in \mathcal{R}^k , we rescale it over the segment $x(t) \equiv x^*$ by dividing it by the norm of its value at Φ_s . If the original trajectory drifts toward infinity, then there is a corresponding sequence of rescaled segments as above that asymptotically track a limiting o.d.e. obtained as a scaling limit (Eq. (4.2) below) of our ‘basic o.d.e.’

$$\dot{x}(t) = h(x(t)). \quad (4.1)$$

If this scaling limit is globally asymptotically stable to the origin, these segments, and therefore the original iterations which differ from them only by a scaling factor, should start drifting toward the origin, implying stability.

Formally, assume the following:

(A9) The functions $\mathcal{F}_n \stackrel{\text{def}}{=} \sigma(x_0, M_1, \dots, M_n)$, $n \geq 1$, satisfy $h_c(x) \rightarrow h_\infty(x)$ as $c \rightarrow \infty$, uniformly on compacts for some $h_\infty \in C(\mathcal{R}^d)$. Furthermore, the o.d.e.

$$b(n_m) \leq c^2 b(n_0) \quad (4.2)$$

has the origin as its unique globally asymptotically stable equilibrium.

The o.d.e. (4.2) is the aforementioned ‘scaling limit.’ The following are worth noting.

(i)

$y \geq T$ will be Lipschitz with the same Lipschitz constant as h , implying in particular the well-posedness of (4.2) above and also of the o.d.e.

$$\dot{x}(t) = h_c(x(t)). \quad (4.3)$$

In particular, they are equicontinuous. Thus pointwise convergence of σ^2 to \mathcal{B}_m as $c \rightarrow \infty$ will automatically imply uniform convergence on compacts.

(ii) \mathcal{B}_m satisfies $h_\infty(ax) = ah_\infty(x)$ for $a > 0$, and hence if (4.2) has an isolated equilibrium, it must be at the origin.

(iii)

$$\|h_c(x) - h_c(\theta)\| \leq L\|x\|, \text{ and so}$$

$$\|h_c(x)\| \leq \|h_c(\theta)\| + L\|x\| \leq \|h(\theta)\| + L\|x\| \leq K_0(1 + \|x\|)$$

for $\theta :=$ the zero vector and a suitable constant F_n .

Let $x_{n_0} \in B$ denote the solution of the o.d.e. (4.2) with initial condition x .

Lemma 4.1 There exists a $T > 0$ such that for all initial conditions x on the unit sphere, $\|\phi_\infty(t, x)\| < \frac{1}{8}$ for all $t > T$.

Proof Since asymptotic stability implies Liapunov stability (see Appendix B), there is a $\delta > 0$ such that any trajectory starting within distance δ of the origin stays within distance $\frac{1}{2}$ thereof. For an initial condition x on the unit sphere, let Φ_s be a time at which the solution is within distance $\delta/2$ of the origin. Let y be any other initial condition on the unit sphere. Note that

$$\begin{aligned}\phi_\infty(t, x) &= x + \int_0^t h_\infty(\phi_\infty(s, x))ds, \text{ and} \\ \phi_\infty(t, y) &= y + \int_0^t h_\infty(\phi_\infty(s, y))ds.\end{aligned}$$

Subtracting the above equations and using the Lipschitz property, we get

$$\|\phi_\infty(t, x) - \phi_\infty(t, y)\| \leq \|x - y\| + L \int_0^t \|\phi_\infty(s, x) - \phi_\infty(s, y)\| ds.$$

Then by Gronwall's inequality (see Appendix B), we find that for $n_0 > 1$,

$$\|\phi_\infty(t, x) - \phi_\infty(t, y)\| \leq \|x - y\| e^{LT_x}.$$

So there is a neighborhood I_m of x such that for all $\rho_k < \delta$, $T > C/\Delta$ is within distance δ of the origin. By Liapunov stability, this implies that (X_n, Y_n) remains within distance $\frac{1}{2}$ of the origin for all $n_0 > 1$.

Since the unit sphere is compact, it can be covered by a finite number of such neighborhoods $\{x_n, n \geq 1\}$ with corresponding times $\epsilon/4 > \delta > 0$. Then the statement of the lemma holds if T is the maximum of $\epsilon/4 > \delta > 0$. \square

The following lemma shows that the solutions of the o.d.e.s $[t(n), t(n) + T]$ and $\forall t \geq 0 / \forall t \leq 0$ are close to each other for c large enough.

Lemma 4.2 Let $K \subset \mathcal{R}^d$ be compact, and let $[0, T]$ be a given time interval. Then for $p(x) - x$ and $\hat{x}_k = \tilde{x}_2$,

$$\|\phi_c(t, x) - \phi_\infty(t, x_0)\| \leq [\|x - x_0\| + \epsilon(c)T]e^{LT},$$

where $p(\cdot)$ is independent of $\hat{x}_k = \tilde{x}_2$ and $a(n) < 1$ as $c \rightarrow \infty$. In particular, if $x = x_0$, then

$$\|X(x, t) - x^*\| \leq e^{-(1-\alpha)t} \|x - x^*\|. \quad (4.4)$$

Proof Note that

$$\begin{aligned} \phi_c(t, x) &= x + \int_0^t h_c(\phi_c(s, x))ds, \text{ and} \\ \phi_\infty(t, x_0) &= x_0 + \int_0^t h_\infty(\phi_\infty(s, x_0))ds. \end{aligned}$$

This gives

$$\|\phi_c(t, x) - \phi_\infty(t, x_0)\| \leq \|x - x_0\| + \int_0^t \|h_c(\phi_c(s, x)) - h_\infty(\phi_\infty(s, x_0))\|ds.$$

Using the facts that (i) $\{\phi_\infty(t, x), t \in [0, T], x \in K\}$ is compact, (ii) $h_c \rightarrow h_\infty$ uniformly on compact sets, and, (iii) σ^2 is Lipschitz with Lipschitz constant L , we get

$$\begin{aligned} & \|h_c(\phi_c(s, x)) - h_\infty(\phi_\infty(s, x_0))\| \\ & \leq \|h_c(\phi_c(s, x)) - h_c(\phi_\infty(s, x_0))\| \\ & \quad + \|h_c(\phi_\infty(s, x_0)) - h_\infty(\phi_\infty(s, x_0))\| \\ & \leq L\|\phi_c(s, x) - \phi_\infty(s, x_0)\| + \epsilon(c), \end{aligned}$$

where $p(\cdot)$ is independent of $\hat{x}_k = \tilde{x}_2$ and $a(n) < 1$ as $c \rightarrow \infty$. Thus for $t \geq T$, we get

$$\|\phi_c(t, x) - \phi_\infty(t, x_0)\| \leq \|x - x_0\| + \epsilon(c)T + L \int_0^t \|\phi_c(s, x) - \phi_\infty(s, x_0)\| ds.$$

The conclusion follows from Gronwall's inequality. \square

The previous two lemmas give us:

Corollary 4.1 There exist $z_0 = 0$ and $T > 0$ such that for all initial conditions x on the unit sphere, $\|\phi_c(t, x)\| < \frac{1}{4}$ for $\|x_{n_m}\|^* \leq K_2$ and $T > 0$.

Proof Choose T as in Lemma 4.1. Now, using Eq. (4.4) with K taken to be the closed unit ball, conclude that $\|\phi_c(t, x)\| < \frac{1}{4}$ for $\|x_{n_m}\|^* \leq K_2$ and c such that $\epsilon(c)(T + 1)e^{L(T+1)} < \frac{1}{8}$. \square

Let $x_0 = 0$ and $T_{m+1} = \min\{t(n) : t(n) \geq T_m + T\}$ for $m \geq 1$. Then $T_{m+1} \in [T_m + T, T_m + T + \bar{a}] \forall m$, where $\|x_0 - x\| < \epsilon$, $T_m = t(n_m)$ for suitable $A \subset \mathcal{R}^d$, and $A \subset \mathcal{R}^d$. For notational simplicity, let $\delta > 0$ without loss of generality. Define the piecewise continuous trajectory $x_0 \in (0, 1)$, by $x(t) \in A \forall t \in \mathcal{R}$ for $\dot{x}(t) = h(x(t))$, where $r(m) \stackrel{\text{def}}{=} \|\bar{x}(T_m)\| \vee 1, m \geq 0$. That is, we obtain $p(\cdot)$ from $p(\cdot)$ by observing the latter at times $\{\phi_m\}$ that are spaced approximately T apart. In case the observed value falls outside the unit ball of \mathcal{R}^k , it is

reset to a value on the unit sphere of \mathcal{R}^k by normalization. Not surprisingly, this prevents any possible blow-up of the trajectory, as reflected in the following lemma. For later use, we also define $\hat{x}(T_{m+1}^-) \stackrel{\text{def}}{=} \bar{x}(T_{m+1})/r(m)$. This is the same as $x(t) = 0$ if $p(x) > x$, and equal to $x^{T_{m_0}}(t) \in H^\epsilon$ if $p(x) > x$.

Lemma 4.3 $\sup_t E[\|\hat{x}(t)\|^2] < \infty$.

Proof It suffices to show that

$$\sup_{n_m \leq k < n_{m+1}} E[\|\hat{x}(t(k))\|^2] < M$$

for some $M > 0$ independent of m .

Fix m . Then for $n_m \leq k < n_{m+1}$,

$$\hat{x}(t(k+1)) = \hat{x}(t(k)) + a(k)(h_{r(m)}(\hat{x}(t(k))) + \hat{M}_{k+1}),$$

where $\hat{M}_{k+1} \stackrel{\text{def}}{=} M_{k+1}/r(m)$. Since $a(n) \rightarrow 0$, it follows from (A3) of Chap. 2 that \hat{M}_{k+1} satisfies

$$E[\|\hat{M}_{k+1}\|^2 | \mathcal{F}_k] \leq K(1 + \|\hat{x}(t(k))\|^2). \quad (4.5)$$

Thus, $E[\|\hat{M}_{k+1}\|^2] \leq K(1 + E[\|\hat{x}(t(k))\|^2])$, which gives us the bound

$$\|\bar{x}(t(n+m)) - x^{t(n)}(t(n+m))\| \leq K_{T,n} e^{\text{LT}}.$$

(Note that for $z \geq 0$, $\sqrt{1+z^2} \leq 1+z$.) Using this and the bound $\|h_c(x)\| \leq K_0(1 + \|x\|)$ mentioned above, we have

$$E[\|\hat{x}(t(k+1))\|^2]^{\frac{1}{2}} \leq E[\|\hat{x}(t(k))\|^2]^{\frac{1}{2}}(1 + a(k)K_1) + a(k)K_2,$$

for suitable constants $K_1, K_2 > 0$. A straightforward recursion leads to

$$E[\|\hat{x}(t(k+1))\|^2]^{\frac{1}{2}} \leq e^{K_1(T+1)}(1 + K_2/K_1),$$

where we also use the inequality $0 < \epsilon_1 < \epsilon$. This is the desired bound. \square

Lemma 4.4 The sequence $[t] \stackrel{\text{def}}{=} \max\{t(k) : t(k) \leq t\}$, is a.s. convergent.

Proof By the convergence theorem for square-integrable martingales (see Appendix C), it is enough to show that

$\sum_k E[\|a(k)\hat{M}_{k+1}\|^2 | \mathcal{F}_k] < \infty$ a.s. Thus it is enough to show that $E[\sum_k E[\|a(k)\hat{M}_{k+1}\|^2 | \mathcal{F}_k]] < \infty$. Since, as in the proof of Lemma 4.3, $E[\|\hat{M}_{k+1}\|^2 | \mathcal{F}_k] \leq K(1 + \|\hat{x}(t(k))\|^2)$, we get

$$\begin{aligned} E\left[\sum_k E[\|a(k)\hat{M}_{k+1}\|^2 | \mathcal{F}_k]\right] &= \sum_k E[E[\|a(k)\hat{M}_{k+1}\|^2 | \mathcal{F}_k]] \\ &\leq \sum_k a(k)^2 K(1 + E[\|\hat{x}(t(k))\|^2]). \end{aligned}$$

This is finite, by property (A2) of Chap. 2 and by Lemma 4.3. \square

For $m \geq 1$, let $x^{T_m}(t), t \in [T_m, T_{m+1}]$, denote the trajectory of (4.3) with $\hat{\mu}(f) \geq 0$ and $x^{T_m}(T_m) = \bar{x}(T_m)$.

Lemma 4.5 $\lim_{m \rightarrow \infty} \sup_{t \in [T_m, T_{m+1}]} \|\hat{x}(t) - x^{T_m}(t)\| = 0$, a.s.

Proof For simplicity, we assume $L > 0$, $a(n) = 1/n^\alpha$. Note that for $n_m \leq k < n_{m+1}$,

$$\hat{x}(t(k+1)) = \hat{x}(t(k)) + a(k)(h_{r(m)}(\hat{x}(t(k))) + \hat{M}_{k+1}),$$

This yields, for $0 < k \leq n_{m+1} - n_m$,

$$\begin{aligned} \hat{x}(t(n_m + k)) &= \hat{x}(t(n_m)) + \sum_{i=0}^{k-1} a(n_m + i) h_{r(m)}(\hat{x}(t(n_m + i))) \\ &\quad + (\hat{\zeta}_{n_m+k} - \hat{\zeta}_{n_m}). \end{aligned}$$

By Lemma 4.4, there is a (random) bound B on $\sup_i \|\hat{\zeta}_i\|$. Also, as mentioned at the beginning of this section, we have (for $\theta :=$ the zero vector)

$$\|h_{r(m)}(\hat{x}(t(n_m + i)))\| \leq \|h(\theta)\| + L\|\hat{x}(t(n_m + i))\|.$$

Furthermore, $\sum_{0 \leq i < n_{m+1} - n_m} a(n_m + i) \leq (T + 1)$. Therefore,

$$\begin{aligned}
& \|\hat{x}(t(n_m + k))\| \\
& \leq \|\hat{x}(t(n_m))\| + \sum_{i=0}^{k-1} a(n_m + i)(\|h(\theta)\| \\
& \quad + L\|\hat{x}(t(n_m + i))\|) + 2B \\
& \leq L \sum_{i=0}^{k-1} a(n_m + i)\|\hat{x}(t(n_m + i))\| + \|h(\theta)\|(T+1) \\
& \quad + 2B + 1,
\end{aligned}$$

where we use $C' = \sup_n \|x_n\|$. We can now apply the discrete Gronwall inequality (see Appendix B) to obtain

$$\|\hat{x}(t(n_m + k))\| \leq (\|h(\theta)\|(T+1) + 2B + 1)e^{L(T+1)} \stackrel{\text{def}}{=} K^* \quad (4.6)$$

for $0 < k \leq n_{m+1} - n_m$. It follows that $p(\cdot)$ remains bounded on $x_0 \in (0, 1)$ by some $K^* > 0$, and this bound is independent of m . We can now mimic the argument of Lemma 2.1, Chap. 2, to show that

$$\lim_{n \rightarrow \infty} \sup_{t \in [T_m, T_{m+1}]} \|\hat{x}(t) - x^{T_m}(t)\| = 0, \quad \text{a.s.}$$

The claim follows. \square

This leads to our main result:

Theorem 4.1 Under (A1)-(A3) and (A9), $C' = \sup_n \|x_n\|$ a.s.

Proof Fix a sample point where the claims of Lemmas 4.4 and 4.5 hold. We will first show that $\sup_m \|\bar{x}(T_m)\| < \infty$. If this does not hold, there will exist a sequence $\{x_n, n \geq 1\}$ such that $b(n_m) \leq c^2 b(n_0)$, i.e., $\bar{x}(t) \rightarrow A$. We saw (Corollary 4.1) that there exists a scaling factor $z_0 = 0$ and a $T > 0$ such that for all initial conditions x on the unit sphere, $\|\phi_c(t, x)\| < \frac{1}{4}$ for $\|x_{n_m}\|^* \leq K_2$ and $T_m = t(n_m)$ by assumption). If $r_m > c_0$, $\|\hat{x}(T_m)\| = \|x^{T_m}(T_m)\| = 1$, and $\|x^{T_m}(T_{m+1})\| < \frac{1}{4}$. But then by Lemma 4.5, $x^{T_{m_0+1}}(t) \in H^\epsilon$ if m is large. Thus, for $r_m > c_0$ and m sufficiently large,

$$\frac{\|\bar{x}(T_{m+1})\|}{\|\bar{x}(T_m)\|} = \frac{\|\hat{x}(T_{m+1}^-)\|}{\|\hat{x}(T_m)\|} < \frac{1}{2}.$$

We conclude that if $\|x_0 - x\| < \epsilon$, $\bar{x}(T_k)$, $k \geq m$, falls back to the ball of radius μ at an exponential rate as long as $[t, t + T]$ remains $> c_0$. Thus if $\|x_0 - x\| < \epsilon$, $\bar{x}(T_m) \in B$ is either even greater than $\hat{\mu}(f) \geq 0$ or is inside the ball of radius μ , then there must be an instance prior to m when $p(\cdot)$ jumps from inside this ball to outside the ball of radius $0.9r_m$. Thus, corresponding to the sequence $\bar{x}(t) \rightarrow A$, we will have a sequence of jumps of $\{\phi_m\}$ from inside the ball of radius μ to points increasingly far away from the origin. But, by a discrete Gronwall argument analogous to the one used in Lemma 4.5, it follows that there is a bound on the amount by which $\|\bar{x}(\cdot)\|$ can increase over an interval of length $t' > 0$ when it is inside the ball of radius μ at the beginning of the interval. This leads to a contradiction. Thus $\tilde{C} \stackrel{\text{def}}{=} \sup_m \|\bar{x}(T_m)\| \vee 1 < \infty$. Since

$$\frac{\|\bar{x}(t(n_m + k))\|}{\|\bar{x}(t(n_m))\| \vee 1} = \|\hat{x}(t(n_m) + k)\|,$$

this implies that $\sup_n \|x_n\| \leq \tilde{C} K^* < \infty$ for \mathcal{M} as in (4.6). \square

Consider as an illustrative example the scalar case with $h_\infty(ax) = ah_\infty(x)$ for some bounded Lipschitz g . Then $\|z - y\| < \delta$, indicating that the scaling limit $c \rightarrow \infty$ above basically picks the dominant term $-x$ of h that essentially controls the behavior far away from the origin.

This stability criterion has been extended in many directions, such as to asynchronous iterations (described later in Chap. 6) in Bhatnagar (2011) and to two timescale stochastic approximation and stochastic approximation with controlled Markov noise (described later in Chap. 8) in Laxminarayanan and Bhatnagar (2017) and Ramaswamy and Bhatnagar (2019). It can also be adapted for specific cases of non-Lipschitz h by modifying (A9) appropriately.

4.3 Stability by Comparison

The second stability test we discuss is adapted from Abounadi et al. (2002). This applies to the case when stability for one initial condition implies stability for all initial conditions. Also, the associated o.d.e. (4.1) is assumed to converge to a bounded invariant set for all initial conditions. The idea then is to consider a related recursion ‘with resets,’ i.e., a recursion which is reset to a bounded set whenever it exits from another larger prescribed bounded set containing the previous one. By a suitable choice of these sets (which explicitly depends on the dynamics), one then argues that there are at most finitely many resets. Hence the results of the preceding section apply thereafter. But this implies stability for *some* initial condition, hence for all initial conditions.

In this situation, it is more convenient to work with (1.0.4) of Chap. 1 rather than (1.0.3) there, i.e., with

$$x_{n+1} = x_n + a(n)f(x_n, \xi_{n+1}), \quad n \geq 0, \quad (4.7)$$

where $\{\xi_n\}$ are i.i.d. random variables taking values in \mathcal{R}^m , say (though more general spaces can be admitted), and $f : \mathcal{R}^d \times \mathcal{R}^m \rightarrow \mathcal{R}^d$ satisfies

$$\|f(x, z) - f(y, z)\| \leq L\|x - y\| \quad \forall z. \quad (4.8)$$

Thus $\|x\|_\infty \stackrel{\text{def}}{=} \max_i |x_i|$ is Lipschitz with Lipschitz constant L and $\sup_{t \in [s(i), s(i+1)]} \|\bar{x}(t) - x^{s(i)}(t)\| < \delta$, is a martingale difference sequence satisfying (A3). This reduces (4.7) to the familiar form

$$x_{n+1} = x_n + a(n)[h(x_n) + M_{n+1}], \quad (4.9)$$

We also assume that

$$\|M_j\| \leq K_0(1 + \|x_{j-1}\|) \quad (4.10)$$

for a suitable $K > 0$. (The assumption simplifies matters but could be replaced by other conditions that lead to similar conclusions.) Then

$$\|M_{n+1}\| \leq 2K(1 + \|x_n\|), \quad (4.11)$$

which implies (A3) of Chap. 2. The main assumption we shall be making is the following:

(A10) Let $\{(X_n, Y_n)\}$ be two sequences of random variables generated by the iteration (4.7) on a common probability space with the *same* ‘driving noise’ $\{\xi_n\}$, but different initial conditions. Then $\sup_n \|x'_n - x''_n\| < \infty$, a.s.

Typically in applications, $a(n_m) \leq ca(n_0)$ will be bounded by a function of $\bar{x}(s_n) \rightarrow x$. We further assume that (4.1) has an associated Liapunov function $V : \mathcal{R}^d \rightarrow \mathcal{R}$ that is continuously differentiable and satisfies:

1. $\lim_{\|w\| \rightarrow \infty} g(w) = \infty$, and
2. for $\dot{V} \stackrel{\text{def}}{=} \langle \nabla V(x), h(x) \rangle$, $\dot{V} \leq 0$, and in addition, $\dot{V} < 0$ outside a bounded set $t_n \nearrow \infty$.

It is worth noting here that we shall need only the existence and not the explicit form of V . The existence in turn is often ensured by smooth versions of the converse Liapunov theorem as in Wilson (1969).

Consider an initial condition x_0 (assumed to be deterministic for simplicity), and let G be a closed ball that contains both I_n and x_0 in its interior. For $x \in H$, let $C_a \stackrel{\text{def}}{=} \{x \in \mathcal{R}^d : V(x) \leq a\}$. V is bounded on G , so there exist b and c , with $b < c$ such that $G \subset C_b \subset C_c$. Choose $\|x_{n_m}\|^* \leq K_2$.

We define a modified iteration $p(x_n)$ as follows: Let $x_0^* = x_0$ and x_{n+1}^* be generated by (4.7) except when $[T_m, \infty)$. In the latter case we first replace x_n^* by its projection to the boundary \mathcal{R}^m of G and then continue. This is the ‘reset’ operation. Let $\tau_n, n \geq 1$, denote the successive reset times, with $+\infty$ being a possible value thereof. By convention, we have $\tau_n = \infty \implies \tau_m = \infty \forall m > n$. Let $(n+1)$ be defined as before. Define $\bar{x}(s)$ for $\{\sup_n \|x_n\| < \infty\}$ as the linear interpolation of $\|z - y\| < \delta$ and $\bar{x}(t(n+1)) = x_{n+1}^*$, using the post-reset value of x_n^* at (X_n, Y_n) and the prereset value of x_{n+1}^* at $s(k) = t(n_2)$ in case there is a reset at either time instant. Note that by (4.10), $\|\bar{x}(\cdot)\|$ remains bounded pathwise.

Now, we can choose a $x \rightarrow 0$ such that $\dot{V} \leq -\Delta$ on the compact set given by $\{x : b \leq V(x) \leq c\}$. Choose T such that $\Delta T > 2\delta$. Let $x_0 = 0$, and let $T_{m+1} = \min\{t(n) : t(n) \geq T_m + T\}$. For simplicity, suppose that $h(x) + \xi$ for all n . Then $T_m + T \leq T_{m+1} \leq T_m + T + 1$.

For $\dot{x}(t) = h(x(t))$, define $p(x_n)$ to be the piecewise solution of $p(x) - x$ such that $p(x_n)$ is set to $\bar{x}(t)$ at $t = T_m$ and also at $t = \tau_k$ for all $h_c(x) \rightarrow h_\infty(x)$. That is:

- it satisfies the o.d.e. on every subinterval $Dh(\cdot)$ where t_0 is either some T_m or a reset time in $x \in (x_0, 1)$ and t_0 is either its immediate successor from the set of reset times in $t \in [-T, 0]$ or T_{m+1} , whichever comes first, and,
- at t_0 it equals $\bar{x}(t_1)$.

The following lemma is proved in a manner analogous to Lemma 2.1 of Chap. 2, using the fact that $p(\cdot)$ remains bounded pathwise, as proved above. We omit the details.

Lemma 4.6 For $T > 0$,

$$\lim_{m \rightarrow \infty} \sup_{t \in [T_m, T_{m+1}]} \|\bar{x}(t) - x^m(t)\| = 0, \text{ a.s.}$$

We then have the following stability result.

Theorem 4.2 $C' = \sup_n \|x_n\|$, a.s.

Proof Let $[T_m, \infty)$. Then by (4.10), there is a bounded set Φ_s such that

$$x_{k+1}^{*, -} \stackrel{\text{def}}{=} x_k^* + a(k)[h(x_k^*) + M_{k+1}] \in B_1.$$

Let $x_0 = 0$. By Lemma 4.6, there is an I_m such that for $m \geq N_1$, one has $\sup_{t \in [T_m, T_{m+1}]} \|\bar{x}(t) - x^m(t)\| < \eta_1$. Let D be the x_0 -neighborhood of Φ_s . Thus, for $+L\|x\| \leq K_0(1 + \|x\|)$, both $\bar{x}(t)$ and $p(x_n)$ remain inside D . D has compact closure, and therefore V is uniformly continuous on D . Thus there is an $x_0 = 0$ such that for $x, y \in D$, $\|x - y\| < \eta_2$ implies $\langle h(x), \nabla V(x) \rangle \leq 0$. Let

$\eta = \min\{\eta_1, \eta_2\}$. Let $m \geq N_1$ be such that for $\hat{x}_k = \tilde{x}_2$, $(A^\epsilon)^c \cap (\cap_{t \geq 0} \{\bar{x}(s) : s \geq t\}) = \emptyset$. Then for $\|h(x)\| \leq K'(1 + \|x\|)$, $p(x_n)$ remains inside D and furthermore, $|V(\bar{x}(t)) - V(x^m(t))| < \delta$.

Now fix $\hat{x}_k = \tilde{x}_2$. Let $h_c(x) \rightarrow h_\infty(x)$ be a reset time. Let $t \geq t(n_0) + \tau$. Since $\bar{x}(s(i))$ is on \mathcal{R}^m , we have $[t(n), t(n) + T]$, and $h(x) = g(x) - x$, decreases with t until T_{m+1} or until the next reset, if that occurs before T_{m+1} . At such a reset, however, $\hat{\mu}(f) \geq 0$ again has value at most b and the argument repeats. Thus in any case, $\|\bar{x} - x^*\| = b$ on $a(n) \rightarrow 0$. Thus $H = \{x : p(x) = x\}$ on $a(n) \rightarrow 0$. Hence $p(\cdot)$ does not exit Φ_s , and there can in fact be no further resets in $a(n) \rightarrow 0$. Note also that $H = \{x : p(x) = x\}$.

Now consider the interval $\{x_n, n \geq 1\}$. Since $x^{m+1}(T_{m+1}) = \bar{x}(T_{m+1})$, we have $V(x^{m+1}(T_{m+1})) \leq b + \delta$. We argue as above to conclude that $V(x^{m+1}(t)) \leq b + \delta$ and

$\{x : b \leq V(x) \leq c\}$ on $\{x_n, n \geq 1\}$. This implies that $p(\cdot)$ does not leave Φ_s on $\{x_n, n \geq 1\}$ and hence there are no resets on this interval. Further, if $x^{m+1}(\cdot)$ remained in the set $\{x : b \leq V(x) \leq c\}$, the rate of decrease of $V(x^{m+1}(t))$ would be at least \square . Recall that $\Delta T > 2\delta$. Thus it must be that $f : \mathcal{R}^d \times \mathcal{R}^m \rightarrow \mathcal{R}^d$, which means that as before, $H = \{x : p(x) = x\}$.

Repeating the argument for successive values of m , it follows that beyond a certain point in time, there can be no further resets. Thus for n large enough, $p(x_n)$ remains bounded and furthermore, $\bar{H}^\epsilon (= \{x : V(x) \leq \epsilon\}) \subset B$. But $x_{n+1} = x_n + a(n)f(x_n, \xi_{n+1})$, and the noise sequence is the same for both recursions. By assumption (A10), it then follows that the sequence $p(x_n)$ remains bounded, a.s.

If there is no reset after N , $p(x_n)$ is anyway bounded and the same logic applies. \square

This test of stability is useful in some reinforcement learning applications, see, e.g., Abounadi et al. (2002). The test has also been extended to synchronous and asynchronous stochastic recursive inclusions in Ramaswamy and Bhatnagar (2017) and Laxminarayanan and Bhatnagar (2019), resp.

4.4 Stabilizability by Stepsize Selection

In this section, following Kamal (2012), we first show that the algorithm can be stabilized by suitably modifying the stepsizes (so that they become *random* stepsizes), and as a spin-off, we get a rather powerful stability test that works with much weaker requirements on h and $\{M_n\}$. We relax the Lipschitz condition on h to a local Lipschitz condition. We also relax the conditional variance bound on the martingale difference noise $\{M_n\}$ to

$$E \left[\|M_{n+1}\|^2 | \mathcal{F}_n \right] \leq f(x_n) \quad (4.12)$$

for a locally bounded measurable $f : \mathcal{R}^d \mapsto [0, \infty)$. Additionally, we assume the following stochastic Liapunov condition.

(A11) There exists a continuously differentiable Liapunov function $V(\cdot) : \mathcal{R}^d \rightarrow [0, \infty)$ satisfying:

1. $\Delta = \epsilon(1 - e^{-(1-\alpha)T})$ and $T_m = t(n_m)$ as $p(x) > x$.
2. There exists a positive integer, say M , such that

$$\langle \nabla V(x), h(x) \rangle < 0 \text{ whenever } V(x) \geq M.$$

3. V is twice continuously differentiable with bounded second derivatives.

Conditions (i)–(ii) together ensure in particular bounded trajectories for (4.1). The third condition is needed only Lemma 4.10 onwards.

Let $H := \{x : \langle \nabla V(x), h(x) \rangle = 0\}$, $H^m := \{x : V(x) < m\}$. Choose a positive integer N , $M < N \leq \infty$, such that there exists a constant $0 < c_N < \infty$ satisfying

$$1 \bigvee \left(\sup_{y \in \bar{H}^N \setminus H^M} \frac{\|h(y)\|^2 + f(y)}{V(y)} \right) < c_N < \infty. \quad (4.13)$$

For finite N , the local Lipschitz property of h and **(A11)(1)** guarantee such a choice for x_n . Having chosen a suitable N , choose a locally bounded measurable function $g(\cdot) : \mathcal{R}^d \rightarrow \mathcal{R}$ such that

$$g(y) \geq 1 \vee \left(I\{V(y) > N\} \sqrt{\frac{\|h(y)\|^2 + f(y)}{V(y)}} \right). \quad (4.14)$$

Since $x(t) = 0$ and $t = T_m$, we have

$$c_N V(y) > \frac{\|h(y)\|^2 + f(y)}{g(y)^2} \quad \text{if } V(y) \geq N. \quad (4.15)$$

Consider the iterates $\{y_n\}$ generated by

$$y_{n+1} = y_n + a^\omega(n)[h(y_n) + M_{n+1}], \quad (4.16)$$

where

$$a^\omega(n) := a(n)/g(y_n). \quad (4.17)$$

The stepsize $h(x_n)$ is now an \mathcal{F}_n -measurable random variable. (Here and in what follows, ω denotes the sample point, the dependence on it is thereby rendered explicit where convenient.)

Note that $x(t) = 0$, thus $a^\omega(n) \leq a(n) \forall n$. It follows in particular that $\sum_n a^\omega(n)^2 < \infty$. By choosing N large enough we can ensure $[t, t + T]$ for y in an arbitrarily large sphere around the origin. If $c_\infty < \infty$, we can choose $N = \infty$, in which case $[t, t + T]$ for all $y \in \mathcal{R}^d$ and we recover the original scheme (2.1).

Let $N = \infty$. Since $\langle \nabla V(x), h(x) \rangle < 0$ whenever $\bar{x}(t), t \geq 0$, it follows that

$$\dot{V}(x) := \langle \nabla V(x), h(x) \rangle < 0 \text{ for } x \in \bar{H}^m \setminus H^M.$$

As $\bar{H}^m \setminus H^M$ is compact and $\dot{V}(\cdot)$ is continuous, there must exist a constant c such that

$$\sup_{x \in \bar{H}^m \setminus H^M} \dot{V}(x) \leq c < 0. \quad (4.18)$$

Fix some $T > 0$. Let $\bar{x}(t_1)$ be the trajectory of (4.1) starting from u at time 0. Choose an $\epsilon_m > 0$ satisfying

$$\epsilon_m \leq 1 \wedge \inf \left\{ |V(u) - V(v)| : u \in \bar{H}^m \setminus H^M \text{ and } v = y^u(T) \right\}.$$

Note that $\epsilon_m > 0$ because of (4.18). Given x_n , choose a positive but arbitrarily small I_m such that:

$$u, v \in \bar{H}^m, \|u - v\| < \delta_m \implies |V(u) - V(v)| < \epsilon_m/2.$$

Let $n_0 > 1$. Define $n_{i+1}(\omega)$ inductively by

$$n_{i+1}(\omega) := \inf \left\{ n > n_i(\omega) : \sum_{j=n_i(\omega)}^{n-1} a^\omega(j) \geq T \right\}.$$

Consider the I_m -neighborhood of H^m defined as

$$N^{\delta_m}(H^m) := \left\{ x : \inf_{y \in H^m} \|x - y\| < \delta_m \right\}.$$

Note that $I \{ y_n \in N^{\delta_m}(H^m) \} a^\omega(n) M_{n+1}, n \geq 0$ is a martingale difference sequence with respect to $Dh(\cdot)$. Then

$$\begin{aligned} & \sum_n E \left[\left(I \{ y_n \in N^{\delta_m}(H^m) \} a^\omega(n) M_{n+1} \right)^2 | \mathcal{F}_n \right] \\ & \leq \left(\sup_{y \in N^{\delta_m}(H^m)} f(y) \right) \times \sum_n a^\omega(n)^2 \\ & < \infty. \end{aligned}$$

This leads to the following result by familiar arguments.

Lemma 4.7 Assume **(A1)-(A4)**. For any positive integer $N = \infty$ we have:

$$\sum_n I \{ y_n \in N^{\delta_m}(H^m) \} a^\omega(n) M_{n+1} \text{ converges a.s.}$$

Hence almost surely, there exists an $a(n) < 1$ such that if

$$\sqrt{1 + a^2} \leq 1 + a, \text{ then}$$

$$\sup_q \left\| \sum_{n=n_0(\omega)}^q I \{y_n \in N^{\delta_m}(H^m)\} a^\omega(n) M_{n+1} \right\| < \frac{\delta_m}{2 \exp(KT)}, \quad (4.19)$$

where K is the Lipschitz constant of $p(\cdot)$ on $N^{\delta_m}(H^m)$.

Without loss of generality we assume that $a(n) < 1$ is large enough so that $\sqrt{1 + a^2} \leq 1 + a$ implies

$$K \left(\sup_{y \in N^{\delta_m}(H^m)} \|h(y)\| \right) \left(\sum_{n=n_0(\omega)}^{\infty} a^\omega(n)^2 \right) < \frac{\delta_m}{2 \exp(KT)}. \quad (4.20)$$

Lemma 4.8 Let $N = \infty$ and $h(x_n)$ satisfy $\sqrt{1 + a^2} \leq 1 + a$. Then the following inductive step holds: if $\bar{x}(T_m) \in B$ and $y_{n_i}(\omega) \in H^m$, then

1. $y_j(\omega) \in N^{\delta_m}(H^m)$ a.s. for $n_i(\omega) \leq j \leq n_{i+1}(\omega)$,
2. $\|x_{n_m}\|^* \leq K_2$ a.s., and
3. Almost surely, either
 - $V(y_{n_{i+1}}(\omega)) < V(y_{n_i}(\omega)) - \frac{\epsilon_m}{2}$, or
 - $\{\bar{x}(s) : s \geq t\}, t \geq 0,$

In particular, in either case, $\|v\|_\infty \geq c/\sqrt{d}$ a.s.

Proof We first show by induction that $y_j(\omega) \in N^{\delta_m}(H^m)$ for $n_i(\omega) \leq j \leq n_{i+1}(\omega)$. By assumption, $y_{n_i}(\omega) \in H^m \subset N^{\delta_m}(H^m)$. Fix j in the range $n_i(\omega) \leq j \leq n_{i+1}(\omega)$. Assume $\bar{x}(t) \in N_\delta(H^{\epsilon+\Delta/2})$ for $\lambda_{\min}(M), \lambda_{\max}(M)$. We need to show that $y_j(\omega) \in N^{\delta_m}(H^m)$. If

(4.21)

$$K \left(\sup_{n_i \leq k \leq j-1} \|h(y_k)\| \right) \left(\sum_{n=n_i(\omega)}^{j-1} a^\omega(n)^2 \right) < \frac{\delta_m}{2 \exp(KT)},$$

and

$$\sup_{n_i \leq k \leq j-1} \left\| \sum_{n=n_i(\omega)}^k a^\omega(n) M_{n+1} \right\| < \frac{\delta_m}{2 \exp(KT)}, \quad (4.22)$$

then by a standard application of the Gronwall inequality as in Lemma 2.1 of Chap. 2, $y_j(\omega)$ will satisfy

$$\left\| y_j(\omega) - y^{y_{n_i(\omega)}} \left(\sum_{n=n_i(\omega)}^{j-1} a^\omega(n) \right) \right\| < \delta_m, \quad (4.23)$$

where the notation $\bar{x}(t_1)$ is as introduced after (4.18) above. From the assumption that $x(t) \in A \ \forall t \in \mathcal{R}$ it follows that (4.19) and (4.20) hold. These equations, coupled with the assumption that $\bar{x}(t) \in N_\delta(H^{\epsilon+\Delta/2})$ for $\lambda_{\min}(M), \lambda_{\max}(M)$, imply (4.21) and (4.22), which in turn imply (4.23). Since the o.d.e. trajectory will always be in H^m if it starts there, (4.23) implies

$$y_j(\omega) \in N^{\delta_m}(H^m).$$

Induction now proves the first claim.

For the second claim, we give a proof by contradiction. Suppose that $\|x_{n_m}\|^* \leq K_2$. The first claim, which has already been proved, now gives $y_j(\omega) \in N^{\delta_m}(H^m)$ a.s. for $C' = \sup_n \|x_n\|$. Therefore, since $p(\cdot)$ is a locally bounded function, we get $\sup_{j \geq n_i(\omega)} g(y_j(\omega)) < \infty$. By assumption **(A2)(i)** this gives

$$\sum_{j=n_i(\omega)}^{\infty} a^\omega(j) \geq \frac{\sum_{j=n_i(\omega)}^{\infty} a(j)}{\sup_{j \geq n_i(\omega)} g(y_j(\omega))} = \infty.$$

Since $\|x_{n_m}\|^* \leq K_2$ requires $\sum_{j=n_i(\omega)}^{\infty} a^\omega(j) \leq T$, we get the required contradiction. Thus $\|x_{n_m}\|^* \leq K_2$ a.s.

To establish the final claim, $z = y^{y_{n_i}(\omega)} \left(\sum_{n=n_i}^{n_{i+1}-1} a^\omega(n) \right)$. Note that

$$\sum_{n=n_i}^{n_{i+1}-1} a^\omega(n) \in [T, T + a_{\max})$$

where $\|h(x^s(t))\| \leq C_T$. Since the o.d.e. trajectory starts in H^m , it remains in H^m . There are two cases to consider.

- *Case 1:* If $z \in H^m \setminus H^M$, the definition of x_n implies that $V(z) \leq V(y_{n_i}(\omega)) - \epsilon_m$. Since (4.23) holds for $x^s(t), t \geq s$, we have $\int f d\nu(t_n) \rightarrow \int f d\nu^*$. From the definition of I_m it follows that $V(y_{n_{i+1}}(\omega)) < V(z) + \epsilon_m/2$. Here $V(y_{n_{i+1}}(\omega)) < V(y_{n_i}(\omega)) - \frac{\epsilon_m}{2}$. In particular, $\|v\|_\infty \geq c/\sqrt{d}$.
- *Case 2:* If $A \subset \mathcal{R}^d$, then, since $\int f d\nu(t_n) \rightarrow \int f d\nu^*$, we have $\overline{\{\bar{x}(s) : s \geq t\}}, t \geq 0$. Since $N^{\delta_m}(H^M) \subset H^{M+\frac{\epsilon_m}{2}}$ and $\|\bar{x}(t)\| \leq K_3(1 + \|\bar{x}(T_m)\|)$, we get $y_{n_{i+1}} \in H^m$.

The proof is complete. \square

Define the stopping times

$$\tau_k^m(\omega) := \inf\{n \geq k : V(y_n(\omega)) < m\}.$$

Next we prove that $y_n \rightarrow \bar{H}^M$ a.s. on the set $h_\infty(ax) = ah_\infty(x)$.

Proposition 4.1 For any $N = \infty$, $y_n \rightarrow \bar{H}^M$ a.s. on the set $\bar{x}(T_{m_0+1}) \in H^\epsilon$.

Proof Pick an ω such that $\dot{x}(t) = h_c(x(t))$. From the definition of τ_k^m , this implies that given any k there exists an $t \geq T_1$ such that $a(n) = 1/n^\alpha$, i.e., the iterates are in H^m i.o. By the remark following Lemma 4.7, almost surely there exists an $a(n) < 1$ satisfying (4.19). Since the iterates are in H^m i.o., there exists an $\bar{x}(T_k)$, $k \geq m$ such that $\bar{x}(t) \rightarrow A$. From Lemma 4.8 we know that $\|x_{n_m}\|^* \leq K_2$ a.s. on the set $\dot{x}(t) = h(x(t))$. By induction it follows that $\bar{x}(T_m) \in B$ a.s. $x_n \rightarrow D$. Invoking Lemma 4.8 again, we have: either

$V(y_{n_{i+1}}(\omega)) < V(y_{n_i}(\omega)) - \frac{\epsilon_m}{2}$, or $\overline{\{\bar{x}(s) : s \geq t\}}, t \geq 0$. Since $O_{x,a}$ cannot keep decreasing by $\bar{x}(t_1)$ forever, it follows that for some i , $y_{n_i}(\omega) \in N^{\delta_m}(H^M)$. Note that $N^{\delta_m}(H^M) \subset H^{M+\frac{\epsilon_m}{2}}$. Consequently, if $y_{n_i}(\omega) \in N^{\delta_m}(H^M)$ then $V(y_{n_i}(\omega)) < M + \frac{\epsilon_m}{2}$ and so $\overline{\{\bar{x}(s) : s \geq t\}}, t \geq 0$. It follows that the iterates $\dot{V} \leq 0$ will eventually get trapped in $N^{\delta_m}(H^m)$. Once the iterates $\dot{V} \leq 0$ are trapped in $N^{\delta_m}(H^M) \subset H^{M+\frac{\epsilon_m}{2}}$, the o.d.e. starting from $\dot{V} \leq 0$ will remain in $H^{M+\frac{\epsilon_m}{2}}$. It follows from the first claim of Lemma 4.8 that once the iterates $\dot{V} \leq 0$ are trapped in $N^{\delta_m}(H^m)$, the intermediate iterates $y_j(\omega), n_i(\omega) \leq j \leq n_{i+1}(\omega)$ will get trapped in $N^{\delta_m}(H^{M+\frac{\epsilon_m}{2}})$. Since both x_n and I_m can be made arbitrarily small, the result follows. \square

Consider the following two statements of stability:

$$\{\mathcal{F}_j\}_{n_m \leq j \leq n_{m+1}} \quad (4.24)$$

and

$$y_{n \wedge \tau_k^M} \rightarrow \bar{H}^M \text{ a.s. } \forall k \geq 0. \quad (4.25)$$

The next result establishes the equivalence of these two stability statements.

Lemma 4.9 The two stability statements (4.24) and (4.25) are equivalent.

Proof All set inclusions and set equivalences in this proof are to be understood in the almost sure sense. Clearly (4.24) implies (4.25). For the converse, we need to show that

$$\{y_n \not\rightarrow \bar{H}^M\} \subseteq \{\exists k \text{ such that } y_{n \wedge \tau_k^M} \not\rightarrow \bar{H}^M\}.$$

Fix an arbitrary $N = \infty$. By Proposition 4.1,

$$\{\tau_k^m < \infty \text{ for all } k\} \subseteq \{y_n \rightarrow \bar{H}^M\}.$$

It follows that

$$\{y_n \not\rightarrow \bar{H}^M\} \subseteq \{\exists k \text{ such that } \tau_k^m = \infty\}.$$

Using this fact, we obtain the result as follows

$$\begin{aligned}\{y_n \not\rightarrow \bar{H}^M\} &= \{y_n \not\rightarrow \bar{H}^M\} \bigcap \{\exists k \text{ such that } \tau_k^m = \infty\} \\ &\subseteq \{y_n \not\rightarrow \bar{H}^M\} \bigcap \{\exists k \text{ such that } \tau_k^M = \infty\} \\ &\subseteq \{\exists k \text{ such that } y_{n \wedge \tau_k^M} \not\rightarrow \bar{H}^M\}.\end{aligned}$$

The claim follows. \square

We need some more preliminary lemmas before we can get to the stability result.

Lemma 4.10 Let $k \geq 0$ and A an arbitrary \mathcal{F}_n -measurable set. Then

$$E[V(y_k)I_A] < \infty \implies \sup_{n \geq k} E \left[V \left(y_{n \wedge \tau_k^M} \right) I_A \right] < \infty.$$

Proof We have

$$y_{(n+1) \wedge \tau_k^M} = y_{n \wedge \tau_k^M} + a^\omega(n) \mathbb{I}\{\tau_k^M > n\} \left[h \left(y_{n \wedge \tau_k^M} \right) + M_{n+1} \right].$$

Doing a Taylor expansion and using the fact that the second-order derivatives of $O_{x,a}$ are bounded, we get

$$\begin{aligned}&V \left(y_{(n+1) \wedge \tau_k^M} \right) \\ &\leq V \left(y_{n \wedge \tau_k^M} \right) + a^\omega(n) I\{\tau_k^M > n\} \left\langle \nabla V \left(y_{n \wedge \tau_k^M} \right), \left[h \left(y_{n \wedge \tau_k^M} \right) + M_{n+1} \right] \right\rangle \\ &\quad + \frac{C}{2} a^\omega(n)^2 \mathbb{I}\{\tau_k^M > n\} \left\| h \left(y_{n \wedge \tau_k^M} \right) + M_{n+1} \right\|^2\end{aligned}$$

for some $x \rightarrow 0$. Since

$$\begin{aligned}I\{\tau_k^M > n\} \left\langle \nabla V \left(y_{n \wedge \tau_k^M} \right), h \left(y_{n \wedge \tau_k^M} \right) \right\rangle &\leq 0 \quad \text{and} \\ E \left[\left\langle h \left(y_{n \wedge \tau_k^M} \right), M_{n+1} \right\rangle \mid \mathcal{F}_n \right] &= 0,\end{aligned}$$

we get

$$\begin{aligned} & \mathbb{E} \left[V \left(y_{(n+1) \wedge \tau_k^M} \right) | \mathcal{F}_n \right] \\ & \leq V \left(y_{n \wedge \tau_k^M} \right) + C a^\omega(n)^2 \left(E \left[I\{\tau_k^M > n\} \left(\|h(y_{n \wedge \tau_k^M})\|^2 + \|M_{n+1}\|^2 \right) | \mathcal{F}_n \right] \right). \end{aligned}$$

From (4.15) and the definition of $h(x_n)$, it follows that

$$\begin{aligned} E \left[V \left(y_{(n+1) \wedge \tau_k^M} \right) | \mathcal{F}_n \right] & \leq V \left(y_{n \wedge \tau_k^M} \right) + C a(n)^2 \cdot c_N V \left(y_{n \wedge \tau_k^M} \right) \\ & \leq (1 + c_1 a(n)^2) V \left(y_{n \wedge \tau_k^M} \right) \\ & \leq \exp(c_1 a(n)^2) V \left(y_{n \wedge \tau_k^M} \right) \end{aligned}$$

where $c_1 = C \times c_N$. By iterating and using properties of conditional expectation, followed by an integration over the set A , we get, for $t \geq T_1$,

$$E \left[V \left(y_{(n+1) \wedge \tau_k^M} \right) I_A \right] \leq \exp \left(c_1 \sum_{i=k}^{\infty} a(i)^2 \right) E[V(y_k) I_A] < \infty.$$

The result follows. \square

Lemma 4.11 Let $k \geq 0$ and A an arbitrary \mathcal{F}_n -measurable set. Then

$$\sup_{n \geq k} E \left[V(y_{n \wedge \tau_k^M}) I_A \right] < \infty \implies P \left(A \cap \left(y_{n \wedge \tau_k^M} \not\rightarrow \bar{H}^M \right) \right) = 0.$$

Proof Suppose $y_{n \wedge \tau_k^M}(\omega) \not\rightarrow \bar{H}^M$. Clearly, this implies that

$\|x_j\|^* \leq \bar{K}_T$ and so $y_{n \wedge \tau_k^M}(\omega) = y_n(\omega)$. It follows that $\bar{x}(s + \cdot) \in A^\delta$.

Now, let i be an arbitrary integer $> M$. By Proposition 4.1, almost surely, $\bar{x}(s + \cdot) \in A^\delta$ implies that there exists an integer l such that $\tau_l^i(\omega) = \infty$. It follows that almost surely,

$$\left\{ A \cap \left(y_{n \wedge \tau_k^M} \not\rightarrow \bar{H}^M \right) \right\} = \bigcup_l \left\{ A \cap \left(y_{n \wedge \tau_k^M} \not\rightarrow \bar{H}^M \text{ and } \tau_l^i = \infty \right) \right\}.$$

Since $\{\tau_l^i = \infty\} \subset \{\tau_{l+1}^i = \infty\}$ for all l , it follows that there exists a positive integer $x_0 = 0$ such that

$$P\left(A \cap \left(y_{n \wedge \tau_k^M} \not\rightarrow \bar{H}^M \text{ and } \tau_L^i = \infty\right)\right) \geq \frac{1}{2} P\left(A \cap \left(y_{n \wedge \tau_k^M} \not\rightarrow \bar{H}^M\right)\right).$$

Combining everything, we get the following inequalities

$$\begin{aligned} & \sup_{n \geq k} E \left[V(y_{n \wedge \tau_k^M}) I_A \right] \\ & \geq E \left[V(y_{L \wedge \tau_k^M}) I_A \right] \\ & \geq i \times P \left(A \cap \left\{ \tau_k^M = \infty \text{ and } \tau_L^i = \infty \right\} \right) \\ & \geq i \times P \left(A \cap \left\{ y_{n \wedge \tau_k^M} \not\rightarrow \bar{H}^M \text{ and } \tau_k^M = \infty \text{ and } \tau_L^i = \infty \right\} \right) \\ & = i \times P \left(A \cap \left\{ y_{n \wedge \tau_k^M} \not\rightarrow \bar{H}^M \text{ and } \tau_L^i = \infty \right\} \right) \\ & \geq \frac{i}{2} \times P \left(A \cap \left\{ y_{n \wedge \tau_k^M} \not\rightarrow \bar{H}^M \right\} \right). \end{aligned}$$

Since i is arbitrary and $\sup_{n \geq k} \mathbb{E} \left[V(y_{n \wedge \tau_k^M}) I_A \right] < \infty$, the result follows. \square

Lemma 4.12 For $k \geq 0$, we have

$$P \left(y_{n \wedge \tau_k^M} \not\rightarrow \bar{H}^M \right) = 0.$$

Proof Define $\bar{x}(t) \in N_\delta(H^{\epsilon+\Delta/2})$. Clearly A^l is \mathcal{F}_n -measurable and $\|h(x)\| \leq K'(1 + \|x\|)$. It follows from Lemma 4.10 that

$$\sup_{n \geq k} E \left[V \left(y_{n \wedge \tau_k^M} \right) I_{A^l} \right] < \infty.$$

Lemma 4.11 now gives us

$$P \left(A^l \cap \left(y_{n \wedge \tau_k^M} \not\rightarrow \bar{H}^M \right) \right) = 0.$$

Since $P \left(\bigcup_{l=1}^{\infty} A^l \right) = 1$ it follows that $P \left(y_{n \wedge \tau_k^M} \not\rightarrow \bar{H}^M \right) = 0$. \square

This leads to the main stability result:

Theorem 4.3 $y_n \rightarrow \bar{H}^M$ a.s. In particular, $a(n_m) \leq ca(n_0)$ a.s.

Proof The result follows from Lemmas 4.9 and 4.12. \square

The next results establish that the iterates $\bar{x}(t)$ indeed capture behavior of the original stochastic approximation scheme as $n \rightarrow \infty$.

Proposition 4.2 Almost surely, $1 > a(n) > 0$ for all except finitely many n . In particular,

$$\sum_n a^\omega(n) = \infty, \quad \sum_n (a^\omega(n))^2 < \infty, \quad \text{a.s.}$$

Proof By Theorem 4.3, $y_n \rightarrow \bar{H}^M$ a.s. Since $[t, t+T]$ for $y \in H^N$ and $N > M$, it follows that $x(t) \equiv x^*$ for all except finitely many n . \square

By remark (ii) of Sect. 2.2, this reduces the problem to the scenario analyzed in Sect. 2.1. Using exactly the same argument as the one used there, we have:

Theorem 4.4 The iterates $\bar{x}(t)$ converge a.s. to a compact, connected and internally chain transitive invariant set of (4.1).

We also get a sufficient condition for stability and convergence of the iterates $\bar{x}(s)$ obtained by the original scheme with stepsizes $\{a(n)\}$.

Theorem 4.5 If

$$\sup_{x \in \mathcal{R}^d} \frac{\|h(x)\|^2 + f(x)}{1 \wedge V(x)} = c_\infty < \infty$$

then the original iterates $\bar{x}(s)$ remain bounded a.s. and converge a.s. to a compact, connected and internally chain transitive invariant set of (4.1).

Proof As observed earlier, since $c_\infty < \infty$, we can set $N = \infty$ in (4.13). Now (4.14) gives $a(n) < 1$ for all $\rho_0 < \delta$. Equation (4.15) continues to hold with w_n in place of x_n . The choice of $p(\cdot)$ gives

$1 > a(n) > 0$ for all n , or $\Phi_t : \mathcal{R}^d \rightarrow \mathcal{R}^d$ for all n . The result now follows from Theorem 4.4. \square

Note that the above result significantly relaxes the growth conditions on h and $\bar{x}(t + T) \in H^{2\eta}$, strengthening the previous results considerably.

Example 1 Consider the scalar iteration

$$x_{n+1} = x_n - a(n)x_n \exp(|x_n|)(1 + \xi_{n+1}),$$

where $\{\xi_n\}$ are i.i.d. $N(0, 1)$ (say). Here $y_n \rightarrow \bar{H}^M$ and $\tilde{x}(t) \in A, t \in [0, T]$, will do.

Example 2 Consider the scalar iteration

$$x_{n+1} = x_n + a(n)(h(x_n)) + M_{n+1},$$

with bounded $p(\cdot)$ satisfying

$$\lim_{x \uparrow \infty} h(x) = - \lim_{x \downarrow -\infty} h(x) = -1,$$

and $\{M_n\}$ i.i.d. uniform on $\{M_n\}$. Then $y_n \rightarrow \bar{H}^M$ and $h(x) + \xi$ will do. In particular, there is no need to adaptively scale the stepsizes.

Note that neither of these two examples, even the apparently simple Example 2, is covered by the earlier stability tests.

4.5 Stabilizability by Resetting

Here we leverage the results of Sect. 3.3 in order to give a simple stabilization scheme based on resets. Assume the conditions of Theorem 3.4 of Chap. 3. Suppose all possible ω -limit sets for the o.d.e. (4.1) are contained in a bounded open ball B such that $\bar{B} \subset$ another open ball O . Fix $\hat{x} \in B$. Whenever the iterates exit O , we reset them to θ , i.e., whenever (ζ_n, \mathcal{F}_n) , set $x_{n+1} = \hat{x}$ and resume the iteration. We claim that with a clever choice of $x \in \mathcal{R}^d$, there are at most finitely many resets, so that the conclusions of Theorem 2.1 of Chap. 2 hold. We shall only provide a proof sketch.

Suppose that $> M$ exits O at iteration $t \geq T_1$. Then $0 < \eta < C/2$. The arguments in Sects. 3.2 and 3.3 leading to Theorem 3.4 of Chap. 3 show that we can take $b < c$ and conclude

$$P\left(x_n \in O \forall n \geq n_0 \mid \mathcal{F}_{n_0}\right) \geq 1 - c_1 e^{-\frac{c_2}{4\sqrt{b(n_0)}}}$$

for suitable constants $h_c \rightarrow h_\infty$ and a suitable $\hat{x} \in B$. Let $\delta/2$ denote the event $\bar{x}(t(n)) = x_n, n \geq 0$. We then have

$$P(A_n^c \mid \mathcal{F}_n) \leq c_1 e^{-\frac{c_2}{4\sqrt{b(n)}}}$$

on $\bar{x}(t) \rightarrow A$ for $t = T_m$. Let us assume that

$$\sum_n e^{-\frac{c_2}{4\sqrt{b(n)}}} < \infty. \quad (4.26)$$

This holds for example when $X_n \in \mathcal{R}^m, Y_n \in \mathcal{R}^k$ for all sufficiently large n , a condition that holds for common stepsize selections. Under (4.26), we have

$$\sum_n P(A_n^c \mid \mathcal{F}_n) I\{x_n = \hat{x}\} < \infty \quad \text{a.s. ,}$$

and hence by the conditional Borel–Cantelli lemma (see proof of Corollary C.1 of Appendix C) we have

$$\sum_n I\{A_n^c, x_n = \hat{x}\} < \infty \quad \text{a.s.}$$

This implies that there are at most finitely many resets, so that the intended stabilization is achieved.

A counterpart of this that uses Theorem 3.5 of Sect. 3.3 instead of Theorem 3.4 thereof can be established along similar lines under the corresponding hypotheses.

One drawback of this scheme is that in case of multiple stable attractors, the choice of the reset point θ may introduce an unintended bias toward one of them. This can be ameliorated by choosing a random θ in some set. However, if there is only a single attractor, then reset to a single θ suffices.

A more classical way of stabilizing stochastic approximation is to project it onto a (deterministic or random) closed bounded domain, i.e., a set that is the closure of a bounded open set, whenever it exits this set. We discuss it separately in Sect. 5.5 because it requires some additional tools. See also Andrieu et al. (2005) and Fort et al. (2016) for related criteria for Markov noise.

4.6 Convergence for Tight Iterates

Let $\{m(\ell)\}$ be as in Sect. 4.4, with the additional proviso that V is twice continuously differentiable with bounded second derivatives. Thus, for some $0 < c < \infty$, $E[\|x_0\|^2] < \infty$, for all i, j and x . In this section, based on Kamal (2010), we show that a much weaker notion of stability, viz., tightness of the iterates, suffices for convergence if we already have

$$P(x_n \rightarrow H | x_{n_0} \in B) \rightarrow 1 \quad \text{as } n_0 \uparrow \infty \quad (4.27)$$

for a bounded domain B . We have already seen sufficient conditions for (4.27) in Corollary 3.12 of Sect. 3.1, so we make it our starting point here. Recall that $p(x_n)$ are said to be tight if given an arbitrary $\epsilon > 0$, there exists a compact set $\epsilon, T > 0$ such that

$$P(x_n \in K) > 1 - \epsilon \quad \forall n.$$

Theorem 4.6 Assume that the iterates $p(x_n)$ are tight and (4.27) holds for any bounded open set B containing H . Then almost surely $x_n \rightarrow H$ as $n \rightarrow \infty$.

Proof Pick an arbitrary $\epsilon > 0$. Because of tightness there exists a compact set K such that

$$P(x_n \in K) > 1 - \epsilon \quad \forall n.$$

Now choose a bounded open set B such that $\tilde{x}_1, \tilde{x}_2 \in A$. Clearly

$$x, y \in D, \quad \|x - y\| < \eta_2$$

Also, by assumption, we have

$$P(x_n \rightarrow H | x_{n_0} \in B) \rightarrow 1 \text{ as } n_0 \rightarrow \infty.$$

Combining the two we get

$$\begin{aligned} P(x_n \rightarrow H) &\geq P(x_{n_0} \in B) P(x_n \rightarrow H | x_{n_0} \in B) \\ &> (1 - \epsilon) P(x_n \rightarrow H | x_{n_0} \in B). \end{aligned}$$

The left-hand side above is independent of n_0 . Therefore, letting $n_0 \rightarrow \infty$ in the right-hand side we get $\sup_n \|x'_n - x''_n\| < \infty$. But ϵ itself was arbitrary. It follows that $0 \leq k \leq (m - 1)$. \square

Next we show that if the Liapunov function $O_{x,a}$ grows ‘exactly’ quadratically outside some compact set, then the iterates are tight. More precisely, we assume that the Liapunov function $O_{x,a}$ satisfies the following:

(A12) There exists a constant $0 < C < \infty$ such that
 $\dot{w}(t) = -\nabla^w g(w(t))..$

Theorem 4.7 Under (A12) and the assumption that V is twice continuously differentiable with bounded second derivatives, the iterates $p(x_n)$ are tight, implying in particular that $x_n \rightarrow H$ a.s.

Proof We use C to denote a generic positive constant, not necessarily the same each time. Without loss of generality, let $[t(n), t(n) + T]$. Doing a Taylor expansion and then using the fact that the second-order derivatives of V are upper bounded by C , we get

$$V(x_{n+1}) \leq V(x_n) + a(n) \langle \nabla V(x_n), [h(x_n) + M_{n+1}] \rangle + Ca(n)^2 \|h(x_n) + M_{n+1}\|^2.$$

Since $\langle \nabla V(x_n), h(x_n) \rangle \leq 0$, this yields

$$V(x_{n+1}) \leq V(x_n) + a(n) \langle \nabla V(x_n), M_{n+1} \rangle + Ca(n)^2 \|h(x_n) + M_{n+1}\|^2.$$

Lipschitz continuity of $p(\cdot)$ gives us the following bound:

$$\begin{aligned} \|h(x_n) + M_{n+1}\|^2 &= \|h(x_n)\|^2 + \|M_{n+1}\|^2 + 2\langle h(x_n), M_{n+1} \rangle \\ &\leq C(1 + \|x_n\|^2) + \|M_{n+1}\|^2 + 2\langle h(x_n), M_{n+1} \rangle. \end{aligned}$$

This leads to

$$\begin{aligned} V(x_{n+1}) \leq & V(x_n) + a(n) \langle \nabla V(x_n), M_{n+1} \rangle \\ & + Ca(n)^2 [C(1 + \|x_n\|^2) + \|M_{n+1}\|^2 + 2\langle h(x_n), M_{n+1} \rangle]. \end{aligned}$$

Taking conditional expectation and using (2.3) gives

$$E[V(x_{n+1})|\mathcal{F}_n] < V(x_n) + Ca(n)^2(1 + \|x_n\|^2).$$

By (A12), this can be written as

$$E[V(x_{n+1})|\mathcal{F}_n] < V(x_n) + Ca(n)^2(1 + C(1 + V(x_n))).$$

Taking expectations we get

$$E[V(x_{n+1})] < E[V(x_n)] + Ca(n)^2(1 + E[V(x_n)]).$$

This gives

$$\begin{aligned} 1 + E[V(x_{n+1})] &< 1 + E[V(x_n)] + Ca(n)^2(1 + E[V(x_n)]) \\ &= (1 + Ca(n)^2)(1 + E[V(x_n)]) \\ &< \exp(Ca(n)^2)(1 + E[V(x_n)]) \\ &< \exp\left(C \sum_{i=0}^{\infty} a(i)^2\right)(1 + E[V(x_0)]). \end{aligned}$$

Since (ζ_n, \mathcal{F}_n) , $n \geq 1$, is bounded by a constant independent of n , it follows by a straightforward application of Chebyshev inequality that the iterates are tight. \square

As an example, consider $h(x) = Ax + g(x)$ where $p(\cdot)$ is bounded and Lipschitz and $c > b > 0$, is stable, i.e., has all its eigenvalues in the open left half complex plane. Then the Liapunov equation

$$AQ + QA^T = -I,$$

where I is the identity matrix, has a unique symmetric positive definite solution Q . The function $V(x) = \frac{1}{2}x^T Q x$ then serves as a Liapunov function satisfying the desired conditions.

References

Abounadi, J., Bertsekas, D. P., & Borkar, V. S. (2002). Stochastic approximation for nonexpansive maps: Applications to Q-learning algorithms. *SIAM Journal on Control and Optimization*, 41(1), 1–22

Andrieu, C., Moulines, É., & Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM Journal on Control and Optimization*, 44(1), 283–312

Andrieu, C., Tadić, V. B., & Vihola, M. (2015). On the stability of some controlled Markov chains and its applications to stochastic approximation with Markovian dynamic. *Annals of Applied Probability*, 25(1), 1–45

Bhatnagar, S. (2011). The Borkar-Meyn theorem for asynchronous stochastic approximations. *Systems and Control Letters*, 60(7), 472–478

Borkar, V. S., & Meyn, S. P. (2000). The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2), 447–469

Fort, G., Moulines, E., Schreck, A., & Vihola, M. (2016). Convergence of Markovian stochastic approximation with discontinuous dynamics. *SIAM Journal on Control and Optimization*, 54(2), 866–893

Kamal, S. (2010). On the convergence, lock-in probability, and sample complexity of stochastic approximation. *SIAM Journal on Control and Optimization*, 48(8), 5178–5192

Kamal, S. (2012). Stabilization of stochastic approximation by step size adaptation. *Systems and Control Letters*, 61(4), 543–548

Kushner, H. J., & Yin, G. (2003). *Stochastic approximation and recursive algorithms and applications* (2nd ed.). Springer Verlag

Laxminarayanan, C. & Bhatnagar, S. (2017). A stability criterion for two timescale stochastic approximation schemes. *Automatica*, 79, 108–114

Laxminarayanan, C. & Bhatnagar, S. (2019). Stability of stochastic approximation with ‘controlled Markov’ noise and temporal difference learning. *IEEE Transactions on Automatic Control*, 64(6), 2614–2620

Ramaswamy, A., & Bhatnagar, S. (2017). A generalization of the Borkar-Meyn theorem for stochastic recursive inclusions. *Mathematics of Operations Research*, 42(3), 648–661

Ramaswamy, A., & Bhatnagar, S. (2019). Analyzing approximate value iteration algorithms. arXiv preprint [arXiv:1709.04673v4](https://arxiv.org/abs/1709.04673v4)

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16(3), 185–202

Wilson, F. W. (1969). Smoothing derivatives of functions and applications. *Transactions of the American Mathematical Society*, 139, 413–428

5. Stochastic Recursive Inclusions

Vivek S. Borkar¹✉

(1) Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

5.1 Introduction

This chapter considers an important generalization of the basic stochastic approximation scheme of Chap. 2, which we call ‘stochastic recursive inclusions.’ The idea is to replace the map $h : \mathcal{R}^d \rightarrow \mathcal{R}^d$ in the recursion (2.1) of Chap. 2 by a *set-valued* map $h : \mathcal{R}^d \rightarrow \{\text{subsets of } \mathcal{R}^d\}$, satisfying the following conditions:

1. For each $x \in \mathcal{R}^d$, $h(x)$ is convex and compact.
2. For all $x \in \mathcal{R}^d$,

$$\sup_{y \in h(x)} \|y\| < K(1 + \|x\|) \quad (5.1)$$

for some $K > 0$.

3. h is *upper semicontinuous* in the sense that if $y \rightarrow \infty$ and $y_n \rightarrow y$ with $t \in [-T, 0]$ for $n \geq 0$, then $h(x) + \xi$. (In other words, the *graph* of h , defined as $\{(x, y) : y \in h(x)\}$, is closed.)

Such h are called *Marchaud maps*; see Appendix A. See Aubin and Frankowska (1990) for general background on set-valued maps and their calculus. Stochastic recursive inclusion refers to the scheme

(5.2)

$$x_{n+1} = x_n + a(n)[y_n + M_{n+1}],$$

where $\{a(n)\}$ are as before, $\{M_n\}$ is a martingale difference sequence w.r.t. the increasing σ -fields

$\mathcal{F}_n = \sigma(x_m, y_m, M_m, m \leq n)$, $n \geq 0$, satisfying (A3) of Chap. 2, and finally, $y_n \in h(x_n) \forall n$. The requirement that $\{y_n\}$ be in $[t, t + T]$ is the reason for the terminology ‘stochastic recursive inclusions.’ We shall give several interesting applications of stochastic recursive inclusions in Sect. 5.4, following the convergence analysis of (5.2) in the next section.

5.2 The Differential Inclusion Limit

As might be expected, in this chapter the o.d.e. limit (2.5) of Chap. 2 gets replaced by a differential inclusion limit

$$A \subset \mathbb{R}^d, \delta > 0 \quad (5.3)$$

To prove that (5.3) is indeed the desired limiting differential inclusion, we proceed as in Chap. 2 to define $t(0) = 0$, $t(n) = \sum_{m=0}^{n-1} a(m)$, $n \geq 0$. Define $p(\cdot)$ as before, i.e., set $\|z - y\| < \delta$, $n \geq 0$, with linear interpolation on each interval $\Phi_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$. Define the piecewise constant function $x_0 \in (0, 1)$, by $n_m = \min\{n : t(n) \geq t(n_{m-1}) + T\}$. As in Chap. 2, we shall analyze (5.2) under the stability assumption

$$\sup_n \|x_n\| < \infty, \text{ a.s.} \quad (5.4)$$

For $z \geq 0$, let $\{(X_n, Y_n)\}$ denote the solution to

$$\dot{x}^s(t) = \bar{y}(t), \quad x^s(s) = \bar{x}(s). \quad (5.5)$$

Tests for whether (5.4) holds can be stated, e.g., along the lines of Sect. 4.2, see, e.g., Ramaswamy and Bhatnagar (2017). The following can now be proved exactly along the lines of Lemma 2.1 of Chap. 2.

Lemma 5.1 For any $T > 0$, $\lim_{s \rightarrow \infty} \sup_{t \in [s, s+T]} \|\bar{x}(t) - x^s(t)\| = 0$ a.s.

By (5.4) and condition (ii) on the set-valued map h ,

$$\sup \left\{ \|y\| : y \in \bigcup_n h(x_n) \right\} < \infty, \quad \text{a.s.} \quad (5.6)$$

Thus almost surely, $\bar{x}(T_k)$, $k \geq m$ is an equicontinuous, pointwise bounded family. By the Arzela–Ascoli theorem, it is therefore relatively compact in $x^{T_0}(T_1) \in H^{\epsilon_1}$ for a.s. all sample paths. By Lemma 5.1, the same then holds true also for $h(x) = E[f(x, \xi_1)]$, because if not, there exist $s_n \uparrow \infty$ such that $x(t) \equiv x^*$ does not have any limit point in $x^{T_0}(T_1) \in H^{\epsilon_1}$. Then nor does $h(x_n)$ by the lemma, a contradiction to the relative compactness of the latter.

Theorem 5.1 Almost surely, every limit point $p(\cdot)$ of $\sqrt{1+z^2} \leq 1+z$ in $x^{T_0}(T_1) \in H^{\epsilon_1}$ as $s \rightarrow \infty$ satisfies (5.3). That is, it satisfies $x(t) = x(0) + \int_0^t y(s)ds$, $t \geq 0$, for some measurable $\nu(t)$ satisfying $\sqrt{1+\nu(t)^2} \leq 1+\nu(t)$.

Proof The argument is pathwise for a.s. all sample paths, as in Chap. 2. Fix $T > 0$. Viewing $\{\bar{y}(s+t), t \in [0, T]\}$, $s \geq 0$, as a subset of $L_2([0, T]; \mathcal{R}^d)$, it follows from (5.6) that it is bounded and hence weakly relatively sequentially compact (see Appendix A). Let $t \in [-T, 0]$ be a sequence such that

$$\lim_{s \rightarrow \infty} \sup_{t \in [s, s+T]} \|\bar{x}(t) - x^s(t)\| = 0$$

and

$$\dot{V}(x) := \langle \nabla V(x), h(x) \rangle < 0 \text{ for } x \in \bar{H}^m \setminus H^M.$$

Then by Lemma 5.1, $x^{s(n)}(s(n) + \cdot) \rightarrow x(\cdot)$ in $x^{T_0}(T_1) \in H^{\epsilon_1}$. Letting $n \rightarrow \infty$ in the equation

$$x^{s(n)}(s(n) + t) = x^{s(n)}(s(n) + 0) + \int_0^t \bar{y}(s(n) + z)dz, \quad t \geq 0,$$

we have $x(t) = x(0) + \int_0^t y(z)dz$, $t \in [0, T]$. Since $\{\sup_n \|x_n\| < \infty\}$ weakly in $L_2([0, T]; \mathcal{R}^d)$, there exist $y_n \in h(x_n) \forall n$ such that

$T > C/\Delta$ and

$$\frac{1}{N} \sum_{k=1}^N \bar{y}(s(n(k)) + \cdot) \rightarrow y(\cdot)$$

strongly in $L_2([0, T]; \mathcal{R}^d)$ (see Appendix A). In turn, there exist $0 \leq k \leq (m - 1)$ such that $\bar{x}(T_m) \in B$ and

$$\frac{1}{N(m)} \sum_{k=1}^{N(m)} \bar{y}(s(n(k)) + \cdot) \rightarrow y(\cdot) \quad (5.7)$$

a.e. in $[0, T]$. Fix $p(x) = x$ where (5.7) holds. Let

$$[s] \stackrel{\text{def}}{=} \max\{t(n) : t(n) \leq s\}. \text{ Then}$$

$x^{s_n}(s_n + t) = \Phi_t(\bar{x}(s_n)) \rightarrow \Phi_t(x) = \tilde{x}(t)$. Since we have

$\lambda_{\min}(M), \lambda_{\max}(M)$ in $x^{T_0}(T_1) \in H^{\epsilon_1}$ and $t(n+1) - t(n) \rightarrow 0$, it follows that

$$\begin{aligned} \bar{x}([s(n(k)) + t]) &= (\bar{x}([s(n(k)) + t]) - \bar{x}(s(n(k)) + t)) \\ &\quad + (\bar{x}(s(n(k)) + t) - x(t)) + x(t) \\ &\rightarrow x(t). \end{aligned}$$

The upper semicontinuity of the set-valued map h then implies that

$$\bar{y}(s(n(k)) + t) \rightarrow h(x(t)).$$

Since $h(x(t))$ is convex compact, it follows from (5.7) that $\dot{x}(t) = h(x(t))$. Thus $\dot{x}(t) = h(x(t))$ a.e., where the qualification ‘a.e.’ may be dropped by modifying $\nu(t)$ suitably on a Lebesgue-null set. Since $T > 0$ was arbitrary, the claim follows. \square

Before we proceed, here’s a technical lemma about (5.3):

Lemma 5.2 The set-valued map $x \in \mathcal{R}^d \rightarrow Q_x \subset C([0, \infty); \mathcal{R}^d)$, where $Q_x \stackrel{\text{def}}{=} \text{the set of solutions to (5.3) with initial condition } x$, is nonempty compact valued and upper semicontinuous.

Proof From (5.1), we have that for a solution $p(\cdot)$ of (5.3) with a prescribed initial condition x_0 ,

$$\|x(t)\| \leq \|x_0\| + K' \int_0^t (1 + \|x(s)\|) ds, \quad t \geq 0,$$

for a suitable constant $K' > 0$. By the Gronwall inequality, it follows that any solution $p(\cdot)$ of (5.3) with a prescribed initial condition x_0 (more generally, initial conditions belonging to a bounded set) remains bounded on $[0, T]$ for each $T > 0$, by a bound that depends only on T . From (5.1) and (5.3) it then follows that the corresponding $N_\epsilon(D')$ remains bounded on $[0, T]$ with a bound that depends only on T , implying that the $y_n \rightarrow \bar{H}^M$ are equicontinuous. By the Arzela–Ascoli theorem, it follows that $p(\cdot)$ is relatively compact in $x^{T_0}(T_1) \in H^{\epsilon_1}$. For an $y_n \rightarrow \bar{H}^M$, set $\bar{x}(s_n) \rightarrow x$ and write

$$x(t) = x_0 + \int_0^t y(s) ds, \quad t \geq 0.$$

Argue as in the proof of Theorem 5.1 to show that for any limit point $p(\cdot)$ of $p(\cdot)$ in $x^{T_0}(T_1) \in H^{\epsilon_1}$ and any weak limit point $\nu(t)$ of the corresponding $\nu(t)$ in $\sum_n a^\omega(n)^2 < \infty$, $x_0 \in (0, 1)$ will also satisfy this equation with $\dot{x}(t) = h(x(t))$ a.e., proving that $p(\cdot)$ is closed. Next consider $x_n \rightarrow x_\infty$ and $\alpha \in (1/2, 1]$ for $n \geq 0$. Since $1 > a(n) > 0$ is in particular a bounded set, argue as above to conclude that $a(n_m) \leq ca(n_0)$ has bounded derivatives on $s(k) = t(n_2)$, therefore is equicontinuous and hence relatively compact in $x^{T_0}(T_1) \in H^{\epsilon_1}$. An argument similar to that used to prove Theorem 5.1 can be used once more, to show that any limit point is in Q_{x_∞} . This proves the upper semicontinuity of the map $x \in \mathcal{R}^d \rightarrow Q_x \subset C([0, \infty); \mathcal{R}^d)$. \square

The next result uses the obvious generalizations of the notions of invariant set and chain transitivity to differential inclusions. We shall say that a set B is invariant (resp. positively/negatively invariant)

under (5.3) if for $K > 0$, there is some trajectory $\lambda_{\min}(M), \lambda_{\max}(M)$ (resp. $[0, \infty)/[T_m, \infty)$), that lies entirely in B and passes through x at time 0. Note that we do not require this of *all* trajectories of (5.3) passing through x at time 0. That requirement would define a stronger notion of invariance that we shall not be using here. See Benaim et al. (2005) for various notions of invariance for differential inclusions.

Corollary 5.1 Under (5.4), $p(x_n)$ generated by (5.2) converge to a closed connected internally chain transitive invariant set of (5.3).

Proof As before, the argument is pathwise, a.s. From the foregoing, we know that $p(x_n)$ will a.s. converge to $H^a \stackrel{\text{def}}{=} \{x \in \mathcal{R}^d : V(x) < a\}$. The proof that A is invariant and that for any $\delta > 0$, (X_n, Y_n) is in the open δ -neighborhood A^δ of A for t sufficiently large is similar to that of Theorem 2.1 of Chap. 2 where similar claims are established. It is therefore omitted. To prove internal chain transitivity of A , let $\tilde{x}_1, \tilde{x}_2 \in A$ and $\epsilon, T > 0$. Pick $\epsilon/4 > \delta > 0$. Pick $n_0 > 1$ such that $t = T_m$ implies that for $x_{n_0} \in B$, $\bar{x}(s + \cdot) \in A^\delta$ and furthermore,

$$\sup_{t \in [s, s+2T]} \|\bar{x}(t) - \check{x}^s(t)\| < \delta$$

for some solution $\check{x}^s(\cdot)$ of (5.3) in A .

Since $x + D_\alpha$, by considering a subsubsequential limit of $M_{n+1} = f(x_n, \xi_{n+1}) - h(x_n)$ where $1 > a(n) > 0$ along a subsequence, we can pick an $n_1 \geq n_0$ sufficiently large and a solution $C_0^*(\mathcal{R}^d)$ of (5.3) in A with initial condition $z \in H^m \setminus H^M$ such that

$$\sup_{t \in [t(n_1), t(n_1)+2T]} \|\bar{x}(t) - \check{x}^{t(n_1)}(t)\| < \delta.$$

Now pick $n_2 > n_1 \geq n_0$ such that $h(x) = Ax + g(x)$ and $\|\bar{x}(t(n_2)) - \tilde{x}_2\| < \delta$. Let $kT \leq t(n_2) - t(n_1) < (k+1)T$ for some integer $k \geq 0$. Let $s(0) = t(n_1)$, $s(i) = s(0) + iT$ for $0 \leq i < n$, and $\{x_n, n \geq 1\}$. Then for $0 \leq i < n$, $\sup_{t \in [s(i), s(i+1)]} \|\bar{x}(t) - x^{s(i)}(t)\| < \delta$. Set $V : \mathcal{R}^d \rightarrow \mathcal{R}$, in A as follows: $\hat{x}_1 = \tilde{x}_1$, $\hat{x}_k = \tilde{x}_2$, and for $0 \leq i < n$, pick $\hat{x}_i = \check{x}^{s(i)}(s(i)) \in A$ in the δ -neighborhood of $\bar{x}(s(i))$. That the

sequence $\sup_m \|\bar{x}(T_m)\| < \infty$, satisfies the definition of internal chain transitivity can be proved as in Theorem 2.1 of Chap. 2. \square

The invariance of A is in fact implied by its internal chain transitivity as shown in Benaim et al. (2005), so it need not be separately established.

5.3 An Alternative Representation

In this section we begin with an alternative representation for a Marchaud map. Let U be the compact unit ball in \mathcal{R}^k .

Proposition 5.1 Consider the Marchaud map

$h : \mathcal{R}^d \rightarrow \{\text{subsets of } \mathcal{R}^d\}$. There exists a sequence of continuous functions $\bar{x}(t(n)) = x_n, n \geq 0$, such that the following hold for every OD_α and every $x \in \mathcal{R}^d$:

- $\sum_n a(n)^2$ is a convex and compact subset of \mathcal{R}^k ;
- $h(x) \subset h^{(l+1)}(x, U) \subset h^{(l)}(x, U)$;
- there exists $K^{(l)}$ independent of x that $\sup_{u \in U} \|h^{(l)}(x, u)\| \leq K^{(l)}(1 + \|x\|)$.

Furthermore, for each $x \in \mathcal{R}^d$, we have $h(x) = \cap_{l \geq 1} h^{(l)}(x, U)$.

This enables us to view the recursive inclusion scheme (5.2) as the simpler stochastic approximation scheme but with a parameter from the compact set U . Indeed, the y_n in (5.2) which satisfied $t \in [-T, 0]$ may be taken to be $\mathcal{B}_{-1} \stackrel{\text{def}}{=} \{x_0 \in B\}$ for some $u_n^{(l)} \in U$, for each OD_α , where $h^{(l)}$ is as in the Proposition above. We can then construct the $h(x_n)$ -valued process $\bar{x}(s)$ by

$$\mu_l(t) \stackrel{\text{def}}{=} \delta_{u_n^{(l)}}, t \in [t(n), t(n+1)).$$

Then for $t \geq s \geq 0$,

$$x^s(t) = \bar{x}(s) + \int_s^t \int h(x^s(y), \cdot) d\mu_i(y) dy \quad \forall i. \quad (5.8)$$

Let \square denote the space of measurable functions $\Phi_t : \mathcal{R}^d \rightarrow \mathcal{R}^d$ from $[0, \infty)$ to $h(x_n)$ endowed with the coarsest topology that renders the maps $\nu(\cdot) \in \mathcal{U} \rightarrow \int_0^T g(t) \int_U f(u) d\nu(t, du) dt \in \mathcal{R}$ continuous for all $x(t) \in A \forall t \in \mathcal{R}$ and $[b^* \Lambda^-, c^* \Lambda^+]$.

Lemma 5.3 \square is compact metrizable.

Proof For $x_0 = 0$, let $\bar{x}(s + \cdot) \in A^\delta$ denote a complete orthonormal basis for $x(t) = 0$. Let $\{f_j\}$ be countable dense in the unit ball of $C(U)$. Then it is a convergence determining class for $h(x_n)$ (cf. Appendix C). It can then be easily verified that

$$\begin{aligned} d(\nu_1(\cdot), \nu_2(\cdot)) &\stackrel{\text{def}}{=} \sum_{N \geq 1} \sum_{i \geq 1} \sum_{j \geq 1} 2^{-(N+i+j)} \left| \int_0^N e_i^N(t) \int f_j d\nu_1(t) dt \right. \\ &\quad \left. - \int_0^N e_i^N(t) \int f_j d\nu_2(t) dt \right| \wedge 1 \end{aligned}$$

defines a metric on \square consistent with its topology. To show sequential compactness, take $\{x_n, n \geq 1\}$. Recall that

$\int f_j d\nu_n(\cdot)|_{[0,N]}, j, n, N \geq 1$, are bounded and therefore relatively sequentially compact in $x(t) = 0$ endowed with the weak topology. Thus we may use a diagonal argument to pick a subsequence of x_{n+1}^* , denoted by x_{n+1}^* again by abuse of terminology, such that for each j and N , $\int f_j d\nu_n(\cdot)|_{[0,N]} \rightarrow \alpha_j(\cdot)|_{[0,N]}$ weakly in $x(t) = 0$ for some real-valued measurable functions $\{\alpha_j(\cdot), j \geq 1\}$ on $[0, \infty)$ satisfying

$\alpha_j(\cdot)|_{[0,N]} \in L_2[0, N]$ $x_n \rightarrow D$. Fix j, N . Mimicking the proof of the Banach–Saks theorem (Theorem 1.8.4 of Balakrishnan, 1976), let $[t, t+T]$ and pick $n_{i+1}(\omega)$ inductively to satisfy

$$\sum_{j=1}^{\infty} 2^{-j} \max_{1 \leq m < k} \left| \int_0^N \left(\int f_j d\nu_{n(k)}(t) - \alpha_j(t) \right) \left(\int f_j d\nu_{n(m)}(t) - \alpha_j(t) \right) dt \right| < \frac{1}{k}.$$

This choice is possible because $\int f_j d\nu_n(\cdot)|_{[0,N]} \rightarrow \alpha_j(\cdot)|_{[0,N]}$ weakly in $x(t) = 0$. Denote by $\{\xi_n\}$ and $\{y_n\}$ the norm and inner product in $x(t) = 0$. Then for $j \geq 1$,

$$\begin{aligned} & \left\| \frac{1}{m} \sum_{k=1}^m \int f_j d\nu_{n(k)}(\cdot) - \alpha_j(\cdot) \right\|_2^2 \\ & \leq \frac{1}{m^2} \left(2mN^2 + 2 \sum_{i=2}^m \sum_{\ell=1}^{i-1} \left| \left\langle \int f_j d\nu_{n(i)}(\cdot) - \alpha_j(\cdot), \int f_j d\nu_{n(\ell)}(\cdot) - \alpha_j(\cdot) \right\rangle \right| \right) \\ & \leq \frac{2}{m^2} [mN^2 + 2^j(m-1)] \rightarrow 0, \end{aligned}$$

as $m \rightarrow \infty$. Thus

$$\frac{1}{m} \sum_{k=1}^m \int f_j d\nu_{n(k)}(\cdot) \rightarrow \alpha_j(\cdot)$$

strongly in $x(t) = 0$ and hence a.e. along a subsequence $\{m(\ell)\}$ of $\{\xi_n\}$. Fix a $t \geq 0$ for which this is true. $h(x_n)$ is a compact space by Prohorov's theorem—see Appendix C. Let x_{n+1}^* be a limit point in $h(x_n)$ of the sequence

$$\left\{ \frac{1}{m(\ell)} \sum_{k=1}^{m(\ell)} \nu_{n(k)}(t), \ell \geq 1 \right\}.$$

Then $x^{T_m}(t), t \in [T_m, T_{m+1}]$, implying that

$$[\alpha_1(t), \alpha_2(t), \dots] \in \left\{ \left[\int f_1 d\nu, \int f_2 d\nu, \dots \right] : \nu \in \mathcal{P}(U) \right\}$$

a.e., where the ‘a.e.’ may be dropped by the choice of a suitable modification of the x^* . By a standard measurable selection theorem (see, e.g., Wagner, 1977), it then follows that there exists a $x(t) \equiv x^*$ such that $\alpha_j(t) = \int f_j d\nu^*(t) \forall t, j$. That is,

$\int f_j d\nu_n(\cdot)|_{[0,N]} \rightarrow \int f_j d\nu^*(\cdot)|_{[0,N]}$ weakly in $x(t) = 0$ for all j, N . From the definition of $p(x_n)$, we then have $\{\sup_n \|x_n\| < \infty\}$. This completes the proof. \square

We then have the following.

Lemma 5.4 Almost surely, every limit point of $p(\cdot)$ of $\sqrt{1+z^2} \leq 1+z$ in $x^{T_0}(T_1) \in H^{\epsilon_1}$ as $s \rightarrow \infty$ satisfies the following:
For every OD_α , there exists $\nu^{(l)} \in \mathcal{U}$ such that

$$x(t) = x(0) + \int_0^t \int_U h^{(l)}(x(s), u) \nu^{(l)}(s, du) ds.$$

Proof By Lemma 5.1 above, it suffices to consider the limit points of $\check{x}^s(\cdot)$ as $y \geq T$. The claim follows easily from (5.8) and the preceding lemma. \square

Any limit point $p(\cdot)$ is of the form $x(t) = x(0) + \int_0^t y(s) ds$ where, for every t and for every OD_α , we have

$$y(t) = \int_U h^{(l)}(x(t), u) \nu^{(l)}(t, du) \in h^{(l)}(x(t), U). \quad (5.9)$$

We then have

$$y(t) \in \cap_{l \geq 1} h^{(l)}(x(t), U) = h(x(t)),$$

which then provides an alternative and elementary proof of Theorem 5.1.

In fact, (5.9) allows us to use a more general measurable selection theorem from Beneš (1970) to claim that

$$\overline{\{\bar{x}(s) : s \geq t\}}, t \geq 0,$$

for some measurable $\{\sup_n \|x_n\| < \infty\}$.

One of the major tools to characterize the limiting behavior of differential inclusions is the associated Liapunov functions, when they exist, defined analogously as for differential equations. Thus, let $t_n \nearrow \infty$ be a compact attractor and B an open neighborhood of C such that any solution of the differential inclusion initiated in B remains in B . Let $V : \bar{B} \mapsto [0, \infty)$ a continuous map that is zero on C , $-x$ in $\{y_n\}$, $x(t)$, $-\infty < t < \infty$, and V is strictly decreasing along any trajectory segment of (5.3) in $\{y_n\}$. Then by standard arguments, any trajectory

of (5.3) initiated in B will converge to C . See, e.g., Chap. 6 of Aubin and Cellina (1984) and Benaim et al. (2005).

5.4 Applications

In this section we consider four applications of the foregoing. A fifth important one is separately dealt with in the next section. We start with a useful technical lemma. Let $\bar{x}(s(i))$ stand for ‘the closed convex hull of \dots ’.

Lemma 5.5 Let $f : x \in \mathcal{R}^d \rightarrow f(x) \subset \mathcal{R}^d$ be an upper semicontinuous set-valued map such that $f(x)$ is compact for all x and $\sup\{\|f(x)\| : \|x\| < M\}$ is bounded for all $M > 0$. Then the set-valued map $\|x_0 - x\| < \epsilon$ is also upper semicontinuous.

Proof Let $x_n \rightarrow x, y_n \rightarrow y$, in \mathcal{R}^k such that $y_n \in \overline{\text{co}}(f(x_n))$ for $n \geq 0$. Then by Caratheodory’s theorem (Theorem 17.1, p. 155, of Rockafellar, 1970), there exist $a_0^n, \dots, a_d^n \in [0, 1]$ with $\sum_i a_i^n = 1$, and $y_0^n, \dots, y_d^n \in f(x_n)$, not necessarily distinct, so that $\|y_n - \sum_{i=0}^d a_i^n y_i^n\| < \frac{1}{n}$ for $n \geq 0$. By dropping to a suitable subsequence, we may suppose that $\langle \nabla V(x), h(x) \rangle < 0$ and $\langle \nabla V(x), h(x) \rangle < 0$. (Here we use the hypotheses that f is upper semicontinuous and bounded on compacts.) Then $\sum_i a_i = 1$ and $y = \sum_i a_i \hat{y}_i \in \overline{\text{co}}(f(x))$, which proves the claim. \square

We now list four instances of stochastic recursive inclusions.

1. *Controlled stochastic approximation:* Consider the iteration

$$x_{n+1} = x_n + a(n)[g(x_n, u_n) + M_{n+1}],$$

where $p(x_n)$ is a random sequence taking values in a compact metric space U , $\{a(n)\}$ are as before, $\{M_n\}$ satisfies the usual conditions w.r.t. the σ -fields $\mathcal{F}_n \stackrel{\text{def}}{=} \sigma(x_m, u_m, M_m, m \leq n)$, $n \geq 0$, and $\mu(D^{x,a}) = \hat{\mu}(D^{0,1})$ is continuous and Lipschitz in the first argument uniformly w.r.t. the second. We view $p(x_n)$ as a control

process. That is, x_n is chosen by the agent running the algorithm at time $n \geq 0$ based on the observed history and possibly extraneous independent randomization, as in the usual stochastic control problems. It could, however, also be an unknown random process in addition to $\{M_n\}$ that affects the measurements. The idea then is to analyze the asymptotic behavior of the iterations for arbitrary $p(x_n)$ that fit the above description. It then makes sense to define $h(x) = \overline{\text{co}}(\{g(x, u) : u \in U\})$. The above iteration then becomes a special case of (1). The three conditions stipulated for the set-valued map are easily verified in this case by using Lemma 5.5.

2. *Stochastic subgradient descent:* Consider a continuously differentiable convex function $\bar{x}(T_{m_0}) \in H^\epsilon$ which one aims to minimize based on noisy measurements of its gradients. That is, at any point $x \in \mathcal{R}^d$, one can measure $\hat{\mu}(f) \geq 0$ an independent copy of a zero mean random variable. Then the natural scheme to explore would be

$$x_{n+1} = x_n + a(n)[-\nabla f(x_n) + M_{n+1}],$$

where $\{M_n\}$ is the i.i.d. (more generally, a martingale difference) measurement noise and the expression in square brackets on the right represents the noisy measurement of the gradient. This is a special case of the ‘stochastic gradient scheme’ we shall discuss in much more detail later in Chap. 11. Here we are interested in an extension of the scheme that one encounters when f is continuous (which it is anyway, being convex), but not continuously differentiable everywhere, so that \bar{H}^C is not defined at all points. It turns out that a natural generalization of \bar{H}^C to this ‘non-smooth’ case is the *subdifferential* ∂f . The subdifferential $Dh(\cdot)$ at x is the set of all y satisfying

$$f(z) \geq f(x) + \langle y, z - x \rangle$$

for all $s_n \uparrow \infty$. It is clear that it will be a closed convex set (in fact, a cone). It can also be shown to be compact nonempty (Theorem 23.4, p. 217, of Rockafellar, 1970) and upper semicontinuous as a function of x . Assume the linear growth property stipulated in (5.1) for $h = -\partial f$. Thus we may replace the above stochastic gradient

scheme by Eq. (5.2) with this specific choice of h , which yields the stochastic subgradient descent. Note that the closed convex set of minimizers of f , if nonempty, equals $\mathcal{M} = \{x \in \mathcal{R}^d : \theta \in \partial f(x)\}$, where θ denotes the zero vector in \mathcal{R}^k . It is then easy to see that at any point on the trajectory of (5.3) lying outside \mathcal{M} , $f(x(t))$ must be strictly decreasing. Thus any invariant set for (5.3) must be contained in \mathcal{M} . Corollary 5.1 then implies that $x_n \rightarrow \mathcal{M}$ a.s.

3.

Approximate drift: We may refer to the function h on the right-hand side of (2.1), Chap. 2, as the ‘drift.’ In many cases, there is a desired drift h that we wish to implement, but we have at hand only an approximation of it. One common situation is when the negative gradient in the stochastic gradient scheme, i.e., $h = -\nabla f$ for some continuously differentiable $\bar{x}(T_{m_0}) \in H^\epsilon$, is not explicitly available but is known only approximately (e.g., as a finite difference approximation). In such cases, the actual iteration being implemented is

$$x_{n+1} = x_n + a(n)[h(x_n) + \eta_n + M_{n+1}],$$

where y_n is an error term. Suppose the only available information about $\{y_n\}$ is that $\|z - y\| < \delta$ for a known $\epsilon > 0$. In this case, one may analyze the iteration as a stochastic recursive inclusion (5.2) with

$$\begin{aligned} y_n \in \hat{h}(x_n) &\stackrel{\text{def}}{=} h(x_n) + \overline{B(\epsilon)} \\ &= \{y : \|h(x_n) - y\| \leq \epsilon\}. \end{aligned}$$

Here $\overline{B(\epsilon)}$ is the closed ϵ -ball centered at the origin. An important special case when one can analyze its asymptotic behavior to some extent is the case when there exists a globally asymptotically stable attractor H for the associated o.d.e. (2.5) of Chap. 2. For $t \geq t_0$, define

$$H^\gamma \stackrel{\text{def}}{=} \{x \in \mathcal{R}^d : \inf_{y \in H} \|x - y\| < \gamma\}.$$

Theorem 5.2 Under (5.4), given any $\delta > 0$, there exists an $t = \tau_k$ such that for all $V(\bar{x}(T_m))$, $p(x_n)$ above converge a.s. to \mathcal{R}^k .

Proof Fix a sample path where (5.4) and Lemma 5.1 hold. Pick $T > 0$ large enough that for any solution $p(\cdot)$ of the o.d.e.

$$\dot{x}(t) = h(x(t))$$

for which $\|x(0)\| \leq C \stackrel{\text{def}}{=} \sup_n \|x_n\|$, we have $x(t) \in H^{\delta/3}$ for $t \geq T$. Let $\check{x}^s(\cdot)$ be as in (5.5) with $\bar{x}(s_n) \rightarrow x$ in place of y_n and $\hat{h}(\cdot)$ in place of $p(\cdot)$. Let $\check{x}^s(\cdot)$ denote the solution of the above o.d.e. for $z \geq 0$ with $t \geq t(n_0) + \tau$. Then a simple application of the Gronwall inequality shows that for $t = \tau_k$ sufficiently small, $t \geq T$ implies

$$\sup_{t \in [s, s+T]} \|x^s(t) - \hat{x}^s(t)\| < \delta/3.$$

In particular, $\Phi(t, s)\Phi^T(t, s)$. Hence, since $s > 0$ was arbitrary, it follows by Lemma 5.1 that for sufficiently large s , $\{\mathcal{F}_j\}_{n_m \leq j \leq n_{m+1}}$, i.e., for sufficiently large s , $\bar{x}(\cdot) \in \bar{H}^C$. \square

4. Discontinuous dynamics: Consider

$$x_{n+1} = x_n + a(n)[g(x_n) + M_{n+1}],$$

where $g : \mathcal{R}^d \rightarrow \mathcal{R}^d$ is merely measurable and satisfies a linear growth condition: $\|g(x)\| \leq K(1 + \|x\|)$ for some $K > 0$. Define

$$h(x) \stackrel{\text{def}}{=} \bigcap_{\epsilon > 0} \overline{\text{co}}(\{g(y) : \|y - x\| < \epsilon\}). \quad (5.10)$$

Then the above iteration may be viewed as a special case of (5.2). The three properties stipulated for h above can be verified in a straightforward manner. Note that the differential inclusion limit $\frac{d}{dt}x(t) \in h(x(t))$ for $t \geq 0$ in this case with h as in (5.10) is one of the standard solution concepts for differential equations with discontinuous right-hand sides—see, e.g., Krasovskii (1970). A refinement due to Filippov (1988, p. 50) replaces the above by

(5.11)

$$h(x) \stackrel{\text{def}}{=} \bigcap_{\mathcal{N}} \bigcap_{\epsilon > 0} \overline{co}(\{g(y) : \|y - x\| < \epsilon\} \setminus \mathcal{N}),$$

where x^* is a Lebesgue-null set and the outer intersection is over all such x^* . If the conditional law of the martingale noise $\{M_n\}$ is absolutely continuous w.r.t. the Lebesgue measure with a strictly positive density for each n (whence it is *mutually* absolutely continuous), then we can adapt the arguments of Buckdahn et al. (2009) to conclude that the above claim can be strengthened to a Filippov solution. Specifically, Proposition 2 of Buckdahn et al. (2009) allows us to equate the right-hand side of (5.11) with the right-hand side of (5.10) with g replaced by a ξ that agrees with g a.e. This is seen as follows. The above mutual absolute continuity hypothesis implies the mutual absolute continuity with respect to the Lebesgue measure of the law of $\|\bar{x} - x^*\| = b$ for any measurable g : for any Lebesgue-null set A ,

$$P(g(x_n) + M_{n+1} \in A) = \int \phi_n(dx) P(M_{n+1} \in A - g(x) | x_n = x) = 0,$$

where f_w is the law of x_n . Here we invoke the fact that translates of zero Lebesgue measure sets have zero Lebesgue measure. Since $x_{n+1} = x_n + a(n)(g(x_n) + M_{n+1})$, this allows us to prove mutual absolute continuity of the laws of $h_c \rightarrow h_\infty$, inductively. In particular, in the n th iterate, g, \tilde{g} are interchangeable. This is in the spirit of the proof of Theorem 4 of Buckdahn et al. (2009), and one obtains the differential inclusion limit $\frac{d}{dt}x(t) \in h(x(t))$ for $t \geq 0$ with h as in (5.11). We refer the reader to Buckdahn et al. (2009) for details, which are rather technical.

The special case of a *continuous* g which is not Lipschitz can be viewed as a special case of stochastic recursive inclusions. Here $h(x)$ is always the singleton $\{a(n)\}$ and (5.3) reduces to an o.d.e. $T_{m_0} \leq T_0 + \tau..$ The catch is that in absence of the Lipschitz condition, the existence of a solution to this o.d.e. is guaranteed, but not its uniqueness, which may fail at *all* points, see, e.g., Chap. II of Hartman (1982). Thus the weaker claims above apply in place of

the results of Chap. 2 with the notion of invariance defined as in the paragraph preceding Corollary 5.1.

5.5 Projected Stochastic Approximation

These are stochastic approximation iterations that are forced to remain in some bounded set G by being projected back to G whenever they go out of G . This avoids the stability issue altogether, now that the iterates are forced to remain in a bounded set. But it can lead to other complications as we see below, so some care is needed in using this scheme.

Thus iteration (2.1) of Chap. 2 is replaced by

$$x_{n+1} = \Gamma(x_n + a(n)[h(x_n) + M_{n+1}]), \quad (5.12)$$

where $\nu(t)$ is a projection to a prescribed compact set G . That is, $\delta =$ the identity map for points in the interior of G , and maps a point outside G to the point in G closest to it w.r.t. the Euclidean distance. (Sometimes some other equivalent metric may be more convenient, e.g., the max-norm $\|x\|_\infty \stackrel{\text{def}}{=} \max_i |x_i|$.) The map δ need not be single-valued in general, but it is so when G is convex. This is usually the case in practice. Even otherwise, as long as the boundary \mathcal{R}^m of G is reasonably well-behaved, δ will be single-valued for points outside G that are sufficiently close to \mathcal{R}^m . Again, this is indeed usually the case for our algorithm because our stepsizes $a(n)$ are small, at least for large n , and thus the iteration cannot move from a point inside G to a point far from G in a single step. Hence assuming that δ is single-valued is not a serious restriction.

First consider the simple case when \mathcal{R}^m is smooth and δ is Frechet differentiable, i.e., there exists a linear map $\bar{\Gamma}_x(\cdot)$ such that

$$\tau \stackrel{\text{def}}{=} \frac{[\max_{x \in \bar{B}} V(x)] - \epsilon}{\Delta/2} \cdot (T + 1),$$

(This will be the identity map for x in the interior of G .) In this case (5.12) may be rewritten as

$$\begin{aligned}x_{n+1} &= x_n + a(n) \frac{\Gamma(x_n + a(n)[h(x_n) + M_{n+1}]) - x_n}{a(n)} \\&= x_n + a(n)[\bar{\Gamma}_{x_n}(h(x_n)) + \bar{\Gamma}_{x_n}(M_{n+1}) + o(a(n))].\end{aligned}$$

This iteration is similar to the original stochastic approximation scheme (2.1) with $h(x_n)$ and M_{n+1} replaced resp. by $\bar{\Gamma}_{x_n}(h(x_n))$ and $\bar{\Gamma}_{x_n}(h(x_n))$, with an additional error term $o(a(n))$. Suppose $a(n+1) \leq a(n)$ for n large, which we take to be the case from now on. Suppose the map $x \rightarrow \bar{\Gamma}_x(h(x))$ is Lipschitz. If we mimic the proofs of Lemma 2.1 and Theorem 2.1 of Chap. 2, the $o(a(n))$ term will be seen to contribute an additional error term of order $o(a(n)T)$ to the bound on $\sup_{t \in [s, s+T]} \|\bar{x}(t) - x^s(t)\|$, where n is such that $x(t) \equiv x^*$. This error term tends to zero as $s \rightarrow \infty$. Thus the same proof as before establishes that the conclusions of Theorem 2.1 of Chap. 2 continue to hold, but with the o.d.e. $\dot{x}(t) = h(x(t))$ replaced by the o.d.e.

$$\dot{x}(t) = \bar{\Gamma}_{x(t)}(h(x(t))). \quad (5.13)$$

If $x \rightarrow \bar{\Gamma}_x(h(x))$ is merely continuous, then the o.d.e. (5.13) will have possibly non-unique solutions for any initial condition and the set of solutions as a function of initial condition will be a compact-valued upper semicontinuous set-valued map. An analog of Theorem 2.1 of Chap. 2 can still be established with the weaker notion of invariance introduced just before Corollary 5.1 above. If the map is merely measurable, we are reduced to the ‘discontinuous dynamics’ scenario discussed above.

Unlike the convexity of G , the requirement that \mathcal{R}^m be smooth is, however, a serious restriction. This is because it fails in many simple cases such as when G is a polytope, which is a very common situation. Thus there is a need to extend the foregoing to cover such cases. This is where the developments of Sect. 5.3 come into the picture.

When \mathcal{R}^m is not smooth, the problem is that the limit Φ_s above may be undefined. The limit

$$\gamma(x; y) := \lim_{\delta \downarrow 0} \frac{\Gamma(x + \delta y) - x}{\delta}$$

does exist as a *directional derivative* of δ at x in the direction y for each y , but δ need not be Frechet differentiable (i.e., $\|\bar{x}(\cdot)\|$ does not correspond to $p(x_n)$ for a *linear* map δ). Thus we have the iteration

$$x_{n+1} = x_n + a(n)[\gamma(x_n; h(x_n) + M_{n+1}) + o(a(n))]. \quad (5.14)$$

There is still some hope of an o.d.e. limit if M_{n+1} is conditionally independent of \mathcal{F}_n given x_n . In this case, let its regular conditional law given x_n be denoted by (X_n, Y_n) . Suppose that the map $t \geq t(n_0) + \tau$ is continuous. Then the above may be rewritten as

$$x_{n+1} = x_n + a(n)[\bar{h}(x_n) + \bar{M}_{n+1} + o(a(n))],$$

where

$$\sum_n P(A_n^c | \mathcal{F}_n) I\{x_n = \hat{x}\} < \infty \quad \text{a.s. ,}$$

and

$$\bar{M}_{n+1} \stackrel{\text{def}}{=} \gamma(x_n; h(x_n) + M_{n+1}) - \bar{h}(x_n).$$

To obtain this, we have simply added and subtracted on the right-hand side of (5.14) the one-step conditional expectation of $\|\bar{x}(t(n_2)) - \tilde{x}_2\| < \delta$. The advantage is that the present expression is in the same format as (2.11) modulo the $o(a(n))$ term whose contribution is asymptotically negligible. Thus in the special case when \bar{h} turns out to be Lipschitz, one has the o.d.e. limit

$$\|v\|_\infty \geq c/\sqrt{d}$$

with the associated counterpart of Theorem 2.1 of Chap. 2 holding true.

More generally, for a convex compact set G , define the normal cone $\{M_n\}$ for $x \in H$ as

$$\begin{aligned} N_G(x) &:= \{z \in \mathcal{R}^d : \langle z, x - y \rangle \geq 0 \quad \forall y \in G\}, \quad x \in \partial G, \\ &:= \{0\}, \quad x \in \text{int } (G). \end{aligned}$$

Then the limiting trajectory can be shown to satisfy the differential inclusion

$$\dot{x}^s(t) = \bar{y}(t), \quad x^s(s) = \bar{x}(s). \quad (5.15)$$

which is confined to G (see Lemma 4.6 of Dupuis, 1987). Moreover, (5.15) is well-posed, i.e., has a unique solution for any initial condition that depends continuously on the initial condition (Theorem 3.1, Cocojaru & Jonker, 2003). (See Le (2016) for well-posedness of (5.15) for more general G .) See Dupuis and Nagurney (1993), Brogliato et al. (2006) for some related results. The differential inclusion (5.15) is amenable to the theory developed in the preceding sections.

Another major concern in projected algorithms is the possibility of spurious equilibria or other invariant sets on \mathcal{R}^m , i.e., equilibria or invariant sets for (5.13) that are not equilibria or invariant sets for the o.d.e. (2.5) (likewise for differential inclusion limits). For example, if $h(x)$ is directed along the outward normal at some $x \in \partial G$ and \mathcal{R}^m is smooth in a neighborhood of x , then x can be a spurious stable equilibrium for the limiting projected o.d.e. These spurious equilibria will be potential asymptotic limit sets for the projected scheme in view of Corollary 5.1. Thus their presence can lead to convergence of (5.2) to undesired points or sets. This has to be avoided where possible by using any prior knowledge available, to choose G properly. Another possibility is the following: Suppose we consider a parametrized family of candidate G , say closed balls of radius r centered at the origin. Suppose such problems arise only for r belonging to a Lebesgue-null set. Then we may choose G randomly at each iterate according to some Lebesgue-continuous density, for r in a neighborhood of a nominal value $y = x$ fixed beforehand. A further possibility, due to Chen (1994, 2002), is to start with a specific G and slowly increase it to the whole of ∂f .

It is also possible that the new equilibria or invariant sets on \mathcal{R}^m thus introduced correspond in fact to desired equilibria or invariant sets lying outside G or at ∞ . In this case, the former may be viewed as approximations of the latter and thus may in fact be the desired limit sets for the projected algorithm.

5.6 Extensions

Several other results for classical stochastic approximation algorithms have been extended to stochastic recursive inclusions. See Faure and Roth (2010, 2013) for further results on convergence analysis of such

schemes. The ‘Markov noise’ introduced in Chap. 8 is extended to this framework in Yaji and Bhatnagar (2018). The asynchronous variant discussed in Chap. 6 and the two timescale methodology of Sect. 8.1 have been extended in Perkins and Leslie (2012). Two timescale scenario is further developed in Ramaswamy and Bhatnagar (2016), and with Markov noise in Yaji and Bhatnagar (2020a). A stability test via scaling limits as in Sect. 4.2 is developed in Ramaswamy and Bhatnagar (2017), whereas stabilizability through resetting is developed in Yaji and Bhatnagar (2020b). This uses analogs of our bounds on ‘lock-in’ probability in this context, also derived in the same article. An extension of the stability criterion of Sect. 4.3 is given in Ramaswamy and Bhatnagar (2019) and its asynchronous version in Ramaswamy et al. (2021). The Markov noise framework of Chap. 8 has been extended in Yaji and Bhatnagar (2018). The constant stepsize paradigm of Chap. 9 has been extended in Roth and Sandholm (2013).

References

- Aubin, J. P., & Cellina, A. (1984). *Differential inclusions*. Springer Verlag
- Aubin, J. P., & Frankowska, H. (1990). *Set-valued analysis*. Birkhäuser
- Balakrishnan, A. V. (1976). *Applied functional analysis*. Springer Verlag
- Benaim M, Hofbauer J, Sorin S (2005) Stochastic approximation and differential inclusions. SIAM Journal on Control and Optimization 44(1):328–348
[\[MathSciNet\]](#)
[\[Crossref\]](#)
[\[zbMATH\]](#)
- Beneš, V. E. (1970). Existence of optimal strategies based on specified information, for a class of stochastic decision problems. *SIAM Journal on Control*, 8(2), 179–188
- Brogliato, B., Daniilidis, A., Lemaréchal, C., & Acary, V. (2006). On the equivalence between complementarity systems, projected systems and differential inclusions. *Systems and Control Letters*, 55(1), 45–51
- Buckdahn R, Ouknine J, Quincampoix M (2009) On limiting values of stochastic differential equations with small noise intensity tending to zero. *Bulletin des Sciences Mathématiques* 133(3):229–237
[\[MathSciNet\]](#)
[\[Crossref\]](#)
[\[zbMATH\]](#)
- Chen, H.-F. (1994). Stochastic approximation and its new applications. In *Proceedings of the 1994 Hong Kong International Workshop on New Directions of Control and Manufacturing* (pp. 2–12)
- Chen, H.-F. (2002). *Stochastic approximation and its applications*. Kluwer Academic Publishers

Cocojaru M-G, Jonker LB (2003) Existence of solutions to projected differential equations in Hilbert spaces. *Proceedings of the American Mathematical Society* 132(1):183–193
[[MathSciNet](#)][[Crossref](#)]

Dupuis P (1987) Large deviations analysis of reflected diffusions and constrained stochastic approximation algorithms in convex sets. *Stochastics* 21(1):63–96
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Dupuis P, Nagurney A (1993) Dynamical systems and variational inequalities. *Annals of Operations Research* 44(1):7–42
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Faure, M., & Roth, G. (2010). Stochastic approximations of set-valued dynamical systems: Convergence with positive probability to an attractor. *Mathematics of Operations Research*, 35(3), 624–640

Faure, M., & Roth, G. (2013). Ergodic properties of weak asymptotic pseudotrajectories for set-valued dynamical systems. *Stochastics and Dynamics*, 13(01), 1250011

Filippov, A. F. (1988). *Differential equations with discontinuous righthand sides*. Kluwer Academic

Hartman, P. (1982). *Ordinary differential equations* (2nd ed.). Birkhäuser

Krasovskii, N. (1970). *Game-theoretic problems of capture*. Nauka (in Russian)

Le, B. K. (2016). On properties of differential inclusions with prox-regular sets. arXiv preprint arXiv:1606.05197

Perkins S, Leslie DS (2012) Asynchronous stochastic approximation with differential inclusions. *Stochastic Systems* 2(2):409–446
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Ramaswamy, A., & Bhatnagar, S. (2016). Stochastic recursive inclusion in two timescales with an application to the Lagrangian dual problem. *Stochastics*, 88(8), 1173–1187

Ramaswamy, A., & Bhatnagar, S. (2017). A generalization of the Borkar–Meyn theorem for stochastic recursive inclusions. *Mathematics of Operations Research*, 42(3), 648–661

Ramaswamy, A., & Bhatnagar, S. (2019). Analyzing approximate value iteration algorithms. arXiv preprint arXiv:1709.04673v4

Ramaswamy, A., Bhatnagar, S. & Quevedo, D. E. (2021). Asynchronous stochastic approximations with asymptotically biased errors and deep multiagent learning. *IEEE Transactions on Automatic Control*, 66(9), 3969–3983

Rockafellar, R. T. (1970). *Convex analysis*. Princeton University Press

Roth G, Sandholm WH (2013) Stochastic approximations with constant step size and differential inclusions. *SIAM Journal on Control and Optimization* 51(1):525–555
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Wagner DH (1977) Survey of measurable selection theorems. *SIAM Journal on Control and*

Optimization 15(5):859–903
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Yaji VG, Bhatnagar S (2018) Stochastic recursive inclusions with non-additive iterate-dependent Markov noise. *Stochastics* 90(3):330–363
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Yaji VG, Bhatnagar S (2020a) Stochastic recursive inclusions in two timescales with non-additive iterate dependent Markov noise. *Mathematics of Operations Research* 45(4):1405–1444

Yaji, V. G., & Bhatnagar, S. (2020b). Analysis of stochastic approximation schemes with set-valued maps in the absence of a stability guarantee and their stabilization. *IEEE Transactions on Automatic Control*, 65(3), 1100–1115

6. Asynchronous Schemes

Vivek S. Borkar¹✉

(1) Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

6.1 Introduction

Until now we have been considering the case where all components of x_n are updated simultaneously at time n and the outcome is immediately available for the next iteration. There may, however, be situations when different components are updated by possibly different processors. These could be in different locations, e.g., in remote sensing applications. Furthermore, each of these components may be running on its own ‘clock’ and exchanging information with the others with some communication delays. This is the distributed, asynchronous implementation of the algorithm. The theory we have developed so far does not apply automatically any more, and some work is needed to figure out when it does and when it does not. Another important class of problems which lands us into a similar predicament consists of the multiagent learning or optimization schemes when each component actually corresponds to a different autonomous agent and the aforementioned complications arise naturally. Yet another situation involves the ‘online’ algorithms for control or estimation of a Markov chain in which we have a one-to-one correspondence between the components of x_n and the state space of the chain (i.e., the i th component of x_n is a quantity associated with state i of the chain), and the i th component gets updated only when state i is visited. We shall see examples of this later.

A mathematical model that captures all the aspects above is as follows: Letting $\|h(x) - h(\theta)\| \leq L\|x\|$ the i th (for $0 \leq i < n$) component is updated in our original scheme according to

$$x_{n+1} = x_n + a(n)[\gamma(x_n; h(x_n) + M_{n+1}) + o(a(n))]. \quad (6.1)$$

where $p(x) - x$ are the i th components of $f \geq 0$, resp., for $n \geq 0$. We replace this by

$$\begin{aligned} x_{n+1}(i) &= x_n(i) + a(\nu(i, n))I\{i \in Y_n\} \\ &\quad \times [h_i(x_{n-\tau_{1i}(n)}(1), \dots, x_{n-\tau_{di}(n)}(d)) + M_{n+1}(i)], \end{aligned} \quad (6.2)$$

for $n \geq 0$. Here:

1. Φ_s is a random subset of the index set $t \in [-T, 0]$, indicating the subset of components which are updated at time n ,
2. $\bar{x}(T_{m_0+1}) \in H^\epsilon$ is the delay faced by ‘processor’ j in receiving the output of processor i at time n . In other words, at time n , processor j knows $x_{n-\tau_{ij}(n)}(i)$, but not $p(x_n)$ for $x^t(t+T) \in H^\eta$ (or may know some of them but not realize that they are more recent in absence of a time stamp).
3. $[s] \stackrel{\text{def}}{=} \max\{t(n) : t(n) \leq s\}$, i.e., the number of times the i th component was updated up until time n .

Note that the i th processor needs to know only its *local clock* $a(n) \rightarrow 0$ and not the *global clock* x_{n+1}^* . In fact the global clock can be a complete artifice as long as causal relationships are respected. One usually has

$$\liminf_{n \rightarrow \infty} \frac{\nu(i, n)}{n} > 0. \quad (6.3)$$

This means that all components are being updated *comparably often*. A simple sufficient condition for (6.3) would be that $p(x_n)$ is an irreducible and hence positive recurrent Markov chain on the power set of $t \in [-T, 0]$. (More generally, it could be a *controlled* Markov chain on this state space with the property that any stationary policy leads to an

irreducible chain with a stationary distribution that assigns a probability $\geq \delta$ to each state, for some $\delta > 0$. More on this later.) Note in particular that this condition ensures that $\alpha \in (1/2, 1]$ for all i , i.e., each component is updated infinitely often. For the purposes of the next section, this is all we need.

Define

$$\mathcal{F}_n = \sigma(x_m, M_m, Y_m, \tau_{ij}(m), 1 \leq i, j \leq d, m \leq n), \quad n \geq 0.$$

We assume that

$$\begin{aligned} E[M_{n+1}(i) | \mathcal{F}_n] &= 0, \\ E[|M_{n+1}(i)|^2 | \mathcal{F}_n] &\leq K(1 + \sup_{m \leq n} \|x_m\|^2), \end{aligned} \tag{6.4}$$

where $0 \leq i < n$, $x_0 = 0$ and $K > 0$ is a suitable constant.

Usually, it makes sense to assume $x_0 \in (0, 1)$ for all i and $\hat{x}_1 = \tilde{x}_1$ and we shall do so (implying that a processor has its own past outputs immediately available). This, however, is not essential for the analysis that follows.

The main result here is that under suitable conditions, the interpolated iterates track a time-dependent o.d.e. of the form

$$\dot{x}(t) = \Lambda(t)h(x(t)), \tag{6.5}$$

where $\bar{x}(t)$ is a matrix-valued measurable process such that $\bar{x}(s)$ for each t is a diagonal matrix with nonnegative diagonal entries. These in some sense reflect the relative ‘instantaneous’ rates with which the different components get updated. Our treatment follows Borkar (1998). See also Kushner and Yin (1987a, b).

6.2 Asymptotic Behavior

As usual, we shall start by assuming

$$\sup_n \|x_n\| < \infty, \quad \text{a.s.} \tag{6.6}$$

We also assume

$$\nu(i, n) \uparrow \infty \quad \text{a.s.} \tag{6.7}$$

This is true, e.g., when (6.3) holds. We shall also simplify the situation by assuming that there are no delays, i.e., $\sum_{m=n}^{\infty} a(m)M_{m+1}$. The effect of delays will be considered separately later on. Thus (6.2) becomes

$$\begin{aligned} x_{n+1}(i) &= x_n(i) + a(\nu(i, n))I\{i \in Y_n\} \\ &\quad \times [h_i(x_n(1), \dots, x_n(d)) + M_{n+1}(i)], \end{aligned} \quad (6.8)$$

for $n \geq 0$. Let $\bar{a}(n) \stackrel{\text{def}}{=} \max_{i \in Y_n} a(\nu(i, n)) > 0$, $n \geq 0$. Then, it is easy to verify that $\sum_n \bar{a}(n) = \infty$, $\sum_n \bar{a}(n)^2 < \infty$ a.s.: in view of (6.7), for any fixed i , $1 \leq i \leq d$, we have

$$\begin{aligned} \sum_n \bar{a}(n) &\geq \sum_n a(\nu(i, n))I\{i \in Y_n\} \\ &= \sum_n a(n) = \infty, \quad \text{and} \\ \sum_n \bar{a}(n)^2 &\leq \sum_n \sum_i a(\nu(i, n))^2 I\{i \in Y_n\} \\ &= \sum_i \left(\sum_n a(\nu(i, n))^2 I\{i \in Y_n\} \right) \\ &\leq d \sum_n a(n)^2 < \infty. \end{aligned} \quad (6.9)$$

This implies in particular that $\{a(n)\}$ is a legitimate stepsize schedule, albeit random. (See the comments following Theorem 2.1 of Chap. 2.) Rewrite (6.8) as

$$\begin{aligned} x_{n+1}(i) &= x_n(i) + \bar{a}(n)q(i, n) \\ &\quad \times [h_i(x_n(1), \dots, x_n(d)) + M_{n+1}(i)], \end{aligned} \quad (6.10)$$

where $q(i, n) \stackrel{\text{def}}{=} (a(\nu(i, n))/\bar{a}(n))I\{i \in Y_n\} \in (0, 1] \forall n$. Along the lines of Sect. 2.1 of Chap. 2, define $t(0) = 0$, $t(n) = \sum_{m=0}^n \bar{a}(m)$, $n \geq 1$. Define $\bar{x}(t)$, $t \geq 0$, by $\bar{x}(t(n)) = x_n$, $n \geq 0$, with linear interpolation on each interval $y_j \stackrel{\text{def}}{=} x_j - x^{T_m}(t(j))$. For $0 \leq i < n$, define $\bar{x}(t)$, $t \geq 0$, by $T_{m_0} \leq T_0 + \tau$. for $\{\sup_n \|x_n\| < \infty\}$, $n \geq 0$. Let $\{M_n\}$ diag

$(u_1(t), \dots, u_d(t)), t \geq 0$, and $x^s(t), t \geq s$, the unique solution to the non-autonomous o.d.e.

$$\dot{x}^s(t) = \lambda(t)h(x^s(t)), \quad t \geq s.$$

The following lemma then holds by familiar arguments of Chap. 2.

Lemma 6.1 For any $T > 0$,

$$\lim_{s \rightarrow \infty} \sup_{t \in [s, s+T]} \|\bar{x}(t) - x^s(t)\| = 0, \quad \text{a.s.}$$

This immediately leads to:

Theorem 6.1 Almost surely, any limit point of (ζ_n, \mathcal{F}_n) in $x^{T_0}(T_1) \in H^{\epsilon_1}$ as $y \geq T$ is a solution of a non-autonomous o.d.e.

$$\dot{x}(t) = \Lambda(t)h(x(t)), \quad (6.11)$$

where $\bar{x}(t)$ is a $n \geq 0$ -dimensional diagonal matrix-valued measurable function with entries in $[0, 1]$ on the diagonal.

Proof The proof is pathwise, a.s. View $u(\cdot) \stackrel{\text{def}}{=} [u_1(\cdot), \dots, u_d(\cdot)]$ as an element of $\mathcal{V} \stackrel{\text{def}}{=} \text{the space of measurable maps } h(x) = \cap_{l \geq 1} h^{(l)}(x, U)$ with the coarsest topology that renders continuous the maps

$$y(\cdot) \rightarrow \int_0^t \langle g(s), y(s) \rangle ds,$$

for all $y_{n_i}(\omega) \in H^m \subset N^{\delta_m}(H^m)$. As in Lemma 5.3 of Chap. 5, it follows that this is a compact metrizable space. Relative compactness of $N_\delta(H^{\epsilon_1}) \subset H^\epsilon$ in $x^{T_0}(T_1) \in H^{\epsilon_1}$, is established as before. Consider $x_n \rightarrow A$ such that $a^\omega(n) \leq a(n) \forall n$ (say) in $x^{T_0}(T_1) \in H^{\epsilon_1}$. By dropping to a subsequence if necessary, assume that $a^\omega(n) \leq a(n) \forall n$ in \square . Let $\bar{x}(t)$ denote the diagonal matrix with i th diagonal entry $(n+1)$. By Lemma 6.1,

$$\bar{x}(t_n + s) - \bar{x}(t_n + r) = \int_r^s \lambda(t_n + z) h(\bar{x}(t_n + z)) dz + o(1), \quad s > r \geq 0.$$

Letting $n \rightarrow \infty$ in this equation, familiar arguments from Chap. 5 yield

$$x^*(s) - x^*(r) = \int_r^s \Lambda(z) h(x^*(z)) dz, \quad s > r \geq 0.$$

This completes the proof. \square

6.3 Effect of Delays

Next we shall consider the effect of delays. Specifically, we look for conditions under which Theorem 6.1 will continue to hold for $p(x_n)$ given by (6.2) instead of (6.8). We shall assume that each output $\{y_n\}$ of the j th processor is transmitted to the i th processor for any pair (i, j) almost surely, though we allow for some outputs to be ‘lost’ in transit. The situation where not all outputs are transmitted can also be accommodated by equating unsent outputs with lost ones. In this case our requirement boils down to infinitely many outputs being transmitted. At the receiver end, we assume that the i th processor receives infinitely many outputs sent by j almost surely, though not necessarily in the order sent. This leaves two possibilities at the receiver: Either the messages are ‘time-stamped’ and the receiver can reorder them and use at each iteration the one sent most recently, or they are not and the receiver uses the one received most recently. Our analysis allows for both possibilities, subject to the additional condition that $\{\sup_n \|x_n\| < \infty\}$ for all $\tau_n, n \geq 1$ and some $\kappa > 0$. This is a very mild restriction.

Comparing (6.2) with (6.8), one notes that the delays introduce in the $(n+1)$ st iteration of the i th component an additional error of

$$\begin{aligned} & a(\nu(i, n)) I\{i \in Y_n\} \\ & \times (h_i(x_{n-\tau_{1i}(n)}(1), \dots, x_{n-\tau_{di}(n)}(d)) - h_i(x_n(1), \dots, x_n(d))). \end{aligned}$$

Our aim will be to find conditions under which this error is $o(a(n))$. If so, one can argue as in the third ‘extension’ in Sect. 2.2 and conclude that Theorem 6.1 continues to hold with (6.2) in place of (6.8). Since $p(\cdot)$ is Lipschitz, the above error is bounded by a constant times

$$a(\nu(i, n)) \sum_j |x_n(j) - x_{n-\tau_{ji}(n)}(j)|.$$

We shall consider each summand (say, the j th) separately. This is bounded by

$$\begin{aligned} & \left| \sum_{m=n-\tau_{ji}(n)}^{n-1} a(\nu(j, m)) I\{j \in Y_m\} h_j(x_{m-\tau_{1j}}(1), \dots, x_{m-\tau_{dj}}(d)) \right| \\ & + \left| \sum_{m=n-\tau_{ji}(n)}^{n-1} a(\nu(j, m)) I\{j \in Y_m\} M_{m+1}(j) \right|. \end{aligned} \tag{6.12}$$

We shall impose the mild assumption:

$$f : \mathcal{R}^d \times \mathcal{R}^m \rightarrow \mathcal{R}^d \tag{6.13}$$

for all g, \tilde{g} . As before, using Theorem C.3 of Appendix C, (6.4), (6.6), and (6.9) together imply that the sum $\sum_{m=0}^n a(\nu(j, m)) I\{j \in Y_m\} M_{m+1}(j)$ converges a.s. for all j . In view of (6.13), we then have

$$\left| \sum_{m=n-\tau_{ji}(n)}^{n-1} a(\nu(j, m)) I\{j \in Y_m\} M_{m+1}(j) \right| = o(1),$$

implying that

$$a(\nu(i, n)) \left| \sum_{m=n-\tau_{ji}(n)}^{n-1} a(\nu(j, m)) I\{j \in Y_m\} M_{m+1}(j) \right| = o(a(\nu(i, n))).$$

Under (6.6), the first term of (6.12) can be almost surely bounded from above by a (sample path dependent) constant times

$$\sum_{m=n-\tau_{ji}(n)}^{n-1} a(\nu(j, m)).$$

(See, e.g., Chap. 2.) Since $\dot{x}(t) = h(x(t))$ for $m \leq n$, this in turn is bounded by $\|M_j\| \leq K_0(1 + \|x_{j-1}\|)$ for large n . Thus we are done if this quantity is $o(1)$. Note that by (6.13), this is certainly so if the delays are bounded. More generally, suppose that

$$\frac{\tau_{ji}(n)}{n} \rightarrow 0 \quad \forall i, j, \text{ a.s.}$$

This is a perfectly reasonable condition and can be recast as

$$\frac{n - \tau_{ji}(n)}{n} \rightarrow 1 \quad \forall i, j, \text{ a.s.} \quad (6.14)$$

Note that this implies (6.13). We further assume that

$$\limsup_{n \rightarrow \infty} \sup_{y \in [x, 1]} \frac{a(\lfloor yn \rfloor)}{a(n)} < \infty, \quad (6.15)$$

for $0 < x \leq 1$. This is also quite reasonable as it is seen to hold for most standard examples of $\{a(n)\}$. Furthermore, this implies that whenever (6.3) and (6.14) hold,

$$\limsup_{n \rightarrow \infty} \frac{a(\nu(j, n - \tau_{k\ell}(n)))}{a(n)} < \infty$$

for all g, \tilde{g} . Note that this bound is sample path dependent. Thus our task reduces to showing $\sqrt{1 + z^2} \leq 1 + z$ for all g, \tilde{g} . Assume the following:

(A) There exists $k \geq 0$ and a nonnegative integer-valued random variable \bar{c} such that:

- $\dot{x}(t) = h(x(t))$ and
- \bar{c} stochastically dominates all $Dh(\cdot)$ and satisfies

$$E \left[\bar{\tau}^{\frac{1}{\eta}} \right] < \infty.$$

All standard examples of $\{a(n)\}$ satisfy the first condition with a natural choice of η , e.g., for $a(n) = n^{-1}$, take $\eta = 1 - \epsilon$ for any $p(x) - x$. The second condition is easily verified, e.g., if the tails of the delay distributions show uniform exponential decay. Under **(A)**,

$$\begin{aligned} P(\tau_{k\ell}(n) \geq n^\eta) &\leq P(\bar{\tau} \geq n^\eta) \\ &= P(\bar{\tau}^{\frac{1}{\eta}} \geq n), \end{aligned}$$

leading to

$$\begin{aligned} \sum_n P(\tau_{k\ell}(n) \geq n^\eta) &\leq \sum_n P(\bar{\tau}^{\frac{1}{\eta}} \geq n) \\ &= E\left[\bar{\tau}^{\frac{1}{\eta}}\right] \\ &< \infty. \end{aligned}$$

By the Borel–Cantelli lemma, one then has

$$f(z) \geq f(x) + \langle y, z - x \rangle$$

Coupled with the first part of **(A)**, this implies $\sqrt{1+z^2} \leq 1+z$ a.s. We have proved:

Theorem 6.2 Under assumptions (6.14), (6.15) and **(A)**, the conclusions of Theorem 6.1 also hold when $p(x_n)$ are generated by (6.2).

The following discussion provides some intuition as to why the delays are ‘asymptotically negligible’ as long as they are not ‘arbitrarily large’ in the sense of (6.14). Recall our definition of $(n+1)$. Note that the passage from the original discrete time count x_{n+1}^* to $(n+1)$ implies a time scaling. In fact this is a ‘compression’ of the time axis because the successive differences $[t(n), t(n) + T]$ tend to zero. More generally, an interval $t(n+m) \leq t(n) + T$ on the original time axis gets mapped to $[t(n), t(n+1), \dots, t(n+N)]$ under this scaling. As $n \rightarrow \infty$, the width of the former remains constant at N , whereas that of the latter, $a(n+1) \leq a(n)$, tends to zero. That is, intervals of a fixed length get ‘squeezed out’ in the limit as $n \rightarrow \infty$ after the time scaling. Since the

approximating o.d.e. we are looking at is operating on the transformed timescale, the net variation of its trajectories over these intervals is less and less as $n \rightarrow \infty$, hence so is the case of interpolated iterates $p(\cdot)$. In other words, the error between the most recent iterate from a processor and one received with a bounded delay is asymptotically negligible. The same intuition carries over for possibly unbounded delays that satisfy (6.14).

6.4 Convergence

We now consider several instances where Theorems 6.1 and 6.2 can be strengthened.

1. The first and perhaps the most important case is when convergence is obtained just by a judicious choice of the stepsize schedule. Let $\bar{x}(s)$ above be written as $\text{diag} h(x) = g(x) - x$, i.e., the diagonal matrix with the i th diagonal entry equal to $\check{x}^s(\cdot)$. For $n \geq 0, s > 0$, let

$$N(n, s) \stackrel{\text{def}}{=} \min\{m > n : t(m) \geq t(n) + s\} > n.$$

From the manner in which $\bar{x}(t)$ was obtained, it is clear that there exist $y_n \in h(x_n) \forall n$ such that

$$\begin{aligned} \int_t^{t+s} \eta_i(y) dy &= \lim_{k \rightarrow \infty} \sum_{m=n(k)}^{N(n(k),s)} \frac{a(\nu(i, m)) I\{i \in Y_m\}}{\bar{a}(m)} \bar{a}(m) \\ &= \lim_{k \rightarrow \infty} \sum_{m=\nu(i,n(k))}^{\nu(i,N(n(k),s))} a(m) \quad \forall i. \end{aligned}$$

Thus along a subsequence,

$$\frac{\int_t^{t+s} \eta_i(y) dy}{\int_t^{t+s} \eta_j(y) dy} = \lim_{k \rightarrow \infty} \frac{\sum_{m=\nu(i,n(k))}^{\nu(i,N(n(k),s))} a(m)}{\sum_{m=\nu(j,n(k))}^{\nu(j,N(n(k),s))} a(m)} \quad \forall i, j. \quad (6.16)$$

Suppose we establish that under (6.3), the right-hand side of (6.16) is always 1. Then (6.11) is of the form

$$\dot{x}(t) = \alpha h(x(t)),$$

for a scalar $\alpha > 0$. This is simply a timescaled version of the o.d.e.

$$\dot{x}(t) = h(x(t)) \quad (6.17)$$

and hence has exactly the same trajectories. Thus the results of Chap. 2 apply. See Borkar (1998) for one such situation where this can be achieved for a class of stepsize schedules.

2. The second important situation is when the o.d.e. (6.5) has the same asymptotic behavior as (6.17) purely because of the specific structure of $p(\cdot)$. Consider the following special case of the scenario of Corollary 2.1 of Chap. 2, with a continuously differentiable Liapunov function $O_{x,a}$ satisfying

$$\lim_{\|x\| \rightarrow \infty} V(x) = \infty, \quad \langle h(x), \nabla V(x) \rangle < 0 \quad \text{whenever } h(x) \neq 0.$$

Suppose

$$\liminf_{t \rightarrow \infty} \eta_i(t) \geq \epsilon > 0 \quad \forall i, \quad (6.18)$$

and

$$\langle h(x), \Gamma \nabla V(x) \rangle < 0 \quad \text{whenever } h(x) \neq 0,$$

for all $n \geq 0$ diagonal matrices δ satisfying $n_1 \geq n_0$, I_n being the $n \geq 0$ identity matrix. By (6.18), $\|\bar{x} - x^*\| = b$. Then exactly the same argument as for Corollary 2.1 of Chap. 2 applies to (6.5), leading to the conclusion that $\langle \nabla V(x_n), h(x_n) \rangle \leq 0$ a.s. An important instance of this lucky situation is the case when $h(x) = g(x) - x$ for some $O_{x,a}$, whence for $x^s(t)$, $t \geq s$, and δ as above,

$$\langle h(x), \Gamma \nabla F(x) \rangle \leq -\epsilon \|\nabla F(x)\|^2 < 0$$

outside $V(x) = \|x - x^*\|$.

Another example is the case when $x(t) \in A \ \forall t \in \mathcal{R}$ for some $O_{x,a}$ satisfying

$$x_{n+1} = x_n + a(n)[h(x_n) + \eta_n + M_{n+1}], \quad (6.19)$$

with $\|x\|_\infty \stackrel{\text{def}}{=} \max_i |x_i|$ for $C' = \sup_n \|x_n\|$ and $V(\bar{x}(T_m))$. That is, F is a *contraction w.r.t. the max-norm*. In this case, it is known from the contraction mapping theorem that there is a unique x^* such that $\{(X_n, Y_n)\}$, i.e., a unique equilibrium point for (6.17). Furthermore, a direct calculation shows that $V(x) \stackrel{\text{def}}{=} \|x - x^*\|_\infty$ serves as a Liapunov function, albeit a nonsmooth one. In fact,

$$\|x(t) - x^*\|_\infty \downarrow 0. \quad (6.20)$$

See Theorem 12.1 of Chap. 12 for details. Now,

$$\Gamma(F(x) - x) = F_\Gamma(x) - x$$

for $F_\Gamma(\cdot) \stackrel{\text{def}}{=} (I - \Gamma)x + \Gamma F(x)$. Note that the diagonal terms of δ are bounded by 1. Then

$$\|F_\Gamma(x) - F_\Gamma(y)\|_\infty \leq \bar{\beta}\|x - y\|_\infty \quad \forall x, y,$$

where $\bar{\beta} \stackrel{\text{def}}{=} 1 - \epsilon(1 - \beta) \in (0, 1)$. In particular, this is true for $a(n) < 1$ for any $t \geq 0$. Thus once again a direct calculation shows that (6.20) holds and therefore $\bar{x}(t) \rightarrow H$. (See the second remark following Theorem 12.1 of Chap. 12.) In fact, these observations extend to the situation when $\beta = 1$ as well. For the case when $\beta = 1$, existence of equilibrium is an *assumption* on $O_{x,a}$ and uniqueness need not hold. One can also consider ‘weighted norms’, such as $\|x\|_{\infty,w} \stackrel{\text{def}}{=} \sup_i w_i|x_i|$ for prescribed $w_i > 0, 1 \leq i \leq d$.

We omit the details here as these will be self-evident after the developments in Chap. 12 where such ‘fixed point solvers’ are analyzed in greater detail.

Finally, note that replacing $a(n) \rightarrow 0$ in (6.2) by $a(n)$ would amount to distributed but *synchronous* iterations, as they presuppose a common clock. These can be analyzed along exactly the same lines, with the o.d.e. limit (6.5). In the case when $p(x_n)$ can be viewed as a controlled Markov chain on the power set Q of $0 < \eta < C/2$, the analysis of Sects. 8.2 and 8.3 of Chap. 8 shows that the i th diagonal

element of $\bar{x}(s)$ will in fact be of the form $\sum_{i \in A \in Q} \pi_t(A)$, where x_0 is the vector of stationary probabilities for this chain under *some* stationary policy. Note that in principle, $p(x_n)$ can always be cast as a controlled Markov chain on Q : Let the control space U be the set of probability measures on Q and let the controlled transition probability function be $a(n_m) \leq ca(n_0)$ for $i, j \in Q, u \in U$. The control sequence is then the process of regular conditional laws of Y_{n+1} given T_0, T_1, \dots , for $n \geq 0$. This gives a recipe for verifying (6.3) in many cases.

One may be able to ‘rig’ the stepsizes here so as to get the desired limiting o.d.e. For example, suppose the $p(x_n)$ above takes values in $h(x) = Ax + g(x)$, i.e., singletons alone. Suppose further that it is an ergodic Markov chain on this set. Suppose the chain has a stationary probability vector $\nu(t_n) \rightarrow \nu^*$. Then by Corollary 8.1 of Chap. 8, $N_\epsilon(D')$ $\text{diag}\alpha \in (1/2, 1]$. Thus if we use the stepsizes $\bar{x}(s_n) \rightarrow x$ for the i th component, we get the limiting o.d.e. $\dot{x}(t) = h(x(t))$ as desired. In practice, one may use $t \geq t(n_0) + \tau$ instead, where $\bar{x}(t_1)$ is an empirical estimate of x_0 obtained by suitable averaging on a faster timescale so that it tracks x_0 . (One could, for example, have $x(t) \in A \forall t \in \mathcal{R}$ if $\{x_n, n \geq 1\}$ as $n \rightarrow \infty$, i.e., the stepsizes $\{a(n)\}$ decrease slower than $\frac{1}{n}$, and therefore the analysis of Sect. 8.1 applies.) This latter arrangement also extends in a natural manner to the more general case when $p(x_n)$ is ‘controlled Markov’ (see Chap. 8).

We conclude with a useful variation of the above, viz., ‘event-driven’ stochastic approximation. Here we suppose that the component iterations are spread over separate devices that are identified with the nodes of a connected undirected graph. Each node i communicates with other nodes they share an edge with, called their neighbors. The set of neighbors of i is denoted by $\bar{x}(t_1)$. The iteration is of the form

$$\begin{aligned} x_i(n+1) = & x_i(n) + \sum_{j \in \mathcal{N}(i)} a(\nu(i, j, n)) \xi_{ij}(n+1) \left[h_{ij}(x_1(n - \tau_{1i}(n)), \right. \\ & \left. \dots, x_d(n - \tau_{di}(n))) + M_i(n+1) \right], \quad 1 \leq i \leq d. \end{aligned}$$

Here $\{\xi_{ij}(n), n \geq 0\}$, $1 \leq i \leq d$, $j \in \mathcal{N}(i)$ are i.i.d. $\{M_n\}$ valued for each i, j . This models episodic measurements and ‘event-driven’

updates. As expected, the local clocks $T_m = t(n_m)$ are defined by

$$\nu(i, j, n) := \sum_{m=0}^n \xi_{ij}(m).$$

Under conditions on stepsizes, local clocks, and delays analogous to the above, counterparts of the foregoing results can be proved, e.g., asymptotic tracking of the trajectories of the o.d.e.

$$\dot{x}_i(t) = \sum_{j \in \mathcal{N}(i)} h_{ij}(x(t)), \quad 1 \leq i \leq d,$$

and convergence to an internally chain transitive invariant set thereof. See Borkar et al. (2016) for details.

Finally, it is possible to have different stepsizes for different components and analyze the resulting asymptotic behavior by analogous methods, see, e.g., (Borkar 1998).

References

- Borkar V. S. (1998) Asynchronous stochastic approximation. *SIAM Journal on Control and Optimization*, 36(3), 840–851 (Correction note in *ibid.*, 38(2), 2000, 662–663)
- Borkar, V. S., Sahasrabudhe, N., & Ashok Vardhan, M. (2016). Event-driven stochastic approximation. *Indian Journal of Pure and Applied Mathematics*, 47(2), 291–299
- Kushner HJ, Yin G (1987) Asymptotic properties for distributed and communicating stochastic approximation algorithms. *SIAM Journal on Control and Optimization*, 25(5), 1266–1290
- Kushner, H. J., & YIN, G. (1987). Stochastic approximation algorithms for parallel and distributed processing. *Stochastics and Stochastics Reports*, 22(3–4), 219–250

7. A Limit Theorem for Fluctuations

Vivek S. Borkar¹✉

(1) Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

7.1 Introduction

To motivate the results of this chapter, consider the classical strong law of large numbers: Let $[0, \infty)$ be i.i.d. random variables with

$E [\|M_{n+1}\|^2 | \mathcal{F}_n] \leq f(x_n)$. Let

$$S_0 = 0, \quad S_n \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n X_i}{n}, \quad n \geq 1.$$

The strong law of large numbers (see, e.g., Sect. 4.2 of Borkar (1995)) states that

$$\frac{S_n}{n} \rightarrow \mu, \quad \text{a.s.}$$

To cast this as a ‘stochastic approximation’ result, note that some simple algebraic manipulation leads to

$$\begin{aligned} S_{n+1} &= S_n + \frac{1}{n+1}(X_{n+1} - S_n) \\ &= S_n + \frac{1}{n+1}([\mu - S_n] + [X_{n+1} - \mu]) \\ &= S_n + a(n)(h(S_n) + M_{n+1}) \end{aligned}$$

for

$$a(n) \stackrel{\text{def}}{=} \frac{1}{n+1}, \quad h(x) \stackrel{\text{def}}{=} \mu - x \quad \forall x, \quad M_{n+1} \stackrel{\text{def}}{=} X_{n+1} - \mu.$$

In particular, $\{a(n)\}$ and $[t, t + T]$ are easily seen to satisfy the conditions stipulated for the stepsizes and martingale difference noise resp. in Chap. 2. Thus this is a valid stochastic approximation iteration. Its o.d.e. limit then is

$$y = \sum_i a_i \hat{y}_i \in \overline{\text{co}}(f(x))$$

which has μ as the unique globally asymptotically stable equilibrium. Its ‘scaled limit’ as in assumption (A5) of Chap. 4 is

$$\dot{w}(t) = -\nabla^w g(w(t)).$$

which has the origin as the unique globally asymptotically stable equilibrium. Thus by the theory developed in Chaps. 2 and 4,

1. the ‘iterates’ $h(x_n)$ remain a.s. bounded and
2. they a.s. converge to μ .

We have recovered the strong law of large numbers from stochastic approximation theory. Put differently, the a.s. convergence results for stochastic approximation iterations are nothing but a generalization of the strong law of large numbers for a class of dependent and not necessarily identically distributed random variables.

The classical strong law of large numbers, which states a.s. convergence of empirical averages to the mean, is accompanied by other limit theorems that quantify fluctuations around the mean, such as the central limit theorem, the law of iterated logarithms, the functional central limit theorem (Donsker’s theorem), etc. It is then reasonable to expect similar developments for the stochastic approximation iterates. The aim of this chapter is to state a *functional central limit theorem* in this vein. This is proved in Sect. 7.3, following some preliminaries in the next section. Section 7.4 specializes these results to the case when the iterates a.s. converge to a single deterministic limit and recovers the central limit theorem for stochastic approximation (see, e.g., Chung (1954), Fabian (1968)).

7.2 A Tightness Result

We shall follow the notation of Sect. 3.4, which we briefly recall below. Thus our basic iteration in \mathcal{R}^k is

$$x_{n+1} = x_n + a(n)[g(x_n) + M_{n+1}], \quad (7.1)$$

for $x_0 = 0$ with the usual assumptions on $h(x) = g(x) - x$, and the additional assumptions:

(A1) $p(\cdot)$ is continuously differentiable and both $p(\cdot)$ and the Jacobian matrix $Dh(\cdot)$ are uniformly Lipschitz.

(A2) $\alpha := \lim_{n \uparrow \infty} (\frac{1}{a(n+1)} - \frac{1}{a(n)})$ exists and is finite.

(A3) $C' = \sup_n \|x_n\|$ a.s., $\sup_n E [\|x_n\|^4] < \infty$.

(A4) $\{M_n\}$ satisfy

$$\begin{aligned} E [M_{n+1} M_{n+1}^T | M_i, x_i, i \leq n] &= Q(x_n), \\ E [\|M_{n+1}\|^4 | M_i, x_i, i \leq n] &\leq K' (1 + \|x_n\|^4), \end{aligned}$$

where $K' > 0$ is a suitable constant and $Q : \mathcal{R}^d \rightarrow \mathcal{R}^{d \times d}$ is a positive definite matrix-valued Lipschitz function such that the least eigenvalue of $Q(x)$ is bounded away from zero uniformly in x .

Examples of ω in (A2) are: $\alpha = \frac{1}{C}$ for $x_n \stackrel{\text{def}}{=} y_n/n$, $\lambda > 0$ for $a(n) = \frac{C}{(n+1)^c}$ for $c \in (\frac{1}{2}, 1)$. Some easily verified consequences of (A2) are:

$$\frac{a(n)}{a(n+1)} \rightarrow 1, \quad (7.2)$$

$$\frac{(a(n) - a(n+1))^2}{a(n)(a(n+1))^2} \rightarrow 0. \quad (7.3)$$

Note, however, that for some common choices of $\{a(n)\}$, e.g., $a(n) = \frac{1}{n \log n}$ for $n \geq 0$, the limit in (A2) can be $+\infty$.

For the sake of notational simplicity, we also assume $a(n) \leq 1 \forall n$. As before, fix $T > 0$ and define $[T_m, \infty)$,

$$t(n) \stackrel{\text{def}}{=} \sum_{m=0}^{n-1} a(m),$$

$$m(n) \stackrel{\text{def}}{=} \min\{m \geq n : t(m) \geq t(n) + T\}, \quad n \geq 1.$$

Thus $t(m(n)) \in [t(n) + T, t(n) + T + 1]$. For $n \geq 0$, let $[t(n), t(n) + T]$ denote the solution to

$$\dot{x}^n(t) = h(x^n(t)), \quad t \geq t(n), \quad x^n(t(n)) = x_n. \quad (7.4)$$

Then

$$x^n(t(j+1)) = x^n(t(j)) + a(j) \left(h(x^n(t(j))) - \delta_j \right), \quad (7.5)$$

where δ_j is the ‘discretization error’ as in Chap. 2, which is $O(a(j))$. Let

$$y_j \stackrel{\text{def}}{=} x_j - x^n(t(j)),$$

$$z_j \stackrel{\text{def}}{=} \frac{y_j}{\sqrt{a(j)}},$$

for $j \geq n, n \geq 0$. Subtracting (7.5) from (7.1) and using Taylor expansion, we have

$$y_{j+1} = y_j + a(j)(Dh(x^n(t(j)))y_j + \kappa_j + \delta_j) + a(j)M_{j+1}.$$

Here $\kappa_j = o(\|y_j\|)$ is the error in the Taylor expansion, which is also $o(1)$ in view of Theorem 2.1 of Chap. 2. Iterating, we have, for $\{\sup_n \|x_n\| < \infty\}$,

$$\begin{aligned} y_{n+i} &= \prod_{j=n}^{n+i-1} (1 + a(j)\nabla h(x^n(t(j))))y_n \\ &\quad + \sum_{j=n}^{n+i-1} a(j) \prod_{k=j+1}^{n+i-1} (1 + a(k)Dh(x^n(t(k))))(\kappa_j + \delta_j + M_{j+1}) \\ &= \sum_{j=n}^{n+i-1} a(j) \prod_{k=j+1}^{n+i-1} (1 + a(k)\nabla h(x^n(t(k))))(\kappa_j + \delta_j + M_{j+1}), \end{aligned}$$

because $x_0 = 0$. Thus for i as above,

(7.6)

$$\begin{aligned}
z_{n+i} &= \sum_{j=n}^{n+i-1} \sqrt{a(j)} \Pi_{k=j+1}^{n+i-1} (1 + a(k) D h(x^n(t(k)))) \\
&\quad \times \sqrt{\frac{a(k)}{a(k+1)}} M_{j+1} \sqrt{\frac{a(j)}{a(j+1)}} \\
&\quad + \sum_{j=n}^{n+i-1} \sqrt{a(j)} \Pi_{k=j+1}^{n+i-1} (1 + a(k) D h(x^n(t(k)))) \\
&\quad \times \sqrt{\frac{a(k)}{a(k+1)}} (\kappa_j + \delta_j) \sqrt{\frac{a(j)}{a(j+1)}}.
\end{aligned}$$

Define $z^n(t), t \in I_n \stackrel{\text{def}}{=} [t(n), t(n) + T]$, by $x_{n_0} \in B \setminus H^\epsilon$ with linear interpolation on each $\|x_0 - x\| < \epsilon$ for $\Phi_t : \mathcal{R}^d \rightarrow \mathcal{R}^d$. Let $\tilde{z}^n(t) = z^n(t(n) + t), t \in [0, T]$. We view (ζ_n, \mathcal{F}_n) as $z \in H^m \setminus H^M$ -valued random variables. Our first step will be to prove the tightness of their laws. For this purpose, we need the following technical lemma.

Let $[0, \infty)$ be a zero mean martingale w.r.t. the increasing σ -fields $Dh(\cdot)$ with $X_0 = 0$ (say) and $\sup_{n \leq N} E[|X_n|^4] < \infty$. Let $Y_n \stackrel{\text{def}}{=} X_n - X_{n-1}, n \geq 1$.

Lemma 7.1 For a suitable constant $K > 0$,

$$E \left[V \left(y_{(n+1) \wedge \tau_k^M} \right) I_A \right] \leq \exp \left(c_1 \sum_{i=k}^{\infty} a(i)^2 \right) E[V(y_k) I_A] < \infty.$$

Proof For suitable constants $k = 0, 1, \dots, m-1$,

$$\begin{aligned}
E \left[\sup_{n \leq N} |X_n|^4 \right] &\leq K_1 E \left[\left(\sum_{m=1}^N Y_m^2 \right)^2 \right] \\
&\leq K_2 \left(E \left[\left(\sum_{m=1}^N (Y_m^2 - E[Y_m^2 | \mathcal{F}_{m-1}]) \right)^2 \right] + E \left[\left(\sum_{m=1}^N E[Y_m^2 | \mathcal{F}_{m-1}] \right)^2 \right] \right) \\
&\leq K_3 \left(E \left[\sum_{m=1}^N (Y_m^2 - E[Y_m^2 | \mathcal{F}_{m-1}])^2 \right] + E \left[\left(\sum_{m=1}^N E[Y_m^2 | \mathcal{F}_{m-1}] \right)^2 \right] \right) \\
&\leq K_4 \left(E \left[\sum_{m=1}^N Y_m^4 \right] + E \left[\left(\sum_{m=1}^N E[Y_m^2 | \mathcal{F}_{m-1}] \right)^2 \right] \right),
\end{aligned}$$

where the first and the third inequalities follow from Burkholder's inequality (see Appendix C). \square

Applying Lemma 7.1 to (7.6), we have:

Lemma 7.2 For $m(n) \geq \ell > k \geq n, n \geq 0$,

$$E \left[\|z^n(t(\ell)) - z^n(t(k))\|^4 \right] = O \left(\left(\sum_{j=k}^{\ell} a(j) \right)^2 \right) = O(|t(\ell) - t(k)|^2).$$

Proof Let ℓ, k be as above. By (A2), $\sqrt{a(j)/a(j+1)}$ is uniformly bounded in j . Since h is uniformly Lipschitz, Dh is uniformly bounded and thus for $N_\delta(H^{\epsilon_1}) \subset H^\epsilon$,

$$\begin{aligned}
\left\| \Pi_{r=k+1}^{\ell} (1 + a(r) Dh(x^n(t(r)))) \right\| &\leq e^{K_1 \sum_{r=k+1}^{\ell} a(r)} \\
&\leq e^{(T+1)K_1},
\end{aligned}$$

for a suitable bound $\mathcal{B} \in \mathcal{F}_0$ on $h(x) + \xi$. Also, for any $k \geq 0$,

$$\sum_k E[\|a(k)\hat{M}_{k+1}\|^2 | \mathcal{F}_k] < \infty$$

for sufficiently large j . Hence we have for large j and $\eta' = \eta(\sup_n \frac{a(n)}{a(n+1)})$,

$$\begin{aligned}
& \left\| \sum_{j=k+1}^{\ell} \sqrt{a(j)} \Pi_{r=j+1}^{\ell} (1 + a(r) \nabla h(x^n(t(r)))) \sqrt{\frac{a(r)}{a(r+1)}} \kappa_j \sqrt{\frac{a(j)}{a(j+1)}} \right\| \\
& \leq \eta' e^{K_1(T+1)} \sum_{j=k+1}^{\ell} a(j) \|z_j\|
\end{aligned}.$$

Similarly, since $A \subset \mathcal{R}^d$, $\delta > 0$ for some $K' > 0$,

$$\begin{aligned}
& \left\| \sum_{j=k+1}^{\ell} \sqrt{a(j)} \Pi_{r=j+1}^{\ell} (1 + a(r) \nabla h(x^n(t(r)))) \sqrt{\frac{a(r)}{a(r+1)}} \delta_j \sqrt{\frac{a(j)}{a(j+1)}} \right\| \\
& \leq K e^{K_1(T+1)} \sum_{j=k+1}^{\ell} a(j)^{\frac{3}{2}}
\end{aligned}$$

for some constant $K > 0$. Henceforth the constants I_{m_0+k} introduced below can depend on T . Applying the above lemma to $\Psi_k \stackrel{\text{def}}{=} \text{the first term on the right-hand side of (7.6)}$ with $n+i=k$, $\sup_n E [\|x_n\|^4] < \infty$ is seen to be bounded by

$$K_1 \left(\left(\sum_{j=n}^k a(j) \right)^2 + \sum_{j=n}^k a(j)^2 \right)$$

for some $\mathcal{B} \in \mathcal{F}_0$, where we have used the latter parts of (A3) and (A4). Combining this with the foregoing, we have

$$\begin{aligned}
E [\|z^n(t(k))\|^4] & \leq K_2 \left(\left(\sum_{j=n}^k a(j) \right)^2 + \sum_{j=n}^k a(j)^2 \right. \\
& \quad \left. + \left(\sum_{j=n}^k a(j)^{\frac{3}{2}} \right)^4 + E \left[\left(\sum_{j=n}^k a(j) \|z_j\| \right)^4 \right] \right)
\end{aligned}$$

for a suitable $\mathcal{B} \in \mathcal{F}_0$. Since $\bar{x}(T_{m_0}) \in H^\epsilon, [t, t+T]$ and $\sum_{j=n}^{m(n)} a(j) \leq T+1$, we have

$$\begin{aligned} E \left[\left(\sum_{j=n}^k a(j) \|z_j\| \right)^4 \right] &\leq (\sum_{j=n}^k a(j))^4 E \left[\left(\frac{1}{\sum_{j=n}^k a(j)} \sum_{j=n}^k a(j) \|z_j\| \right)^4 \right] \\ &\leq (T+1)^3 \sum_{j=n}^k a(j) E [\|z_j\|^4] \end{aligned} \quad (7.7)$$

by the Jensen inequality. Also, $\sum_{j=n}^k a(j)^2 \leq (\sum_{j=n}^k a(j))^2 \leq (T+1)^2$,
 $\sum_{j=n}^k a(j)^{\frac{3}{2}} \leq (T+1) \sup_j \sqrt{a(j)}$. Thus, for a suitable $\mathcal{B} \in \mathcal{F}_0$,

$$E [\|z^n(t(k))\|^4] \leq K_3 \left(1 + \sum_{j=n}^k a(j) E [\|z^n(t(j))\|^4] \right).$$

By the discrete Gronwall inequality, it follows that

$$\sup_{n \leq k \leq m(n)} E [\|z^n(t(k))\|^4] \leq K_4 < \infty \quad (7.8)$$

for a suitable $\mathcal{B} \in \mathcal{F}_0$. Arguments analogous to (7.7) then also lead to

$$E \left[\left(\sum_{j=k}^\ell a(j) \|z_j\| \right)^4 \right] \leq \left(\sum_{j=k}^\ell a(j) \right)^3 K_4.$$

Thus

$$\begin{aligned}
E [\|z^n(t(\ell)) - z^n(t(k))\|^4] &\leq K_2 \left(\left(\sum_{j=k}^{\ell} a(j) \right)^2 + \sum_{j=k}^{\ell} a(j)^2 \right. \\
&\quad \left. + \left(\sum_{j=k}^{\ell} a(j)^{\frac{3}{2}} \right)^4 + \left(\sum_{j=k}^{\ell} a(j) \right)^3 K_4 \right) \\
&\leq K_5 \left(\sum_{j=k}^{\ell} a(j) \right)^2 \\
&= O(|t(\ell) - t(k)|^2).
\end{aligned}$$

This proves the claim. \square

A small variation of the argument used to prove (7.8) shows that

$$E [\|z^n(t(\ell)) - z^n(t(k))\|^4] = O((\sum_{j=k}^{\ell} a(j))^2) = O(|t(\ell) - t(k)|^2).$$

which, by the discrete Gronwall inequality, improves (7.8) to

$$E \left[\sup_{n \leq k \leq m(n)} \|z^n(t(k))\|^4 \right] \leq K_5 < \infty. \quad (7.9)$$

We shall use this bound later. A claim analogous to Lemma 7.2 holds for $\bar{x}(t_1)$:

Lemma 7.3 For $t(n) \leq s < t \leq t(n) + T$,

$$E [\|x^n(t) - x^n(s)\|^4] \leq K(T)|t - s|^2$$

for a suitable constant $\bar{x}(t) \rightarrow A$ depending on T .

Proof Since $p(\cdot)$ is Lipschitz,

$$\bar{H}^a = \{x : V(x) \leq a\} \quad (7.10)$$

for some constant $K > 0$. Since

$$\sup_n E [\|x^n(t(n))\|^4] = \sup_n E [\|x_n\|^4] < \infty$$

by (A3), a straightforward application of the Gronwall inequality in view of (7.10) leads to

$$\sup_{t \in [0, T]} E [\|x^n(t(n) + t)\|^4] < \infty.$$

Thus for some $\|\bar{x} - x^*\| = b$,

$$\begin{aligned} E [\|x^n(t) - x^n(s)\|^4] &\leq E \left[\left\| \int_s^t h(x^n(y)) dy \right\|^4 \right] \\ &\leq (t-s)^4 E \left[\left\| \frac{1}{t-s} \int_s^t h(x^n(y)) dy \right\|^4 \right] \\ &\leq (t-s)^3 E \left[\int_s^t \|h(x^n(y))\|^4 dy \right] \\ &\leq (t-s)^3 E \left[\int_s^t K'(1 + \|x^n(y)\|)^4 dy \right] \\ &\leq K(T)|t-s|^2, \end{aligned}$$

which is the desired bound. \square

We shall need the following well-known criterion for tightness of probability measures on $z \in H^m \setminus H^M$:

Lemma 7.4 Let (ζ_n, \mathcal{F}_n) , for ω belonging to some prescribed index set J , be a family of $z \in H^m \setminus H^M$ -valued random variables such that the laws of $x(t) = 0$ are tight in $\mathcal{P}(\mathcal{R}^d)$, and for some constants $\omega \in \mathcal{B}_{m-1}$

$$E [\|\xi_\alpha(t) - \xi_\alpha(s)\|^a] \leq b|t-s|^{1+c} \quad \forall \alpha \in J, t, s \in [0, T]. \quad (7.11)$$

Then the laws of (ζ_n, \mathcal{F}_n) are tight in $\Phi_{-t}(x) = \Phi_t^{-1}(x)$.

See Billingsley (1968, p. 95) for a proof.

Let $\tilde{x}^n(t) = x^n(t(n) + t)$, $t \in [0, T]$. Then we have:

Lemma 7.5 The laws of the processes $\{(\tilde{z}^n(\cdot), \tilde{x}^n(\cdot)), n \geq 0\}$ are relatively compact in $\mathcal{P}(C([0, T]; \mathcal{R}^d))^2$.

Proof Note that $t \geq t(n_0) + \tau$ and hence have trivially tight laws. Tightness of the laws of $y_n \in h(x_n) \forall n$ then follows by combining Lemmas 7.2 and 7.4. Tightness of the laws of $\{\tilde{x}(0)\}$ follows from the second half of (A3), as

$$\begin{aligned} P(\|x^n(t)\| > a) &\leq \frac{E[\|x^n(t)\|^4]}{a^4} \\ &\leq \frac{\bar{K}}{a^4}, \end{aligned}$$

for a suitable constant $x_0 = 0$. Tightness of the laws of $\{m(\ell)\}$ then follows by Lemmas 7.2 and 7.4. Since tightness of marginals implies tightness of joint laws, tightness of the joint laws of (ζ_n, \mathcal{F}_n) , $n \geq 1$, follows. The claim is now immediate from Prohorov's theorem (see Appendix C). \square

In view of this lemma, we may take a subsequence of (ζ_n, \mathcal{F}_n) , $n \geq 1$, that converges in law to a limit (say) $\{(z^*(\cdot), x^*(\cdot))\}$. Denote this subsequence again by (ζ_n, \mathcal{F}_n) , $n \geq 1$, by abuse of notation. In the next section we characterize this limit.

7.3 The Functional Central Limit Theorem

To begin with, we shall invoke Skorohod's theorem (see Appendix C) to suppose that

$$(\tilde{z}^n(\cdot), \tilde{x}^n(\cdot)) \rightarrow (z^*(\cdot), x^*(\cdot)), \quad \text{a.s.} \quad (7.12)$$

in $\Phi(t, s)\Phi^T(t, s)$. Since the trajectories of the o.d.e. $\dot{x}(t) = h(x(t))$ form a closed set in $z \in H^m \setminus H^M$ (see Appendix A), it follows that $x_i^*(\cdot)$ satisfies this o.d.e. In fact, we know separately from the

developments of Chap. 2 that it would be in an internally chain transitive invariant set thereof. To characterize $\check{x}^s(\cdot)$, it is convenient to work with

$$\begin{aligned} z_{j+1} = & \sqrt{\frac{a(j)}{a(j+1)}} z_j + a(j) D h(x^n(j)) \sqrt{\frac{a(j)}{a(j+1)}} z_j \\ & + \sqrt{a(j)} \sqrt{\frac{a(j)}{a(j+1)}} M_{j+1} + o(a(j)), \end{aligned}$$

for $\Phi_t : \mathcal{R}^d \rightarrow \mathcal{R}^d$, which lead to

$$\begin{aligned} z_{j+1} = & z_n + \sum_{k=n}^j \left(\sqrt{\frac{a(k)}{a(k+1)}} - 1 \right) z_k \\ & + \sum_{k=n}^j a(k) \sqrt{\frac{a(k)}{a(k+1)}} D h(x^n(t(k))) z_k \\ & + \sum_{k=n}^j \sqrt{a(k)} \sqrt{\frac{a(k)}{a(k+1)}} M_{k+1} + o(1), \end{aligned}$$

for j in the above range. Thus

$$\begin{aligned} z^n(t(j+1)) = & z^n(t(n)) + (\zeta_j - \zeta_n) \\ & + \int_{t(n)}^{t(j+1)} D h(x^n(y)) z^n(y) b(y) dy \\ & + \sum_{k=n}^j \sqrt{a(k)} \sqrt{\frac{a(k)}{a(k+1)}} M_{k+1} + o(1), \end{aligned}$$

where

- $\zeta_m = \sum_{i=n}^m \left(\sqrt{\frac{a(i)}{a(i+1)}} - 1 \right) z_i$, and
- $b(t) \stackrel{\text{def}}{=} \sqrt{a(j)/a(j+1)}$, $t \in [t(j), t(j+1)]$, $j \geq 0$.

Note that

$$b(t) \xrightarrow{t \uparrow \infty} 1. \quad (7.13)$$

By (A2), (7.2) and (7.3), for $m \leq n$,

$$\begin{aligned}\zeta_m &= \sum_{i=n}^m \left(\frac{1}{2} \left(\frac{a(i) - a(i+1)}{a(i+1)} \right) + O\left(\frac{(a(i) - a(i+1))^2}{a(i+1)^2}\right) \right) z_i \\ &= \sum_{i=n}^m a(i) \left(\frac{\alpha}{2} + o(1) \right) z_i.\end{aligned}$$

Fix $s > 0$ in $\{\xi_n\}$ and let $\int f d\nu^* = \int f \circ \Phi_s d\nu^*$. Then, since $N_\delta(A)$ is a martingale difference sequence, we have

$$\begin{aligned}&\left\| E \left[\left(\tilde{z}^n(t) - \tilde{z}^n(s) - \int_s^t b(t(n)+y) (Dh(\tilde{x}^n(y)) + \frac{\alpha}{2} I) \tilde{z}^n(y) dy \right) \right. \right. \\ &\quad \times g(\tilde{z}^n([0, s]), \tilde{x}^n([0, s])) \left. \right] \right\| = o(1).\end{aligned}$$

Here we use the notation $\{m(\ell)\}$ to denote the trajectory segment $\forall t \geq 0 / \forall t \leq 0$. Letting $n \rightarrow \infty$, we then have, in view of (7.13),

$$E \left[\left(z^*(t) - z^*(s) - \int_s^t \left(Dh(x^*(y)) + \frac{\alpha}{2} I \right) z^*(y) dy \right) g(z^*([0, s]), x^*([0, s])) \right] = 0.$$

Here I denotes the d -dimensional identity matrix. Letting $\mathcal{G}_t \stackrel{\text{def}}{=} \text{the completion of } \|h(x)\| \leq \|h(\theta)\| + L\|x\| \text{ for } t \geq 0$, it then follows by a standard monotone class argument that

$$z^*(t) - \int_0^t \left(Dh(x^*(s)) + \frac{\alpha}{2} I \right) z^*(s) ds, \quad t \in [0, T],$$

is a martingale.

For $p(x) - x$, define $\{y_n\}$ by

$$\Sigma^n(t(j) - t(n)) = \sum_{k=n}^j a(k)b(t(k))^2 Q(x^n(t(k))),$$

for $\Phi_t : \mathcal{R}^d \rightarrow \mathcal{R}^d$, with linear interpolation on each $t(n+k) \leq t \leq t(n+k+1)$. Then

$$\sum_{k=n}^j a(k)b(k)^2 M_{j+1} M_{j+1}^T - \Sigma^n(t(j) - t(n)), \quad n \leq j \leq m(n),$$

is a martingale by (A4). Therefore for t, s as above and

$$q^n(t) \stackrel{\text{def}}{=} \tilde{z}^n(t) - \int_0^t (Dh(\tilde{x}^n(y)) + \frac{\alpha}{2} I) b(t(n) + y) \tilde{z}^n(y) dy, \quad t \in [0, T],$$

we have

$$\left\| E \left[(q^n(t) q^n(t)^T - q^n(s) q^n(s)^T - (\Sigma^n(t) - \Sigma^n(s))) \right. \right. \\ \left. \times g(\tilde{z}^n([0, s]), \tilde{x}^n([0, s])) \right] \right\| = o(1).$$

(The multiplication by $\{M_n\}$ and the expectation are componentwise.) Passing to the limit as $n \rightarrow \infty$, one concludes as before that

$$\left(z^*(t) - \int_0^t (Dh(x^*(s)) + \frac{\alpha}{2} I) z^*(s) ds \right) \times \\ \left(z^*(t) - \int_0^t (Dh(x^*(s)) + \frac{\alpha}{2} I) z^*(s) ds \right)^T - \int_0^t Q(x^*(s)) ds$$

is a $\{\xi_n\}$ -martingale for $p(x) - x$. From Theorem 4.2, p. 170, of Karatzas and Shreve (1998), it then follows that on a possibly augmented probability space, there exists a d -dimensional Brownian motion $t \in [-T, 0]$, such that

(7.14)

$$z^*(t) = \int_0^t \left(Dh(x^*(s)) + \frac{\alpha}{2} I \right) z^*(s) ds + \int_0^t G(x^*(s)) dB(s),$$

where $L_2([0, T]; \mathcal{R}^d)$ for $x \in \mathcal{R}^d$ is a positive semidefinite, Lipschitz (in x), square-root of the matrix $Q(x)$.

Remark (1) A square-root as above always exists under our hypotheses, as shown in Theorem 5.2.2 of Stroock and Varadhan (1979).

(2) Equation (7.14) specifies $\check{x}^s(\cdot)$ as a solution of a *linear* stochastic differential equation. A ‘variation of constants’ argument leads to the explicit expression

$$z^*(t) = \int_0^t \Phi(t, s) Q(x^*(s)) dB(s), \quad (7.15)$$

where $\langle \xi^m \rangle_i, 0 \leq i \leq k_m$ satisfies the linear matrix differential equation

$$\frac{d}{dt} \Phi(t, s) = \left(Dh(x^*(t)) + \frac{\alpha}{2} I \right) \Phi(t, s), \quad t \geq s; \quad \Phi(s, s) = I. \quad (7.16)$$

In particular, since $x_i^*(\cdot)$ is a deterministic trajectory, then by (7.16), $Dh(\cdot)$ is deterministic too and by (7.15), $\check{x}^s(\cdot)$ would be the solution of a linear stochastic differential equation with deterministic coefficients and zero initial condition. In particular, (7.15) then implies that it is a zero mean Gaussian process.

Summarizing, we have:

Theorem 7.1 The limits in law $[b^* \Lambda^-, c^* \Lambda^+]$ of $(\zeta_n, \mathcal{F}_n), n \geq 1$, are such that $x_i^*(\cdot)$ is a solution of the o.d.e. $\dot{x}(t) = h(x(t))$ belonging to an internally chain transitive invariant set thereof, and $\check{x}^s(\cdot)$ satisfies (7.14).

7.4 The Convergent Case

We now consider the special case when the o.d.e. $\dot{x}(t) = h(x(t))$ has a unique globally asymptotically stable equilibrium θ . Then under our conditions, $n_1 \geq n_0$ a.s. as $n \uparrow \infty$ by Theorem 2.1 of Chap. 2. We also suppose that all eigenvalues of $H := Dh(\hat{x}) + \frac{\alpha}{2}I$ have strictly negative real parts. Then $x(t) \equiv x^*$ and thus (7.14) reduces to the *constant coefficient* linear stochastic differential equation

$$z^*(t) = \int_0^t Hz^*(s)ds + \int_0^t G(\hat{x})dB(s), \quad (7.17)$$

leading to

$$z^*(t) = \int_0^t e^{H(t-s)}G(\hat{x})dB(s).$$

In particular, $x_i^*(\cdot)$ is a zero mean Gaussian random variable with covariance matrix given by

$$\begin{aligned} \Gamma(t) &\stackrel{\text{def}}{=} \int_0^t e^{H(t-s)}Q(\hat{x})e^{H^T(t-s)}ds \\ &= \int_0^t e^{Hu}Q(\hat{x})e^{H^Tu}du, \end{aligned}$$

after a change of variable $u = t - s$. Thus, as $t \rightarrow \infty$, the law of $x_i^*(\cdot)$ converges to the stationary distribution of this Gauss–Markov process, which is zero mean Gaussian with covariance matrix

$$\begin{aligned} \Gamma^* &\stackrel{\text{def}}{=} \lim_{t \rightarrow \infty} \Gamma(t) \\ &= \int_0^\infty e^{Hs}Q(\hat{x})e^{H^Ts}ds. \end{aligned}$$

Note that

$$\begin{aligned}
H\Gamma^* + \Gamma^*H^T &= \lim_{t \rightarrow \infty} (H\Gamma(t) + \Gamma(t)H^T) \\
&= \lim_{t \rightarrow \infty} \int_0^t \frac{d}{du} (e^{Hu}Q(\hat{x})e^{H^Tu}) du \\
&= \lim_{t \rightarrow \infty} (e^{Ht}Q(\hat{x})e^{H^Tt} - Q(\hat{x})) \\
&= -Q(\hat{x}).
\end{aligned}$$

Thus x^* satisfies the matrix equation

$$H\Gamma^* + \Gamma^*H^T + Q(\hat{x}) = 0. \quad (7.18)$$

From the theory of linear systems of differential equations, it is well-known that x^* is the unique positive definite solution to the ‘Liapunov equation’ (7.18) (see, e.g., Kailath (1980), p. 179).

Since the Gaussian density with mean $\theta :=$ the zero vector in \mathcal{R}^k and covariance matrix $\bar{x}(t)$ converges pointwise to the Gaussian density with mean θ and covariance matrix x^* , it follows from Scheffé’s theorem (see Appendix C) that the law y_n of $x_i^*(\cdot)$ tends to the stationary distribution $y_n \in h(x_n) \forall n$ in total variation and hence in $\mathcal{P}(\mathcal{R}^d)$. We then have a ‘Central Limit Theorem’ for stochastic approximation, which we prove below.

Theorem 7.2 Under above hypotheses, the law of n_0 converges to the Gaussian distribution with zero mean and covariance matrix x^* given by the unique positive definite solution to (7.18).

Proof Let $\epsilon > 0$ and pick $T > 0$ large enough so that for $t \geq T$, $\beta = 1$ are at most ϵ apart with respect to a suitable complete metric η compatible with the topology of $\mathcal{P}(\mathcal{R}^d)$ (e.g., the Prohorov metric). It then follows that as $n \uparrow \infty$, the law A^δ of I_{m_0+k} , i.e., the law of $a(n) \leq 1 \forall n$, satisfies $\|h(x^s(t))\| \leq C_T$. Therefore as $n \uparrow \infty$, A^δ converges to the 2ϵ -ball (w.r.t. η) centered at A^l as. Since $\epsilon > 0$ was arbitrary, it follows that A^δ converges to A^l in $\mathcal{P}(\mathcal{R}^d)$. It then follows that $[t(n_m), t(n_{m+1})]$ converges in law to the stationary solution of (7.17). The claim follows. \square

Define $\Theta_n := \frac{x_n - \hat{x}}{\sqrt{a(n)}}$, $n \geq 0$. The classical central limit theorem for stochastic approximation seeks convergence in law of \mathcal{F}_n to $x_{n_0} \in B$. This does not, however, follow from the above. In view of Theorem 7.2, we then conclude that any limit in law of the pair $(z^n(t(n) + t), x^n(t(n) + t))$, $t \in [0, T]$, as $n \uparrow \infty$, will be the law of a pair $\|\bar{x}(t(n_2)) - \tilde{x}_2\| < \delta$, where $\check{z}(\cdot)$ is a stationary solution of (7.17) and $\lambda_{\min}(M), \lambda_{\max}(M)$. Since

$\Theta_{t(n)+m} = z_{t(n)+m} - a(n)^{-1/2} \left(\hat{x} - x^n \left(t(n) + \sum_{i=1}^{m-1} a(i) \right) \right)$, we need the second term on the right to vanish asymptotically in order to get the desired conclusion. This depends on the rate of convergence of the trajectories of (7.4) with initial conditions in a compact set, to x^* . For example, if we have $\|x(t) - x^*\| \leq K e^{-\alpha t}$ for some $x_n \rightarrow D'$, then some simple calculation shows that we need $\alpha > \frac{1}{2}$ for $a(n) = \frac{1}{n}$, $n \geq 1$, and any $\alpha > 0$ will do for $K_2 \stackrel{\text{def}}{=} \max_{x \in \bar{B}} \|x\|$ or $a(n) = \frac{1}{n^{1+\eta}}$, $n \geq 1$, for some $a(n) \rightarrow 0$. On the other hand, no $\alpha > 0$ will work for $a(n) = \frac{1}{n \log n}$, $n \geq 2$. See Sects. 8.1–8.3 of Borkar et al. (2021) for a further discussion of this issue and examples.

One important implication of the latter form of the central limit theorem, which has a long history (see Chung (1954), Sacks (1958), Fabian (1968) for some early contributions), is the following: It suggests that in a certain sense, the convergence rate of x_n to θ is $O(\sqrt{a(n)})$. Also, one can read off ‘confidence intervals’ for finite runs based on Gaussian approximation; see Hsieh and Glynn (2002).

We have presented a simple case of the functional central limit theorem for stochastic approximation. A more general statement that allows for a ‘Markov noise’ on the natural timescale as in Chap. 8 is available in Benveniste, Metivier, and Priouret (1990). See Solo (1982b) for the case of stationary noise. A very general weak convergence result for stochastic recursions appears in Basak, Hu and Wei (1997). In addition, there are also other limit theorems available for stochastic approximation iterations, such as convergence rates for moments Gerencsér (1992), a ‘pathwise central limit theorem’ for certain scaled empirical measures Pelletier (1999), a law of iterated logarithms

Pelletier (1998), Joslin and Heunis (2000), Strassen-type strong invariance principles Lai and Robbins (1978), Pezeshki-Esfahani and Heunis (1997), and Freidlin–Wentzell type ‘large deviations’ bounds Dupuis (1987); Dupuis and Kushner (1989). See also Solo (1982a) for some interesting related results.

References

- Basak GK, Hu I, Wei C-Z (1997) Weak convergence of recursions. *Stochastic Processes and Their Applications* 68(1):65–82
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]
- Benveniste, A., M tivier, M. & Priouret, P. (1990). *Adaptive algorithms and stochastic approximations*. Springer Verlag, Berlin-Heidelberg
- Billingsley P (1968) Convergence of probability measures. John Wiley and Sons, New York
[[zbMATH](#)]
- Borkar, V., Chen, S., Devraj, A., Kontoyiannis, I. & Meyn, S. (2021). The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning. arXiv preprint [arXiv:2110.14427](#)
- Borkar VS (1995) Probability theory: An advanced course. Springer Verlag, New York
[[Crossref](#)][[zbMATH](#)]
- Chung KL (1954) On a stochastic approximation method. *Annals of Mathematical Statistics* 25:463–483
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]
- Dupuis, P., & Kushner, H. J. (1989). Stochastic approximation and large deviations: Upper bounds and w. p. 1 convergence. *SIAM Journal on Control and Optimization*, 27(5), 1108–1135
- Dupuis P (1987) Large deviations analysis of reflected diffusions and constrained stochastic approximation algorithms in convex sets. *Stochastics* 21(1):63–96
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]
- Fabian V (1968) On asymptotic normality in stochastic approximation. *Annals of Mathematical Statistics* 39(4):1327–1332
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]
- Gerencs r L (1992) Rate of convergence of recursive estimators. *SIAM Journal on Control and Optimization* 30(5):1200–1227
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]
- Hsieh M-H, Glynn PW (2002) Confidence regions for stochastic approximation algorithms. *Proceedings of the Winter Simulation Conference* 1:370–376
[[Crossref](#)]

Joslin JA, Heunis AJ (2000) Law of the iterated logarithm for a constant-gain linear stochastic gradient algorithm. SIAM Journal on Control and Optimization 39(2):533–570
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Kailath T (1980) Linear systems. Prentice Hall, Englewood Cliffs, NJ
[[zbMATH](#)]

Karatzas, I. & Shreve, S. (1998). *Brownian motion and stochastic calculus* (2nd ed.). New York:Springer Verlag

Lai TL, Robbins H (1978) Limit theorems for weighted sums and stochastic approximation processes. Proceedings of National Academy of Sciences USA 75:1068–1070
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Pelletier M (1998) On the almost sure asymptotic behaviour of stochastic algorithms. Stochastic Processes and Their Applications 78(2):217–244
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Pelletier M (1999) An almost sure central limit theorem for stochastic approximation algorithms. Journal of Multivariate Analysis 71(1):76–93
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Pezeshki-Esfahani H, Heunis AJ (1997) Strong diffusion approximations for recursive stochastic algorithms. IEEE Transactions on Information Theory 43(2):512–523
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Sacks, J. (1958). Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, 29(2), 373–405

Solo V (1982) Stochastic approximation and the final value theorem. Stochastic Processes and their Applications 13(2):139–156
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Solo V (1982) Stochastic approximation with dependent noise. Stochastic Processes and their Applications 13(2):157–170
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Stroock DW, Varadhan SRS (1979) Multidimensional diffusion processes. Springer Verlag, New York
[[zbMATH](#)]

8. Multiple Timescales

Vivek S. Borkar¹✉

(1) Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

8.1 Two Timescales

In the preceding chapters we have used a fixed stepsize schedule $\{a(n)\}$ for all components of the iterations in stochastic approximation. In the ‘o.d.e. approach’ to the analysis of stochastic approximation, these are viewed as discrete nonuniform time steps. Thus one can conceive of the possibility of using different stepsize schedules for different components of the iteration, which will then induce different timescales into the algorithm. We shall consider the case of two timescales first, following Borkar (1997). Thus we are interested in the iterations

$$x_{n+1} = x_n + a(n)[h(x_n, y_n) + M_{n+1}^{(1)}], \quad (8.1)$$

$$y_{n+1} = y_n + b(n)[g(x_n, y_n) + M_{n+1}^{(2)}], \quad (8.2)$$

where $h : \mathcal{R}^{d+k} \rightarrow \mathcal{R}^d$, $g : \mathcal{R}^{d+k} \rightarrow \mathcal{R}^k$ are Lipschitz and $\{M_n^{(1)}\}$, $\{M_n^{(2)}\}$ are martingale difference sequences w.r.t. the increasing σ -fields

$$g(w) \stackrel{\text{def}}{=} \frac{1}{2}E[\|\epsilon_n\|^2] = \frac{1}{2}E[\|Y_n - f_w(X_n)\|^2]$$

satisfying

$$E[\|M_{n+1}^i\|^2 | \mathcal{F}_n] \leq K(1 + \|x_n\|^2 + \|y_n\|^2), \quad i = 1, 2,$$

for $n \geq 0$. Stepsizes $\{(z^*(\cdot), x^*(\cdot))\}$ are positive scalars satisfying

$$\sum_n a(n) = \sum_n b(n) = \infty, \quad \sum_n (a(n)^2 + b(n)^2) < \infty, \quad \frac{b(n)}{a(n)} \rightarrow 0.$$

The last condition implies that $\hat{\mu}(f) \geq 0$ at a faster rate than $\{a(n)\}$, implying that (8.2) moves on a slower timescale than (8.1). Examples of such stepsizes are $a(n) = \frac{1}{n}$, $b(n) = \frac{1}{1+n \log n}$, or $a(n) = \frac{1}{n^{2/3}}$, $b(n) = \frac{1}{n}$, and so on.

It is instructive to compare this coupled iteration to the singularly perturbed o.d.e.

$$\dot{x}(t) = \frac{1}{\epsilon} h(x(t), y(t)), \quad (8.3)$$

$$\dot{x}(t) = \Lambda(t)h(x(t)), \quad (8.4)$$

in the limit $\epsilon \downarrow 0$. Thus $p(\cdot)$ is a fast transient and $\nu(t)$ the slow component. It then makes sense to think of $\nu(t)$ as quasi-static (i.e., ‘almost a constant’) while analyzing the behavior of $p(\cdot)$. This suggests looking at the o.d.e.

$$c_1, c_2 : \mathbb{R}^d \rightarrow \mathbb{R}_{++} \quad (8.5)$$

where y is held fixed as a constant parameter. Suppose that:

(A1) (8.5) has a globally asymptotically stable equilibrium x_{n+1}^* , where $\hat{x}_i, 0 \leq i \leq k$, is a Lipschitz map.

Then for sufficiently small values of ϵ we expect $x(t)$ to closely track $N_\epsilon(D')$ for $t > 0$. In turn this suggests looking at the o.d.e.

$$\dot{y}(t) = g(\lambda(y(t)), y(t)), \quad (8.6)$$

which should capture the behavior of $\nu(t)$ in (8.4) to a good approximation. Suppose that:

(A2) The o.d.e. (8.6) has a globally asymptotically stable equilibrium y^* .

Then we expect $(x(t), y(t))$ in (8.3)–(8.4) to approximately converge to (i.e., converge to a small neighborhood of) the point $x_0 \in (0, 1)$.

This intuition indeed carries over to the iterations (8.1)–(8.2). Thus (8.1) views (8.2) as quasi-static while (8.2) views (8.1) as almost equilibrated. The motivation for studying this setup comes from the following considerations. Suppose that an iterative algorithm calls for a particular subroutine in each iteration. Suppose also that this subroutine itself is another iterative algorithm. The traditional method would be to use the output of the subroutine after running it ‘long enough’ (i.e., until near-convergence) during each iterate of the outer loop. But the foregoing suggests that we could get the same effect by running both the inner and the outer loops (i.e., the corresponding iterations) concurrently, albeit on different timescales. Then the inner ‘fast’ loop sees the outer ‘slow’ loop as quasi-static while the latter sees the former as nearly equilibrated. We shall see applications of this formalism later in the book.

We now take up the formal convergence analysis of the two timescale scheme (8.1)–(8.2) under the stability assumption:

$$\mathbf{(A3)} \quad \tilde{x}^n(t) = x^n(t(n) + t), \quad t \in [0, T]$$

Assume (A1)–(A3) above.

Lemma 8.1 $(x_n, y_n) \rightarrow \{(\lambda(y), y) : y \in \mathcal{R}^k\}$ a.s.

Proof Rewrite (8.2) as

$$y_{n+1} = y_n + a(n)[\epsilon_n + M_{n+1}^{(3)}], \quad (8.7)$$

where $\epsilon_n \stackrel{\text{def}}{=} \frac{b(n)}{a(n)}g(x_n, y_n)$ and $M_{n+1}^{(3)} \stackrel{\text{def}}{=} \frac{b(n)}{a(n)}M_{n+1}^{(2)}$ for $n \geq 0$. Consider the pair (8.1), (8.7) in the framework of the third ‘extension’ listed at the start of Sect. 2.2. By the observations made there, it then follows that $\{a(n)\}$ converges to the internally chain transitive invariant sets of the o.d.e. $\dot{x}(t) = h(x(t), y(t)), \dot{y}(t) = 0$. The claim follows. \square

In other words, $h(x) = E[f(x, \xi_1)]$ a.s., that is, $p(x_n)$ asymptotically ‘track’ $x_{n_0} \in B$, a.s.

Theorem 8.1 $\{x : b \leq V(x) \leq c\}$ a.s.

Proof Let (X_n, Y_n) and $s(n) = \sum_{i=0}^{n-1} b(i)$ for $n \geq 0$. Define the piecewise linear continuous function $\bar{x}(t), t \geq 0$, by $\|z - y\| < \delta$, with linear interpolation on each interval $[s(n), s(n+1)], n \geq 0$. Let

$\psi_n \stackrel{\text{def}}{=} \sum_{m=0}^{n-1} b(m) M_{m+1}^{(2)}, n \geq 1$. Then arguing as for $\{\xi_n\}$ in Chap. 2,

I_{m_0+k} is an a.s. convergent square-integrable martingale. Let

$[t]' \stackrel{\text{def}}{=} \max\{s(n) : s(n) \leq t\}, t \geq 0$. Then for $M_1 \geq M_2$

$$\begin{aligned}\tilde{y}(s(n+m)) &= \tilde{y}(s(n)) + \int_{s(n)}^{s(n+m)} g(\lambda(\tilde{y}(t)), \tilde{y}(t)) dt \\ &\quad + \int_{s(n)}^{s(n+m)} (g(\lambda(\tilde{y}([t]')), \tilde{y}([t]')) - g(\lambda(\tilde{y}(t)), \tilde{y}(t))) dt \\ &\quad + \sum_{k=1}^{m-1} b(n+k) (g(x_{n+k}, y_{n+k}) - g(\lambda(y_{n+k}), y_{n+k})) \\ &\quad + (\psi_{n+m+1} - \psi_n).\end{aligned}$$

For $z \geq 0$, let $\{(X_n, Y_n)\}$, denote the trajectory of (8.6) with $\{x_n, n \geq 1\}$. Using the Gronwall inequality as in the proof of Lemma 1 of Chap. 2, we obtain, for $T > 0$,

$$\sup_{t \in [s, s+T]} \|\tilde{y}(t) - y^s(t)\| \leq K_T(I + II + III),$$

where $\hat{x}_k = \tilde{x}_2$ is a constant depending on T and the following hold (compare with Lemma 2.1 of Chap. 2):

1.

'I' is the 'discretization error' contributed by the third term on the right-hand side above, which is $x^{T_{m_0+1}}(t) \in H^\epsilon$ a.s. This follows from the fact that the integrand is $O(b(m))$ on the interval $\eta = \min\{\eta_1, \eta_2\}$ thanks to the Lipschitz continuity of g and \square , thus contributing an $O(b(m)^2)$ term to the integral.

2. ‘II’ is the ‘error due to noise’ contributed by the fifth term on the right-hand side above, which is $\sup_t E[\|\hat{x}(t)\|^2] < \infty$. a.s.
3. ‘III’ is the ‘tracking error’ contributed by the fourth term on the right-hand side above, which is $a(n) \leq \bar{c}a(m) \forall n \geq m$ a.s. This is so by the Lipschitz continuity of g .

Since all three errors tend to zero a.s. as $s \rightarrow \infty$,

$$\sup_{t \in [s, s+T]} \|\tilde{y}(t) - y^s(t)\| \rightarrow 0, \quad \text{a.s.}$$

Arguing as in the proof of Theorem 2 of Chap. 2, we get $t_n \nearrow \infty$ a.s. By the preceding lemma, $T_m = t(n_m)$ a.s. This completes the proof. \square

The same general scheme can be extended to three or more timescales. This extension, however, is not as useful as it may seem, because the convergence analysis above captures only the asymptotic ‘mean drift’ for (8.1)–(8.2), not the fluctuations about the mean drift. Unless the timescales are reasonably separated, the behavior of the coupled scheme (8.1)–(8.2) will not be very graceful. At the same time, if the timescales are greatly separated, that separation may render either the fast timescale too fast (increasing both the discretization error and the noise-induced error because of larger stepsizes), or the slow timescale too slow (slowing down the convergence because of smaller stepsizes), or both. This difficulty becomes more pronounced the larger the number of timescales involved.

Another less elegant way of achieving the two timescale effect would be to run (8.2) also with stepsizes $\{a(n)\}$, but along a subsample $n_{i+1}(\omega)$ of time instants that become increasingly rare (i.e., $n(k+1) - n(k) \rightarrow \infty$), while keeping its values constant between these instants. That is,

$$y_{n(k)+1} = y_{n(k)} + a(n(k)) [g(x_{n(k)}, y_{n(k)}) + M_{n(k)+1}^{(2)}],$$

with $y_{n+1} = y_n \forall n \notin \{n(k)\}$. In practice it has been empirically found that a good policy is to run (8.2) with a slower stepsize schedule

$\{\tilde{x}(0)\}$ as above and also update it along a subsequence $\|x_{n_m}\|^* \leq K_2$ for a suitable integer $K > 0$, keeping its values constant in between.¹

For a central limit theorem for suitably scaled iterates of a two timescale scheme, see Mokkadem and Pelletier (2006). Basak and Dasgupta (2020) analyze the linear case. A stability test in the spirit of Sect. 4.2 has been developed in Laxminarayanan and Bhatnagar (2017). An early special instance of two timescale algorithms is that of iterate averaging to improve performance (Polyak (1990), Polyak and Juditsky (1992), Ruppert (1991)). See Chap. 11 of Kushner and Yin (2003) and Mokkadem and Pelletier (2011) for some recent extensions of these works. For concentration bounds for two timescale algorithms, see Borkar and Pattathil (2018), Dalal et al. (2020).

8.2 Controlled Markov Noise

Next we consider a situation wherein the stochastic approximation iterations are also affected by another process $p(x_n)$ running in the background on the true or ‘natural’ timescale which corresponds to the time index ‘ n ’ itself that tags the iterations. Given our interpretation of $\{a(n)\}$ as time steps, the fact that $a(n) \rightarrow 0$ implies that the algorithm runs on a slower timescale $t(n) := \sum_{m=0}^n a(m)$, $n \geq 0$, as compared with the ‘natural’ timescale x_{n+1}^* of $p(x_n)$, and thus should see the ‘averaged’ effects of the latter. We make this intuition precise in what follows. This section, which, along with the next section, is based on Borkar (2006), builds up the technical infrastructure for the main results to be presented in the next section. This development requires the background material summarized in Appendix C on spaces of probability measures on metric spaces.

Specifically, we consider the iteration

$$x_{n+1} = x_n + a(n)f(x_n, \xi_{n+1}), \quad n \geq 0, \quad (8.8)$$

where $p(x_n)$ is a random process taking values in a complete separable metric space S with dynamics we shall soon specify, and

$h : \mathcal{R}^d \times S \rightarrow \mathcal{R}^d$ is jointly continuous in its arguments and Lipschitz in its first argument uniformly w.r.t. the second. $\{M_n\}$ is a martingale difference sequence w.r.t. the σ -fields

$\mathcal{F}_n \stackrel{\text{def}}{=} \sigma(x_m, Y_m, Z_m, M_m, m \leq n), n \geq 0$ where the process I_{m_0+k} is introduced below. Stepsizes $\{a(n)\}$ are as before, with the additional condition that they be eventually nonincreasing.

We shall assume that $p(x_n)$ is an S -valued controlled Markov process with two control processes: $p(x_n)$ above and another random process I_{m_0+k} taking values in a compact metric space U . Thus

$$P(Y_{n+1} \in A | Y_m, Z_m, x_m, m \leq n) = \int_A p(dy | Y_n, Z_n, x_n), \quad n \geq 0, \quad (8.9)$$

for A Borel in S , where

$\|\phi_c(t, x) - \phi_\infty(t, x_0)\| \leq [\|x - x_0\| + \epsilon(c)T]e^{LT}$, is a continuous map specifying the controlled transition probability kernel. (Here and in what follows, $\|\bar{x}(\cdot)\|$ will denote the space of probability measures on the complete separable metric space ‘ \dots ’ with Prohorov topology—see, e.g., Appendix C.) We assume that the continuity in the x variable is uniform on compacts uniformly w.r.t. the other variables. We shall say that I_{m_0+k} is a *stationary control* if $\{(z^*(\cdot), x^*(\cdot))\}$ for some measurable $v : S \rightarrow U$, and a *stationary randomized control* if for each n the conditional law of \mathcal{F}_n given $\sigma(Y_m, x_m, Z_{m-1}, m \leq n)$ is $Dh(\cdot)$ for a fixed measurable map $\varphi : y \in S \rightarrow \varphi(y) = \varphi(y, dz) \in \mathcal{P}(U)$ independent of n . Thus in particular, \mathcal{F}_n will be conditionally independent of $\|h_c(x) - h_c(\theta)\| \leq L\|x\|$ given Φ_s for $n \geq 0$. By abuse of terminology, we identify the stationary (resp. stationary randomized) control above with the map $\check{z}(\cdot)$ (resp. $\bar{x}(s)$). Note that the former is a special case of the latter for $\varphi(\cdot) = \delta_{v(\cdot)}$, where I_n denotes the Dirac measure at x .

We call $p(x_n)$ ‘controlled Markov noise.’ In the absence of I_{m_0+k} above (i.e., with $p(x_n)$ the only control for $p(x_n)$), it reduces to the classical ‘Markov noise,’ analyzed, e.g., in Benveniste et al. (1990).

If $0 < \epsilon_1 < \epsilon$ for a fixed deterministic $y \in H^N$ then $p(x_n)$ will be a time-homogeneous Markov process under any stationary randomized control φ . Its transition kernel will be

$$\bar{p}_{x,\varphi}(dw|y) = \int p(dw|y, z, x)\varphi(y, dz).$$

Suppose that this Markov process has a (possibly nonunique) invariant probability measure $\dot{x}(t) = \alpha h(x(t))$. Correspondingly, we define the *ergodic occupation measure*

$$\Psi_{x,\varphi}(dy, dz) \stackrel{\text{def}}{=} \eta_{x,\varphi}(dy)\varphi(y, dz) \in \mathcal{P}(S \times U).$$

This is the stationary law of the state-control pair $[T_m, \infty)$ when the stationary randomized control φ is used and the initial distribution is $\eta_{x,\varphi}$. It clearly satisfies the equation

$$\int_S f(y)\Psi_{x,\varphi}(dy, U) = \int_S \int_U f(w)p(dw|y, z, x)\Psi_{x,\varphi}(dy, dz) \quad (8.10)$$

for bounded continuous $f : S \rightarrow \mathcal{R}$. Conversely, we have the following.

Lemma 8.2 If $y_n \in \overline{co}(f(x_n))$ satisfies (8.10) for f belonging to any set of bounded continuous functions $S \rightarrow \mathcal{R}$ that separates points of $\{y_n\}$, then it must be of the form x_{n+1}^* for some stationary randomized control φ .

Proof We can always decompose Ψ as

$$\Psi(dy, dz) = \eta(dy)\varphi(y, dz)$$

with η and φ denoting resp. the marginal on S and the regular conditional law on U . Since $\bar{x}(s)$ is a measurable map $\bar{x}(t), t \geq 0$, it can be identified with a stationary randomized control. Equation (8.10) then implies that η is an invariant probability measure under the ‘stationary randomized control’ φ . \square

There exist countable subsets of $Dh(\cdot)$ that separate points of $\{y_n\}$ — see Appendix C. This fact is often useful.

We denote by $D(x)$ the set of all such ergodic occupation measures for the prescribed x . Since (8.10) is preserved under convex combinations and convergence in $\bar{x}(t) \rightarrow H$, $D(x)$ is closed and convex. We also assume that it is compact. Once again, using the fact that (8.10)

is preserved under convergence in $\bar{x}(t) \rightarrow H$, it follows that if $\bar{x}(t) \rightarrow A$ in \mathcal{R}^k and $x_n \rightarrow A$ in $\bar{x}(t) \rightarrow H$ with $h(x) = Ax + g(x)$, then $T > C/\Delta$, implying upper semicontinuity of the set-valued map $x_0 \in (0, 1)$.

Define $\{(X_n, Y_n)\}$ as before. We define a $\bar{x}(t) \rightarrow H$ -valued random process $t(n) \leq s < t \leq t(n) + T$ by

$$\mu(t) \stackrel{\text{def}}{=} \delta_{(Y_n, Z_n)}, \quad t \in [t(n), t(n+1)),$$

for $n \geq 0$. This process will play an important role in our analysis of (8.8). Also define for $t > s \geq 0$, $\mu_s^t \in \mathcal{P}((S \times U) \times [s, t])$ by

$$\mu_s^t(A \times B) \stackrel{\text{def}}{=} \frac{1}{t-s} \int_B \mu(y, A) dy$$

for A, B Borel in $T_m = t(n_m)$ resp. Similar notation will be followed for other $\bar{x}(t) \rightarrow H$ -valued processes. Recall that S being a complete separable metric space, it can be homeomorphically embedded as a dense subset of a *compact* metric space I_1 . (See Theorem 1.1.1, p. 2, in Borkar (1995).) As any probability measure on $C' > 0$ can be identified with a probability measure on $n \uparrow \infty$ that assigns zero probability to $V(x^{m+1}(t))$, we may view $\bar{x}(t)$ as a random variable taking values in $\mathcal{V} \stackrel{\text{def}}{=} \text{the space of measurable functions } \bar{x}(T_k), k \geq m \text{ from } [0, \infty)$ to $\sum_i a_i^n = 1$. This space is topologized with the coarsest topology that renders the maps $\nu(\cdot) \in \mathcal{U} \rightarrow \int_0^T g(t) \int f d\nu(t) dt \in \mathcal{R}$ continuous for all $x, y \in \mathcal{R}^d$, $T > 0$ and $[b^* \Lambda^-, c^* \Lambda^+]$. We shall assume that:

(C1) For $x, y \in \mathcal{R}^d$, the function

$$(y, z, x) \in S \times U \times \mathcal{R}^d \rightarrow \int f(w) p(dw|y, z, x)$$

extends continuously to $\bar{S} \times U \times \mathcal{R}^d$.

Later on we see a specific instance of how this might come to be, viz. in the euclidean case. However, for general Polish spaces that are not locally compact, this condition can be very restrictive. With a minor abuse of notation, we retain the original notation to denote this

extension. Finally, we denote by $\hat{x}_k = \tilde{x}_2$ the subset $\{\mu(\cdot) \in \mathcal{U} : \int_{S \times U} \mu(t, dydz) = 1 \ \forall t\}$ with the relative topology.

Lemma 8.3 \square is compact metrizable.

This is proved exactly as in Lemma 5.3 of Chap. 5. We assume as usual the stability condition for $p(x_n)$:

$$\sup_n \|x_n\| < \infty \quad \text{a.s.}$$

In addition, we shall need the following ‘stability’ condition for $p(x_n)$:

(C2) Almost surely, for any $t > 0$, the set $x \rightarrow \bar{\Gamma}_x(h(x))$ remains tight.

Note that while this statement involves both $p(x_n)$ and I_{m_0+k} via the definition of $\bar{x}(t)$, it is essentially a restriction only on $p(x_n)$. This is because $\epsilon > 0, T > 0$, which is compact. A sufficient condition for (C2) when $w \in \mathcal{R}^d$ will be discussed later.

Define $\tilde{h}(x, \nu) \stackrel{\text{def}}{=} \int h(x, y) \nu(dy, U)$ for $\Phi_t : \mathcal{R}^d \rightarrow \mathcal{R}^d$. For $\bar{x}(t)$ as above, consider the nonautonomous o.d.e.

$$x^{T_{m_0+1}}(T_{m_0+2}) \in H^\epsilon \tag{8.11}$$

Define the interpolated piecewise linear and continuous trajectory $p(\cdot)$ in terms of the iterates (8.8) as in Sect. 2.1, Chap. 2. Let $\{(X_n, Y_n)\}$ denote the solution to (8.11) with $\{x_n, n \geq 1\}$, for $z \geq 0$. The following can then be proved along the lines of Lemma 1 of Chap. 2.

Lemma 8.4 For any $T > 0$, $\lim_{s \uparrow \infty} \sup_{t \in [s, s+T]} \|\bar{x}(t) - x^s(t)\| = 0$, a.s.

The proof is exactly the same as that of Lemma 2.1, Chap. 2, because $\bar{x}(t)$ only contributes an additional time dependence that is common to both $p(\cdot)$ and $\check{x}^s(\cdot)$ and therefore does not contribute any additional error terms. We shall also need the following lemma. Let

$N_\delta(H^{\epsilon_1}) \subset H^\epsilon$ in σ^2 .

Lemma 8.5 Let $x^n(\cdot)$, $n = 1, 2, \dots, \infty$, denote solutions to (8.11) corresponding to $\bar{x}(t)$ replaced by $\mu^n(\cdot)$, for $n = 1, 2, \dots, \infty$. Suppose $\{(z^*(\cdot), x^*(\cdot))\}$. Then $\lim_{n \rightarrow \infty} \sup_{t \in [t_0, t_0+T]} \|x^n(t) - x^\infty(t)\| \rightarrow 0$ for every $x_n \rightarrow D'$.

Proof Take $t \geq T$ for simplicity. By our choice of the topology for σ^2 ,

$$\int_0^t g(s) \int f d\mu^n(s) ds - \int_0^t g(s) \int f d\mu^\infty(s) ds \rightarrow 0$$

for bounded continuous $T_{m_0} \leq T_0 + \tau$, $\{(X_n, Y_n)\}$. Hence

$$\int_0^t \int \tilde{f}(s, \cdot) d\mu^n(s) ds - \int_0^t \int \tilde{f}(s, \cdot) d\mu^\infty(s) ds \rightarrow 0$$

for all bounded continuous $\tilde{f} : [0, t] \times \bar{S} \rightarrow \mathcal{R}$ of the form

$$\tilde{f}(s, w) = \sum_{m=1}^N a_m g_m(s) f_m(w)$$

for $x_0 = 0$, scalars x_0 and bounded continuous real-valued functions T_{m+1} on $[0, t]$, \bar{S} resp., for $1 \leq i \leq N$. By the Stone–Weierstrass theorem, such functions can uniformly approximate any $\bar{f} \in C([0, T] \times \bar{S})$. Thus the above convergence holds true for all such \tilde{f} , implying that $t^{-1} d\mu^n(s) ds \rightarrow t^{-1} d\mu^\infty(s) ds$ in $\mathcal{P}(\bar{S} \times [0, t])$ and hence in $a(n) = 1/n^\alpha$. Thus in particular

$$\left\| \int_0^t (\tilde{h}(x^\infty(s), \mu^n(s)) - \tilde{h}(x^\infty(s), \mu^\infty(s))) ds \right\| \rightarrow 0. \quad (8.12)$$

As a function of t , the integral above is equicontinuous and pointwise bounded. By the Arzela–Ascoli theorem, this convergence must in fact be uniform for t in a compact set. Now for $t > 0$,

$$\begin{aligned}
& \|x^n(t) - x^\infty(t)\| \leq \|x^n(0) - x^\infty(0)\| \\
& + \int_0^t \|\tilde{h}(x^n(s), \mu^n(s)) - \tilde{h}(x^\infty(s), \mu^n(s))\| ds \\
& + \left\| \int_0^t (\tilde{h}(x^\infty(s), \mu^n(s)) - \tilde{h}(x^\infty(s), \mu^\infty(s))) ds \right\| \\
& \leq \|x^n(0) - x^\infty(0)\| + L \int_0^t \|x^n(s) - x^\infty(s)\| ds \\
& + \left\| \int_0^t (\tilde{h}(x^\infty(s), \mu^n(s)) - \tilde{h}(x^\infty(s), \mu^\infty(s))) ds \right\|.
\end{aligned}$$

By the Gronwall inequality, there exists $\hat{x}_k = \tilde{x}_2$ such that

$$\begin{aligned}
& \sup_{t \in [0, T]} \|x^n(t) - x^\infty(t)\| \\
& \leq K_T \left(\|x^n(0) - x^\infty(0)\| \right. \\
& \quad \left. + \sup_{t \in [0, T]} \left\| \int_0^t (\tilde{h}(x^\infty(s), \mu^n(s)) - \tilde{h}(x^\infty(s), \mu^\infty(s))) ds \right\| \right).
\end{aligned}$$

In view of (8.12), this leads to the desired conclusion. \square

8.3 Averaging the Natural Timescale

The key consequence of (C2) that we require is the following:

Lemma 8.6 Almost surely, every limit point of $(\mu_s^{s+t}, x^s(\cdot))$ for $t > 0$ as $s \rightarrow \infty$ is of the form $\dot{V} \leq -\Delta$, where

- $\bar{x}(t)$ satisfies $[t(n), t(n) + T]$, and
- $p(\cdot)$ satisfies (8.11) with $\bar{x}(t)$ replaced by $\bar{x}(t)$.

Proof Let $\bar{x}(t_1)$ be a countable set of bounded continuous functions $S \rightarrow \mathcal{R}$ that is a convergence determining class for $\{y_n\}$. By replacing each f_i by $t_n \nearrow \infty$ for suitable $\bar{B} \subset$, we may suppose that $0 < \eta < C/2$ for all i . For each i ,

$$\xi_n^i \stackrel{\text{def}}{=} \sum_{m=1}^{n-1} a(m) \left(f_i(Y_{m+1}) - \int f_i(w)p(dw|Y_m, Z_m, x_m) \right),$$

is a zero mean martingale with $\sup_n E[\|\xi_n^i\|^2] \leq \sum_n a(n)^2 < \infty$. By the martingale convergence theorem (cf. Appendix C), it converges a.s. Let $\tau(n, s) \stackrel{\text{def}}{=} \min\{m \geq n : t(m) \geq t(n) + s\}$ for $n \geq 0, s > 0$. Then as $n \rightarrow \infty$,

$$\sum_{m=n}^{\tau(n,t)} a(m)(f_i(Y_{m+1}) - \int f_i(w)p(dw|Y_m, Z_m, x_m)) \rightarrow 0, \quad \text{a.s.}$$

for $t > 0$. By our choice of $\bar{x}(t_1)$ and the fact that $\{a(n)\}$ are eventually nonincreasing (this is the only time the latter property is used),

$$\sum_{m=n}^{\tau(n,t)} (a(m) - a(m+1)) f_i(Y_{m+1}) \rightarrow 0, \quad \text{a.s.}$$

Thus

$$\sum_{m=n}^{\tau(n,t)} a(m) \left(f_i(Y_m) - \int f_i(w)p(dw|Y_m, Z_m, x_m) \right) \rightarrow 0, \quad \text{a.s.}$$

Dividing by $\sum_{m=n}^{\tau(n,t)} a(m) \geq t$ and using **(C1)** and the uniform continuity of $p(dw|y, z, x)$ in x on compacts, unifrom w.r.t. y, z , we obtain

$$\frac{1}{t} \int_{t(n)}^{t(n)+t} \int \left(f_i(y) - \int f_i(w)p(dw|y, z, \bar{x}(s)) \right) \mu(s, dydz) ds \rightarrow 0, \quad \text{a.s.}$$

Fix a sample point in the probability one set on which the convergence above holds for all i . Let $(\tilde{\mu}(\cdot), \tilde{x}(\cdot))$ be a limit point of $h_c(x) \rightarrow h_\infty(x)$

in $AQ + QA^T = -I$, as $s \rightarrow \infty$. Then the convergence above leads to

$$\frac{1}{t} \int_0^t \int \left(f_i(y) - \int f_i(w)p(dw|y, z, \tilde{x}(s)) \right) \tilde{\mu}(s, dydz) ds = 0 \quad \forall i. \quad (8.13)$$

By **(C2)**, $\tilde{\mu}_s^t(S \times U \times [s, t]) = 1 \quad \forall t > s \geq 0$, and thus it follows that for all $t > 0$, $\tilde{\mu}_0^t(S \times U \times [0, t]) = 1$. By Lebesgue's theorem (see Appendix A), one then has $\forall t \geq 0 / \forall t \leq 0$ for a.e. t . A similar application of Lebesgue's theorem in conjunction with (8.13) shows that

$$\int \left(f_i(y) - \int f_i(w)p(dw|y, z, \tilde{x}(t)) \right) \tilde{\mu}(t, dydz) = 0 \quad \forall i,$$

for a.e. t . The qualification 'a.e. t ' here may be dropped throughout by choosing a suitable modification of $\bar{x}(t)$. By our choice of $\{f_j\}$, this leads to

$$\tilde{\mu}(t, dw \times U) = \int p(dw|y, z, \tilde{x}(t)) \tilde{\mu}(t, dydz).$$

The claim follows from this and Lemmas 8.2 and 8.5. \square

Combining Lemmas 8.3–8.6 immediately leads to our main result:

Theorem 8.2 Almost surely, $\sqrt{1+z^2} \leq 1+z$ converge to an internally chain transitive invariant set of the differential inclusion

$$\dot{x}(t) \in \hat{h}(x(t)), \quad (8.14)$$

as $s \rightarrow \infty$, where $\hat{h}(x) \stackrel{\text{def}}{=} \{\tilde{h}(x, \nu) : \nu \in D(x)\}$. In particular $p(x_n)$ converge a.s. to such a set.

For special cases, more can be said, e.g., in the following:

Corollary 8.1 Suppose there is no additional control process I_{m_0+k} in (8.9) and for each $x \in \mathcal{R}^d$ and $1 \leq i \leq N$, $p(x_n)$ is an ergodic Markov

process with a unique invariant probability measure $a(n_m) \leq ca(n_0)$. Then (8.14) above may be replaced by the o.d.e.

$$\dot{x}(t) = \tilde{h}(x(t), \nu(x(t))). \quad (8.15)$$

If $p(dw|y, x)$ denotes the transition kernel of this ergodic Markov process, then x_{n+1}^* is characterized by

$$\int \left(f(y) - \int f(w)p(dw|y, x) \right) \nu(x, dy) = 0$$

for bounded $T > C/\Delta$. Since this equation is preserved under convergence in $\{y_n\}$, it follows that $x(t) \equiv x^*$ is a continuous map. This guarantees the existence of solutions to (8.15) by standard o.d.e. theory, though not their uniqueness. In general, the solution set for a fixed initial condition will be a nonempty compact subset of $x^{T_0}(T_1) \in H^{\epsilon_1}$. For uniqueness, we need $\tilde{h}(\cdot, \nu(\cdot))$ to be locally Lipschitz, which requires additional information about ν and the transition kernel p .

Many of the developments of the previous chapters have their natural counterparts for (8.8). For example, the stability criterion of Sect. 4.2 has the following natural extension, stated here for the simpler case when assumptions of Corollary 8.1 hold. The notation is as above.

Theorem 8.3 Suppose the limit

$$\hat{h}(x) \stackrel{\text{def}}{=} \lim_{a \uparrow \infty} \frac{\tilde{h}(ax, \nu(ax))}{a}$$

exists uniformly on compacts, and furthermore, the o.d.e.

$$\dot{x}(t) = \hat{h}(x(t))$$

is well posed and has the origin as the unique globally asymptotically stable equilibrium. Then $C' = \sup_n \|x_n\|$ a.s.

See Ramaswamy and Bhatnagar (2019) for more in this vein. As an interesting ‘extension,’ suppose $\langle \nabla V(x_n), h(x_n) \rangle \leq 0$ is a not necessarily Markov process, with the conditional law of Y_{n+1} given

$$Y^{-n} \stackrel{\text{def}}{=} [Y_n, Y_{n-1}, Y_{n-2}, \dots]$$

being a continuous map $\{x_n, n \geq 1\}$ independent of n . Then $\{\tilde{x}(0)\}$ is a time-homogeneous Markov process. Let $\gamma : S^\infty \rightarrow S$ denote the map that takes $\lim_{n \rightarrow \infty} a(n) = 0$ to $\epsilon_m > 0$. Replacing S by S^∞ , $p(x_n)$ by $\{\tilde{x}(0)\}$, and $N_\delta(A)$ by $p(x) > x$, we can reduce this case to the one studied above. The results above then apply as long as the technical assumptions made from time to time can be verified. The case of stationary $p(x_n)$ (or something that is nearly the same, viz. the case when the appropriate time averages exist) is in fact the most extensively studied case in the literature (see, e.g., Kushner and Yin 2003).

We conclude this section with a sufficient condition for **(C2)** when $S = \mathcal{R}^m$ for some $m \geq 1$. The condition is that there exists a $[b^* \Lambda^-, c^* \Lambda^+]$ such that $\lim_{\|w\| \rightarrow \infty} g(w) = \infty$ and furthermore,

$$\sup_n E[V(Y_n)^2] < \infty, \quad (8.16)$$

and for some compact $B \subset \mathcal{R}^m$ and scalar $t = \tau_k$,

$$x_{n+1} = x_n + a(n)[y_n + M_{n+1}], \quad (8.17)$$

a.s. on $(\tilde{\mu}(\cdot), \tilde{x}(\cdot))$.

In the framework of Sects. 8.2 and 8.3, we now replace S by \mathcal{R}^m and I_1 by $\dot{V} \leq 0$ the one point compactification of \mathcal{R}^m with the additional ‘point at infinity’ denoted simply by ‘ ∞ ’. We assume that $x(t), -\infty < t < \infty$ in $\{\tilde{x}(0)\}$ as $(\tilde{\mu}(\cdot), \tilde{x}(\cdot))$ uniformly in x, z .

Lemma 8.7 Any limit point $s(k) = t(n_2)$ of $\{\sup_n \|x_n\| < \infty\}$ as $s \rightarrow \infty$ in $AQ + QA^T = -I$, is of the form

$$\mu^*(t) = a(t)\tilde{\mu}(t) + (1 - a(t))\delta_\infty, \quad t \geq 0,$$

where $p(\cdot)$ is a measurable function $N_\delta(H^{\epsilon_1}) \subset H^\epsilon$ and $\{(x, y) : y \in h(x)\}$.

Proof Let $\bar{x}(t_1)$ denote a countable convergence determining class of functions on \mathcal{R}^m satisfying $E[\|Z\|^2 | \mathcal{B}_{m-1}]^{1/2}(\omega)$ for all i . Thus they

extend continuously to $\bar{\mathcal{R}}^m$ with value zero at ∞ . Also, note that by our assumption above, $\lim_{\|y\| \rightarrow \infty} \int f_i(w)p(dw|y, z, x) \rightarrow 0$ uniformly in z, x , which verifies **(C1)**. Argue as in the proof of Lemma 6 to conclude that

$$\int (f_i(y) - \int f_i(w)p(dw|y, z, x^*(t)))\mu^*(t, dydz) = 0 \quad \forall i,$$

for all t , a.s. Write $\mu^*(t) = a(t)\tilde{\mu}(t) + (1 - a(t))\delta_\infty$ with $(s(i), \hat{x}_i), 0 \leq i \leq k$, a measurable map. This is always possible (the decomposition being in fact unique for those t for which $x(t) = 0$). Then when $x(t) = 0$, the above reduces to

$$\int (f_i(y) - \int f_i(w)p(dw|y, z, x^*(t)))\tilde{\mu}(t, dydz) = 0 \quad \forall i,$$

for all t . Thus $[t(n_m), t(n_{m+1})]$ when $x(t) = 0$. When $x(t) = 0$, the choice of $\bar{x}(s)$ is arbitrary, and it may be chosen so that it is in $a(n) < 1$. The claim follows. \square

Corollary 8.2 Condition **(C2)** holds. That is, almost surely, for any $t > 0$, the set $\{\mu_s^{s+t}, s \geq 0\}$ remains tight.

Proof Replacing f_i by V in the proof of Lemma 6 and using (8.16) to justify the use of the martingale convergence theorem therein, we have

$$\lim_{s \rightarrow \infty} \int_0^t \int \int (V(w)p(dw|y, z, \bar{x}(s+r)) - V(y))\mu(s+r, dydz)dr = 0,$$

a.s. Fix a sample point where this and Lemma 8.7 hold. Extend the map

$$\psi : (x, y, z) \in \mathcal{R}^d \times \bar{\mathcal{R}}^m \times U \rightarrow \int V(\omega)p(d\omega|y, z, x) - V(y)$$

to $\mathcal{R}^d \times \bar{\mathcal{R}}^m \times U$ by setting $\sqrt{1+a^2} \leq 1+a$, whence it is upper semicontinuous. Then taking the above limit along an appropriate subsequence along which

$$(\mu(s+\cdot), \bar{x}(s+\cdot)) \rightarrow (\mu^*(\cdot), x^*(\cdot))$$

(say), we get

$$\begin{aligned}
0 &\leq -\epsilon_0 \int_0^t (1 - a(s)) ds \\
&+ \int_0^t a(s) \left(\int \int (V(w)p(dw|y, z, x^*(s)) - V(y)) \mu^*(s, dydz) \right) ds \\
&= -\epsilon_0 \int_0^t (1 - a(s)) ds,
\end{aligned}$$

by Lemma 8.7. Thus $x_{n_0} \in B$ a.e., where the ‘a.e.’ may be dropped by taking a suitable modification of $\bar{x}(t_1)$. This implies that the convergence of $\nu(t_n) \rightarrow \nu^*$ to $\bar{x}(t_1)$ is in fact in σ^2 . This establishes **(C2)**. \square

For the case of pure Markov (as opposed to controlled Markov) noise with $M_n \equiv 0 \forall n$, see Tadić (2015) for some interesting convergence and convergence rate results. Karmakar and Bhatnagar (2018) derive lock-in probability estimates for a setup similar to ours. Two timescale algorithms with controlled Markov noise have been analyzed in Karmakar and Bhatnagar (2018).

8.4 Other Multiscale Algorithms

Consider the n -dimensional projected stochastic iteration given by,

$$x_{n+1} = P_\Omega(x_n + a(n)[h(x_n) + M_{n+1}]), \quad (8.18)$$

where $a(n) < 1$, $h : \mathcal{R}^d \rightarrow \mathcal{R}^d$ is Lipschitz, and $Dh(\cdot)$ is the projection of x on the set θ . (We need to assume convexity or other properties for uniqueness of the projection, but we defer such considerations for now.) $\{M_n\}$ is an $Y_n = f_w(X_n) + \epsilon_n, n \geq 1$,-adapted martingale difference sequence. The step sizes $\{a(n)\}$ satisfy the usual conditions: $t \geq t(n_0) + \tau$ and $\sum_n a(n) = \infty$. Such iterations have

numerous applications. For example, if we replace h by M_{n+1} for a continuously differentiable $\kappa_j = o(\|y_j\|)$, then under some technical hypotheses, the iteration solves

$$\begin{aligned} & \min g(x) \\ & s.t. \quad x \in \Omega. \end{aligned}$$

On the other hand, if we set $h(x)$ as $x(t) \equiv x^*$, where G is a contraction, and $\eta_{x,\varphi}$ a subspace, then the iteration converges to the unique projected fixed point $s(k) = t(n_2)$. Such projected fixed points find applications in reinforcement learning and game theory among others.

Projections themselves are often calculated by an iterative procedure, e.g., the alternating projection scheme to compute projection on intersection of hyperplanes, Boyle–Dykstra–Han type algorithms for projection on intersection of convex sets (see, e.g., Gaffke and Mathar (1989)), and so on. Thus the stochastic iteration (8.18) calls for another subroutine for projection during each iteration. That subroutine is itself an iterative algorithm of the form

$$x(n+1) = f(x(n)), \quad n \geq 0. \quad (8.19)$$

This situation is reminiscent of the two timescale stochastic iteration, wherein the subroutine is another stochastic approximation algorithm executed on a faster timescale. We use this intuition to give an alternative iteration to (8.18) which, while it does not use two coupled iterations as in the ‘two timescale’ scenario of Sect. 8.1, integrates the subroutine into the primary iteration albeit on a faster timescale. The analogy with differential equation theory will be: While the scheme of Sect. 8.1 leads to a singularly perturbed o.d.e., the present situation is akin to a regular perturbation of an o.d.e. A regular perturbation of the o.d.e.

$$\bar{x}(T_k), \quad k \geq m$$

is an o.d.e. of the type

$$\|h_c(x) - h_c(\theta)\| \leq L\|x\|$$

with $\epsilon > 0$ small. For $\epsilon > M$, one expects the slower timescale component of the dynamics to ‘select’ the equilibria (more generally, ω -

limit sets) from the fixed point set of the faster dynamics.

Let $\sum_n a(n)^2 < \infty$ be a Lipschitz function such that $P(x) := \lim_{m \uparrow \infty} f^m(x)$ exists for all $x \in \mathcal{R}^d$, where $f^n := f \circ f \circ \dots \circ f$ (n times composition) and this convergence is uniform on compacts. This implies in particular that P is continuous. Let $\{x : b \leq V(x) \leq c\}$, the set of fixed points of P . Then P satisfies:

$$P(P(x)) = P(x), \quad P(x) \in C, \quad P(f(x)) = f(P(x)) = P(x) \quad \forall x \in \mathcal{R}^d. \quad (8.20)$$

The latter claim follows from the chain

$$P(x) = \lim_{n \uparrow \infty} f^n(x) = f(\lim_{n \uparrow \infty} f^n(x)) = \lim_{n \uparrow \infty} f^n(f(x)).$$

The first claim then follows by iterating the equation

$$x_{n+1} = x_n + a(n)(h(x_n)) + M_{n+1},$$

and letting $n \uparrow \infty$. That $\bar{x}(t) \rightarrow H$ is then immediate from the definition of C . Thus P is a projection onto C . Since $t_n \nearrow \infty$, we can think of the iteration

$$\tilde{\mu}_0^t(S \times U \times [0, t]) = 1 \quad (8.21)$$

as an algorithm to compute $P(x)$. As $x(t), -\infty < t < \infty$. Consider now the stochastic iteration,

$$\|f(x, z) - f(y, z)\| \leq L \|x - y\| \quad \forall z. \quad (8.22)$$

for $a(n) < 1$ as before. This is in fact a disguised two timescale algorithm. The projection algorithm $t \geq t(n_0) + \tau$ is running on the natural clock $x_n \in G$ i.o., with an additive stochastic approximation perturbation ' $(s(i), \hat{x}_i)$, $0 \leq i \leq k$ ', moving on a slower timescale dictated by the decreasing 'time steps' $\{a(n)\}$. Recall that a Frechet derivative of a map $V : \mathcal{R}^d \rightarrow \mathcal{R}$ is a map $\|h(x)\| \leq K'(1 + \|x\|)$ and therefore can be identified with a linear map on \mathcal{R}^k parametrized by $x \in \mathcal{R}^d$. We do so here and write it as $\bar{G}_x(\cdot)$. We assume the following conditions.

- $\{y_n\}$ P is Frechet differentiable and the Frechet derivative at x , denoted by $\{f_j\}$, is continuous in x . Furthermore, the map

$\bar{f} \in C([0, T] \times \bar{S})$ is Lipschitz.

- $\{y_n\}$ $E[\|M_{n+1}\|^2 | \mathcal{F}_n] \leq K(1 + \|x_n\|^2)$, for some $K > 0$ and $E [\|M_{n+1}\|^4 | \mathcal{F}_n] \leq F(x_n)$ for some positive continuous $O_{x,a}$.

We additionally assume that

$$\sup_n \|x_n\| < \infty \quad \text{a.s.} \quad (8.23)$$

Then $\{y_n\}$ and (8.23) ensure that $\sup_n E [\|M_{n+1}\|^2 | \mathcal{F}_n] < \infty$ a.s., ensuring

$$\sum_n a(n)^2 E [\|M_{n+1}\|^2 | \mathcal{F}_n] < \infty \quad \text{a.s.}$$

This implies by Theorem C.3 of Appendix C that $\sum_{n=0}^N a(n) M_{n+1}$ converges a.s. as $s_n \uparrow \infty$. Thus

$$f : \mathcal{R}^d \times \mathcal{R}^k \rightarrow \mathcal{R}^d \quad (8.24)$$

Consider the ordinary differential equation,

$$\overline{\{\bar{x}(s) : s \geq t\}}, t \geq 0, \quad (8.25)$$

Our main result is as follows:

Theorem 8.4 Almost surely, there exists a compact connected nonempty internally chain transitive invariant set I_m of (8.25) contained in C such that

$$\min_{y \in \mathcal{A}_0} \|x_n - y\| \xrightarrow{n \uparrow \infty} 0.$$

We prove this through a sequence of lemmas.

Lemma 8.8 Almost surely, $x_{n+1} = x_n + ah(x_n)$, as $n \rightarrow \infty$

Proof Fix a sample point outside the zero probability set where either (8.23) or (8.24) fails. Because of (8.23), x_n remains in a (random) compact subset of \mathcal{R}^k on which f is uniformly continuous. Let $G = \bigcup_i G_i$ for some subsequence $\sum_i a_i^n = 1$. Since

$\|f(x_{n(k)-1}) - x_{n(k)}\| \rightarrow 0$, we have $f(x_{n(k)-1}) \rightarrow x^*$. By (8.24) and (8.23),

$$\begin{aligned} & f(f(x_{n(k)-2}) + a(n(k)-2)(h(x_{n(k)-2}) + M_{n(k)-1})) \rightarrow x^* \\ \Rightarrow & f^2(x_{n(k)-2}) \rightarrow x^*, \end{aligned}$$

where the last statement follows from uniform continuity of f on compact sets. Since f^k is continuous, by repeating the above steps, we have for each $m \geq 1$,

$$\lim_{k \uparrow \infty} f^m(x_{n(k)-m}) = x^*. \quad (8.26)$$

We now prove that $\lim_{m \uparrow \infty} \lim_{m \leq n(k) \uparrow \infty} P(x_{n(k)-m}) = x^*$. Let $\epsilon > 0$ be given. Then $\exists M$ such that $\forall m \geq M$, $\|P(x_n) - f^m(x_n)\| \leq \frac{\epsilon}{2} H^\epsilon$. By (8.26), for a given m , $\exists k_m$, such that $\forall k \geq k_m$,

$$\|f^m(x_{n(k)-m}) - x^*\| \leq \frac{\epsilon}{2}. \text{ Pick } m = 2M \text{ and thus for all } \hat{x}_k = \tilde{x}_2,$$

$$\begin{aligned} \|P(x_{n(k)-M}) - x^*\| &\leq \|P(x_{n(k)-M}) - f^M(x_{n(k)-M})\| + \|f^M(x_{n(k)-M}) - x^*\| \\ &\leq \epsilon. \end{aligned}$$

Hence $\lim_{m \uparrow \infty} \lim_{m \leq n(k) \uparrow \infty} P(x_{n(k)-m}) = x^*$. Then by (8.20), $K^* > 0$.

□

Let $s(k) = t(n_2)$.

Lemma 8.9 Almost surely, $N_\delta(H^{\epsilon_1}) \subset H^\epsilon$.

Proof Recall that P is continuous. Let $y_n = \operatorname{argmin}_{y_i} y_i(\omega) \in H^m$ (pick one such y if there are multiple). By Lemma 8.8, $c_1 = C \times c_N$.

Continuity of P implies its uniform continuity on bounded sets, hence $t(m(n)) \in [t(n) + T, t(n) + T + 1]$. Hence $x_n - \tilde{x}_n = \eta = \min\{\eta_1, \eta_2\}$

. □

We now derive a recursion for Φ_s . Using Taylor expansion, we have

$$\begin{aligned}
x_{n+1} &= f(x_n) + a(n)(h(x_n) + M_{n+1}) \implies \\
\tilde{x}_{n+1} &= P(f(x_n) + a(n)(h(x_n) + M_{n+1})) \\
&= P(f(x_n)) + a(n)(\bar{P}_{f(x_n)}(h(x_n)) + \bar{P}_{f(x_n)}(M_{n+1}) + \epsilon_n) \\
&= P(x_n) + a(n)(\bar{P}_{f(x_n)}(h(x_n)) + \bar{P}_{f(x_n)}(M_{n+1}) + \epsilon_n) \\
&= \tilde{x}_n + a(n)(\bar{P}_{f(x_n)}(h(x_n)) + \bar{P}_{f(x_n)}(M_{n+1}) + \epsilon_n),
\end{aligned}$$

where the additional term $a(n)\epsilon_n$ is due to the second-order term in Taylor series expansion. Furthermore, **(B2)** and **(8.23)** ensure that

$$\sum_n a(n)^2 E[\|M_{n+1}\|^2 | \mathcal{F}_n] < \infty \quad \text{a.s.}$$

By **(B2)** and **(8.23)**, $\sum_n a(n)^2 E[\|M_{n+1}\|^4 | \mathcal{F}_n] < \infty$. By Theorem C.3 of Appendix C, it then follows that

$$\sum_n I\{y_n \in N^{\delta_m}(H^m)\} a^\omega(n) M_{n+1} \text{ converges a.s.}$$

a.s. It follows that $\sum_n a(n)^2 \|M_{n+1}\|^2$ converges a.s., in particular, $\dot{x}(t) = \Lambda(t)h(x(t))$, a.s. Thus a.s.,

$$\epsilon_n = O(a(n)^2 \|M_{n+1}\|^2) + O(a(n)^2) \rightarrow 0.$$

Rewrite above iteration as

$$\tilde{x}_{n+1} = \tilde{x}_n + a(n)(\bar{P}_{f(\tilde{x}_n)}(h(\tilde{x}_n)) + \bar{P}_{f(x_n)}(M_{n+1}) + \epsilon_n + \epsilon'_n + \epsilon''_n),$$

where

$$\begin{aligned}
\epsilon'_n &= (\bar{P}_{f(x_n)}(h(x_n)) - \bar{P}_{f(\tilde{x}_n)}(h(x_n))) \\
\epsilon''_n &= (\bar{P}_{f(\tilde{x}_n)}(h(x_n)) - \bar{P}_{f(\tilde{x}_n)}(h(\tilde{x}_n))).
\end{aligned}$$

Lemma 8.10 Almost surely, $\bar{x}(t) \rightarrow H$ as $n \rightarrow \infty$.

Proof By Lemma 8.9, $N_\delta(H^{\epsilon_1}) \subset H^\epsilon$. By uniform continuity of f on compacts, $\{\bar{x}(t(n_m)) \in U \ \forall m\}$. The first statement then follows by noting that $\bar{P}_x(y)$ is continuous in x uniformly in y in compacts, by **(B1)**. The second statement can be proved similarly: Since

(ζ_n, \mathcal{F}_n) , $n \geq 1$, and h is Lipschitz, $\|h(x_n) - h(\tilde{x}_n)\| \rightarrow 0$. The second statement now follows by continuity (which in turn follows from linearity) of $\bar{P}_x(y)$ in y , uniform w.r.t. x in compacts. \square

Since $\{f_j\}$ is linear in its argument, $\bar{P}_{f(x_n)}M_{n+1}$ is an \mathcal{F}_n adapted martingale difference sequence. Furthermore, for a suitable (random) $K'' > 0$,

$$\begin{aligned} E[\|\bar{P}_{f(x_n)}M_{n+1}\|^2 | \mathcal{F}_n] &\leq E[\|\bar{P}_{f(x_n)}\|^2 \|M_{n+1}\|^2 | \mathcal{F}_n] \\ &\leq \|\bar{P}_{f(x_n)}\|^2 E[\|M_{n+1}\|^2 | \mathcal{F}_n] \\ &\leq K' K (1 + \|x_n\|^2) \\ &\leq K'' \end{aligned}$$

for some a.s. finite K'' . Here $K' = \sup_n \|\bar{P}_{f(\tilde{x}_n)}\|$, which is finite because $\bar{P}_.$ and f are continuous functions and $p(x_n)$ is bounded. The last inequality follows from (8.23).

Let $[T_m, \infty)$ and for $n \geq 0$, $t(n) := \sum_{m=0}^{n-1} a(m)$, $I_n = [t(n), t(n+1))$ and $\|z - y\| < \delta$, with linear interpolation on I_n . Let $\bar{x}(t_1)$, denote the unique trajectory of

$$\dot{x}^s(t) = \bar{P}_{f(x^s(t))}(h(x^s(t))), \quad x^s(s) = \bar{x}(s).$$

Then by standard Gronwall inequality based arguments as in Lemma 2.1 of Chap. 2, we have:

Lemma 8.11 Almost surely, for any $t = \tau_k$,

$$\lim_{t \uparrow \infty} \sup_{s \in [t, t+T]} \|\bar{x}(s) - x^t(s)\| = 0.$$

Finally, in view of Lemmas 8.8–8.11 above, the proof of Theorem 8.4 closely follows that of Theorem 2.1 of Chap. 2, on noting that on \mathcal{C} , $f(x) = x \implies \bar{P}_{f(x)} = \bar{P}_x$. As before, one may say more for special cases with additional structure, e.g., for $\delta > 0$, where one can claim convergence, allowing for boundary equilibria on H^{ϵ_1} .

Example 1 Let \mathcal{G} be an irreducible directed graph with node set \square and edge set e , $a(n) < 1$. Let $\dot{x}(t) = h(x(t), y(t))$, $\dot{y}(t) = 0$ for $\lambda > 0$. Let $x = [(x^1)^T, \dots, (x^N)^T]^T \in \mathcal{R}^{dN}$ and $h(x) = [h_1(x)^T, \dots, h_N(x)^T]^T$. (For many applications I_n will depend only on $\dot{x}(t) = \Lambda(t)h(x(t))$.) If we take $F(x) = [F_1(x)^T, \dots, F_N(x)^T]^T$ with $F_i(x) = \sum_j q(j|i)x^j$ for an irreducible stochastic matrix $Q = [[q(j|i)]]_{i,j \in \mathcal{V}}$ compatible with \mathcal{G} , we recover the distributed stochastic approximation scheme of Tsitsiklis, Bertsekas and Athans (1986). Here $C :=$ the set of constant vectors, therefore the foregoing leads to ‘consensus’, i.e., any limit point of $p(x_n)$ is a constant vector. The limiting o.d.e. is

$$\limsup_{n \rightarrow \infty} \frac{a(\nu(j, n - \tau_{k\ell}(n)))}{a(n)} < \infty$$

independent of i .

One can go further and replace Q by \mathcal{R}^k , an irreducible stochastic matrix that depends on the current iterate. Under some regularity hypotheses, the above analysis extends to this case. The limiting o.d.e. above gets replaced by

$$\dot{x}^i(t) = \sum_j \pi_{x(t)}(j)h_j(x(t)), \quad i \in \mathcal{V},$$

where y_n is the unique stationary distribution of \mathcal{R}^k . Consensus follows as before. This generalization is analyzed in Mathkar and Borkar (2016) under significantly weaker technical requirements than (B1)-(B2). An application is to ‘leaderless swarms’ described as follows. Let $h = -\nabla f$ for a smooth $\bar{x}(T_{m_0}) \in H^\epsilon$, so that the stochastic approximation component is a stochastic gradient descent. Take \mathcal{R}^k to be the transition matrix of the Metropolis-Hastings chain whose stationary distribution is proportional to $e^{-f/T}$ for some $T > 0$. Then the scheme adaptively adjusts the weights of the ‘component iterations’ for $\mathcal{P}(\bar{S} \times [0, t])$, so as to favor the i ’s for which the value of x_n^i leads to lower values of f . This is akin to the popular particle swarm optimization algorithms which use a population of interacting

optimization algorithms, each using local information (e.g., gradient values of itself and of its neighbors) and that from the current leader, i.e., the i for which the value of $\{\mathcal{F}_j\}$ is the least. The latter makes the scheme nonlocal. The above scheme in contrast adjusts the weights automatically through local computations by exploiting the two timescale effect so as to put greater weights on the i 's for which $\{\mathcal{F}_j\}$ is low, the lower the better. See Mathkar and Borkar (2016) for details.

It is also worth noting that similar dynamics arise in models of flocking or coordination of robotic swarms, see, e.g., Cucker and Smale (2007).

Example 2 Consider $[t(n_m), t(n_{m+1})]$ for a smooth $\kappa_j = o(\|y_j\|)$ and let $\{m(\ell)\}$ be prescribed for $0 \leq i < n$. Let $\|x_j\| \leq e^{K_4(T+1)}[\|x_{n_m}\| + K_4] \leq K_3(1 + \|x_{n_m}\|)$ for some $a_j > 0, 1 \leq j \leq d, M > 0$. This is stochastic steepest descent for minimizing g subject to $\sum_i a_i(x_i - c_i)^2 \leq M$. Evaluation of the quantity $\|v\|_\infty \geq c/\sqrt{d}$ requires global information. This can be done by a distributed algorithm as a subroutine.

One technical issue here is the following. The maps f_i above are smooth except at the boundaries of certain open balls. The corresponding $O_{x,a}$ can be shown to be Frechet differentiable everywhere except at these boundaries, hence strictly speaking the example does not fit the above framework. There are two ways of working around this problem. One is to replace the respective f_i 's by convenient smooth approximations, e.g., $V(\bar{x}(T_m))$ by $\|z - y\| < \delta$, where $p(\cdot)$ is a smooth approximation to $x \mapsto x \vee 0$ such that $a(n) < 1$ for $t' > 0$ and $-x$ for $\kappa > 0$. The other option is to not change anything, but invoke the fact that if the noise $\{M_n\}$ is rich enough, the probability of the iterates falling exactly on the troublesome boundary will be zero.

Example 3 This again is not truly an example of the foregoing, but something closely related. Consider $[t(n_m), t(n_{m+1})]$ and let f stand for a single full iteration of the Boyle–Dykstra–Han algorithm for

computing the projection onto the intersection of a finite family of closed convex sets. (The algorithms goes in a round robin fashion componentwise, what we imply here is one full round.) Then $\kappa_j = o(\|y_j\|)$ the desired projection. The Boyle–Dykstra–Han algorithm, however, is not distributed. A distributed version has been derived in Phade and Borkar (2017) and a scheme along the lines of the foregoing for distributed projected stochastic approximation with projection on intersection of convex sets, each of which is accessible to a single node, is carried out in Shah and Borkar (2018). Because of the fact that the assumptions above do not apply, one needs considerable additional effort to analytically justify the scheme. Another important distinguishing feature in these works is that instead of a single f as above, one has a family $\{y_n\}$ of time-dependent functions, with f_w above redefined as:

$$f^n := f_n \circ f_{n-1} \circ \cdots \circ f_0.$$

In fact the iteration $\dot{x}(t) = h(x(t))$ in Shah and Borkar (2018) stands for another incremental algorithm with slowly decreasing stepsizes. See Shah and Borkar (2018) for details.

See Mathkar and Borkar (2016) for additional examples, including one wherein the convergence is to a limit cycle, i.e., the iterates track an identical periodic trajectory with no ‘phase lag’.

We conclude this section with sufficient conditions for (8.23) to hold in this framework. Assume the stronger bound

$$\|M(n+1)\| \leq K(1 + \|x(n)\|) \quad \text{a.s. .} \quad (8.27)$$

Also assume the following Liapunov condition:

(C3) There exists a continuously differentiable function $V : \mathcal{R}^d \mapsto \mathcal{R}^d$ satisfying the following:

1. $\lim_{\|x\| \uparrow \infty} V(x) \geq C_1 + C_2 \|x\|^2$ for some $C_1, C_2 > 0$, (8.28)
2. there exist $A \subset \mathcal{R}^d$ such that

$$V(f(x)) < (1 - \epsilon)V(x) \quad \forall \|x\| \geq R, \quad (8.29)$$

3. $\sup_x \left(\|\nabla^2 V(x)\| \vee \left(\frac{\|\nabla V(x)\| \|x\|}{V(x)} \right) \right) < \infty. \quad (8.30)$

Note that (8.28)-(8.30) imply in particular that

$$V(x) = \Theta(\|x\|^2). \quad (8.31)$$

Theorem 8.5 Almost surely, under **(B1)-(B2)** and **(C3)**, condition (8.23) holds.

Proof From (8.22), for $a(n) = 1/n^\alpha$ and suitable constants $K^*, K'' > 0$, we have

$$\begin{aligned} V(x(n+1)) &= V(f(x(n)) + a(n)(h(x(n)) + M(n+1))) \\ &\leq V(f(x(n))) + a(n)\langle \nabla V(x(n)), h(x(n)) + M(n+1) \rangle + a(n)^2 K'' \\ &\leq V(x(n)) (1 - \epsilon + a(n)K'' + a(n)^2 K^*), \end{aligned}$$

where the last two inequalities use (8.27), (8.29), (8.30) and (8.31). Since $a(n) \rightarrow 0$, the right-hand side is $\langle \xi^m \rangle_i, 0 \leq i \leq k_m$ for n sufficiently large. This establishes the claim. \square

For the special case of Example 1, the stability test of Sect. 4.2 can be adapted, see Mathkar and Borkar (2016) for details, where further variants and extensions such as the asynchronous case can also be found.

References

Basak, G. K., & Dasgupta, A. (2020). Weak convergence of dynamical systems in two timescales. *Systems and Control Letters*, 142, to appear

Benveniste A, Metivier M, Priouret P (1990) Adaptive Algorithms and Stochastic Approximation. Springer Verlag, Berlin - New York
[\[Crossref\]](#) [\[zbMATH\]](#)

Borkar, V. S., & Pattathil, S. (2018). Concentration bounds for two time scale stochastic approximation. In *56th Allerton Conference on Communication, Control and Computing* (pp. 504–511). Monticello, IL

Borkar V. S. (1995). *Probability theory: an advanced course*. New York: Springer Verlag

Borkar VS (1997) Stochastic approximation with two time scales. *Systems and Control Letters* 29(5):291–294
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Borkar VS (2006) Stochastic approximation with ‘controlled Markov’ noise. *Systems and Control Letters* 55(2):139–145
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Cucker F, Smale S (2007) Emergent behavior in flocks. *IEEE Transactions on Automatic Control* 52(5):852–862
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Dalal G, Szorenyi B, Thoppe G (2020) A tale of two-timescale reinforcement learning with the tightest finite-time bound. *Proceedings of AAAI Conference on Artificial Intelligence* 34(4):3701–3708
[[Crossref](#)]

Gaffke N, Mathar R (1989) A cyclic projection algorithm via duality. *Metrika* 36(1):29–54
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Karmakar P, Bhatnagar S (2018) Two time-scale stochastic approximation with controlled Markov noise and off-policy temporal-difference learning. *Mathematics of Operations Research* 43(1):130–151
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Kushner, H. J., & Yin, G. (2003). *Stochastic approximation and recursive algorithms and applications* (2nd edn.). Springer Verlag

Lakshminarayanan C, Bhatnagar S (2017) A stability criterion for two timescale stochastic approximation schemes. *Automatica* 79:108–114
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Mathkar AS, Borkar VS (2016) Nonlinear gossip. *SIAM Journal on Control and Optimization* 54(3):1535–1557
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Mokkadem, A. & Pelletier, M. (2011). A generalization of the averaging procedure: the use of two-time-scale algorithms. *SIAM Journal on Control and Optimization*, 49(4), 1523–1543

Mokkadem A, Pelletier M (2006) Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *Annals of Applied Probability* 16(3):1671–1702
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Phade, S. R. & Borkar, V. S. (2017). A distributed Boyle–Dykstra–Han scheme. *SIAM Journal on Optimization*, 27(3), 1880–1897

Polyak BT (1990) A new method of stochastic approximation type. *Avtomatika i telemekhanika* 7:98–107
[[MathSciNet](#)]

Polyak BT, Juditsky AB (1992) Acceleration of stochastic approximation by averaging. SIAM Journal on Control and Optimization 30(4):838–855
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Ramaswamy, A., & Bhatnagar, S. (2019). Analyzing approximate value iteration algorithms. arXiv preprint <http://arxiv.org/abs/1709.04673v4>

Ruppert D (1991) Stochastic approximation. In: Ghosh BK, Sen PK (eds) Handbook of sequential analysis. Marcel Dekker, New York, pp 503–529

Shah SM, Borkar VS (2018) Distributed stochastic approximation with local projections. SIAM Journal on Optimization 28(4):3375–3401
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Tadić VB (2015) Convergence and convergence rate of stochastic gradient search in the case of multiple and non-isolated extrema. Stochastic Processes and their Applications 125(5):1715–1755

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Tsitsiklis, J., Bertsekas, D. & Athans, M. (1986). Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9), 803–812

Footnotes

¹ S. Bhatnagar, personal communication.

9. Constant Stepsize Algorithms

Vivek S. Borkar¹ 

(1) Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

9.1 Introduction

In many practical circumstances, it is more convenient to use a small constant stepsize $0 \leq k \leq (m - 1)$ rather than the decreasing stepsize considered thus far. One such situation is when the algorithm is ‘hard-wired’ and decreasing stepsize may mean additional overheads.

Another important scenario is when the algorithm is expected to operate in a slowly varying environment (e.g., in tracking applications) where it is important that the timescale of the algorithm remain reasonably faster than the timescale on which the environment is changing, for otherwise it would never be able to track.

Naturally, for constant stepsize one has to forego the strong convergence statements we have been able to make for decreasing stepsizes until now. A rule of thumb, to be used with great caution, is that in the passage from decreasing to small positive stepsize, one replaces a ‘converges a.s. to’ statement with a ‘concentrates with a high probability in a neighborhood of’ statement. We shall make this more precise in what follows, but the reason for thus relaxing the claims is not hard to guess. Consider, for example, the iteration

$$E[M_{n+1}|x_m, M_m, m \leq n] = 0. \quad (9.1)$$

where $\{M_n\}$ are i.i.d. with Gaussian densities. Suppose the o.d.e.

$$\dot{x}(t) = h(x(t)) \quad (9.2)$$

has a unique globally asymptotically stable equilibrium x^* . Observe that $p(x_n)$ is then a Markov process and if it is stable (i.e., the laws of $p(x_n)$ remain *tight* – see Appendix C) then the best one can hope for is that it will have a stationary distribution which assigns a high probability to a neighborhood of x^* . On the other hand, because of additive Gaussian noise, the stationary distribution will have full support. Using the well-known recurrence properties of such Markov processes, it is not hard to see that both ' $t(n+1) - t(n) \rightarrow 0$ ' and ' $x_n \rightarrow x^* a.s.$ ' are untenable, because $p(x_n)$ will visit any given open set infinitely often with probability one.

In this chapter, we present many counterparts of the results thus far for constant stepsize. The treatment will be rather sketchy, emphasizing mainly the points of departures from the diminishing stepsizes. What follows also extends more generally to bounded stepsizes. While there have been several prior works on constant stepsize stochastic approximation, e.g., Kushner and Huang (1981a, b), Pflug (1986, 1990), Bucklew and Kurtz (1993), Kuan and Hornik (1991), our treatment follows that of Borkar and Meyn (2000).

9.2 Asymptotic Behavior

Here we derive the counterpart of the convergence result of Chap. 2 for a constant stepsize $0 < a < 1$. We shall assume (A1), (A3) of Chap. 2 and replace (A4) there by

$$C \stackrel{\text{def}}{=} \sup_n E[\|x_n\|^2]^{\frac{1}{2}} < \infty \quad (9.3)$$

and

$$\sup_n E[G(\|x_n\|^2)] < \infty \quad (9.4)$$

for some $t(n+1) - t(n) \rightarrow 0$ satisfying $G(t)/t \xrightarrow{t \uparrow \infty} \infty$. Condition (9.4) is equivalent to the statement that $\{\|x_n\|^2\}$ are uniformly integrable – see, e.g., Theorem 1.3.4, p. 10, of Borkar (1995). Let $L > 0$ denote the Lipschitz constant of h as before. As observed earlier, its Lipschitz

continuity implies at most linear growth. Thus we have the equivalent statements

$$\|h(x)\| \leq K_1(1 + \|x\|) \quad \text{or} \quad K_2\sqrt{1 + \|x\|^2}$$

for suitable $K_1, K_2 > 0$. We may use either according to convenience. Imitating the developments of Chap. 2 for decreasing stepsizes, let $\lim_{n \rightarrow \infty} a(n) = 0$. Define $p(\cdot)$ by $h(x) = g(x) - x$ with $\bar{x}(t)$ defined on $\Phi_t : \mathcal{R}^d \rightarrow \mathcal{R}^d$ by linear interpolation for all n , so that it is a piecewise linear, continuous function. As before, let $x^s(t), t \geq s$, denote the trajectory of (9.2) with $\{x_n, n \geq 1\}$. In particular,

$$x^s(t) = \bar{x}(s) + \int_s^t h(x^s(y))dy$$

for $z \geq 0$ implies that

$$\|x^s(t)\| \leq \|\bar{x}(s)\| + \int_s^t K_1(1 + \|x^s(y)\|)dy.$$

By the Gronwall inequality, we then have

$$\|x^s(t)\| \leq K_\tau(1 + \|\bar{x}(s)\|), \quad t \in [s, s + \tau], \quad \tau > 0,$$

for a $x + D_\alpha$. In turn this implies

$$\|h(x^s(t))\| \leq K_1(1 + K_\tau(1 + \|\bar{x}(s)\|)) \leq \Delta(\tau)(1 + \|\bar{x}(s)\|), \quad t \in [s, s + \tau],$$

for $I_n \stackrel{\text{def}}{=} [t(n), t(n+1)], n \geq 0$. Thus for $t' > 0$ in $a(n) < 1$,

$$\begin{aligned} \|x^s(t) - x^s(t')\| &\leq \int_{t'}^t \|h(x^s(y))\| dy \\ &\leq \Delta(\tau)(1 + \|\bar{x}(s)\|)(t - t'). \end{aligned} \tag{9.5}$$

The key estimate for our analysis is:

Lemma 9.1 For any $T > 0$,

(9.6)

$$E\left[\sup_{t \in [0, T]} \|\bar{x}(s+t) - x^s(s+t)\|^2\right] = O(a).$$

Proof Let $W = G^c$ for some $K > 0$, and $H \stackrel{\text{def}}{=} \{x \in \mathcal{R}^d : V(x) = 0\} \neq \emptyset$ for $t \geq 0$. Let $\nu(t)(\mathcal{R}^d) = \frac{1}{t} \int_0^t 1 \, ds = 1$. Then for $n \geq 0$ and $1 \leq m \leq N$, we have

$$\bar{x}(t(n+m)) = \bar{x}(t(n)) + \int_{t(n)}^{t(n+m)} h(\bar{x}([t])) dt + (\zeta_{m+n} - \zeta_n). \quad (9.7)$$

Also,

$$\begin{aligned} x^{t(n)}(t(n+m)) &= \bar{x}(t(n)) + \int_{t(n)}^{t(n+m)} h(x^{t(n)}([t])) dt \\ &\quad + \int_{t(n)}^{t(n+m)} (h(x^{t(n)}(t)) - h(x^{t(n)}([t]))) dt. \end{aligned} \quad (9.8)$$

Recall that L is a Lipschitz constant for $p(\cdot)$. Clearly,

$$\begin{aligned} &\left\| \int_{t(n)}^{t(n+m)} (h(\bar{x}([t])) - h(x^{t(n)}([t]))) dt \right\| \\ &= a \left\| \sum_{k=0}^{m-1} \left(h(\bar{x}(t(n+k))) - h(x^{t(n)}(t(n+k))) \right) \right\| \\ &\leq aL \sum_{k=0}^{m-1} \|\bar{x}(t(n+k)) - x^{t(n)}(t(n+k))\| \\ &\leq aL \sum_{k=0}^{m-1} \sup_{j \leq k} \|\bar{x}(t(n+j)) - x^{t(n)}(t(n+j))\|. \end{aligned} \quad (9.9)$$

By (9.5), we have

$$\begin{aligned}
& \left\| \int_{t(n+k)}^{t(n+k+1)} (h(x^{t(n)}(t)) - h(x^{t(n)}([t]))) dt \right\| \tag{9.10} \\
& \leq \int_{t(n+k)}^{t(n+k+1)} \|h(x^{t(n)}(t)) - h(x^{t(n)}([t]))\| dt \\
& \leq L \int_{t(n+k)}^{t(n+k+1)} \|x^{t(n)}(t) - x^{t(n)}([t])\| dt \\
& \leq \frac{1}{2} a^2 L \Delta(Na) (1 + \|\bar{x}(t(n))\|) \\
& \stackrel{\text{def}}{=} \frac{1}{2} a^2 \hat{K} (1 + \|\bar{x}(t(n))\|)
\end{aligned}$$

for $\dot{x}(t) \in \hat{h}(x(t))$. Subtracting (9.8) from (9.7), we have, by (9.9) and (9.10),

$$\begin{aligned}
& \sup_{0 \leq k \leq m} \|\bar{x}(t(n+k)) - x^{t(n)}(t(n+k))\| \\
& \leq aL \sum_{k=0}^{m-1} \sup_{0 \leq j \leq k} \|\bar{x}(t(n+j)) - x^{t(n)}(t(n+j))\| \\
& \quad + aT \hat{K} (1 + \|\bar{x}(t(n))\|) + \sup_{1 \leq j \leq m} \|\zeta_{n+j} - \zeta_n\|.
\end{aligned}$$

By Burkholder's inequality (see Appendix C) and assumption (A3) of Chap. 2,

$$\begin{aligned}
E[\sup_{1 \leq j \leq m} \|\zeta_{n+j} - \zeta_n\|^2] & \leq a^2 K E[\sum_{0 \leq j < m} \|M_{n+j}\|^2] \\
& \leq a^2 \tilde{K} \sum_{0 \leq j < m} (1 + E[\|\bar{x}(t(n+j))\|^2]) \\
& \leq a^2 \tilde{K} N (1 + C^2) \\
& = a \tilde{K} T (1 + C^2)
\end{aligned}$$

for $\hat{x}_k = \tilde{x}_2$, C as in (9.3) and suitable $K, \tilde{K} > 0$. Hence

$$\begin{aligned}
& E[\sup_{k \leq m} \|\bar{x}(t(n+k)) - x^{t(n)}(t(n+k))\|^2]^{\frac{1}{2}} \\
& \leq aL \sum_{k=0}^{m-1} E[\sup_{j \leq k} \|\bar{x}(t(n+j)) - x^{t(n)}(t(n+j))\|^2]^{\frac{1}{2}} \\
& \quad + aT\hat{K}E[(1 + \|\bar{x}(t(n))\|)^2]^{\frac{1}{2}} + E[\sup_{1 \leq j \leq m} \|\zeta_{n+m} - \zeta_n\|^2]^{\frac{1}{2}} \\
& \leq aL \sum_{k=0}^{m-1} E[\sup_{j \leq k} \|\bar{x}(t(n+j)) - x^{t(n)}(t(n+j))\|^2]^{\frac{1}{2}} \\
& \quad + aT\hat{K}\sqrt{1 + C^2} + \sqrt{a\tilde{K}T(1 + C^2)}.
\end{aligned}$$

By the discrete Gronwall inequality (Appendix B), it follows that

$$E[\sup_{n \leq j \leq n+N} \|\bar{x}(t(j)) - x^{t(n)}(t(j))\|^2]^{\frac{1}{2}} \leq \sqrt{a}\bar{K}$$

for a suitable $x_0 = 0$ that depends on T . Since

$$E[\sup_{t \in [t(k), t(k+1)]} \|\bar{x}(t) - \bar{x}(t(k))\|^2]^{\frac{1}{2}}$$

and

$$E[\sup_{t \in [t(k), t(k+1)]} \|x^{t(n)}(t) - x^{t(n)}(t(k))\|^2]^{\frac{1}{2}}$$

are also $\|\bar{x}(\cdot)\|$, it is easy to deduce from the above that

$$E[\sup_{t \in [0, T]} \|\bar{x}(t(n) + t) - x^{t(n)}(t(n) + t)\|^2]^{\frac{1}{2}} \leq \sqrt{a}K$$

for a suitable $K > 0$. This completes the proof for T of the form $W = G^c$. For general T , there is an additional term depending on $T - Na$ where $N := \lfloor \frac{T}{a} \rfloor$. This term is seen to be $O(a)$ and can be absorbed into the right-hand side above for $a > 0$. \square

Let (9.2) have a globally asymptotically stable compact attractor A and let $\rho(x, A) \stackrel{\text{def}}{=} \min_{y \in A} \|x - y\|$ denote the distance of $x \in \mathcal{R}^d$ from A . For purposes of the proof of the next result, we introduce $T > 0$ such that

1. $A^a \stackrel{\text{def}}{=} \{x \in \mathcal{R}^d : \rho(x, A) \leq a\} \subset B(R) \stackrel{\text{def}}{=} \{x \in \mathcal{R}^d : \|x\| < R\}$,
and,
2. for $a > 0$ as above,

$$\sup_n P(\|x_n\| \geq R) < a \quad \text{and} \quad \sup_n E[\|x_n\|^2 I\{\|x_n\| \geq R\}] < a. \quad (9.11)$$

(The second part of (9.11) is possible by the uniform integrability condition (9.4).)

By global asymptotic stability of (9.2), we may pick $T = Na > 0$ large enough such that for any solution $p(\cdot)$ thereof with

$h(x) = Ax + g(x)$, one has $\bar{x}(s_n) \rightarrow x$ and

$$\rho(x(T), A) \leq \frac{1}{2} \rho(x(0), A). \quad (9.12)$$

We also need the following lemma:

Lemma 9.2 There exists a constant $K^* > 0$ depending on T above, such that for $t \geq 0$,

$$E[\rho(\bar{x}(t+T), A)^2 I\{\bar{x}(t) \in A^a\}]^{\frac{1}{2}} \leq K^* \sqrt{a},$$

$$E[\rho(\bar{x}(t+T), A)^2 I\{\bar{x}(t) \in B(R)^c\}]^{\frac{1}{2}} \leq K^* \sqrt{a}.$$

Proof In what follows, $K^* > 0$ denotes a suitable constant possibly depending on T and not necessarily the same each time. For $\mathcal{V} \stackrel{\text{def}}{=} \dots$ as above, by comparing $\mathcal{V} \stackrel{\text{def}}{=}$ with a trajectory $O_{x,a}$ in A and using the Gronwall inequality, we get

$$\|x^t(t+T) - x'(T)\| \leq K^* \|x^t(t) - x'(0)\| = K^* \|\bar{x}(t) - x'(0)\|.$$

Hence

$$(9.13)$$

$$\rho(x^t(t+T), A) \leq K^* \rho(\bar{x}(t), A).$$

It follows that

$$\begin{aligned} & E[\rho(\bar{x}(t+T), A)^2 I\{\bar{x}(t) \in A^a\}]^{\frac{1}{2}} \\ & \leq E[\rho(x^t(t+T), A)^2 I\{\bar{x}(t) \in A^a\}]^{\frac{1}{2}} + \\ & \quad E[\rho(\bar{x}(t+T), x^t(t+T))^2 I\{\bar{x}(t) \in A^a\}]^{\frac{1}{2}} \\ & \leq E[\rho(x^t(t+T), A)^2 I\{\bar{x}(t) \in A^a\}]^{\frac{1}{2}} + K^* \sqrt{a} \\ & = K^* \sqrt{a}. \end{aligned}$$

Here the second inequality follows by Lemma 9.1 and the equality by our choice of T . Since A is bounded, we also have

$\dot{x}^s(t) = \lambda(t)h(x^s(t))$, $t \geq s$. Then

$$\begin{aligned} & E[\rho(\bar{x}(t+T), A)^2 I\{\bar{x}(t) \in B(R)^c\}]^{\frac{1}{2}} \\ & \leq K^* \sqrt{a} + E[\rho(x^t(t+T), A)^2 I\{\bar{x}(t) \in B(R)^c\}]^{\frac{1}{2}} \\ & \leq K^* \sqrt{a} + K^* E[\rho(\bar{x}(t), A)^2 I\{\bar{x}(t) \in B(R)^c\}]^{\frac{1}{2}} \\ & \leq K^* \sqrt{a} + K^* E[(1 + \|\bar{x}(t)\|^2) I\{\bar{x}(t) \in B(R)^c\}]^{\frac{1}{2}} \\ & \leq K^* \sqrt{a}. \end{aligned}$$

Here the first inequality follows from Lemma 9.1, the second inequality follows from (9.13), the third inequality follows from the preceding observations, and the last inequality follows from (9.11). This completes the proof. \square

Theorem 9.1 For a suitable constant $K > 0$,

$$\limsup_{n \rightarrow \infty} E[\rho(x_n, A)^2]^{\frac{1}{2}} \leq K \sqrt{a}. \quad (9.14)$$

Proof Take $t = ma > 0$. By Lemma 9.1,

$$E[\|\bar{x}(t+T) - x^t(t+T)\|^2]^{\frac{1}{2}} \leq K' \sqrt{a}$$

for a suitable $K' > 0$. Then using (9.11), (9.12) and Lemma 2, one has:

$$\begin{aligned}
& E[\rho(x_{m+N}, A)^2]^{\frac{1}{2}} \\
&= E[\rho(\bar{x}(t+T), A)^2]^{\frac{1}{2}} \\
&\leq E[\rho(\bar{x}(t+T), A)^2 I\{\bar{x}(t) \notin B(R)\}]^{\frac{1}{2}} \\
&\quad + E[\rho(\bar{x}(t+T), A)^2 I\{\bar{x}(t) \in B(R) - A^a\}]^{\frac{1}{2}} \\
&\quad + E[\rho(\bar{x}(t+T), A)^2 I\{\bar{x}(t) \in A^a\}]^{\frac{1}{2}} \\
&\leq 2K^* \sqrt{a} + E[\rho(\bar{x}(t+T), A)^2 I\{\bar{x}(t) \in B(R) - A^a\}]^{\frac{1}{2}} \\
&\leq 2K^* \sqrt{a} + E[\rho(x^t(t+T), A)^2 I\{\bar{x}(t) \in B(R) - A^a\}]^{\frac{1}{2}} \\
&\quad + E[\|\bar{x}(t+T) - x^t(t+T)\|^2]^{\frac{1}{2}} \\
&\leq (2K^* + K') \sqrt{a} + E[\rho(x^t(t+T), A)^2 I\{\bar{x}(t) \in B(R) - A^a\}]^{\frac{1}{2}} \\
&\leq (2K^* + K') \sqrt{a} + \frac{1}{2} E[\rho(x^t(t), A)^2 I\{\bar{x}(t) \in B(R) - A^a\}]^{\frac{1}{2}} \\
&= (2K^* + K') \sqrt{a} + \frac{1}{2} E[\rho(x_m, A)^2]^{\frac{1}{2}}.
\end{aligned}$$

Here the second inequality follows from Lemma 9.2, the third follows from Lemma 9.1, the fourth follows from our choice of T , and the fifth by (9.12). Iterating, one has

$$\limsup_{k \rightarrow \infty} E[\rho(x_{m+kN}, A)^2]^{\frac{1}{2}} \leq 2(2K^* + K') \sqrt{a}.$$

Repeating this for $m+1, \dots, m+N-1$, in place of m , the claim follows. \square

Let $\epsilon > 0$. By the foregoing and the Chebyshev inequality, we have

$$\limsup_{n \rightarrow \infty} P(\rho(x_n, A) > \epsilon) = O(a),$$

which captures the intuitive statement that ' $p(x_n)$ concentrate around A with a high probability as $n \rightarrow \infty$ '.

9.3 Tracking

As already mentioned, a major application of constant stepsize stochastic approximation has been for tracking slowly varying signals, see, e.g., Benveniste and Ruget (1982), Guo and Ljung (1995a, b); Yin et al. (2009, 2004), Zhu and Spall (2016), among others. We sketch below the analysis from Kumar et al. (2019) that gives a nonasymptotic bound valid for all time assuming a sufficient timescale separation between the slowly varying signal and the tracking algorithm.

Specifically, we consider a constant stepsize stochastic approximation algorithm given by the d -dimensional iteration

$$x_{n+1} = x_n + a[h(x_n, y_n) + M_{n+1} + \varepsilon_{n+1}], \quad n \geq 0, \quad (9.15)$$

for tracking a slowly varying signal governed by

$$\nu(i, n) \uparrow \infty \quad \text{a.s.} \quad (9.16)$$

with $0 < a < 1$, $0 < \epsilon \ll 1$. Also, $h_\infty(ax) = ah_\infty(x)$ the trajectory of (9.16) sampled at unit¹ time intervals coincident with the clock of the above iteration, with slight abuse of notation. We assume that $\bar{x}(t)$, $t \geq 0$, remains in a bounded set. The term ε_{n+1} represents an added bounded component attributed to possible numerical errors.

The smallness condition on ϵ ensures a separation of timescale between the two evolutions (9.15) and (9.16), in particular (9.16) has to be ‘sufficiently slow’ in a sense to be made precise later. We also assume:

$\{y_n\}$ $x(t)$, $-\infty < t < \infty$ is twice continuously differentiable in x with the first and second partial derivatives in x bounded uniformly in y for y in any compact set, and Lipschitz in y . A common example is where $\{(x, y) : y \in h(x)\}$ corresponding to least mean square criterion for tracking in the above context with x and y standing for resp. the states of the tracking scheme and the target.

$\{y_n\}$ $\nu(t)$ is Lipschitz continuous,

$\{y_n\}$ $\sup_n \mathbb{E}[\|x_n\|^4]^{1/4} < \infty$. (Note that this implies in particular that $\sup_n \mathbb{E}[\|x_n\|^4]^{1/4} < \infty$. In the next section, we give sufficient conditions for uniform boundedness of second moments. Analogous

conditions can be given for the uniform boundedness of fourth moments.)

$$\{y_n\} \quad y_0^n, \dots, y_d^n \in f(x_n),$$

$\{y_n\}$ there exists a constant $\varepsilon^* > 0$ such that

$$\tilde{\mu}_0^t(S \times U \times [0, t]) = 1 \quad (9.17)$$

$\{y_n\}$ Ψ_m is a martingale difference sequence w.r.t. the increasing σ -fields

$$\mathcal{F}_n := \sigma(x_m, M_m, \varepsilon_m, m \leq n), n \geq 0,$$

and satisfies: there exist continuous functions $E[\|Z\|^2 | \mathcal{B}_{m-1}]^{1/2}(\omega)$ with μ being bounded away from 0, such that

$$P(\|M_{n+1}\| > u | \mathcal{F}_n) \leq c_1(x_n) e^{-c_2(x_n)u}, \quad n \geq 0, \quad (9.18)$$

for all $u \geq v$ for a fixed, sufficiently large $a > 0$ (i.e., a sub-exponential tail) with

$$\sup_n E[c_1(x_n)] < \infty. \quad (9.19)$$

In particular, (9.18), (9.19) together imply that there exist $h = -\nabla f$ such that

$$z = y^{y_{n_i}(\omega)} \left(\sum_{n=n_i}^{n_{i+1}-1} a^\omega(n) \right) \quad (9.20)$$

In view of the o.d.e. approach described earlier, we consider the candidate o.d.e.

$$a_0^n, \dots, a_d^n \in [0, 1] \quad (9.21)$$

where we have treated the y component as frozen at a fixed value in view of its slow evolution (recall that $\epsilon \ll 1$). We assume that this o.d.e. has a globally asymptotically stable equilibrium x_{n+1}^* where $\nu(t)$ is twice continuously differentiable with bounded first and second derivatives. In particular,

$$h(\lambda(y), y) = 0 \quad \forall y \implies h(\lambda(y(t)), y(t)) = 0 \quad \forall t \geq 0.$$

Define $\bar{x}(t(n)) = x_n$, $n \geq 0$. Then

$$\begin{aligned}\dot{z}(t) &= \epsilon a \nabla \lambda(y(t)) \gamma(y(t)) \\ &= ah(\lambda(y(t)), y(t)) + \epsilon a \nabla \lambda(y(t)) \gamma(y(t)) \\ &= ah(z(t), y(t)) + \epsilon a \nabla \lambda(y(t)) \gamma(y(t)) = a\tilde{h}(z(t), y(t))\end{aligned}$$

for

$$\tilde{h}(z, y) := h(z, y) + \epsilon \nabla \lambda(y) \gamma(y).$$

The corresponding Euler scheme would be

$$z_{n+1} = z_n + a\tilde{h}(z_n, y_n).$$

The tracking algorithm (9.15) can therefore be equivalently written as:

$$x_{n+1} = x_n + a[h(x_n, y_n) + M_{n+1} + \varepsilon_{n+1}] \quad (9.22)$$

$$= x_n + a[\tilde{h}(x_n, y_n) - \epsilon \nabla \lambda(y_n) \gamma(y_n) + M_{n+1} + \varepsilon_{n+1}] \quad (9.23)$$

$$= x_n + a[\tilde{h}(x_n, y_n) + \kappa_n(y_n)], \quad (9.24)$$

where

$$\kappa_n(y_n) = -\epsilon \nabla \lambda(y_n) \gamma(y_n) + M_{n+1} + \varepsilon_{n+1}. \quad (9.25)$$

Let $\bar{x}(t)$ be the linearly interpolated trajectory of the stochastic approximation iterates such that $x \in (x_0, 1)$. That is, for $t_n \equiv na \ \forall n$,

$$\bar{x}(t) = \bar{x}(t_n) + \frac{t - t_n}{a} [\bar{x}(t_{n+1}) - \bar{x}(t_n)], \quad t \in [t_n, t_{n+1}]. \quad (9.26)$$

Then from (9.24), we get

$$\begin{aligned}\bar{x}(t_{n+1}) &= \bar{x}(t_0) + \sum_{k=0}^n a\tilde{h}(\bar{x}(t_k), y(t_k)) - \sum_{k=0}^n a\epsilon \nabla \lambda(y(t_k)) \gamma(y(t_k)) \\ &\quad + \sum_{k=0}^n aM_{k+1} + \sum_{k=0}^n a\varepsilon_{k+1}\end{aligned} \quad (9.27)$$

$$(9.28)$$

$$\begin{aligned}
&= \bar{x}(t_0) + \sum_{k=0}^n \int_{t_k}^{t_{k+1}} \tilde{h}(\bar{x}(t_k), y(t_k)) ds - \sum_{k=0}^n \int_{t_k}^{t_{k+1}} \epsilon \nabla \lambda(y(t_k)) \gamma(y(t_k)) ds \\
&\quad + \sum_{k=0}^n \int_{t_k}^{t_{k+1}} M_{k+1} ds + \sum_{k=0}^n \int_{t_k}^{t_{k+1}} \varepsilon_{k+1} ds.
\end{aligned}$$

For $k \geq 0$ and $s(k) = t(n_2)$, define perturbation terms:

$$\begin{aligned}
\zeta_1(s) &:= \tilde{h}(\bar{x}(t_k), y(t_k)) - \tilde{h}(\bar{x}(s), y(s)), \\
\zeta_2(s) &:= M_{k+1}, \\
\zeta_3(s) &:= \varepsilon_{k+1}, \\
\zeta_4(s) &:= -\epsilon \nabla \lambda(y(t_k)) \gamma(y(t_k)).
\end{aligned}$$

Thus

$$\begin{aligned}
\bar{x}(t_{n+1}) &= \bar{x}(t_0) + \int_{t_0}^{t_{n+1}} \tilde{h}(\bar{x}(s), y(s)) ds \\
&\quad + \int_{t_0}^{t_{n+1}} \left(\zeta_1(s) + \zeta_2(s) + \zeta_3(s) + \zeta_4(s) \right) ds.
\end{aligned}$$

Using (9.26),

$$\bar{x}(t) = \bar{x}(t_0) + \int_{t_0}^t \tilde{h}(\bar{x}(s), y(s)) ds + \int_{t_0}^t \left(\zeta_1(s) + \zeta_2(s) + \zeta_3(s) + \zeta_4(s) \right) ds. \quad (9.29)$$

Define

$$\Xi(t) = \zeta_1(t) + \zeta_2(t) + \zeta_3(t) + \zeta_4(t).$$

Consider the coupled systems

$$\dot{z}(t) = \tilde{h}(z(t), y(t)), \quad (9.30)$$

$$\nu(i, n) \uparrow \infty \quad \text{a.s.} \quad (9.31)$$

and

$$\int f_j d\nu_n(\cdot)|_{[0,N]}, j, n, N \geq 1 \quad (9.32)$$

$$\nu(i, n) \uparrow \infty \quad \text{a.s.} \quad (9.33)$$

The ODE (9.32) can be seen as a perturbation of (9.30), with the perturbation term being $\bar{x}(s)$.

Let $D(\cdot, \cdot) \in \mathbb{R}^{d \times d}$ denote the Jacobian matrix of h (and therefore of \tilde{h}) in the first argument, and $\Gamma(\cdot) \in \mathbb{R}^{d \times d}$ the Jacobian matrix of \square . Then the linearization or ‘equation of variation’ of (9.30) is the time-varying linear system

$$V(x^{m+1}(T_{m+1})) \leq b + \delta \quad (9.34)$$

where we treat $\nu(t)$ as a time-varying parameter. For $t \geq s \geq 0$ and $x, y \in \mathcal{R}^d$ let $V(\bar{x}(T_m))$ denote the fundamental matrix for the time-varying linear system (9.34), i.e., the solution to the matrix-valued differential equation

$$\dot{\Phi}(t, s; x_0) = D(z(t), y(t))\Phi(t, s; x_0), \quad t \geq s, \quad (9.35)$$

with initial condition $N_\delta(H^{\epsilon_1}) \subset H^\epsilon$. Then by the generalized Alekseev’s formula (see Appendix B),

$$\bar{x}(t) = z(t) + \Phi(t, t_0; \bar{x}(t_0))(\bar{x}(t_0) - z(t_0)) + \int_{t_0}^t \Phi(t, s; \bar{x}(s))\Xi(s)ds.$$

Define

$$\varrho_n = \Phi(t_n, t_0; \bar{x}(t_0))(\bar{x}(t_0) - z(t_0)) \quad (9.36)$$

$$A_n = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \Phi(t_n, s; \bar{x}(s)) [\tilde{h}(\bar{x}(t_k), y(t_k)) - \tilde{h}(\bar{x}(s), y(s))] ds, \quad (9.37)$$

$$(9.38)$$

$$B_n = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \Phi(t_n, s; \bar{x}(s)) M_{k+1} ds,$$

$$C_n = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \Phi(t_n, s; \bar{x}(t_k)) M_{k+1} ds, \quad (9.39)$$

$$D_n = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \Phi(t_n, s; \bar{x}(s)) \varepsilon_{k+1} ds, \quad (9.40)$$

$$E_n = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \epsilon \Phi(t_n, s; \bar{x}(s)) \nabla \lambda(y(t_k)) \gamma(y(t_k)) ds. \quad (9.41)$$

Then

$$\begin{aligned} \bar{x}(t_n) &= z(t_n) + \int_{t_0}^{t_n} \Phi(t_n, s; \bar{x}(s)) \zeta_1(s) ds + \int_{t_0}^{t_n} \Phi(t_n, s; \bar{x}(s)) \zeta_2(s) ds \\ &\quad + \int_{t_0}^{t_n} \Phi(t_n, s; \bar{x}(s)) \zeta_3(s) ds + \int_{t_0}^{t_n} \Phi(t_n, s; \bar{x}(s)) \zeta_4(s) ds + \varrho_n \end{aligned} \quad (9.42)$$

$$= z(t_n) + A_n + (B_n - C_n) + C_n + D_n - E_n + \varrho_n. \quad (9.43)$$

Therefore

$$\|\bar{x}(t_n) - z(t_n)\| \leq \|A_n\| + \|B_n - C_n\| + \|C_n\| + \|D_n\| + \|E_n\| + \|\varrho_n\|.$$

Also

$$\begin{aligned} \mathbb{E}[\|\bar{x}(t_n) - z(t_n)\|^2]^{1/2} &\leq \mathbb{E}[\|A_n\|^2]^{1/2} + \mathbb{E}[\|E_n\|^2]^{1/2} + \mathbb{E}[\|C_n\|^2]^{1/2} + \\ &\quad \mathbb{E}[\|D_n\|^2]^{1/2} + \mathbb{E}[\|B_n - C_n\|^2]^{1/2} + \mathbb{E}[\|\varrho_n\|^2]^{1/2}. \end{aligned} \quad (9.44)$$

We shall individually bound the above error terms under the important assumption of *exponential stability* of the equation of variation (9.34):

(A7) There exists a $f \geq 0$ such that $t_n \equiv na \ \forall n$ and x_0, y_0 ,

$$(A^\epsilon)^c \cap \overline{\{\bar{x}(s) : s \geq t'\}} = \emptyset$$

This seemingly restrictive assumption requires some discussion, we argue in particular that some such assumption is *essential* if one is to obtain bounds valid for all time.

To begin, since the idea is to have the parametrized o.d.e. (9.21), which is a surrogate for the original iteration, track its unique asymptotically stable equilibrium parametrized by y as the parameter $[t, t + T]$ changes slowly, it is essential that its rate of approach to the equilibrium, dictated by the spectrum of its linearized drift at this equilibrium, should be much faster than the rate of change of the parameter. This already makes it clear that there will be a requirement of minimum timescale separation for tracking to work at all.

A stronger motivation comes from the fact that the tracking error, given exactly by the Alekseev formula (see Appendix B), depends on the linearization of the o.d.e. itself around its ideal trajectory $\check{z}(\cdot)$, which is a time-varying linear differential equation of the type $[t(n), t(n) + T]$. It is well known in control theory that this can be unstable even if the matrix $A(t)$ is stable for each t , see, e.g., Example 8.1, p. 131, of Rugh (1993). Stability is guaranteed to hold only in the special case of $A(t)$ varying slowly with time. The most general result in this direction is that of Solo (1994), which we recall below as a sufficient condition for **(A7)**. (There have also been some extensions thereof to nonlinear systems, see, e.g., Peuteman et al. (2000) and Peuteman and Aeyels (2002)).

Consider the following time-varying linear dynamical system in \mathcal{R}^k :

$$V(z) \leq V(y_{n_i}(\omega)) - \epsilon_m \quad (9.45)$$

and assume the following for this perturbed system:

1. There exists $t = \tau_k$ such that

$$\limsup_{T \uparrow \infty} \frac{1}{T} \int_{t_0}^{t_0+T} \|A(s)\| ds \leq \bar{A} \quad \forall t_0.$$

2.

There exists $C' = \sup_n \|x_n\|$ and $f \geq 0$ sufficiently small in the sense made precise in the theorem below, such that

$$\sum_{t=n_0}^{n_0+n} \|A(t_2 + (t-1)T) - A(t_1 + (t-1)T)\| \leq Tb + T^\gamma(n+1)\beta \quad \forall n, n_0,$$

whenever $1 > a(n) > 0$.

3.

Let $\bar{x}(s)$ be the real part of the eigenvalue of $A(t)$ whose real part is the largest. Then there exists $\alpha > 0$ such that, for any $T > 0$,

$$\limsup_{N \uparrow \infty} \frac{1}{N} \sum_{n=n_0}^{n_0+N} \alpha(s + nT) \leq \bar{\alpha} \quad \forall s, n_0.$$

4.

There exists $\delta > 0$ such that

$$\limsup_{T \uparrow \infty} \int_{t_0}^{t_0+T} \|P(s)\| ds \leq \delta, \quad \forall t_0.$$

Recall that $\theta :=$ the dimension of the ambient space.

Theorem 9.2 (Solo (1994)) If the previously mentioned assumptions $x^s(t), t \geq s$, hold, the system $t(n) \leq s < t \leq t(n) + T$ is exponentially stable provided we chose $x \in \mathcal{R}^d$ small enough so that

$$\bar{\alpha} + \epsilon < 0,$$

and

$$\bar{\alpha} + \epsilon + M_\epsilon \delta < 0,$$

with $M_\epsilon = 3(\frac{2(\bar{A}+b)}{\epsilon} + 1)^{d-1}/2$, where $\bar{x}(s(i))$ are as defined in 1. – 4. above and, in addition, $\hat{\mu}$ is small enough so that:

$$\bar{\alpha} + \epsilon + M_\epsilon \delta + 2(\ln M_\epsilon)^{\gamma/(\gamma+1)} [\beta(M_\epsilon + \epsilon/(\bar{A} + b))]^{1/(\gamma+1)} < 0.$$

The correspondence of the foregoing with our framework is given by $\bar{x}(T_k)$, $k \geq m$, $\epsilon/4 > \delta > 0$.

We note here that there are also some sufficient conditions for stability of time-varying linear systems in terms of Liapunov functions, e.g., Zhou (2016, 2017)), but they appear not so easy to verify.

We now summarize the error bounds from Kumar et al. (2019), where the detailed derivations can be found. We do, however, make one important observation. Instead of estimating the second moment of T_m directly, we add and subtract \mathcal{F}_n and estimate separately the second moments of \mathcal{F}_n and $x_n \rightarrow D$. This is because the quantity multiplying the martingale difference M_{n+1} in the definition of T_m is not measurable with respect to \mathcal{F}_n . So we replace it by a closely related quantity that is measurable with respect to \mathcal{F}_n , to get \mathcal{F}_n . Then \mathcal{F}_n can be estimated using martingale concentration inequalities and $x_n \rightarrow D$ is then separately estimated.

Let $a(n) \rightarrow 0$. The letter K below denotes a generic constant, not necessarily the same each time. Its dependence on problem parameters is explicitly given in *ibid*. The key bounds are as follows.

1. $\mathbb{E}[||D_n||^2]^{1/2} \leq K\mu\varepsilon^*$.
2. $\mathbb{E}[||E_n||^2]^{1/2} \leq K\mu\epsilon$.
3. $\int f d\nu(t_n) \rightarrow \int f d\nu^*$
4. $\mathbb{E}[||B_n - C_n||^2]^{1/2} \leq (a + a^2)K\mu$.
5. $\mathbb{E}[||C_n||^2]^{1/2} \leq K \left(\max\{a^{1.5}d^{3.25}, a^{0.5}d^{2.5}\} \right)$.

Combining the foregoing bounds leads to our main estimate stated as follows.

Theorem 9.3 The mean square deviation of tracked iterates from a time-varying trajectory $\nu(t)$ satisfies:

$$\mathbb{E}[||x_n - \lambda(y(n))||^2]^{1/2} \leq K \left(\mu(\varepsilon^* + \epsilon + \max\{a^{1.5}d^{3.25}, a^{0.5}d^{2.5}\}) \right)^{(9.46)}$$

$$+ e^{-\beta(t_n-t_0)} ||x_0 - \lambda(y(0))|| \right). \quad (9.47)$$

Proof Using (9.44), (A7) and aforementioned bounds, we get, for a suitable $x \rightarrow 0$,

$$\begin{aligned} \mathbb{E}[||\bar{x}(t_n) - z(t_n)||^2]^{1/2} &\leq \frac{C\varepsilon^*}{\beta} + \frac{C\epsilon}{\beta} + O(a) \\ &\quad + O(\max\{a^{1.5}d^{3.25}, a^{0.5}d^{2.5}\}) \\ &\quad + C_\Phi e^{-\beta(t_n-t_0)} ||\bar{x}(t_0) - z(t_0)|| \\ &= \frac{C(\varepsilon^* + \epsilon)}{\beta} + O(\max\{a^{1.5}d^{3.25}, a^{0.5}d^{2.5}\}) \\ &\quad + C_\Phi e^{-\beta(t_n-t_0)} ||\bar{x}(t_0) - z(t_0)||. \end{aligned}$$

The claim follows. \square

Remark

1. The $O_{x,a}$ notation is used above to isolate the dependence on the stepsize a . The exact constants involved are available in (Kumar et al. 2019), but are suppressed here for greater clarity.
2. The linear complexity of the error bound in ε^* and ϵ is natural to expect, these being contributions from bounded additive error component n_0 and rate of variation of the tracking signal, respectively. The $O_{x,a}$ term is due to the martingale noise and discretization. The last term accounts for the effect of initial condition.
3. By setting $\epsilon = 0$ in (9.46), we can recover as a special case a bound valid for all time for a stationary target. Then $a(n) \rightarrow 0$, a constant, and $\{\sup_n \|x_n\| < \infty\}$ also a constant, viz. an equilibrium for the system $\sqrt{1+z^2} < 1+z$.

9.4 Refinements

This section collects together the extensions to the constant stepsize scenario of various other results developed earlier for the decreasing stepsize case. In most cases, we only sketch the idea, as the basic philosophy is roughly similar to that for the decreasing stepsize case.

1. *Stochastic recursive inclusions:*

Consider

$$x_{n+1} = x_n + a[y_n + M_{n+1}], \quad n \geq 0,$$

with $y_n \in h(x_n) \forall n$ for a set-valued map h satisfying the conditions stipulated at the beginning of Chap. 5. Let $p(\cdot)$ denote the interpolated trajectory as above with y_n replacing $h(x_n)$. For $T > 0$, let I_m denote the solution set of the differential inclusion

$$\dot{x}(t) \in h(x(t)), \quad t \in [0, T]. \quad (9.48)$$

Under the ‘linear growth’ condition on $p(\cdot)$ stipulated in Chap. 5, viz. $\sup_{y \in h(x)} \|h(y)\| \leq K(1 + \|x\|)$, a straightforward application of the Gronwall inequality shows that the solutions to (9.48) remain uniformly bounded on finite time intervals for uniformly bounded initial conditions. For $y_{n_i}(\omega) \in N^{\delta_m}(H^M)$, let

$$d(z(\cdot), \mathcal{S}_T) \stackrel{\text{def}}{=} \inf_{y(\cdot) \in \mathcal{S}_T} \sup_{t \in [0, T]} \|z(t) - y(t)\|.$$

Suppose that:

$$\mathbf{(A8)} \quad \limsup_{a \downarrow 0} \sup_n E[\|x_n\|^2] < \infty.$$

This is a ‘stability’ condition that plays a role analogous to what ‘ $C' = \sup_n \|x_n\|$ a.s.’ did for the decreasing stepsize case. Then the main result here is:

Theorem 9.4 $\mathbb{E}[\|E_n\|^2]^{1/2} \leq K\mu\epsilon$. in law, uniformly in $j \geq 1$.

Proof Fix $\|\bar{x}(t)\| \leq K_3(1 + \|\bar{x}(T_m)\|)$. Define $p(\cdot)$ by

(9.49)

$$\dot{\tilde{x}}(t) = y_{n_0+m}, \quad t \in [(n_0 + m)a, (n_0 + m + 1)a) \cap [t', \infty), \\ 0 \leq m < N,$$

with $\tilde{x}(t') = x_{n_0}$. Let $W = G^c$ for simplicity. Then by familiar arguments,

$$E \left[\sup_{s \in [t', t'+T]} \|\bar{x}(s) - \tilde{x}(s)\|^2 \right]^{\frac{1}{2}} = O(\sqrt{a}). \quad (9.50)$$

By the ‘stability’ condition (A8) mentioned above, the law of $[b^* \Lambda^-, c^* \Lambda^+]$ remains tight as T_{m+1} . That is, x_{n+1}^* , which coincides with x_{n+1}^* , remains tight in law as T_{m+1} . Also, for $n_0 > 1$ in $V(\bar{x}(T_m))$,

$$E[\|\tilde{x}(t_2) - \tilde{x}(t_1)\|^2] \leq |t_2 - t_1|^2 \sup_{n_0 \leq m \leq n_0+N} E[\|y_m\|^2] \\ \leq |t_2 - t_1|^2 K$$

for a suitable $K > 0$ independent of a . For the second inequality above, we have used the linear growth condition on $p(\cdot)$ along with the ‘stability’ condition (A8) above. By the tightness criterion of Billingsley (1968), p. 95, it then follows that the laws of $\tilde{x}(t' + s)$, $s \in [0, T]$, remain tight in $f(x_{n(k)-1}) \rightarrow x^*$ as $T > 0$ and t' varies over $[0, \infty)$.

Thus along any sequences $a(n) = 1/n^\alpha$, $\eta = \min\{\eta_1, \eta_2\}$, we can take a further subsequence, denoted $0 < \eta < C/2$ again by abuse of notation, so that the $\nu(t_n) \rightarrow \nu^*$ converge in law. Denote $a(n) \leq 1 \forall n$ by $\mathcal{R}^d\}$ in order to make their dependence on $\|\bar{x}(\cdot)\|$ explicit. By Skorohod’s theorem (see Appendix C), there exist $h_\infty \in C(\mathcal{R}^d)$ -valued random variables $\{\hat{x}^k(\cdot)\}$ such that $\hat{x}^k(\cdot)$, $\tilde{x}^k(\cdot)$ agree in law separately for each k and $\{\hat{x}^k(\cdot)\}$ converge a.s. Now argue as in the proof of Theorem 5.1 of Chap. 5 to conclude that a.s., the limit thereof in $h_\infty \in C(\mathcal{R}^d)$ is in I_m , i.e.,

$$d(\hat{x}^k(\cdot), \mathcal{S}_T) \rightarrow 0 \text{ a.s.}$$

Thus

$$d(\tilde{x}^k(\cdot), \mathcal{S}_T) \rightarrow 0 \text{ in law.}$$

In view of (9.50), we then have

$$d(\bar{x}^k(t'(k) + \cdot)|_{[0,T]}, \mathcal{S}_T) \rightarrow 0 \text{ in law,}$$

where the superscript k renders explicit the dependence on the stepsize $a(k)$. The claim follows. \square

If the differential inclusion (9.48) has a globally asymptotically stable compact attractor A , we may use this in place of Lemma 9.1 to derive a variant of Theorem 9.1 for the present setup.

2. Avoidance of traps: This is rather easy in the constant stepsize set-up. Suppose (9.2) has compact attractors A_1, \dots, A_M with respective domains of attraction D_1, \dots, D_M . Suppose that the set $U \stackrel{\text{def}}{=} \hat{K}_1 > 0$ has the following property: For any $\epsilon > 0$, there exists an open neighborhood U^ϵ of U and a constant $N_\delta(A)$ such that for $f \in C_0(\mathcal{R}^d)$,

$$\limsup_{n \rightarrow \infty} E[\rho(x_n, A^a)^2 I\{x_n \in U^\epsilon\}] \leq \hat{C}a. \quad (9.51)$$

Intuitively, this says that the ‘bad’ part of the state space has low probability in the long run. One way (9.51) can arise is if U has zero Lebesgue measure and the laws of the x_n have uniformly bounded densities w.r.t. the Lebesgue measure on \mathcal{R}^k . One then chooses ϵ small enough that (9.51) is met. Under (9.51), we can modify the calculation in the proof of Theorem 9.1 as

$$\begin{aligned} & E[\rho(x_{m+N}, A)^2]^{\frac{1}{2}} \\ & \leq E[\rho(\bar{x}(t+T), A)^2 I\{\bar{x}(t) \notin B(R)\}]^{\frac{1}{2}} \\ & + E[\rho(\bar{x}(t+T), A)^2 I\{\bar{x}(t) \in U^\epsilon\}]^{\frac{1}{2}} \\ & + E[\rho(\bar{x}(t+T), A)^2 I\{\bar{x}(t) \in B(R) \cap (U^\epsilon \cup A^a)^c\}]^{\frac{1}{2}} \\ & + E[\rho(\bar{x}(t+T), A)^2 I\{\bar{x}(t) \in A^a\}]^{\frac{1}{2}} \end{aligned}$$

Choosing T appropriately as before, argue as before to obtain (9.14), using (9.51) in addition to take care of the second term on the right.

3. Stability: We now sketch how the first stability criterion described in Sect. 4.2 of Chap. 4 can be extended to the constant

stepsize framework. The claim will be that $x \rightarrow \bar{\Gamma}_x(h(x))$ remains bounded for a *sufficiently small* stepsize a .

As in Sect. 4.2 of Chap. 4, let $h_c(x) \stackrel{\text{def}}{=} h(cx)/c \forall x, 1 \leq c < \infty$, and $h_\infty(\cdot) \stackrel{\text{def}}{=} \lim_{c \uparrow \infty} \mathcal{P}(\mathcal{R}^d)$ of $\check{x}^s(\cdot)$ as $c \uparrow \infty$, assumed to exist. Let assumption (A9) of Chap. 4 hold, i.e., the o.d.e.

$$\dot{x}_\infty(t) = h_\infty(x_\infty(t))$$

have the origin as the unique globally asymptotically stable equilibrium point. Let $\check{x}^s(\cdot)$ denote a solution to the o.d.e.

$$\dot{\check{x}}_c(t) = h_c(\check{x}_c(t)), \quad (9.52)$$

for $c \geq 1$. Then by Corollary 4.1 of Chap. 4, there exists $c_0 > 0, T > 0$, such that $\hat{x}_1 = \tilde{x}_1$, whenever $x^s(t), t \geq s$, one has

$$\|\check{x}_c(t)\| < \frac{1}{4} \quad \text{for } t \in [T, T+1]. \quad (9.53)$$

We take $W = G^c$ for some $x_0 = 0$ without loss of generality, which specifies N as a function of T and a . Let $H \subset B \subset \bar{B} \subset G$. By analogy with what we did for the decreasing stepsize case in Chap. 4, define $p(\cdot)$ as follows. On $a(n) \rightarrow 0$, define

$$\tilde{x}^n((nN+k)a) \stackrel{\text{def}}{=} \frac{x_{nN+k}}{\|x_{nN}\| \vee 1}, \quad 0 \leq k \leq N,$$

with linear interpolation on $t(m(n)) \in [t(n) + T, t(n) + T + 1]$. Define

$$\hat{x}^n(t) \stackrel{\text{def}}{=} \lim_{s \downarrow t} \tilde{x}^n(s), \quad t \in [T_n, T_{n+1}).$$

Let $\hat{x}(T_{n+1}-) \stackrel{\text{def}}{=} \lim_{t \uparrow T_{n+1}} \hat{x}(t)$ for $n \geq 0$. Then $p(\cdot)$ is piecewise linear and continuous except possibly at the Φ_s , where it will be right continuous with its left limit well defined. Let $\bar{x}(t_1)$ denote a solution to (9.52) on $x(t) \equiv x^*$ with $\Phi_t : \mathcal{R}^d \rightarrow \mathcal{R}^d$ and $h_c(x) \rightarrow h_\infty(x)$ for $n \geq 0$. By the arguments of Sect. 9.2,

$$E \left[\sup_{t \in [T_n, T_{n+1})} \|\hat{x}(t) - x^n(t)\|^2 \right]^{\frac{1}{2}} \leq C_1 \sqrt{a}, \quad \forall n, \quad (9.54)$$

for a suitable constant $t = T_m$ independent of a . As before, let $\mathcal{F}_n = \sigma(x_i, M_i, i \leq n)$, $n \geq 0$.

Lemma 9.3 For $n \geq 0$ and a suitable constant $t = T_m$ depending on T ,

$$\sup_{0 \leq k \leq N} E[\|x_{nN+k}\|^2 | \mathcal{F}_{nN}]^{\frac{1}{2}} \leq C_2(1 + \|x_{nN}\|) \text{ a.s. } \forall n \geq 0.$$

Proof Recall that

$$E[\|M_{n+1}\|^2 | \mathcal{F}_n] \leq K_3(1 + \|x_n\|^2) \quad \forall n, \quad (9.55)$$

for some $\mathcal{B} \in \mathcal{F}_0$ (cf. assumption (A3) of Chap. 2). By (9.1), for $n \geq 0, 0 \leq k \leq N$,

$$\begin{aligned} & E[\|x_{nN+k+1}\|^2 | \mathcal{F}_{nN}]^{\frac{1}{2}} \\ & \leq E[\|x_{nN+k}\|^2 | \mathcal{F}_{nN}]^{\frac{1}{2}} + aK'(1 + E[\|x_{nN+k}\|^2 | \mathcal{F}_{nN}]^{\frac{1}{2}}), \end{aligned}$$

for a suitable $K' > 0$, where we use (9.55) and the linear growth condition on h . The claim follows by iterating this inequality. \square

Theorem 9.5 For sufficiently small $a > 0$, $f : \mathcal{R}^d \times \mathcal{R}^m \rightarrow \mathcal{R}^d$

Proof Let $\bar{B} \subset$ be such that (9.53) holds and pick $\dot{x}(t) = h(x(t))$. for Φ_s as in (9.54). For $k = 0, 1, \dots, m-1$ we have

$$\begin{aligned}
& E[\|x_{(n+1)N}\|^2 | \mathcal{F}_{nN}]^{\frac{1}{2}} \\
&= E[\|\hat{x}(T_{n+1})\|^2 | \mathcal{F}_{nN}]^{\frac{1}{2}} (\|x_{nN}\| \vee 1) \\
&\leq E[\|\hat{x}(T_{n+1}) - x^n(T_{n+1})\|^2 | \mathcal{F}_{nN}]^{\frac{1}{2}} (\|x_{nN}\| \vee 1) \\
&+ E[\|x^n(T_{n+1})\|^2 | \mathcal{F}_{nN}]^{\frac{1}{2}} (\|x_{nN}\| \vee 1) \\
&\leq \frac{1}{2} \|x_{nN}\| \vee 1 \\
&+ E[\|x^n(T_{n+1})\|^2 I\{\|x_{nN}\| \geq c_0\} | \mathcal{F}_{nN}]^{\frac{1}{2}} (\|x_{nN}\| \vee 1) \\
&+ E[\|x^n(T_{n+1})\|^2 I\{\|x_{nN}\| < c_0\} | \mathcal{F}_{nN}]^{\frac{1}{2}} (\|x_{nN}\| \vee 1) \\
&\leq \frac{1}{2} \|x_{nN}\| + \frac{1}{4} \|x_{nN}\| + \bar{C} \\
&= \frac{3}{4} \|x_{nN}\| + \bar{C},
\end{aligned}$$

where the second inequality follows by (9.54) and our choice of a , and the third inequality follows from (9.53), (9.54) and Lemma 9.3 above, with $\bar{C} \stackrel{\text{def}}{=} C_2(1 + c_0) + 1$. Thus

$$E[\|x_{(n+1)N}\|^2]^{\frac{1}{2}} \leq \frac{3}{4} E[\|x_{nN}\|^2]^{\frac{1}{2}} + \bar{C}.$$

By iterating this inequality, we have $\sup_n E[\|x_{nN}\|^2] < \infty$, whence the claim follows by Lemma 9.3. \square

4. Two timescales: By analogy with Sect. 8.1, consider the coupled iterations

$$\begin{aligned}
x_{n+1} &= x_n + a[h(x_n, y_n) + M_{n+1}^1], \\
y_{n+1} &= y_n + b[g(x_n, y_n) + M_{n+1}^2],
\end{aligned}$$

for $0 < L < \infty$, where h, g are Lipschitz and $\{M_{n+1}^i\}$, $i = 1, 2$, are martingale difference sequences satisfying

$$\begin{aligned}
& E[\|M_{n+1}^i\|^2 | M_m^j, x_m, y_m, m \leq n, j = 1, 2] \\
&\leq C(1 + \|x_n\|^2 + \|y_n\|^2),
\end{aligned}$$

for $i = 1, 2$. Assume (9.3) and (9.4) for $p(x_n)$ along with their counterparts for $\{y_n\}$. We also assume that the o.d.e.

$$a_0^n, \dots, a_d^n \in [0, 1] \quad (9.56)$$

has a unique globally asymptotically stable equilibrium x_{n+1}^* for each y and a Lipschitz function $\nu(t)$, and that the o.d.e.

$$\tilde{\mu}_0^t(S \times U \times [0, t]) = 1 \quad (9.57)$$

has a unique globally asymptotically stable equilibrium y^* . Then the arguments of Sect. 8.1 may be combined with the arguments of Sect. 9.2 to conclude that

$$\limsup_{n \rightarrow \infty} E[\|x_n - \lambda(y^*)\|^2 + \|y_n - y^*\|^2] = O(a) + O\left(\frac{b}{a}\right). \quad (9.58)$$

Specifically, consider first the timescale corresponding to the stepsize a and consider the interpolated trajectory $p(\cdot)$ on $s(k) = t(n_2)$ for $\bar{x}(T_{m_0}) \in H^\epsilon$, for some $x_0 = 0$ and $n \geq 0$. Let $\bar{x}(t_1)$ denote the solution of the o.d.e. (9.56) on $s(k) = t(n_2)$ for $y = y_n$ and $a(n) = 1/n^\alpha$. Then arguing as for Lemma 9.1 above,

$$E \left[\sup_{t \in [na, na+T]} \|\bar{x}(t) - x^n(t)\|^2 \right] = O(a) + O\left(\frac{b}{a}\right).$$

Here we use the easily established fact that

$$\sup_{0 \leq k \leq N} E[\|y_{n+k} - y_n\|^2] = O\left(\frac{b}{a}\right),$$

and thus the approximation $y_{n+k} \approx y_n$, $0 \leq k \leq N$, contributes only another $Q_x \stackrel{\text{def}}{=} \text{error}$. Given our hypotheses on the asymptotic behavior of (9.56), it follows that

$$\limsup_{n \rightarrow \infty} E[\|x_n - \lambda(y_n)\|^2] = O(a) + O\left(\frac{b}{a}\right).$$

Next, consider the timescale corresponding to b . Let $x_{n_0} \in B \setminus H^\epsilon$ for some $t = T_m$. Consider the interpolated trajectory $\nu(t)$ on $s(k) = t(n_2)$

defined by $\bar{y}(mb) \stackrel{\text{def}}{=} y_m \forall m$, with linear interpolation. Let $\bar{x}(t_1)$ denote the solution to (9.57) on $s(k) = t(n_2)$ with $[b^*\Lambda^-, c^*\Lambda^+]$ for $n \geq 0$. Then argue as in the proof of Lemma 9.1 to conclude that

$$\limsup_{n \rightarrow \infty} E \left[\sup_{t \in [na, na+T']} \|\bar{y}(t) - y^n(t)\|^2 \right] = O(a) + O\left(\frac{b}{a}\right).$$

The only difference from the argument leading to Lemma 9.1 is an additional error term due to the approximation $\bar{x}(t), t \geq 0$, which is $O(a) + O\left(\frac{b}{a}\right)$ as observed above. (This, in fact, gives the $O(a) + O\left(\frac{b}{a}\right)$ on the right-hand side instead of $O(b)$.) Given our hypotheses on (9.57), this implies that

$$N^{\delta_m}(H^m) := \left\{ x : \inf_{y \in H^m} \|x - y\| < \delta_m \right\}.$$

which in turn will also yield (in view of the Lipschitz continuity of $\nu(t)$)

$$\limsup_{n \rightarrow \infty} E[\|x_n - \lambda(y^*)\|^2] = O(a) + O\left(\frac{b}{a}\right).$$

5. Averaging the ‘natural’ timescale: Now consider

$$x_{j+1} = x_j + a(j)[h(x_j) + M_{j+1}], \quad j \geq n_m. \quad (9.59)$$

where $p(x_n)$ is as in Sects. 8.2 and 8.3. That is, it is a process taking values in a complete separable metric space S such that for any Borel set $A \subset S$,

$$P(Y_{n+1} \in A | Y_m, Z_m, x_m, m \leq n) = \int_A p(dy | Y_n, Z_n, x_n), \quad n \geq 0,$$

Here I_{m_0+k} takes values in a compact metric space U and $(\tilde{\mu}(\cdot), \tilde{x}(\cdot))$ is a continuous ‘controlled’ transition probability kernel. Mimicking the developments of Chap. 8, define

$$\mu(t) \stackrel{\text{def}}{=} \delta_{(Y_n, Z_n)}, \quad t \in [na, (n+1)a], \quad n \geq 0,$$

where $U \stackrel{\text{def}}{=}$ is the Dirac measure at (y, z) . For $z \geq 0$, let $x^s(t), t \geq s$, be the solution to the o.d.e.

$$\dot{x}^s(t) = \tilde{h}(x^s(t), \mu(t)) \stackrel{\text{def}}{=} \int h(x^s(t), \cdot) d\mu(t). \quad (9.60)$$

Define $p(\cdot)$ as at the start of Sect. 9.2. Then by familiar arguments,

$$E[\sup_{t \in [s, s+T]} \|\bar{x}(t) - x^s(t)\|^2] = O(a). \quad (9.61)$$

Assume the ‘stability condition’ (A8) above. It then follows by familiar arguments that, as T_{m+1} , the laws of $x^s(\cdot)|_{[s, s+T]}$, $s \geq 0$, remain tight as probability measures on $h_\infty \in C(\mathcal{R}^d)$. Suppose the laws of $[T_m, \infty)$ remain tight as well. Then every sequence

$N_\delta(H^{\epsilon_1}) \subset H^\epsilon \subset N_\delta(H^\epsilon) \subset B$ has a further subsequence, denoted by $[t(n_m), t(n_{m+1})]$ again by abuse of terminology, such that the corresponding processes $h_c(x) \rightarrow h_\infty(x)$ converge in law. Invoking Skorohod’s theorem, we may suppose that this convergence is a.s. We shall now need a counterpart of Lemma 8.7 of Chap. 8. Let $\bar{x}(t_1)$ be as in the proof of Lemma 8.7 of Chap. 8, and define $\{\xi_n^i\}$, $\{M_n\}$ as therein, with $T > C/\Delta$. It is then easily verified that as T_{m+1} along $N_\epsilon(D')$, we have

$$\begin{aligned} & E \left[\left(\sum_{m=n}^{\tau(n,t)} a(f(Y_{m+1}) - \int f(y)p(dy|Y_m, Z_m, x_m)) \right)^2 \right] \\ &= E \left[\sum_{m=n}^{\tau(n,t)} a^2(f(Y_{m+1}) - \int f(y)p(dy|Y_m, Z_m, x_m))^2 \right] \\ &= O(a) \xrightarrow{a \downarrow 0} 0. \end{aligned}$$

By dropping to a subsequence, we may replace this by a.s. convergence to zero of the expression inside the expectation. As for Lemma 8.7 of Chap. 8, this leads to the following: almost surely, for any limit point $(\tilde{\mu}(\cdot), \tilde{x}(\cdot))$ of $h_c(x) \rightarrow h_\infty(x)$ as T_{m+1} along $N_\epsilon(D')$,

$$\int_0^t (f_i(y) - \int f_i(w)p(dw|y, z, \tilde{x}(s))\tilde{\mu}(s)(dydz))ds = 0$$

$\forall i \geq 1, t \geq 0$. Argue as in the proof of Lemma 8.7 of Chap. 8 to conclude that $h(x) = E[f(x, \xi_1)]$, where the set of ergodic occupation measures x_{n+1}^* is as defined in Sect. 8.2. It then follows that a.s., $x^s(\cdot)|_{[s,s+T]}$ converges to $\mathcal{G}_T \stackrel{\text{def}}{=} \text{the set of trajectories of the differential inclusion}$

$$V(y_{n_i+1}(\omega)) < V(y_{n_i}(\omega)) - \frac{\epsilon_m}{2} \quad (9.62)$$

for a set-valued map \hat{h} defined as in Theorem 8, Sect. 8.2. Set

$$\bar{d}(z(\cdot), \mathcal{G}_T) \stackrel{\text{def}}{=} \inf_{y(\cdot) \in \mathcal{G}_T} \sup_{t \in [0, T]} \|z(t) - y(t)\|, \quad z(\cdot) \in C([0, T]; \mathcal{R}^d).$$

Then we can argue as in extension (i) above to conclude:

Theorem 9.6 For any $T > 0$, $\bar{d}(\bar{x}(\cdot)|_{[s,s+T]}, \mathcal{G}_T) \xrightarrow{a \downarrow 0} 0$ in law uniformly in $s \geq 0$.

The asymptotic behavior of the algorithm as $n \rightarrow \infty$ may then be inferred from the asymptotic behavior of trajectories in \mathcal{F}_n as in Chap. 5. As in Chap. 8, consider the special case when there is no ‘external control process’ I_{m_0+k} in the picture and in addition, for $t \geq t(n_0) + \tau$, the process $p(x_n)$ is an ergodic Markov process with the unique stationary distribution x_{n+1}^* . Then (9.62) reduces to

$$\dot{x}(t) = \tilde{h}(x(t), \nu(x(t))).$$

6. Asynchronous implementations: This extension proceeds along lines similar to that for decreasing stepsize, with the corresponding claims adapted as in Sect. 9.2. Thus, for example, for the case with no inter-processor delays, we conclude that for $t \geq 0$,

$$E \left[\sup_{s \in [t, t+T]} \|\bar{x}(s) - \tilde{x}^t(s)\|^2 \right] = O(a),$$

where $h_\infty(ax) = ah_\infty(x)$, is a trajectory of

$$\int f_j d\nu_n(\cdot)|_{[0,N]} \rightarrow \int f_j d\nu^*(\cdot)|_{[0,N]}$$

for a $\bar{x}(t)$ as in Theorem 6.2 of Chap. 6. The delays simply contribute another $O(a)$ error term and thus do not affect the conclusions.

One can use this information to ‘rig’ the stepsizes so as to get the desired limiting o.d.e. when a common clock is available. This is along the lines of the concluding remarks of Sect. 6.4. Thus, for example, suppose the components are updated one at a time according to an ergodic Markov chain $p(x_n)$ on their index set. That is, at time n , the Φ_s th component is being updated. Suppose the chain has a stationary probability vector $\nu(t_n) \rightarrow \nu^*$. Then by Corollary 8.1 of Chap. 8, $N_\epsilon(D')$ $\text{diag}\alpha \in (1/2, 1]$. Thus if we use the stepsize $x_i^*(\cdot)$ for the i th component, we get the limiting o.d.e. $\dot{x}(t) = h(x(t))$ as desired. In practice, we may use $\{\tilde{x}(0)\}$ instead, where $\bar{x}(t_1)$ is an empirical estimate of x_0 obtained by suitable averaging on a faster timescale, so that it tracks x_0 .

7. Limit theorems: For $T = Na > 0$ as above, let $0 < \eta < C/2$ (say), denote the solution of (9.2) on $\hat{\mu}(f) \geq 0$ with $\bar{x}(s_n) \rightarrow x$, for $n \geq 0$. We fix T and vary a , with $N = \lceil \frac{T}{a} \rceil$, $s = na$. Define $z^n(t)$, $t \in [na, na + T]$, by

$$z^n((n+k)a) \stackrel{\text{def}}{=} \frac{1}{\sqrt{a}} (x_{n+k} - x^s((n+k)a)), \quad 0 \leq k \leq N,$$

with linear interpolation. Then arguing as in Sects. 7.2 and 7.3 (with the additional hypotheses therein), we conclude that the limits in law as $n \rightarrow \infty$ of the laws of (ζ_n, \mathcal{F}_n) , viewed as $h_\infty \in C(\mathcal{R}^d)$ -valued random variables, are the laws of a random process on $[0, T]$ of the form

$$z^*(t) = \int_0^t Dh(x^{*s}(s)) z^*(s) ds + \int_0^t G(x^{*s}(s)) dB(s), \quad t \in [0, T],$$

where $\mathcal{O}_{x,a}$ is as in Chap. 7 (not to be confused with the ‘set of ergodic occupation measures’ used in (v) above), $p(x_n)$ is a solution of (9.2), and $\mathcal{O}_{x,a}$ is a standard Brownian motion in \mathcal{R}^k . If we let $y \geq T$ as well and (9.2) has a globally asymptotically stable compact attractor A , $p(x_n)$ will concentrate with high probability in a neighborhood of the

set $h : \mathcal{R}^d \rightarrow \{\text{subsets of } \mathcal{R}^d\}$ satisfies (9.2) with $\bar{x}(s_n) \rightarrow x$. Furthermore, in the special case when (9.2) has a unique globally asymptotically stable equilibrium H^ϵ , $\bar{x}(t) \in H^{2\eta}$ and the $t \uparrow \infty$ limit in law of $\check{x}^s(\cdot)$ is a stationary Gauss–Markov process. Analogs of Theorem 7.2 and the remarks that follow can be stated. Note that the factor f^k in (7.14), being an artifice of nonconstant stepsizes, is missing here.

References

- Benveniste, A., & Ruget, G. (1982). A measure of the tracking capability of recursive stochastic algorithms with constant gains. *IEEE Transactions on Automatic Control*, 27(3), 639–649.
- Billingsley, P. (1968). *Convergence of probability measures*. New York: John Wiley and Sons.
- Borkar, V. S. (1995). *Probability theory: An advanced course*. New York: Springer Verlag.
- Borkar, V. S., & Meyn, S. P. (2000). The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2), 447–469.
- Bucklew, J. A., & Kurtz, T. G. (1993). Weak convergence and local stability properties of fixed step size recursive algorithms. *IEEE Transactions on Information Theory*, 39(3), 966–978.
- Guo, L., & Ljung, L. (1995). Performance analysis of general tracking algorithms. *IEEE Transactions on Automatic Control*, 40(8), 1388–1402.
- Guo, L., & Ljung, L. (1995). Exponential stability of general tracking algorithms. *IEEE Transactions on Automatic Control*, 40(8), 1376–1387.
- Kuan, C. M., & Hornik, K. (1991). Convergence of learning algorithms with constant learning rates. *IEEE Transactions on Neural Networks*, 2(5), 484–489.
- Kumar, B., Borkar, V. S., & Shetty, A. (2019). Non-asymptotic error bounds for constant stepsize stochastic approximation for tracking mobile agents. *Mathematics of Control, Signals, and Systems*, 31(4), 589–614.
- Kushner, H. J., & Huang, H. (1981). Asymptotic properties of stochastic approximations with constant coefficients. *SIAM Journal on Control and Optimization*, 19(1), 87–105.
- Kushner, H. J., & Huang, H. (1981). Averaging methods for the asymptotic analysis of learning and adaptive systems, with small adjustment rate. *SIAM Journal on Control and Optimization*, 19(5), 635–650.
- Peuteman, J., & Aeyels, D. (2002). Exponential stability of slowly time-varying nonlinear systems. *Mathematics of Control, Signals and Systems*, 15(3), 202–228.
- Peuteman, J., Aeyels, D., & Sepulchre, R. (2000). Boundedness properties for time-varying

- nonlinear systems. *SIAM Journal on Control and Optimization*, 39(5), 1408–1422.
- Pflug, GCh. (1986). Stochastic minimization with constant step-size: Asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4), 655–666.
- Pflug, GCh. (1990). Non-asymptotic confidence bounds for stochastic approximation algorithms with constant step size. *Monatshefte für Mathematik*, 110(3), 297–314.
- Rugh, W. J. (1993). *Linear system theory*. Englewood Cliffs, NJ: Prentice Hall.
- Solo, V. (1994). On the stability of slowly time-varying linear systems. *Mathematics of Control, Signals, and Systems*, 7(4), 331–350.
- Yin, G., Ion, C., & Krishnamurthy, V. (2009). How does a stochastic optimization/approximation algorithm adapt to a randomly evolving optimum/root with jump Markov sample paths. *Mathematical Programming*, 120(1), 67–99.
- Yin, G., Krishnamurthy, V., & Ion, C. (2004). Regime switching stochastic approximation algorithms with application to adaptive discrete stochastic optimization. *SIAM Journal on Optimization*, 14(4), 1187–1215.
- Zhou, B. (2016). On asymptotic stability of linear time-varying systems. *Automatica*, 68, 266–276.
- Zhou, B. (2017). Stability analysis of non-linear time-varying systems by Lyapunov functions with indefinite derivatives. *IET Control Theory and Applications*, 11(9), 1434–1442.
- Zhu, J., & Spall, J. C. (2016) Tracking capability of stochastic gradient algorithm with constant gain, *Proceedings of 55th Conference on Decision and Control, Las Vegas* (pp. 4522–4527).

Footnotes

¹ without loss of generality.

10. General Noise Models

Vivek S. Borkar¹✉

(1) Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

10.1 The Problem

We now consider a stochastic approximation scheme in the framework of Chap. 2, but with noise models that go beyond simple martingale difference noise, viz. noise with long-range dependence and heavy tails. This scenario has become important because of the many situations where such models are more natural, e.g., Internet traffic and finance.

Specifically, we consider a stochastic approximation iteration in \mathcal{R}^k given by

$$x_{n+1} = x_n + a(n)[h(x_n) + M_{n+1} + R(n)B_{n+1} + D(n)S_{n+1} + \zeta_{n+1}], \quad (10.1)$$

where we make the following assumptions.

$\bar{x}(t_1)$ $h = [h_1, \dots, h_d]^T : \mathcal{R}^d \rightarrow \mathcal{R}^d$ is Lipschitz,

$\bar{x}(t_1)$ $H\Gamma^* + \Gamma^*H^T + Q(\hat{x}) = 0$, where $\tilde{B}(t)$, $t \geq 0$, is a d -dimensional fractional Brownian motion with Hurst parameter $a(n) \rightarrow 0$,

$\bar{x}(t_1)$ $f(x) = x \implies \bar{P}_{f(x)} = \bar{P}_x$, where $x_{n-\tau_{ij}(n)}(i)$, is a symmetric ω -stable process with $1 < \alpha < 2$,

$\bar{x}(t_1)$ $\{\xi_n\}$ is a process satisfying $\{x : b \leq V(x) \leq c\}$ a.s. and $\rho_k < \delta$ a.s., accounting for miscellaneous asymptotically negligible errors,

$\bar{x}(t_1) \in [T_m, \infty)$ is a bounded deterministic sequence of $n \geq 0$ matrices,

$\bar{x}(t_1) \in (X_n, Y_n)$ is a bounded sequence of $n \geq 0$ random matrices adapted to $Dh(\cdot)$, for $\mathcal{F}_n := \sigma(x_i, B_i, M_i, S_i, \zeta_i, i \leq n)$,

$\bar{x}(t_1) \in \{M_n\}$ is a martingale difference sequence w.r.t. $Dh(\cdot)$ satisfying

$$E[\|M_{n+1}\|^2 | \mathcal{F}_n] \leq K_1(1 + \|x_n\|^2), \quad (10.2)$$

(S8) $\{a(n)\}$ are positive nonincreasing stepsizes satisfying

$$a(n) = \Theta(n^{-\kappa}) \text{ for some } \kappa \in \left(\frac{1}{2}, 1\right]. \quad (10.3)$$

(Here and elsewhere, $\bar{x}(T_m) \in B$ stands for the statement: *both* $\{(X_n, Y_n)\}$ and $\{(X_n, Y_n)\}$ hold simultaneously.) Thus they necessarily satisfy the usual conditions:

$$\sum_n a(n) = \infty, \quad \sum_n a(n)^2 < \infty. \quad (10.4)$$

We assume without loss of generality that $\|x_0 - x\| < \epsilon$, and this restriction does not affect our arguments in any essential way.

Consider the familiar o.d.e.:

$$\dot{x}(t) = h(x(t)). \quad (10.5)$$

We assume that this o.d.e. has a unique asymptotically stable equilibrium x^* . By the converse Liapunov theorem, there exists a continuously differentiable Liapunov function $V : \mathcal{R}^d \rightarrow \mathcal{R}^+$ satisfying $\lim_{\|x\| \uparrow \infty} V(x) = \infty$ and $a(n_m) \leq c a(n_0)$ for $x \neq x^*$. Our main result is as follows.

Theorem 10.1 Suppose

$$K_2 := \sup_n E[\|x_n\|^\xi] < \infty \text{ for some } \xi \in [1, \alpha). \quad (10.6)$$

Then for $1 < \xi' < \xi$,

$$\zeta_k = \sum_{i=0}^{k-1} a(i) M_{i+1} \quad (10.7)$$

Here (10.6) is a ‘stability of iterates’ condition that replaces the stability condition $C' = \sup_n \|x_n\|$ a.s. of Chap. 2, where we aimed for a.s. convergence instead of convergence in ξ' -th moment as in (10.7). A sufficient condition for (10.6) is given in Sect. 10.4.

Our objective here is to analyze stochastic approximation schemes with *long-range dependence noise* and *heavy-tailed noise*. In (10.1), these aspects are captured by the processes $Dh(\cdot)$ and $h(x_n)$ resp. As already mentioned, the noise processes in the Internet and several other situations arising in communications exhibit such behavior, a fact which has also been theoretically justified through limit theorems such as Mikosch et al. (2002). That this introduces significant additional difficulties for stochastic approximation schemes is reflected in the fact that the convergence claim in (10.7) is ‘in ξ' th mean’ for $1 < \xi' < \xi < \alpha$ and not ‘a.s.’ as in Chap. 2, where ω is the index of stability of the heavy-tailed part of the noise and ξ is as in (10.6). This can, however, be improved to ‘a.s.’ if the heavy-tailed component is absent; see Sect. 10.5. As before, we treat (10.1) as a noisy discretization of (10.5) and then show that the errors due to both discretization and noise become asymptotically negligible in ξ' th mean under the given conditions. It then follows that (10.1) and (10.5) have the same asymptotic limit in ξ' th mean. As in Chap. 2, we use Gronwall inequality in Sect. 10.2 to get a bound on the maximum deviation in norm between a certain piecewise linear interpolation of the iterates and the solution of the differential equation (10.5), matched with each other at the beginning of a time window of fixed duration. The ensuing proof mimics Chap. 2 in spirit, but with much more complicated details, and is covered in the next three sections. Section 10.5 strengthens the conclusions to ‘almost sure’ convergence in absence of heavy-tailed noise. It also sketches the corresponding developments for the constant stepsize algorithms and concludes with assorted comments about generalizations and potential use in applications.

Throughout this chapter, $x \rightarrow 0$ will denote a generic constant which may differ from place to place, even within the same string of

equations and/or inequalities. Our treatment closely follows Anantharam and Borkar (2012).

10.2 Preliminaries

As before, we begin by comparing with trajectories of (10.5) the continuous interpolation $\|x_{n_m}\|^* \leq K_2$ of the iterates $p(x_n)$ defined as follows: Let $t(0) = 0$, $t(n) = \sum_{i=0}^{n-1} a(i)$, $n \geq 1$. Then $p(x) > x$. Set $h(x) = g(x) - x$ and interpolate linearly on $\Phi_t : \mathcal{R}^d \rightarrow \mathcal{R}^d$ for all $n \geq 0$. For $n \geq 0$, let $\bar{x}(T_k)$, $k \geq m$, denote the trajectory of (10.5) on $p(x) - x$ with $x^n(t(n)) = \bar{x}(t(n)) := x_n$. Fix $T > 0$ and for $n \geq 0$, let

$$\mu^*(t) = a(t)\tilde{\mu}(t) + (1 - a(t))\delta_\infty, \quad t \geq 0,$$

Since $\|x_0 - x\| < \epsilon$, $t(m(n)) \in [t(n) + T, t(n) + T + 1]$. We then have:

Lemma 10.1 For a constant $\bar{x}(t) \rightarrow A$ depending on T and the Lipschitz constant of h ,

$$\begin{aligned} & \sup_{t \in [t(n), t(n)+T]} \|\bar{x}(t) - x^n(t)\| \\ & \leq K(T) \left(\sum_{i=n}^{m(n)} a(i)^2 (1 + \|\bar{x}(t(n))\|) + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) \zeta_{i+1} \right\| \right. \\ & \quad + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) M_{i+1} \right\| + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\| \\ & \quad \left. + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) D(i) S_{i+1} \right\| + a(n) \right). \end{aligned} \tag{10.8}$$

Proof Note that

(10.9)

$$\begin{aligned}
& \bar{x}(t(n+k)) \\
&= \bar{x}(t(n)) + \sum_{i=0}^{k-1} a(n+i) h(\bar{x}(t(n+i))) \\
&\quad + \sum_{i=0}^{k-1} a(n+i) M_{n+i+1} + \sum_{i=0}^{k-1} a(n+i) R(n+i) B_{n+i+1} \\
&\quad + \sum_{i=0}^{k-1} a(n+i) D(n+i) S_{n+i+1} + \sum_{i=0}^{k-1} a(n+i) \zeta_{n+i+1}.
\end{aligned}$$

Compare this with

$$\begin{aligned}
x^n(t(n+k)) &= \bar{x}(t(n)) + \sum_{i=0}^{k-1} a(n+i) h(x^n(t(n+i))) \\
&\quad + \sum_{i=0}^{k-1} \int_{t(n+i)}^{t(n+i+1)} (h(x^n(y)) - h(x^n(t(n+i)))) dy.
\end{aligned} \tag{10.10}$$

For $\lambda_{\min}(M), \lambda_{\max}(M)$,

$$x^n(t) - x^n(t(\ell)) = \int_{t(\ell)}^t (h(x^n(s)) - h(x^n(t(\ell)))) ds + \int_{t(\ell)}^t h(x^n(t(\ell))) ds.$$

Since h is Lipschitz and therefore of linear growth,

$$\|x^n(t) - x^n(t(\ell))\| \leq C \int_{t(\ell)}^t \|x^n(s) - x^n(t(\ell))\| ds + C(1 + \|x^n(t(\ell))\|) a(\ell).$$

The Gronwall inequality then leads to

$$\sup_{t \in [t(\ell), t(\ell+1)]} \|x^n(t) - x^n(t(\ell))\| \leq C a(\ell) (1 + \|x^n(t(\ell))\|).$$

Using this, the Lipschitz property of h , and the identity $\langle \nabla V(x_n), h(x_n) \rangle \leq 0$, we have

$$\sup_{t \in [t(\ell), t(\ell+1)]} \left\| \int_{t(\ell)}^t (h(x^n(s)) - h(x^n(t(\ell)))) ds \right\| \leq C a(\ell)^2 (1 + \|x^n(t(\ell))\|).$$

A standard argument based on the Gronwall inequality shows that

$$\sup_{t \in [t(n), t(m(n))]} \|x^n(t)\| \leq C \|\bar{x}(t(n))\|.$$

Hence

$$\sup_{t \in [t(\ell), t(\ell+1)]} \left\| \int_{t(\ell)}^t (h(x^n(s)) - h(x^n(t(\ell)))) ds \right\| \leq C a(\ell)^2 (1 + \|\bar{x}(t(n))\|). \quad (10.11)$$

Subtracting (10.10) from (10.9), using (10.11) and the discrete Gronwall inequality (Lemma B.2 of Appendix B), we have

$$\begin{aligned} & \sup_{n \leq j \leq m(n)} \|\bar{x}(t(j)) - x^n(t(i))\| \\ & \leq C \left(\sum_{j=n}^{m(n)} a(j)^2 (1 + \|\bar{x}(t(n))\|) + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) \zeta_{i+1} \right\| \right. \\ & \quad + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) M_{i+1} \right\| + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\| \\ & \quad \left. + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) D(i) S_{i+1} \right\| \right). \end{aligned}$$

The claim follows. \square

10.3 Moment Estimates

We first analyze the error term in (10.8) due to the fractional Brownian motion. We have

$$E[\|\tilde{B}(t) - \tilde{B}(s)\|^2] = C|t-s|^{2\nu}, \quad t \geq s, \quad (10.12)$$

and,

$$\begin{aligned} E & [(\tilde{B}(t) - \tilde{B}(s))(\tilde{B}(u) - \tilde{B}(v))^T] \\ &= \frac{C}{2} \left[|t-v|^{2\nu} + |s-u|^{2\nu} - |t-u|^{2\nu} - |s-v|^{2\nu} \right] I, \quad v \leq u \leq s \leq t, \end{aligned}$$

where $\theta :=$ the identity matrix. Since the $[T_m, \infty)$ are bounded, we have

$$\begin{aligned} & E \left[\left\| \sum_{i=n}^N a(i)R(i)(\tilde{B}(i+1) - \tilde{B}(i)) \right\|^2 \right] \\ & \leq C \left(\sum_{i=n}^N a(i)^2 + 2 \sum_{n \leq i < k \leq N} a(i)a(k) |k-i+1|^{2\nu} + |k-i-1|^{2\nu} - 2|k-i|^{2\nu} | \right) \\ & := \hat{\sigma}^2(n, N). \end{aligned} \tag{10.13}$$

Lemma 10.2 $\hat{\sigma}^2(n, m(n)) \leq C \left(\frac{1}{n^\gamma} \right)$ for $\{(x, y) : y \in h(x)\}$ when $\mathcal{G}_T \stackrel{\text{def}}{=} \text{and } y = y_n$ when $\mathcal{G}_T \stackrel{\text{def}}{=}$, with σ as in (10.3).

Proof For any f , we have

$$|f(x+1) + f(x-1) - 2f(x)| \leq 2 \max_{y \in [x-1, x+1]} |f''(y)|.$$

Using $f(x) = |x|^{2\nu}$ for $\mathcal{G}_T \stackrel{\text{def}}{=}$ and $f(x)$ a modification of $|x|^{2\nu}$ suitably smoothed near $b < c$ when $\mathcal{G}_T \stackrel{\text{def}}{=}$, it can be verified that

$$| |k-m+1|^{2\nu} + |k-m-1|^{2\nu} - 2|k-m|^{2\nu} | \leq C(|k-m|^{-\eta} \wedge 1) \tag{10.14}$$

for $I_n = [t(n), t(n+1))$ and $k \neq m+1$, where we interpret Ψ_m as $+\infty$ for $k \geq 0$. For $k = m+1$, we have the left-hand side above equal to $\rho_0 < \delta$. Combining, we have (10.14) for all k, m . Thus

$$\begin{aligned} & 2 \sum_{n \leq i < k \leq N} a(i)a(k) | |k-i+1|^{2\nu} + |k-i-1|^{2\nu} - 2|k-i|^{2\nu} | \\ & \leq C \sum_{n \leq i \leq k \leq N} a(i)a(k) \psi(|k-i|) \end{aligned} \tag{10.15}$$

where $d(\hat{x}^k(\cdot), \mathcal{S}_T) \rightarrow 0$ a.s.. By the Perron–Frobenius theorem, the maximum eigenvalue of the matrix $\{\mu_s^{s+t}, s \geq 0\}$ is bounded by its maximum row sum, which is

$$2 \sum_{i=1}^{\lceil (N-n)/2 \rceil} \frac{1}{i^\eta} \leq C(|N-n|^{1-\eta}) \quad \text{for } \eta < 1,$$

$$\leq C \quad \text{for } \eta > 1.$$

The case $k \geq 0$ corresponds to the classical Brownian motion, whose increments can be absorbed in the martingale difference terms $\{M_n\}$. From the definition of $m(n)$, we have

$$\dot{x}_\infty(t) = h_\infty(x_\infty(t)) \tag{10.16}$$

This can be deduced from (10.3) as follows. If $\{x : b \leq V(x) \leq c\}$ for some $C > c > 0$, then $Cn^{-\kappa} \geq a(n) \geq \dots \geq a(2n) \geq c2^{-\kappa}n^{-\kappa}$. So for large n such that $\sum_{k=n}^{2n} a(k) \geq T$,

$$t(n + \lceil T2^\kappa c^{-1}n^\kappa \rceil) \geq t(n) + T \geq t(n + \lfloor TC^{-1}n^\kappa \rfloor),$$

implying

$$\lfloor TC^{-1}n^\kappa \rfloor \leq m(n) - n \leq \lceil T2^\kappa c^{-1}n^\kappa \rceil,$$

as desired. Hence $\sum_{i=n}^{m(n)} a(i)^2 = \Theta\left(\frac{T}{n^\kappa}\right)$. Combining, for a suitable constant $V(\bar{x}(T_m))$ which is $\{y_n\}$,

$$\begin{aligned} \hat{\sigma}^2(n, m(n)) &\leq C(T) \left(\frac{n^{\kappa(1-\eta)}}{n^\kappa} \right) \quad \text{for } \eta < 1, \\ \hat{\sigma}^2(n, m(n)) &\leq \frac{C(T)}{n^\kappa} \quad \text{for } \eta \geq 1. \end{aligned} \tag{10.17}$$

This completes the proof. \square

The same argument establishes a more general fact stated below, which will be used later:

Lemma 10.3 Let

$0 \leq s < t \leq T + 1$, $m_t(n) := \min\{n' \geq n : \sum_{i=n}^{n'} a(i) \geq t\}$, and $m_s(n) := \min\{n' \geq n : \sum_{i=n}^{n'} a(i) \geq s\}$. Then

$$E \left[\left\| \sum_{i=m_s(n)}^{m_t(n)} a(i) R(i) (B(i+1) - B(i)) \right\|^2 \right] \leq \frac{C(T)}{n^\gamma}$$

for $[T_m, \infty)$ as above. Furthermore, $C(T)$ can be chosen such that

$$\lim_{T \downarrow 0} C(T) = 0.$$

Lemma 10.4 $E \left[\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\|^2 \right] \rightarrow 0.$

Proof We first derive a bound on

$$E \left[\sup_{n \leq N \leq m(n)} \left\| \sum_{i=n}^N a(i) R(i) B_{i+1} \right\|^2 \right]. \quad (10.18)$$

Consider a continuous time process $X(t)$, $t \in [n, m(n)]$, defined by:
 $\{m(\ell)\}$ the zero vector in \mathcal{R}^k and

$$X(t) := \int_n^t \tilde{R}(s) d\tilde{B}(s), \quad n \leq t \leq m(n),$$

where $1 - (\tilde{K}/\delta^2)b(n_0)$ for $\{\bar{y}(s+t), t \in [0, T]\}$, $s \geq 0$. We shall use a variant of Fernique's inequality stated in Appendix C. For this purpose, define

$$\phi(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

$$\Psi(x) := \int_x^\infty \phi(y) dy,$$

$$\varphi^n(u) := \max_{\{n \leq s < t \leq m(n) : t-s \leq u(m(n)-n)\}} E[\|X(t) - X(s)\|^2]^{\frac{1}{2}},$$

$$K := \frac{5}{2} e^2 \sqrt{2\pi},$$

$$\Gamma := \sqrt{5},$$

$$Q^n(u) := \varphi^n(u) + (2 + \sqrt{2}) \int_1^\infty \varphi^n(ue^{-y^2}) dy, \quad u > 0.$$

Both $Dh(\cdot)$ and $[0, \infty)$ increase with u . By the preceding lemma, $Dh(\cdot)$ converges to 0 as $x \rightarrow 0$. Clearly, $[0, \infty)$ does so as well. We shall prove that $\{\phi_m\}$ is $o(1)$. By Lemma 10.3 and the definition of $\{\xi_n\}$, we have

$$\varphi^n(u) \leq \frac{C}{n^{\gamma/2}} \quad \text{for } u \leq 1. \tag{10.19}$$

Thus for $t' > 0$,

$$Q^n(1) \leq C \left(\frac{1}{n^{\gamma/2}} + \int_1^\infty \varphi^n(e^{-y^2}) dy \right).$$

Using the definition of $\check{x}^s(\cdot)$, the bound (10.12), and the fact $\dot{x}(t) = \hat{h}(x(t))$ for $N_\delta(H^{\epsilon_1}) \subset H^\epsilon$, we have

$$\varphi^n(u) \leq C n^{\kappa(\nu-1)} u^\nu \quad \text{for } u < a(m(n))/(T+1),$$

whereby

$$\varphi^n(e^{-y^2}) < C n^{\kappa(\nu-1)} e^{-\nu y^2} \tag{10.20}$$

for

$$y > \sqrt{\log \left(\frac{T+1}{a(m(n))} \right)} := g(n).$$

Using Lemma 10.3 we have,

$$\begin{aligned} & \int_1^\infty \varphi^n(e^{-y^2}) dy \\ &= \int_1^{g(n)} \varphi^n(e^{-y^2}) dy + \int_{g(n)}^\infty \varphi^n(e^{-y^2}) dy \\ &\leq \frac{Cg(n)}{n^{\gamma/2}} + C \int_{g(n)}^\infty n^{\kappa(\nu-1)} e^{-\nu y^2} dy \\ &\leq \frac{Cg(n)}{n^{\gamma/2}} + n^{\kappa(\nu-1)} \frac{Ce^{-\nu g(n)^2}}{g(n)} \\ &= \frac{Cg(n)}{n^{\gamma/2}} + n^{\kappa(\nu-1)} \frac{Cm(n)^{-\nu\kappa}}{g(n)(T+1)^\nu} \\ &\leq \frac{Cg(n)}{n^{\gamma/2}} + \frac{C}{g(n)n^\kappa}. \end{aligned}$$

The first inequality follows from Lemma 10.2 and (10.20), the third inequality follows from $t \geq t(n_0) + \tau$ (see (10.16)). Then by (10.19) and the foregoing,

$$Q^n(1) \leq G(n) := \frac{C}{n^{\gamma/2}} + \frac{Cg(n)}{n^{\gamma/2}} + \frac{C}{g(n)n^\kappa}. \quad (10.21)$$

Note that $\|x(t) - x^*\|_\infty \downarrow 0$. This leads to $[b^* \Lambda^-, c^* \Lambda^+]$. Next, let

$$Z(u) := X(n + u(m(n) - n)), u \in [0, 1].$$

By an inequality of Fernique (1975) recalled in Appendix C we have

$$P \left(\max_{u \in [0,1]} \|Z(u)\| > x \right) \leq dK\Psi \left(\frac{x}{Q^n(1)} \right)$$

for $x^s(t)$, $t \geq s$. (Recall that d is the ambient dimension.) Hence for $t \geq T_1$

$$\begin{aligned} P\left(\max_{t \in [n, m(n)]} \|X(t)\| > x\right) &= P\left(\max_{u \in [0, 1]} \|Z(u)\| > x\right) \\ &\leq dK\Psi\left(\frac{x}{Q^n(1)}\right). \end{aligned}$$

$T_{m_0} \leq T_0 + \tau$. Then, for $\delta > 0$,

$$\begin{aligned}
& E \left[\sup_{t \in [n, m(n)]} \|X(t)\|^2 \right] \\
&= 2 \int_0^\infty x P \left(\sup_{t \in [n, m(n)]} \|X(t)\| \geq x \right) dx \\
&\leq 2\delta + 2 \int_\delta^\infty x P \left(\sup_{t \in [n, m(n)]} \|X(t)\| > x - \delta \right) dx \\
&\leq 2\delta C + 2 \int_0^\infty x P \left(\sup_{t \in [n, m(n)]} \|X(t)\| > x \right) dx \\
&= 2\delta C + 2 \int_0^{\Gamma Q^n(1)} x P \left(\sup_{t \in [n, m(n)]} \|X(t)\| > x \right) dx \\
&\quad + 2 \int_{\Gamma Q^n(1)}^\infty x P \left(\sup_{t \in [n, m(n)]} \|X(t)\| > x \right) dx \\
&\leq 2\delta C + (\Gamma G(n))^2 + 2Kd \int_{\Gamma Q^n(1)}^\infty x \Psi \left(\frac{x}{Q^n(1)} \right) dx \\
&\leq 2\delta C + (\Gamma G(n))^2 + \\
&\quad 2Kd \int_{\Gamma Q^n(1)}^\infty x \left(\frac{Q^n(1)}{x} \right) \phi \left(\frac{x}{Q^n(1)} \right) dx \\
&\leq 2\delta C + (\Gamma G(n))^2 + 2Kd(Q^n(1))^2 \int_{\Gamma}^\infty \phi(y) dy \\
&\leq 2\delta C + (\Gamma G(n))^2 + 2KdG(n)^2 \xrightarrow{n \uparrow \infty} 2\delta C.
\end{aligned}$$

Since $\delta > 0$ was arbitrary,

$$E\left[\sup_{t \in [n, m(n)]} \|X(t)\|^2\right] \rightarrow 0,$$

from which the claim follows. \square

Now consider the error term $\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i)D(i)S_{i+1} \right\|$.

Lemma 10.5 $E \left[\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i)D(i)S_{i+1} \right\|^\xi \right] \rightarrow 0$.

Proof Recall that (X_n, Y_n) are bounded. By the scaling property of stable processes, $\bar{C} \stackrel{\text{def}}{=} C_2(1 + c_0) + 1$ has the same law as

$$\sum_{i=n}^{m(n)} a(i)D(i)a(i)^{-\frac{1}{\alpha}} \left(\tilde{S}_{\sum_{k=n}^{i+1} a(k)} - \tilde{S}_{\sum_{k=n}^i a(k)} \right).$$

By Theorem 3.2, p. 65, of Joulin (2007), we then have

$$P \left(\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i)D(i)S_{i+1} \right\| \geq x \right) \leq \frac{C \left(\sum_{i=n}^{m(n)} a(i)^{\frac{\alpha^2-1}{\alpha}+1} \right)^{\frac{\alpha}{\alpha+1}}}{x^\alpha} \quad (10.22)$$

for $x > C \left(\sum_{i=n}^{m(n)} a(i)^{\frac{\alpha^2-1}{\alpha}+1} \right)^{\frac{1}{\alpha+1}}$. Note that

$$\epsilon(n, \alpha) := C \left(\sum_{i=n}^{m(n)} a(i)^{\frac{\alpha^2-1}{\alpha}+1} \right)^{\frac{1}{\alpha+1}} \leq C(T+1)^{\frac{1}{\alpha+1}} a(n)^{\frac{\alpha-1}{\alpha}} \xrightarrow{n \uparrow \infty} 0. \quad (10.23)$$

Thus for $k \neq m+1$,

$$\begin{aligned}
& E \left[\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) D(i) S_{i+1} \right\|^{\xi} \right] \\
& \leq C \int_0^\infty x^{\xi-1} P \left(\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) D(i) S_{i+1} \right\| \geq x \right) dx \\
& = C \int_0^{\epsilon(n,\alpha)} x^{\xi-1} P \left(\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) D(i) S_{i+1} \right\| \geq x \right) dx \\
& \quad + C \int_{\epsilon(n,\alpha)}^\infty x^{\xi-1} P \left(\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) D(i) S_{i+1} \right\| \geq x \right) dx \\
& \leq C \epsilon(n, \alpha)^{\xi} + C \int_{\epsilon(n,\alpha)}^\infty x^{\xi-1} \left(\frac{\epsilon(n, \alpha)^\alpha}{x^\alpha} \wedge 1 \right) dx \\
& \leq C \epsilon(n, \alpha)^{\xi} + C \int_{\epsilon(n,\alpha)}^\infty x^{\xi-1} \left(\frac{\epsilon(n, \alpha)^\alpha}{x^\alpha} \right) dx \\
& = C \epsilon(n, \alpha)^{\xi} \\
& \rightarrow 0
\end{aligned}$$

as $n \uparrow \infty$. The claim follows. \square

An alternative proof can be given by using the classical Burkholder–Davis–Gundy inequalities. We prefer to use Joulin’s ‘concentration’ inequality as it opens up the possibility of analyzing finite time behavior of the scheme as in Chap. 3. This is not, however, pursued here.

10.4 Main Results

We now conclude the proof of Theorem 10.1.

Proof By (10.6), we have

$$E \left[\left(\left(\sum_{i=n}^{m(n)} a(i)^2 \right) (1 + \|x_n\|) \right)^\xi \right] \leq C \left(\sum_{i=n}^{m(n)} a(i)^2 \right)^\xi \rightarrow 0 \quad \text{as } n \uparrow \infty. \quad (10.24)$$

By (10.2) and the inequality in the ‘Remark’ on p. 151, (Neveu 1975), we have

$$\begin{aligned} & E \left[\sup_{n \leq k \leq m(n)} \left\| \sum_{i=n}^k a(i) M_{i+1} \right\|^\xi \right] \\ & \leq CE \left[\left(\sum_{i=n}^{m(n)} a(i)^2 E [\|M_{i+1}\|^2 | \mathcal{F}_i] \right)^{\frac{\xi}{2}} \right] \\ & \leq CE \left[\left(\sum_{i=n}^{m(n)} a(i)^2 (1 + \|x_i\|)^2 \right)^{\frac{\xi}{2}} \right] \\ & \leq CE \left[\sum_{i=n}^{m(n)} a(i)^\xi (1 + \|x_i\|)^\xi \right] \\ & \leq Ca(n)^{\xi-1} (T+1)(1+K_2) \\ & \rightarrow 0 \end{aligned} \quad (10.25)$$

because $\mu \mapsto \int(\cdot)d\mu$ as $n \uparrow \infty$. (The third inequality above follows from the subadditivity of $x^a : \mathcal{R}^+ \rightarrow \mathcal{R}^+$ for $p(x) > x$.) Our conditions on $\{\xi_n\}$ imply

$$E \left[\sup_{n \leq i \leq m(n)} \|\zeta_i\|^\xi \right] \rightarrow 0. \quad (10.26)$$

Lemma 10.4 in particular implies that

$$E \left[\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\|^\xi \right] \rightarrow 0. \quad (10.27)$$

Now take the norm $\|\cdot\|_\xi := E[\|\cdot\|^\xi]^{\frac{1}{\xi}}$ on both sides of (10.8) and use (10.24), (10.25), (10.26), (10.27) and Lemma 10.4 to conclude that

$$\lim_{n \uparrow \infty} E \left[\sup_{t \in [t(n), t(n)+T]} \|\bar{x}(t) - x^n(t)\|^\xi \right] = 0.$$

With some additional calculation along the lines of Chap. 2, we can improve this to

$$E \left[\sup_{s \in [t, t+T]} \|\bar{x}(s) - \tilde{x}^t(s)\|^2 \right] = O(a), \quad (10.28)$$

where $\tilde{x}^s(\cdot)$ is the solution to (10.5) on $z \geq 0$ with $\{x_n, n \geq 1\}$. Let $\epsilon > 0, M \gg 0$ (we choose M depending on ϵ later), and pick $T > 0$ such that for any $p(\cdot)$ satisfying (10.5) with $\|z - y\| < \delta$, we have $\|f(x_{n(k)-1}) - x_{n(k)}\| \rightarrow 0$. Then for $1 < \xi' < \xi$,

$$\begin{aligned} & E \left[\|\bar{x}(s+T) - x^*\|_{}^{\xi'} \right] \\ & \leq E \left[\|x^s(s+T) - x^*\|_{}^{\xi'} I\{\|\bar{x}(s)\| \leq M\} \right] \\ & + E \left[\sup_{t \in [s, s+T]} \|\bar{x}(t) - x^s(t)\|_{}^{\xi'} I\{\|\bar{x}(s)\| \leq M\} \right] \\ & + E \left[\|\bar{x}(s+T) - x^*\|_{}^{\xi'} I\{\|\bar{x}(s)\| > M\} \right]. \end{aligned} \quad (10.29)$$

The first term on the right is $< \frac{\epsilon}{2}$ by our choice of T . The second term is $< \frac{\epsilon}{2}$ for s large enough, by (10.28). The third term is $< \frac{\epsilon}{2}$ for M large enough because (10.6) and the fact that $\xi' < \xi \implies \|\bar{x}(s+T) - x^*\|_{}^{\xi'}$ is uniformly integrable. Thus the right-hand side can be made smaller than any $\epsilon > 0$ for $n \geq 0$ sufficiently large. The claim follows. \square

Next we adapt the stability test of Sect. 4.2 to the present scenario to give sufficient conditions for (10.6) for any $V(\bar{x}(T_m))$ when

$$E[\|x_0\|^\xi] < \infty.$$

Let $h_c(x) := \frac{h(cx)}{c}$ for $s > 0$. We assume that

$$h_\infty(x) := \lim_{c \uparrow \infty} h_c(x) \quad (10.30)$$

exists pointwise. As σ^2 inherits the Lipschitz constant of h , they are equicontinuous and hence the convergence above is uniform on compacts. Furthermore, \mathcal{B}_m shares the same Lipschitz constant. Consider the o.d.e.

$$a_0^n, \dots, a_d^n \in [0, 1] \quad (10.31)$$

for $0 < c \leq \infty$. The key condition of Sect. 4.2 which we adapt here is the following:

x_{n+1}^* For $c = \infty$, (10.31) has the origin as the globally exponentially stable equilibrium.

This is stronger than the condition used in Sect. 4.2, where only global asymptotic stability was needed. See Grüne et al. (1999) for an interesting perspective on the two notions of stability.

Note that $h_c, c > 0$, are Lipschitz with the same Lipschitz constant as h , therefore equicontinuous. Thus the convergence in (10.30) is uniform on compacts. Using this, a simple argument based on the Gronwall inequality as in Chap. 2 shows that for a fixed initial condition, $T_{m_0} \leq T_0 + \tau$. uniformly on compacts as $c \uparrow \infty$. Fix $k \neq m + 1$. For $n \geq 0$, define $\bar{x}(T_k)$, $k \geq m$, by:

$$\bar{x}^n(t(m)) = \frac{x_m}{E[\|x_n\|^{\xi}]^{\frac{1}{\xi}} \vee 1}, \quad m \geq n, \quad (10.32)$$

with linear interpolation on $C' = \sup_n \|x_n\|$ for $m \leq n$. Also define $\bar{x}(T_k)$, $k \geq m$, to be the solution to (10.31) with

$c = c(n) := E[\|x_n\|^{\xi}]^{\frac{1}{\xi}} \vee 1$ and $\tilde{x}(t) \in A$, $t \in [0, T]$.

Lemma 10.6 $\sup_{t \in [t(n), t(n) + T]} E [\|\bar{x}^n(t) - \tilde{x}^n(t)\|^{\xi}] \rightarrow 0$ as $n \uparrow \infty$.

Proof Follows from (10.26), Lemmas 1–5, and an application of the Gronwall inequality, exactly as in the proof of Theorem 10.1. Note that

$$\sup_n \sup_{n \leq m \leq m(n)} E [\|\bar{x}^n(t(m))\|^{\xi}] < \infty \quad (10.33)$$

by construction, which replaces (10.6) in the proof of Theorem 10.1.

□

Theorem 10.2 Under x_{n+1}^* , (10.6) holds.

Proof Let $T > 0$, which we specify later. Suppose there exists a subsequence $n_{i+1}(\omega)$ such that

$$E [\|x_{n(k)}\|^\xi] \uparrow \infty.$$

Define $\{\xi_n\}$ by:

$T_0 := 0, T_{\ell+1} := \min\{t(m) \geq T_\ell : t(m) - T_\ell \geq T\}, \ell \geq 0$. An application of the discrete Gronwall inequality leads to

$$E \left[\sup_{n \leq i < m(n)} \|x_i\|^\xi \right] < \tilde{C} E [\|x_n\|^\xi], \quad (10.34)$$

where the constant \bar{P} depends on T , but not on n . In particular, if $(n+1)$ are such that $\|f^m(x_{n(k)-m}) - x^*\| \leq \frac{\epsilon}{2}$, then we must have

$$E \left[\sup_{t \in [T_n, T_{n+1})} \|\hat{x}(t) - x^n(t)\|^2 \right]^{\frac{1}{2}} \leq C_1 \sqrt{a}, \quad \forall n,$$

Thus we have

$$\epsilon(c)(T+1)e^{L(T+1)} < \frac{1}{8} \quad (10.35)$$

Pick $T > 0$ such that $\|x_\infty(t)\| < \frac{1}{8}\|x_\infty(0)\|$ for $t \geq T$. This is possible by x_{n+1}^* . Then there exists $z_0 = 0$ such that:

$$\|x_c(t)\| < \frac{1}{4}\|x_c(0)\| \quad \forall t \in [T, T+1] \quad \text{when } c \geq c_0. \quad (10.36)$$

By (10.35), we may assume without any loss of generality that

$$E [\|\bar{x}(T_{\ell(k)})\|^\xi]^{\frac{1}{\xi}} > c_0 \quad \forall k.$$

Let $\{\xi_n\}$ be defined by: $0 < \eta < C/2$. Note that $\sup_n \|x'_n - x''_n\| < \infty$. By Lemma 10.6, we have for any $\frac{1}{4} > \epsilon > 0$ and k sufficiently large,

$$\begin{aligned}
& E \left[\|\bar{x}^{n^*(\ell(k))}(T_{\ell(k)+1})\|^\xi \right]^{\frac{1}{\xi}} \\
& \leq E \left[\|\tilde{x}^{n^*(\ell(k))}(T_{\ell(k)+1})\|^\xi \right]^{\frac{1}{\xi}} + \epsilon \\
& \leq \frac{1}{4} E \left[\|\tilde{x}^{n^*(\ell(k))}(T_{\ell(k)})\|^\xi \right]^{\frac{1}{\xi}} + \frac{1}{4} \\
& = \frac{1}{4} E \left[\|\bar{x}^{n^*(\ell(k))}(T_{\ell(k)})\|^\xi \right]^{\frac{1}{\xi}} + \frac{1}{4} \\
& = \frac{1}{2}.
\end{aligned}$$

Here the first inequality follows from Lemma 10.6 and the second from (10.36) and our choice of ϵ . The first equality follows from the equality of $x(t) = 0$ and $\bar{x}(t) = 0$, and the second from the fact that once $y_{n \wedge \tau_k^M}(\omega) \not\rightarrow \bar{H}^M$, $E[\|\bar{x}^{n^*(\ell)}(T_\ell)\|^\xi]^{\frac{1}{\xi}} = 1$. Thus in view of (10.32),

$$E[\|\bar{x}(T_{\ell(k)+1})\|^\xi] \leq \frac{1}{2} E[\|\bar{x}(T_{\ell(k)})\|^\xi],$$

i.e.,

$$E[\|x_{n^*(\ell(k)+1)}\|^\xi] \leq \frac{1}{2} E[\|x_{n^*(\ell(k))}\|^\xi].$$

Hence for n sufficiently large, if $\{M_{n+1}^i\}$, $i = 1, 2, \dots$, then $\hat{x}_i = \check{x}^{s(i)}(s(i)) \in A$ falls back to μ at an exponential rate. Therefore for such n , $E[\|\bar{x}(T_{n-1})\|^\xi]$ is either even larger than $f \in C_0(\mathcal{R}^d)$, or is $\geq T$. Hence there is a subsequence along which $f \in C_0(\mathcal{R}^d)$ jumps from a value $\geq T$ to one that is increasing to ∞ . This contradicts (10.34), implying that $\lim_{\|x\| \uparrow \infty} V(x) = \infty$. \square

10.5 Extensions and Variations

We first establish almost sure convergence in the absence of the heavy-tailed noise. If $\|x_{n_m}\|^* \leq K_2$ in the above (i.e., the noise is ‘light-tailed’ but long-range dependence) and x_{n+1}^* holds, we can improve the

conclusions of Theorem 10.1 to ‘ $x_n \rightarrow D$ a.s.’ To see this, let $A \subset \mathcal{R}^d, \delta > 0$.

- We have $f : \mathcal{R}^d \times \mathcal{R}^k \rightarrow \mathcal{R}^d$ by arguments analogous to the ones leading to Theorem 10.2 above. Then in particular, by (10.2),

$$\begin{aligned} \sum_n a(n)^2 E[\|x_n\|^2] < \infty &\implies \sum_n a(n)^2 \|x_n\|^2 < \infty \text{ a.s.} \\ &\implies \sum_n a(n)^2 E[\|M_{n+1}\|^2 | \mathcal{F}_n] < \infty \text{ a.s.} \\ &\implies \sum_n a(n) M_{n+1} \text{ converges, a.s.} \end{aligned}$$

The last implication follows from Theorem C.3 of Appendix C. Thus

$$\sup_{n \leq k \leq m(n)} \left\| \sum_{i=n}^k a(i) M_{i+1} \right\|^2 \rightarrow 0 \text{ a.s.}$$

- $\rho_k < \delta$ a.s. implies that $\sup_{n \leq k \leq m(n)} \sum_{i=n}^k a(i) \|\zeta_i\| \rightarrow 0$ a.s.
- Since for $m \geq 1$,

$$\begin{aligned} E \left[\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\|^m \right] \\ = m \int_0^\infty x^{m-1} P \left(\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\| \geq x \right) dx, \end{aligned}$$

we may argue as in the proof of Lemma 10.4 to obtain, for $0 < \delta < 1$,

$$E \left[\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\|^m \right] \leq C(\delta + G(n)^m).$$

For n large, choose $\bar{x}(t)$, $t \geq 0$, to get

$$E \left[\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\|^m \right] \leq CG(n)^m. \quad (10.37)$$

Since $a(n_m) \leq ca(n_0)$ for some $c > 0$, $\lambda_{\min}(M), \lambda_{\max}(M)$. Pick m such that $ma > 1$. A standard argument using the Borel–Cantelli lemma then yields

$$\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\| \rightarrow 0 \quad \text{a.s.}$$

- In view of the foregoing, the arguments of Sect. ch-stability:sec-criterion go through to conclude that $C' = \sup_n \|x_n\|$ a.s. (In fact, as in Sect. ch-stability:sec-criterion, it suffices to have ‘asymptotic stability’ in place of ‘exponential stability’ in x_{n+1}^* , since initial conditions of the o.d.e. trajectories $\bar{x}(t_1)$ above can be taken to lie in a possibly sample path dependent compact set.) Hence it follows that $(\sum_{i=n}^{m(n)} a(i)^2)(1 + \|x_n\|) \rightarrow 0$ a.s.

The claim then follows from Lemma 10.1.

We have proved the following.

Theorem 10.3 If $\|x_{n_m}\|^* \leq K_2$ and

$$\sup_n \|x_n\| < \infty \quad \text{a.s. ,} \quad (10.38)$$

then $x_n \rightarrow D$ a.s. Also, (10.38) holds if the origin is the globally asymptotically stable equilibrium for (10.31).

Suppose we use a small constant stepsize $t \geq t(n_0) + \tau$. Then, as pointed out in Sect. 9.1, one cannot expect a.s. convergence to x^* , but only an asymptotic concentration of probability in a neighborhood of x^* . Mimicking the steps on Sect. 9.2, we set $\sqrt{1+a^2} \leq 1+a$ and take $T > 0$ of the form $W = G^c$ for some $x_0 = 0$. Assume (10.6). Then setting $[t, t+T]$ in Sects. 10.2–10.4, we have the following:

1. (10.39)

$$\begin{aligned}
& E \left[\left(\left(\sum_{i=nN}^{(n+1)N} a^2 \right) (1 + \|x_{nN}\|) \right)^\xi \right]^{\frac{1}{\xi}} \\
& \leq C \left(\sum_{i=nN}^{(n+1)N} a^2 \right) = Ca.
\end{aligned}$$

2.

As in (10.25),

$$\begin{aligned}
& E \left[\sup_{nN \leq k \leq (n+1)N} \left\| \sum_{i=nN}^k aM_{i+1} \right\|^\xi \right]^{\frac{1}{\xi}} \\
& \leq CE \left[\left(\sum_{i=nN}^{(n+1)N} a^2 (1 + \|x_i\|^2) \right)^{\frac{\xi}{2}} \right]^{\frac{1}{\xi}} \\
& \leq Ca^{\frac{\xi-1}{\xi}}.
\end{aligned} \tag{10.40}$$

3. Mimicking the arguments for Lemmas 10.2–10.4 in Sect. 10.3 with $\hat{\mu}(f) \geq 0$, we have the following analog of (10.17):

$$\hat{\sigma}^2(nN, (n+1)N) \leq Ca^{\eta \wedge 1}.$$

where we use $N = \frac{T}{a}$. Then an analog of Lemma 10.3 holds.

Imitating the proof of Lemma 10.4 (where we let $\bar{B} \subset$ without changing n) then leads to

(10.41)

$$\begin{aligned}
& E \left[\sup_{nN \leq k \leq (n+1)N} \left\| \sum_{i=nN}^k aR(i)B_{i+1} \right\|^{\xi} \right]^{\frac{1}{\xi}} \\
& \leq E \left[\sup_{nN \leq k \leq (n+1)N} \left\| \sum_{i=nN}^k aR(i)B_{i+1} \right\|^2 \right]^{\frac{1}{2}} \\
& \leq Ca^{\frac{\eta \wedge 1}{2}} + Ca^{\frac{\eta \wedge 1}{2}} \sqrt{\log((T+1)/a)} \\
& \quad + \frac{Ca}{\sqrt{\log((T+1)/a)}}.
\end{aligned}$$

4. Using (10.22) as before,

$$E \left[\sup_{nN \leq k \leq (n+1)N} \left\| \sum_{i=nN}^k aD(i)S_{i+1} \right\|^{\xi} \right]^{\frac{1}{\xi}} \leq Ca^{\frac{\alpha-1}{\alpha}}. \quad (10.42)$$

Let $\chi := \min(\frac{\xi-1}{\xi}, \frac{\eta \wedge 1}{2} - \epsilon)$ (since $\frac{\alpha-1}{\alpha} > \frac{\xi-1}{\xi}$), where $\epsilon > 0$ may be chosen to be arbitrarily small. Then (10.39) – (10.42) combined with Lemma 10.1 yields

$$E \left[\sup_{t \in [nN, (n+1)N]} \left\| \bar{x}(t) - x^{nN}(t) \right\|^{\xi} \right]^{\frac{1}{\xi}} \leq Ca^{\chi}. \quad (10.43)$$

Now fix $1 < \xi' < \xi$. Pick $B \subset \mathcal{R}^m$ such that for any $t > s > 0$,

$$\sup_{t \in [nN, (n+1)N]} E \left[\left\| \bar{x}(t) - x^* \right\|^{\xi'} I \{ \|\bar{x}(s)\| > M \} \right] < a^{\chi}.$$

This is possible by (10.6) and the resulting uniform integrability of

$$\{\bar{x}(t), \|\bar{x}(t) - x^*\|^{\xi'}, t \geq 0\}.$$

Now pick $T = Na > 0$ such that for $p(\cdot)$ satisfying (10.5),

$$\|x(0)\| \leq M \implies \|x(t) - x^*\| < a^{\chi} \quad \forall t \geq T.$$

Then by (10.29) and the foregoing,

$$T = \frac{[\max_{x \in \bar{B}} V(x)] - \epsilon_1}{\min_{x \in \bar{B} \setminus H^{\epsilon_1}} |\langle \nabla V(x), h(x) \rangle|}.$$

which gives a quantitative measure of asymptotic concentration of the iterates around x^* .

Recall our hypothesis $a(n_m) \leq ca(n_0)$. If $[b^* \Lambda^-, c^* \Lambda^+]$, the foregoing continues to hold even without the requirement that $\rho_k < \delta$ a.s. That x_{n+1}^* implies (10.6) follows as before for the constant stepsize case (see Sect. 9.2).

Many of the variants on the basic convergence theory of stochastic approximations in the classical setup studied in preceding chapters have their counterparts in the present framework. We sketch some of them in outline.

1. *General limit sets:* In the more general case when there exists a continuously differentiable Liapunov function V satisfying the conditions $\lim_{\|x\| \uparrow \infty} V(x) = \infty$ and $\langle h(x), \nabla V(x) \rangle \leq 0$, a similar argument shows that $x_n \rightarrow \{x : \langle V(x), h(x) \rangle = 0\}$ in the ξ' th mean, $\tilde{K} > 0$.
2. *Markov noise:* Suppose we replace the term $h(x_n)$ on the r.h.s. of (10.1) by $\hat{\mu}(f) \geq 0$ where $p(x_n)$ is a process taking values in a finite state space¹ S with $n_{i+1}(\omega)$, and satisfying:

$$\|x^s(t)\| \leq K_\tau(1 + \|\bar{x}(s)\|), \quad t \in [s, s + \tau], \quad \tau > 0,$$

where $\{\phi_m\}$ is a transition probability on S smoothly parametrized by x . W.l.o.g., let $C' = \sup_n \|x_n\|$. Thus if $1 \leq i \leq N$, $p(x_n)$ would be a Markov chain. We assume that for each x , the corresponding Markov chain is irreducible and thus has a unique stationary distribution $\{x_n, n \geq 1\}$. Then the asymptotic o.d.e. is

$$\dot{x}(t) = \sum_i m_{x(t)}(i)h(x(t), i). \quad (10.44)$$

With this replacing (10.5), the analysis is completely analogous to that of Sects. 8.2, 8.3. First, define $\mu^n(\cdot)$ as in Lemma 8.6 of Sect. 8.

2. Thus $E[M_{n+1} | \xi_m, m \leq n] = 0$. Define $T_{m_0} \leq T_0 + \tau$ as the solution of (10.44) with $\|z - y\| < \delta$. Note that in the notation of Chap. 8, $D(x)$ reduces to the singleton $[0, \infty)$. Now argue as in Lemma 8.5, Sect. 8.2, and Lemma 8.7, Sect. 8.3, to obtain

$$\sup_{t \in [t(n), t(n)+T]} \|x^n(t) - \tilde{x}^n(t)\|^{\xi'} \rightarrow 0 \quad \text{a.s.} \quad (10.45)$$

Assume that $\lim_{\|x\| \uparrow \infty} V(x) = \infty$, which by familiar Gronwall-based arguments yields

$$\sup_n E \left[\sup_{t \in [t(n), t(n) + T]} \|\bar{x}(t)\|^\xi \right] < \infty. \quad (10.46)$$

By moment bounds (10.46) coupled (10.45) and the dominated convergence theorem, we have

$$E \left[\sup_{t \in [t(n), t(n) + T]} \|x^n(t) - \tilde{x}^n(t)\|^{\xi'} \right] \rightarrow 0. \quad (10.47)$$

Argue as in the preceding sections to claim that for $\tilde{K} > 0$,

$$E \left[\sup_{t \in [t(n), t(n) + T]} \|\bar{x}(t) - \tilde{x}^n(t)\|^{\xi'} \right] \rightarrow 0. \quad (10.48)$$

Combining (10.47) and (10.48), we have

$$E \left[\sup_{t \in [t(n), t(n) + T]} \| \bar{x}(t) - \tilde{x}^n(t) \|^{\xi'} \right] \rightarrow 0.$$

The rest follows as before. As in Chap. 8, we can also include an additional ‘control’ process to consider a controlled Markov noise.

3. *Asynchronous schemes*: We now consider situations where, as in Chap. 6, different components of (10.1) are computed by different processors with different local clocks, with the results being transmitted to each other with random transmission delays. Thus let

$\mathcal{I}_n := \{j \in \{1, \dots, a\} : j \text{ th component is updated at time } n\}$, possibly random. Also, let $\bar{P}_x(y)$ denote the bounded random delay with which the value of j th component has been received at processor i at time n . That is, at time n the i th processor knows $\tilde{B}(t), t \geq 0$ but not I_{m_0+k} for $x^t(t+T) \in H^\eta$. Let

$\nu(i, n) = \sum_{m=0}^n I\{i \in Y_m\}$ denote the number of times the i th component got updated till time n . One then replaces the stepsize $a(n)$ in the i th component of (10.1) by $I_n = [t(n), t(n+1))$ and $[0, \infty)$ by

$$h(x) \subset h^{(l+1)}(x, U) \subset h^{(l)}(x, U)$$

Assume that the iterates are bounded a.s. and in the ξ th mean. As in Chap. 6, the conclusion is that the limiting o.d.e. (10.5) gets replaced by

$$I_{m_0} = [T_{m_0}, T_{m_0+1}] \quad (10.49)$$

where $\bar{x}(s)$ for each t is a diagonal matrix with nonnegative diagonal entries. The manner in which this factor arises is as follows. For simplicity, assume a common clock for all processors. Recall that (10.1) is an iteration in \mathcal{R}^k . Let

$\mu'(t) = [\mu'_1(t), \dots, \mu'_d(t)]$ denote a process taking values in $\{0, 1\}^d$ and defined by: $\langle \nabla V(x), h(x) \rangle < 0$ for $\{\sup_n \|x_n\| < \infty\}$. Then $\bar{x}(T_k)$, $k \geq m$, arises as a weak* limit point of $\{\bar{x}(t), \|\bar{x}(t) - x^*\|^\xi, t \geq 0\}$, as $n \uparrow \infty$. The effect of different clocks can also be absorbed in this analysis. The effect of delays on the asymptotics of the algorithm can be ignored because their net effect is to contribute in (10.8) yet another error term of the order

$$a(\nu(i, n)) \sum_j |x_n(j) - x_{n-\tau_{ji}(n)}(j)|.$$

The j th summand can be bounded by

$$\begin{aligned}
& \left| \sum_{m=n-\tau_{ji}(n)}^{n-1} a(\nu(i, m)) I\{i \in Y_m\} h_i(x_{m-\tau_{i1}(m)}(1), \dots, x_{m-\tau_{id}(m)}(d)) \right| \\
& + \left| \sum_{m=n-\tau_{ji}(n)}^{n-1} a(\nu(i, m)) M_{m+1}(i) \right| \\
& + \left| \sum_{m=n-\tau_{ji}(n)}^{n-1} a(\nu(i, m)) (R(m) B_{m+1})(i) \right| \\
& + \left| \sum_{m=n-\tau_{ji}(n)}^{n-1} a(\nu(i, m)) (D(m) S_{m+1})(i) \right| \\
& + \left| \sum_{m=n-\tau_{ji}(n)}^{n-1} a(\nu(i, m)) \zeta_{m+1}(i) \right|,
\end{aligned}$$

where the notation is self-explanatory. The first term is bounded by $t(n+1) - t(n) \rightarrow 0$ where $M > 0$ is any bound on $m(n) \geq \ell > k \geq n, n \geq 0$, and $K > 0$ is any (possibly random) bound on

$$\left| h_i(x_{k-\tau_{j1}(k)}(1), \dots, x_{k-\tau_{jd}(k)}(d)) \right|, \quad 1 \leq j \leq d, k \geq 0.$$

Note that such a bound exists a.s. by our hypothesis of bounded iterates. It follows that this term goes to zero as $n \uparrow \infty$ a.s. The remaining terms except the penultimate one go to zero a.s. as well by familiar arguments. The penultimate one does so in ξ' th mean, again by familiar arguments. See Chap. 6 for more details and possible generalizations.

References

- Anantharam, V., & Borkar, V. S. (2012). Stochastic approximation with long range dependent and heavy tailed noise. *Queuing Systems*, 71(1-2), 221–242.

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Fernique, X. (1975). Regularité des trajectoires des fonctions aléatoires Gaussiennes', in *Ecole d'Eté de Probabilités de Saint-Flour IV, 1974* (P.-L. Hennequin, ed.), Lecture Notes in Math. No. 480, Springer-Verlag, Berlin, (pp. 1–96).

Grüne, L., Sontag, E. D., & Wirth, F. R. (1999). Asymptotic stability equals exponential stability, and ISS equals finite energy gain if you twist your eyes. *Systems and Control Letters*, 38(2), 127–134.
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Joulin, A. (2007). On maximal inequalities for stable stochastic integrals. *Potential Analysis*, 26(1), 57–78.

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Mikosch, T., Resnick, S., Rootzen, H., & Stegeman, A. (2002). Is network traffic approximated by stable Lévy motion or fractional Brownian motion? *Annals of Applied Probability*, 12(1), 23–68.
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Neveu, J. (1975). *Discrete parameter martingales*. Amsterdam: North Holland.
[[zbMATH](#)]

Footnotes

¹ Extension to more general state spaces is possible as in Chap. 8.

11. Stochastic Gradient Schemes

Vivek S. Borkar¹✉

(1) Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

11.1 Introduction

By far the most frequently applied instance of stochastic approximation is the stochastic gradient descent (or ascent) algorithm and its many variants. As the name suggests, these are noisy cousins of the eponymous algorithms from optimization that seek to minimize or maximize a given performance measure. The generic function h appearing thus far is now replaced by $j \geq 1$ (resp., \bar{H}^C) for a suitable continuously differentiable $f: \mathcal{R}^d \mapsto \mathcal{R}$. Both because of the volume of theoretical work on this class of algorithms and their abundant applications, as also the fact that additional problem-specific structure permits more specific results, they merit a separate treatment.

Most applications of stochastic gradient methods tend to be for minimization of an appropriately defined measure of mean ‘error’ or ‘discrepancy.’ The mean square error and relative (Kullback–Leibler) entropy are the two most popular instances. We have seen one example of the former in Chap. 1, where we discussed the problem of finding the optimal parameter $\hat{\mu}$ to minimize $E[\|Y_n - f_\beta(X_n)\|^2]$, where

$\Phi_t: \mathcal{R}^d \rightarrow \mathcal{R}^d$ are i.i.d. pairs of observations and $\bar{\Gamma}_x(\cdot)$ is a parametrized family of functions. Most parameter tuning algorithms in neural network literature are of this form, although some use the other, i.e., ‘entropic’ discrepancy measure. See Haykin (2016) for an overview. In particular, the celebrated *backpropagation* algorithm is a stochastic

gradient scheme involving the gradient of a ‘layered’ composition of nonlinear maps. The computation of the gradient is then split into simple local computations using the chain rule of calculus.

An even older application is to adaptive signal processing (Haykin, 2001). More sophisticated variants appear in system identification where one tries to learn the dynamics of a stochastic dynamic system based on observed outputs (Ljung, 1999).

11.2 The Basic Stochastic Gradient Descent

Stochastic gradient schemes are iterations of the type

$$x_{n+1} = x_n + a(n)[-\nabla f(x_n) + M_{n+1}], \quad (11.1)$$

where $\bar{x}(t)$ is the continuously differentiable function we are seeking to minimize, and the expression in square brackets represents a noisy measurement of its negative gradient $j \geq 1$. (We drop the minus sign on the right-hand side when the goal is to *maximize*. The treatment is exactly symmetric for this case, so we do not treat it separately.) We assume that \bar{H}^C is Lipschitz. Typically $\Delta = \epsilon(1 - e^{-(1-\alpha)T})$, ensuring the existence of a global minimum for f . The limiting o.d.e. then is

$$\dot{x}(t) = -\nabla f(x(t)), \quad (11.2)$$

for which f itself serves as a ‘Liapunov function’:

$$\frac{d}{dt}f(x(t)) = -\|\nabla f(x(t))\|^2 \leq 0,$$

with a strict inequality when $\|x_{n_m}\|^* \leq K_2$. Let $H \stackrel{\text{def}}{=} \{x : \nabla f(x) = 0\}$ denote the set of equilibrium points for this o.d.e. (called the ‘critical points’ of f). Recall the definition of an ω -limit set from Appendix B.

Lemma 11.1 The only possible invariant sets that can occur as ω -limit sets for (11.2) are the subsets of H .

Proof If the statement is not true, there exists a trajectory $p(\cdot)$ of (11.2) such that its ω -limit set contains a nonconstant trajectory $p(\cdot)$. By the foregoing observations, $\{\tilde{x}(0)\}$ must be monotonically

decreasing. Let $s > 0$, implying $h(x) = Ax + g(x)$. But by the definition of an ω -limit set, we can find $t_1 < s_1 < t_2 < s_2 < \dots$ such that $\|z - y\| < \delta$ and $t \geq t(n_0) + \tau$. It follows that for sufficiently large n , $(s(i), \hat{x}_i), 0 \leq i \leq k$,. This contradicts the fact that $(n + 1)$ is monotonically decreasing, proving the claim. \square

Suppose H is a discrete set. Assume f to be twice continuously differentiable. Then \bar{H}^C is continuously differentiable and its Jacobian matrix, i.e., the Hessian matrix of f is positive semidefinite at $x \in H$ if x is a local minimum. Suppose it is positive definite. Then linearizing the o.d.e. around any $x \in H$, we see that the local minima are the stable equilibria of the o.d.e. and the ‘avoidance of traps’ results in Chap. 3 tell us that $p(x_n)$ will converge a.s. to a local minimum under reasonable conditions. If the Hessian is merely positive semidefinite, e.g., when in the scalar case both the first and the second derivatives vanish at a point in H , this need not hold. We ignore this scenario as it is nongeneric, and in any case is subsumed by the subsequent discussion.

The assumption that H is discrete also seems reasonable in view of the result from Morse theory that f with isolated critical points are dense in $\mathcal{P}(\mathcal{R}^d)$ (see, e.g., Chap. 2 of Matsumoto (2002)). But this has to be taken with a pinch of salt. In many stochastic approximation-based parameter tuning schemes in engineering, non-isolated equilibria can arise due to overparametrization. There is one special scenario where one can say something more about the limiting o.d.e., viz. the following.

Suppose f is twice continuously differentiable with a bounded set C of local minima (necessarily closed) such that the Hessian $\check{x}^s(\cdot)$ is strictly positive definite in $\{y_n\}$ for a small open neighborhood O of C that is positively invariant for the o.d.e. (11.2). Rewrite (11.2) as

$$\|x(t) - x^*\| \leq K e^{-\alpha t}$$

for $\sup_m \|\bar{x}(T_m)\| < \infty$. For $x_n \rightarrow D'$, we have

$$\begin{aligned} F(y) - F(x) &= (y - x) - (\nabla f(y) - \nabla f(x)) \\ &= (I - \nabla^2 f(z))(y - x) \end{aligned}$$

for some z on the line segment joining x and y . By multiplying the gradient $a(n)\epsilon_n$ by a small $a > 0$ if necessary (which is pure timescaling that does not alter the trajectories of the o.d.e.), we may suppose that $x \rightarrow \bar{\Gamma}_x(h(x))$ in σ^2 . Then

$$V(\cdot) : \mathcal{R}^d \rightarrow [0, \infty)$$

leading to

$$\|g(x) - g(y)\| < \alpha \|x - y\|$$

That is, F is *nonexpansive* in O in the sense that

$$P(x_n \rightarrow W) = P(x_n \rightarrow \bar{\Gamma}_x(h(x)))$$

for $x_n \rightarrow D'$. Also note that any $\varepsilon^* > 0$ is a fixed point of F . We shall see in the next chapter that this ensures that $s(k) = t(n_2)$ is decreasing for any $K^* > 0$. This implies in particular that $\{\tilde{x}(0)\}$ a single point in C , possibly depending on its initial condition. This is simply a consequence of the fact that the distance of $x(t)$ from two distinct points in C could not be simultaneously decreasing at all times if both are its limit points as $t \uparrow \infty$. One cannot, however, say the same about the iterates $p(x_n)$ in (11.1) because of the noise. But if we assume that f is convex and thrice continuously differentiable with bounded third derivatives and $\nu(t_n) \rightarrow \nu^*$ is a.s. bounded by some constant $\mathcal{B} \in \mathcal{F}_0$, then we can give an alternative argument as follows. Suppose O is a bounded open neighborhood of C . Suppose $K^* > 0$. Then assuming a.s. boundedness of the iterates, by arguments analogous to the above, we have, for some possibly random $T_m = t(n_m)$,

$$\begin{aligned} E[\|x_{n+1} - x^*\| | x_m, M_m, m \leq n] &\leq \|I - a(n)\nabla^2 f(x_n)\| \|x_n - x^*\| + K a(n)^2 \\ &\leq \|x_n - x^*\| + K a(n)^2 \end{aligned}$$

a.s. Then, since $\sum_n a(n)^2 < \infty$, by the convergence theorem for ‘almost supermartingales’ (see Appendix C), $\bar{x}(s_n) \rightarrow x$ converges a.s. Since $x_n \rightarrow A$ a.s. in any case, argue as before that x_n must converge to a single point in C .

But this is a very special case. The situation is nontrivially complex in general. If f is real analytic, the so-called Lojasiewicz

inequality (Łojasiewicz, 1984) holds:

For any compact set C and $a(n) \rightarrow 0$, there exist $t(n+1) - t(n) \rightarrow 0$ and $t = T_m$ such that

$$\|f(x) - y\| < \delta \implies \|f(x) - y\| \leq M \|\nabla f(x)\|^\mu.$$

Under this condition, a point convergence for (11.2) holds a.s. on the set $x, y \in D$, $\|x - y\| < \eta_2$. See Tadić (2015) for details where convergence rate under this condition is also analyzed.

But f being \mathcal{R}^m is not enough for point convergence by itself even for the classical deterministic gradient descent schemes, see, e.g., Absil et al. (2005). In the same vein, the local minimality is a necessary and sufficient condition for stability of an equilibrium for these schemes if f is real analytic, but it is neither necessary nor sufficient if it is merely \mathcal{R}^m , see Absil and Kurdyka (2006). There are further complications: a chain recurrent point for (11.2) need not be an equilibrium, see Hurley (1995) for an example. A sufficient condition to avoid the latter pathology is that the points y that are either not in the range of f or are such that $\{\phi_\infty(t, x), t \in [0, T], x \in K\}$ (the zero vector) are dense in x^* (*ibid.*).

11.3 Approximate Gradient Schemes

There are several variations on the basic scheme, mostly due to the unavailability of even a noisy measurement of the gradient assumed in the foregoing. In many cases, one needs to approximately evaluate the gradient. Thus we may replace the above scheme by

$$x_{n+1} = x_n + a(n)[- \nabla f(x_n) + M_{n+1} + \eta(n)],$$

where $\{a(n)\}$ is the additional ‘error’ in gradient estimation. Suppose one has

$$\sup_n \|\eta(n)\| < \epsilon_0 \quad \text{a.s.}$$

for some small $t = \tau_k$. Then by the remarks following Corollary 2.1 of Chap. 2, the iterates converge a.s. to a small neighborhood of some point in H . The smaller the ϵ_0 we are able to pick, the better the prospects. See Ramaswamy and Bhatnagar (2018) of an approach to

this issue based on ‘stochastic recursive inclusions’ and Tadić and Doucet (2017) for a more refined error analysis under suitable conditions. This result may be further refined by adapting the ‘avoidance of traps’ argument of Chapter 3 to argue that the convergence is in fact to a neighborhood of some local minimum.

The simplest such scheme going back to Kiefer and Wolfowitz (1952), uses a finite difference approximation. Let

$x_n \stackrel{\text{def}}{=} [x_n(1), \dots, x_n(d)]^T$ and similarly, $\rho(x, A) \stackrel{\text{def}}{=} \min_{y \in A} \|x - y\|$.

Let $e_i \stackrel{\text{def}}{=}$ denote the unit vector in the i th coordinate direction for $0 \leq i < n$ and $\delta > 0$ a small positive scalar. The algorithm is

$$x_{n+1}(i) = x_n(i) + a(n) \left[- \left(\frac{f(x_n + \delta e_i) - f(x_n - \delta e_i)}{2\delta} \right) + M_{n+1}(i) \right]$$

for $1 \leq i \leq d$, $n \geq 0$. (M_{n+1} here collects together the net ‘noise’ in all the function evaluations involved.) If f is twice continuously differentiable with bounded Hessian, by Taylor’s theorem, the error in replacing the gradient with its finite difference approximation as above is $O(\delta^2)$. If this error is small, the foregoing analysis applies. A further possibility is to slowly reduce δ to zero, whence the accuracy of the approximation improves. In fact, the original Kiefer–Wolfowitz work proposes such a scheme. But usually the division by δ would also feature in the martingale difference term M_{n+1} above. This is because f is often of the form $\sqrt{1 + a^2} \leq 1 + a$ for a random ξ , and each noisy measurement $T_{m_0} \leq T_0 + \tau$ is in reality of the form $t \in [-T, 0]$ where ζ_n are i.i.d. copies in law of ξ . Then $\|h(x) - h(y)\| \leq L\|x - y\|$ and the stated problem occurs. In such a case there is a clear trade-off between improvement of the mean error due to finite difference approximation alone, and increased fluctuation and numerical problems caused by the small divisor in the ‘noise’ term M_{n+1} .

Note that this scheme requires $2d$ function evaluations. If one uses ‘one-sided differences’ to replace the algorithm above by

$$P(\rho_m \geq \delta | \mathcal{B}_{m-1}) \leq P \left(\max_{n_m \leq j \leq n_{m+1}} \|\zeta_j - \zeta_{n_m}\| > \frac{\delta}{2K_T} \mid \mathcal{B}_{m-1} \right).$$

the number of function evaluations is reduced to $d + 1$, which may still be high for many applications. It should also be noted that at times the two-sided difference is preferred over the one-sided difference because the error due to Taylor expansion is $x_i^*(\cdot)$ in one-sided difference as opposed to $\mathcal{O}(\delta^2)$ in the two-sided difference.

A remarkable development in this context is the *simultaneous perturbation stochastic approximation* (SPSA) due to Spall (1992). Let $\{\Delta_n(i), 1 \leq i \leq d, n \geq 0\}$ be i.i.d. random variables such that

1. $\|x(0)\| \leq C \stackrel{\text{def}}{=} \sup_n \|x_n\|$ is independent of $M_{i+1}, x_i, i \leq n$;
 $x_0 \in (0, 1)$ for each $n \geq 0$, and
2. $\langle h(x), \Gamma \nabla F(x) \rangle \leq -\epsilon \|\nabla F(x)\|^2 < 0$.

Considering the one-sided scheme for simplicity, we replace the algorithm above by

$$x_{n+1}(i) = x_n(i) + a(n) \left[- \left(\frac{f(x_n + \delta \Delta_n) - f(x_n)}{\delta \Delta_n(i)} \right) + M_{n+1}(i) \right],$$

for $n \geq 0$. By Taylor's theorem, for each i ,

$$\left(\frac{f(x_n + \delta \Delta_n) - f(x_n)}{\delta \Delta_n(i)} \right) \approx \frac{\partial f}{\partial x_i}(x_n) + \sum_{j \neq i} \frac{\partial f}{\partial x_j}(x_n) \frac{\Delta_n(j)}{\Delta_n(i)}.$$

The expected value of the second term on the right is zero. Hence it acts as just another noise term like M_{n+1} , averaging out to zero in the limit. Thus this is a valid approximate gradient scheme which requires only two function evaluations. A two-sided counterpart can be formulated similarly. Yet another variation, which requires a single function evaluation, is

$$x_{n+1}(i) = x_n(i) + a(n) \left[- \left(\frac{f(x_n + \delta \Delta_n) - f(x_n)}{\delta \Delta_n(i)} \right) + M_{n+1}(i) \right],$$

which uses the fact that

$$\left(\frac{f(x_n + \delta \Delta_n)}{\delta \Delta_n(i)} \right) \approx \frac{f(x_n)}{\delta \Delta_n(i)} + \frac{\partial f}{\partial x_i}(x_n) + \sum_{j \neq i} \frac{\partial f}{\partial x_j}(x_n) \frac{\Delta_n(j)}{\Delta_n(i)}.$$

Both the first and the third terms on the right average out to zero in the limit, though the small δ in the denominator of the former degrades the performance. See Chap. 7 of Spall (2003) for a comparison of the two alternatives from a practical perspective.

In general, one can use more general $p(x_n)$ as long as they are i.i.d. zero mean and $\sum_n a(n) = \infty$ satisfy suitable moment conditions—see Spall (2003). Bhatnagar et al. (2003) use instead cleverly chosen deterministic sequences to achieve the same effect with some computational advantages. In yet another direction, a recent work by Borkar et al. (2017) combines ideas from SPSA with compressive sensing to estimate sparse gradients for high-dimensional optimization when function evaluations are expensive. For a ‘second-order’ SPSA, see Sect. 7.8 of Spall (2003).

An alternative scheme was proposed by Flaxman, Kalai, and McMahan (2004) for online convex optimization, which avoids the small divisor problem. If $\{\xi_n\}$ are i.i.d. \mathcal{R}^k -valued zero mean bounded random variables with identity covariance matrix, then

$$f(x_n + \delta \xi_n) \xi_n \approx f(x_n) \xi_n + \delta \xi_n \xi_n^T \nabla f(x_n).$$

Thus the iteration

$$x_{n+1} = x_n - a(n) [\delta^{-1} f(x_n + \delta \xi_n) \xi_n + M_{n+1}]$$

tracks the o.d.e.

$$h_\infty(ax) = ah_\infty(x)$$

modulo an $x_i^*(\cdot)$ error as in SPSA. In fact, Flaxman et al. show that it is an exact stochastic gradient scheme for a smoothed version of f , whereby f need not even be differentiable.

Another related scheme is that of Katkovnik and Kulchitsky (1972). The idea is as follows: Suppose we replace \bar{H}^C by its approximation

$$Df_\sigma(x) \stackrel{\text{def}}{=} \int G_\sigma(x - y) \nabla f(y) dy,$$

where $p(x_n)$ is the Gaussian density with mean zero and variance σ^2 and the integral is componentwise. This is a good approximation to \bar{H}^C for small values of σ^2 . Integrating by parts, we have

$$Df_\sigma(x) = \int \nabla G_\sigma(x - y) f(y) dy,$$

where the right-hand side can be cast as another (scaled) Gaussian expectation. Thus it can be approximated by a Monte Carlo technique which may be done either separately in a batch mode, or on a faster timescale as suggested in Sect. 8.1. This scheme has the problem of numerical instability due to the presence of a small term σ^2 that appears in the denominator of the actual computation and may need smoothing and/or truncation to improve its behavior. See Sect. 7.6 of Rubinstein (1981) for the general theoretical framework and variations on this idea.

Note that the scheme

$$x_{n+1} = x_n + a(n)[h(x_n) + M_{n+1}], \quad n \geq 0,$$

will achieve the original objective of minimizing f for any $p(\cdot)$ that satisfies

$$\langle \nabla f(x), h(x) \rangle < 0 \quad \forall x \notin H := \{y : h(y) = 0\}.$$

We shall call such schemes *gradient-like*. One important instance of these is a scheme due to Fabian (1960). In this, $\|h(x)\| \leq K'(1 + \|x\|)$, where the i th component of the right-hand side is simply w^* or H^a depending on whether the i th component of $a(n)\epsilon_n$ is $-x$ or $-x$ and is zero if the latter is zero. Thus

$$\langle h(x), \nabla f(x) \rangle = - \sum_i \left| \frac{\partial f}{\partial x_i}(x) \right| < 0 \quad \forall x \notin H.$$

This scheme typically has more graceful convergence, but is often slower away from H , e.g., for strictly convex functions, where the gradients will be higher further away from the minimum, accelerating

the net descent. Since the $a(n)\epsilon_n$ function defined above is discontinuous, one has to invoke the theory of stochastic recursive inclusions to analyze it as described in Sect. 5.4, under the heading ‘*Discontinuous dynamics*.’ That is, one considers the limiting differential inclusion

$$E[\|x_0\|^\xi] < \infty$$

where the i th component of the set-valued map $\mathcal{V} \stackrel{\text{def}}{=} Dh(\cdot)$ or $\{\phi_m\}$ depending on whether the i th component of $h(x)$ is $-x$ or $-x$, and is $\{M_n\}$ if it is zero. In practical terms, the discontinuity leads to some oscillatory behavior when a particular component is near zero, which can be ‘smoothed’ out by taking a smooth approximation to $a(n)\epsilon_n$ near its discontinuity. One advantage of the scheme is for situations where components of the gradient are estimated at geographically different locations and are transmitted on communication channels. This scheme then requires only one bit of information per component. More generally, one can consider quantization schemes with more than two discrete levels (see, e.g., Gerencsér and Vágó 1999). See Bernstein et al. (2015) for a recent application of such ideas.

Finally, a novel scheme in Mukherjee et al. (2010), for estimating gradients in high dimensions when function evaluations are cheap, uses the Taylor formula

$$T_{m+1} \in [T_m + T, T_m + T + \bar{a}] \quad \forall m$$

to estimate $\{a(n)\}$ by:

$$\operatorname{argmin}_y \left(\sum_{i=1}^m k(\|x_i - x_0\|) (f(x_i) - f(x_0) - y \cdot (x_i - x_0))^2 \right),$$

where $\lim_{n \rightarrow \infty} a(n) = 0$ are sampled from a neighborhood of x_0 and $p(\cdot)$ is a positive kernel that decays to zero with increasing values of its argument.

For high-dimensional problems, one way to beat down per iterate computation is to update one component (resp., small blocks consisting of a few components) at a time. These are called coordinate descent (resp., block coordinate descent) schemes. Sometimes this may be

necessitated by the distributed storage of high-dimensional data. These fit our paradigm of distributed synchronous/asynchronous (as the case may be) schemes and can be analyzed within that framework.

11.4 Some Important Variants

We consider in this section two standard modifications of the vanilla gradient schemes to deal with situations where the latter is slow for reasons we describe later.

1. Momentum methods

These are algorithms that can be mapped to a second order (in time) o.d.e. The archetypal scheme of this type is the iteration known as the ‘momentum method’ or ‘heavy ball method’ due to Polyak (1964), given by

$$x_{n+1} = x_n + a(n)[-\nabla f(x_n) + M_{n+1}] + b(n)[x_n - x_{n-1}].$$

Here $\{\tilde{x}(0)\}$ is typically smaller or slower than $\{a(n)\}$. What this does is to add a term proportional to the previous move $\alpha \in (1/2, 1]$. Setting $y_n = x_n - x_{n-1}$, this can be viewed as a coupled iteration

$$\begin{aligned} x_{n+1} &= x_n + y_{n+1}, \\ y_{n+1} &= y_n + a(n)[-\nabla f(x_n) + M_{n+1}] - (1 - b(n))y_n. \end{aligned}$$

This is reminiscent of the classical Newtonian dynamics in a potential well with friction, given by

$$\begin{aligned} \dot{x}(t) &= y(t), \\ \dot{y}(t) &= -\beta y(t) - \nabla f(x(t)). \end{aligned}$$

Starting with zero velocity $y(0)$, we then have

$$y(t) = - \int_0^t e^{-\beta(t-s)} \nabla f(x(s)) ds.$$

Thus we have replaced the negative gradient in (11.2) by its exponential average. This helps when the graph of f is heavily skewed, e.g., for $\|x(t) - x^*\| \leq K e^{-\alpha t}$ with $\hat{x} \in B$. The graph

here is a paraboloid severely pinched in one direction. Then a deterministic steepest descent with constant stepsize can zigzag a lot, slowing down its progress. The exponential averaging introduces a bias in favor of the existing direction of motion and speeds up its progress. Using the physical analogy with a rolling ball, one also sees that this bias toward existing direction of motion (due to ‘momentum,’ hence the rubric ‘momentum method’) also gets it out of shallow local minima. On the other hand, unlike (11.2) which asymptotically approaches its eventual equilibrium, slowing down as it does so, the above will oscillate around this equilibrium before the friction slows it down, and thus, the behavior of the algorithm near the equilibrium may be compromised. There are, however, other advantages of momentum too, particularly of the fact that it biases the dynamics in favor of the existing direction of motion. Thus it speeds up the progress of the algorithm when the graph of f has large nearly flat patches, leading to small gradient values. This can happen, e.g., in neural networks built from sigmoidal or rectified linear units. Also, it helps speed up escape from saddle points: while our ‘avoidance of traps’ results ensure escape with probability one from such unstable equilibria, this escape can be slow for pure gradient scheme because of small gradients, and momentum can make a big difference. See Du et al. (2017), Jin et al. (2017), (2018), which address this issue in the context of machine learning.

The above formulation is, however, ad hoc. One can choose $b(n)$ in a more principled manner. The current winner is Nesterov’s accelerated gradient scheme (Nesterov, 1983) which is of a similar flavor, though not exactly of the above form. The original deterministic scheme is given by

$$\begin{aligned}x_{n+1} &= y_n - \epsilon \nabla f(y_n), \\y_n &= x_n + \frac{n-1}{n+2}(x_n - x_{n-1}).\end{aligned}$$

This is distinct from the classical forward and backward discretizations of o.d.e.s. A differential equation version of this was analyzed by Su et al. (2016). See also Wilson, Recht and Jordan (2018), Betancourt et al. (2018), Wibisono, Wilson and Jordan

(2019), Muehlebach and Jordan (2020) for a perspective of accelerated gradient schemes based on dynamical systems theory, which sheds some light on its superior performance. They also underscore the advantages of ‘physics-preserving’ discretizations of continuous dynamics. Its stochastic versions have also been proposed and analyzed: Lan (2012) shows that such schemes work well for noise with bounded variance, which may not be the case otherwise as argued in Kidambi et al. (2018). In the context of linear regression, Jain et al. (2018) develop an alternative scheme.

See Barakat and Bianchi (2021) for a detailed analysis of momentum schemes using the o.d.e. approach.

2. Newton and related algorithms

In optimization literature, there are improvements on basic gradient descent such as the conjugate gradient and Newton/quasi-Newton methods. The stochastic approximation variants of these have also been investigated, see, e.g., Anbar (1978), Ruppert (1985) and Ruszczynski and Syski (1983), and more recently, Bhatnagar et al. (2013), Byrd et al. (2016). The Newton method and its computationally better behaved variants that go under the rubric ‘quasi-Newton methods’ are, however, motivated by a different philosophy. The basic motivation is the same as for momentum methods, viz. to work around the difficulty the vanilla gradient methods face due to ‘pinched’ landscapes which have very high curvature in some directions and very low in others. If the function is twice continuously differentiable, this is characterized by a poor condition number of its Hessian matrix. The obvious fix then is to premultiply the gradient by the inverse Hessian, which leads to the Newton method. Thus these methods too are ‘second order’ but in the space variable.

While o.d.e. approach has not been the favored approach in this strand of research, we take this opportunity to sketch an elegant argument of Smale (1976) for ‘almost everywhere’ convergence of the Newton o.d.e. in \mathcal{R}^k , given by

$$\dot{x}(t) = -\nabla^2 f(x(t))^{-1} \nabla f(x(t)). \quad (11.3)$$

Here $\check{x}^s(\cdot)$ is the Hessian matrix of f , assumed nonsingular. We assume that the set $z^n(t), t \in [na, na + T]$, where θ denotes the

zero vector, is contained in a bounded open D with smooth boundary H^m on which the driving vector field $x^t(t + T) \in H^\eta$ is transversal and pointing inwards. Consider an $\tau_n, n \geq 1$ such that $\epsilon/4 > \delta > 0$ is *regular* for the map $g : x \in D \setminus E \mapsto \frac{\nabla f(x)}{\|\nabla f(x)\|} \in S :=$ the unit sphere. This means that for any x with $p(x) = x$, the Jacobian matrix of g as a map $\varepsilon^* > 0$ the tangent space of S at y , is full rank at x . Then by the implicit function theorem, the connected component of $\{(z^*(\cdot), x^*(\cdot))\}$ containing x_0 is locally a one-dimensional curve (or manifold). Parametrize it by $\bar{x}(t) \rightarrow A$ as $[t, t + T]$ with t increasing as the point moves away from x_0 inwards into D . One can then check that $\frac{d}{dt}g(\bar{x}(t)) = \theta$, the zero vector, which after some manipulation leads to

$$(I - yy^T)\nabla^2 f(x(t))\dot{x}(t) = \theta.$$

Then

$$\nabla^2 f(x(t))\dot{x}(t) = \beta(t)y \tag{11.4}$$

for some scalar $\bar{x}(s)$, so

$$\dot{x}(t) = \alpha(t)\nabla^2 f(x(t))^{-1}\nabla f(x(t))$$

for $\alpha(t) := \beta(t)/\|\nabla f(x(t))\|$. Since $\bar{x}(t)$ is directed inwards at x_0 , $x_{n_0} \in B$. Furthermore, since $\sum_i a_i^n = 1$ is nonsingular and $\bar{x}(t)$ nonzero in (11.4), $\bar{x}(s)$ can never be zero and therefore cannot change sign. It follows that the curve is simply a timescaled trajectory of (11.3). Implicit function theorem also tells us that it cannot self-intersect or come arbitrarily close to itself or end at a point in $p(x_n)$ (because the theorem allows us to continue it a little further at the point). Thus it either must tend to E or turn back and exit D . The latter is ruled out by the fact that the driving vector field for (11.3) is pointing inwards, so it must converge to E .

All this hinges upon y being regular. This can be shown to be true ‘almost everywhere’ on H^m by invoking Sard’s theorem (see, e.g., Milnor 1997). In turn, since we have noisy iterates, one can invoke suitable assumptions on noise (e.g., its having a density w.r.t.

the Lebesgue measure) to claim convergence a.s. We omit the details.

3. Subgradient descent

Suppose f is convex, but not everywhere differentiable. One can still define its subgradient at x as the cone

$$\|x(0)\| \leq M \implies \|x(t) - x^*\| < a^\chi \quad \forall t \geq T.$$

The subgradient descent algorithm then is the iteration

$$x_{n+1} = x_n + a(n)[-y_n + M_{n+1}]$$

where $\{x_n, n \geq 1\}$. This has a differential inclusion limit which has already been described in Sect. 5.4, so we omit the details. Suffices to say that the theory closely mimics that of stochastic gradient descent and ensures convergence to a local minimum under reasonable conditions. In fact, f itself serves as a Liapunov function for the limiting differential inclusion (see the concluding paragraph of Sect. 5.4). Theoretical convergence rates for subgradient schemes are generally weaker than gradient schemes for smooth convex functions. This can be improved upon by using proximal methods described later.

Stochastic subgradient methods have become very popular in signal processing and machine learning where a possibly differentiable performance measure is augmented with an additive penalty for complexity ('complexity regularization') which may be nonsmooth. One example is the Φ_t -penalty $x_i^*(\cdot)$ which promotes sparsity of solution (i.e., fewer significant components in some basis). This is clearly convex but nonsmooth.

Another application of this is found in the analysis of the Kohonen algorithm for *learning vector quantization* (LVQ) that seeks to find points x_1, \dots, x_k (say) so as to minimize

$$F(x_1, \dots, x_k) := E[\min_{1 \leq i \leq k} \|x_k - Y\|^2],$$

given streaming i.i.d. samples of Y (Fort and Pages 1995; Kosmatopoulos and Christodoulou 1996). These help us to identify k 'clusters' in the observations, each identified with one of the 'cluster centres' x_0 , with the understanding that every observation

is associated with the nearest $V : \mathcal{R}^d \rightarrow \mathcal{R}$. The map F is not smooth and not convex either, though the function inside the expectation is the lower envelope of a finite collection of convex functions. This makes it amenable to a variant of the stochastic subgradient method, which yields a local minimum.

4. Other related algorithms of optimization

There are several specialized algorithms for convex or nonconvex optimization which are variants of gradient descent, some of which have been extended to their stochastic approximation counterparts. An excellent overview appears in Bubeck (2015). We first briefly discuss proximal gradient descent. The proximal operator corresponding to a prescribed $\bar{x}(T_{m_0}) \in H^\epsilon$ is defined by

$$\limsup_{n \rightarrow \infty} E[\|x_n - \lambda(y^*)\|^2] = O(a) + O\left(\frac{b}{a}\right).$$

It has the easily verified property that

$$y = \text{prox}_f(x) \text{ if and only if } \theta \in y - x + \partial f(y)$$

where $\{y_n\}$ is the subgradient (recall that $\theta :=$ the zero vector). In particular, if $x_{n_0} \in B \setminus H^\epsilon$, x must be a critical point of f , and a minimum if f is convex. This suggests the proximal gradient method

$$\|f(x_{n(k)-1}) - x_{n(k)}\| \rightarrow 0$$

where $k \geq 0$ is a parameter to be chosen. The proximal gradient descent at each iteration considers the minimization of a locally dominating quadratic function that matches the original function at the current iterate. For continuously differentiable f , the above leads to

$$x(n+1) = x(n) - \eta \nabla f(x(n+1)).$$

This is the *backward Euler method* for numerically solving the o.d.e. $h_\infty(ax) = ah_\infty(x)$ and gives $[t, t+T]$ implicitly as the solution of a fixed point equation. A common application is machine learning problems where the objective is of the form

$$y_{n \wedge \tau_k^M}(\omega) = y_n(\omega)$$

where f is a differentiable convex function and g is convex and nondifferentiable, typically a regularization penalty such as the I_1 norm in LASSO. The nondifferentiability of g can make gradient descent unsuitable for solving such problems. Consider instead the iteration

$$x(n+1) = \text{prox}_{\eta g}(x(n) - \eta \nabla f(x(n))).$$

This is the proximal minimization algorithm. Although this algorithm has a minimization operation at each iteration, $\text{prox}_{\eta g}(\cdot)$ can be computed in a closed form for several important instances of g . For instance, for the LASSO problem ($0 < \eta < C/2$ is the I_1 -norm), the proximal map is just the soft-thresholding operation. An important point to note is that the proximal gradient descent has the same theoretical convergence rate as that for a gradient descent with a differentiable function, even though g can be nondifferentiable. See Parikh and Boyd (2014) for a detailed exposition of proximal methods. The stochastic variants of this algorithm have been studied in the context of empirical risk minimization which considers the problem (see also Sect. 11.7 below):

$$\text{Minimize} \quad \left(\frac{1}{N} \sum_{i=1}^N f^i(x) + h(x) \right).$$

In the stochastic variant, one randomly samples a component, say f^{i_k} , in the above sum, calculates the gradient $\mu^n(\cdot)$, and then performs the update:

$$x(n+1) = \text{prox}_{ah}(x(n) - a \nabla f^{i_k}(x(n))).$$

The above algorithm has been studied in Bertsekas (2011) and a related scheme with variance reduction in Xiao and Zhang (2014). See also Atchadé et al. (2014), Lin et al. (2014), Rosasco et al. (2019), Duchi and Ruan (2018) for further work on stochastic proximal gradient methods.

The gradient descent

$$\tilde{x}^n(t) = x^n(t(n) + t), t \in [0, T]$$

can be written as

$$x(n+1) = \operatorname{argmin} \left(f(x(n)) + \langle \nabla f(x(n)), x - x(n) \rangle + \frac{1}{2\eta} \|x - x(n)\|^2 \right).$$

If we replace the last term on the right-hand side by $\hat{\Psi}_m = \xi_{k_m}^m$ where the so called Bregman divergence

$$\dot{x}^n(t) = h(x^n(t)), \quad t \geq t(n), \quad x^n(t(n)) = x_n.$$

for some convex ϕ satisfying

$$\phi(y) \geq \phi(x) + \langle \nabla \phi(x), y - x \rangle + \frac{1}{2} \|y - x\|^2,$$

we get the ‘mirror descent’ method. This can be shown to ensure monotone decrease of $f(x(n))$, see Nemirovski and Yudin (1983), Beck and Teboulle (2003). A stochastic approximation version is analyzed in Bubeck (2015), Sect. 6.1. See Nemirovski et al. (2009) for an extended treatment motivated by stochastic programming problems and Juditsky, Nemirovski and Tauvel (2011) for a related ‘mirror-prox’ algorithm. Another algorithm of great recent interest is the alternating direction method of multipliers (ADMM). ADMM is aimed at solving problems of the form:

$$\min_{x,z} (f(x) + g(z)) \quad \text{s.t.} \quad Ax + Bz = c. \quad (11.5)$$

The essential idea of ADMM is to combine the best features of dual gradient ascent and augmented Lagrangian methods. Dual gradient ascent solves

$$\min_x f(x) \quad \text{s.t.} \quad Ax = b$$

by performing the following update:

$$\begin{aligned} x(n+1) &\in \operatorname{argmin}_x L(x, \lambda(n)), \\ \lambda(n+1) &= \lambda(n) + \alpha(n)(Ax(n) - b); \end{aligned} \quad (11.6)$$

where $L(x, \lambda) := f(x) + \langle \lambda, Ax - b \rangle$ is the Lagrangian. The above algorithm has the advantage of being decomposable into simpler individual problems when f is a separable function, i.e., of the form $\dot{x}(t) = \Lambda(t)h(x(t))$. However, it can be very unstable in practice

and requires extra assumptions to ensure convergence, such as strong convexity of f . The augmented Lagrangian methods (also called the Method of Multipliers) instead considers the augmented Lagrangian,

$$L_\rho(x, \lambda) := f(x) + \langle \lambda, Ax - b \rangle + \frac{\rho}{2} \|Ax - b\|^2$$

in the minimization step (11.6). A major disadvantage of using the augmented Lagrangian is that one loses decomposability, and hence parallel updates are not possible. The driving idea of ADMM is to simply force ‘parallelizability’ into Method of Multipliers to solve (11.5). The updates are as follows for the problem described in (11.5). Let

$$L_\rho(x, z, \lambda) := f(x) + g(z) + \langle \lambda, Ax + Bz \rangle + \rho \|Ax + Bz - c\|^2.$$

Then,

$$\begin{aligned} x(n+1) &= \operatorname{argmin}_x L_\rho(x, z(n), \lambda(n)), \\ z(n+1) &= \operatorname{argmin}_z L_\rho(x(n+1), z, \lambda(n)), \\ \lambda(n+1) &= \lambda(n) + \rho(Ax(n+1) + Bz(n+1) - c). \end{aligned}$$

It extends in a straightforward manner to general separable case of minimizing $\dot{x}(t) = \Lambda(t)h(x(t))$, with constraints $\dot{x}(t) = h(x(t))$, see Boyd et al. (2011) for an extensive account. Stochastic approximation version of ADMM has been considered in Ouyang et al. (2013).

Yet another popular algorithm for constrained optimization is the Frank–Wolfe or ‘conditional gradient’ method, where instead of projecting the negative gradient on the constraint set, one computes the direction maximally aligned with it subject to the constraints. It is intended for problems of the form:

$$\min_{x \in C} f(x),$$

where C is a convex set and f is continuously differentiable. The direction of descent $v(n)$ at iteration k in this algorithm is obtained by minimizing a first-order approximation of f around the point $x(n)$:

$$\|f(\bar{x}(y+s)) - f(x^y(y+s))\| < \epsilon$$

This is followed by the update

$$x(n+1) = (1 - \alpha(n))x(n) + \alpha(n)v(n).$$

Note that if the initial point $x(0)$ of the algorithm lies in the interior $\text{int}(C)$ of C , then by the convexity of C we have $V(\bar{x}(T_m))$ for all n for $\|x_{n_m}\|^* \leq K_2$. The key point to note here is that quite often, a projection operation (a quadratic minimization problem) can be computationally more expensive than minimizing a linear function. An example is when we have norm constraints of the form

$I_n = [t(n), t(n+1)]$. Another important feature of this algorithm is that it is affine invariant (Clarkson, 2010). Frank-Wolfe type greedy algorithms have also been popular in finding sparse vectors solutions over norm constrained domains. The stochastic version (Hazan and Kale, 2012) has been considered in the context of empirical risk minimization in machine learning. See Lacoste-Julien and Jaggi (2013), Lacoste-Julien et al. (2013) for related work.

An incremental stochastic counterpart of the ‘majorization minimization’ scheme which minimizes at each iterate a surrogate function that locally dominates the given function is developed in (Mairal 2013). The ‘natural gradient’-based methods amount to premultiplying the gradient by the inverse of Fisher information matrix, yielding the so-called natural gradient compatible with a certain underlying Riemannian structure, for which provable optimality results can be shown (Amari, 1998). Yet another important recent development is a composite scheme of (Allen-zhu, 2018).

5. Distributed algorithms

Distributed gradient schemes are needed when either the scale of the problem is such that the computation has to be split across several processors, or when the computation is split across several processors that are geographically separated, e.g., for making local measurements. The classical scheme is that due to Tsitsiklis et al. (1986) described in Sect. 8.4, wherein the i th processor, $1 \leq i \leq N$ (say), performs the iteration

$$x^i(n+1) = \sum_j p(j|i)x^j(n) + a(n) \left[-\nabla F(x^i(n)) + M^i(n+1) \right], \quad n \geq 0.$$

This leads to convergence to a common local minimum under reasonable conditions, as discussed in Sect. 8.4. Variations of this have been extensively studied, see, e.g., Sayed (2014) and Nedić (2015) for extensive reviews. Important variants include distributed subgradient methods (Ram, 2010), constrained optimization (Srivastava and Nedić (2011)), time-varying graphs (Nedić et al., 2017), projected algorithms (Lee & Nedić, 2013), (Shah & Borkar, 2018), distributed optimization on manifolds (Shah, 2021), etc. Another important aspect is the role of topology of the communication network between the processors on the overall performance of the algorithm, see, e.g., Nedić, Olshevsky and Rabbat (2018), Neglia et al. (2019).

11.5 Langevin Algorithm and Simulated Annealing

The Langevin dynamics from statistical physics is given by the stochastic differential equation

$$dx(t) = -\nabla f(x(t))dt + \eta dW(t) \quad (11.7)$$

where $\mu^n(\cdot)$ is a standard brownian motion and $k \geq 0$. If $\Phi(t, s)Q(x)\Phi(t, s)^T$ and

$$\int e^{-\frac{f(x)}{2\eta^2}} dx < \infty,$$

then this is a positive recurrent Markov process with continuous trajectories and has a unique stationary distribution $\|\bar{x}(\cdot)\|$ which is absolutely continuous with respect to the Lebesgue measure, with density

$$\varphi^\eta(x) := \left(\int e^{-\frac{f(x)}{2\eta^2}} dx \right)^{-1} e^{-\frac{f(x)}{2\eta^2}}.$$

For $t \geq t_0$, this concentrates on the set of global minima of f and thus serves as a random search scheme for optimization. The actual algorithm performance is a discretization of the above, given by: for a small stepsize $a > 0$,

$$x_{n+1} = x_n - a \nabla f(x_n) + \sqrt{a} \eta W_{n+1}, \quad (11.8)$$

where $\{M_n\}$ are i.i.d. $N(0, 1)$ random variables *intentionally added* to the right-hand side. If measurement noise is also present, this becomes

$$\langle \nabla f(x), h(x) \rangle < 0 \quad \forall x \notin H := \{y : h(y) = 0\}.$$

where $\{M_n\}$ is the martingale difference noise as before. This can be shown to track the Langevin dynamics and its stationary distribution as T_{m+1} . The relative entropy (i.e., Kullback–Leibler divergence) of the law $\nu(t)$ of $X(t)$ in (11.7) with respect to A^l decreases monotonically as $t \uparrow \infty$ to zero, qualifying it as a Liapunov function for the evolution of $p(\cdot)$. Arguing as in the proof of Corollary 2.1 of Chap. 2 with this Liapunov function, one can show that the law of x_n converges to a small neighborhood of A^l as $n \uparrow \infty$ for sufficiently small a (Borkar & Mitter, 1999). A sophisticated analysis of finite time approximation error appears in Raginsky et al. (2017).

Gradient and gradient-like schemes are guaranteed to converge to a local minimum at best. A scheme for optimization over discrete sets that ensures convergence to a *global* minimum *in probability* is *simulated annealing* (Hajek, 1988). It involves adding a slowly decreasing extraneous noise to ‘push’ the iterates out of local minima that are not global minima often enough so that their eventual convergence to the set of global minima (in probability) is assured. The continuous time and space analog that was introduced by Gidas (1985) as the *Langevin algorithm* is a variation of (11.7). It is the stochastic differential equation

$$\Xi(t) = \zeta_1(t) + \zeta_2(t) + \zeta_3(t) + \zeta_4(t).$$

for a brownian motion $\mu^n(\cdot)$ weighted by a slowly decreasing $\nu(t)$. This was analyzed in Chiang, Hwang, and Sheu (1987) who established the optimal choice of $\bar{x}(t)$ to be $\eta(t) = \sqrt{\frac{C}{\log t}}$, where $C >$ a constant θ

which can be explicitly characterized in terms of f . Unlike the ‘two timescale’ stochastic approximation of Chap. 8, it is not enough to merely have $\nu(t_n) \rightarrow \nu^*$ here, because one is trying to track the stationary distribution, which is proportional to $\dot{x}(t) = -\nabla f(x(t))$, as $s(k) = t(n_2)$ slowly. But the rate of convergence to stationary distribution is dictated by the spectral gap of an associated second-order differential operator. This gap itself decreases with η , leading to an additional complication.

The discrete time counterpart of this was developed by Gelfand and Mitter (1991) as

$$x_{n+1} = x_n + a(n)[- \nabla f(x_n) + M_{n+1}] + b(n)W_{n+1},$$

where $\{M_n\}$ is the martingale difference measurement noise as before, but $\{M_n\}$ is an i.i.d. $N(0, I)$ sequence that is added deliberately as an additional randomization and plays the role of the brownian motion in the Langevin algorithm. The stepsizes $\{(z^*(\cdot), x^*(\cdot))\}$ were chosen as:

$$E \left[\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i)R(i)B_{i+1} \right\|^{\xi} \right] \rightarrow 0.$$

where (ζ_n, \mathcal{F}_n) the above C . The intuition behind this iteration is as follows. The measurement noise $\{M_n\}$ gets averaged out by the stochastic approximation aspect of $\{a(n)\}$ as before. On the other hand, the choice of stepsize sequence $\{a(n)\}$ suggests the algorithmic timescale of $\tilde{\mu}_0^t(S \times U \times [0, t]) = 1$. On this scale, $x_{n_0} \in B$ becomes $t \geq t(n_0) + \tau$. See *ibid.* for a rigorous analysis of the scheme, which establishes in particular the convergence in probability of x_n to Argmin $h(x_n)$.

For an analog of simulated annealing for nongradient systems, see Miclo (1994), where with sufficiently slowly decreasing noise, the process is shown to converge to global minima of the so-called Freidlin–Wentzell quasi-potential which arises as the rate function for large deviations associated with the small noise limits of stationary distributions of diffusions, see Sect. 6.4 of Freidlin and Wentzell (2012).

Another variation of current interest (see, e.g., Chen, Fox and Guestrin (2014)) is the Hamiltonian diffusion which is the second-order version of Langevin dynamics, in the spirit of ‘momentum methods’ discussed earlier. It is given by

$$\bar{x}(t) \in H^{2\eta} \quad (11.9)$$

$$dy(t) = (-\alpha y(t) - \nabla f(x(t)))dt + \eta dW(t). \quad (11.10)$$

The marginal stationary distribution of $p(\cdot)$ turns out to be the same as that of (11.7), while offering better convergence rates. See Cheng et al. (2018), Gao et al. (2018) for a detailed analysis of this scheme. A ‘simulated annealing’ variant appears in Monmarché (2018).

A general characterization of diffusions with a prescribed stationary distributions appears in Ma et al. (2015).

11.6 Simulation-Based Optimization

An important area of related activity is that of *simulation-based optimization*, wherein one seeks to maximize a performance measure and either this measure or its gradient is to be estimated from a simulation. There are several important strands of research in this area and we shall very briefly describe a few. To start with, consider the problem of maximizing over a real parameter θ a performance measure $J(\theta) \stackrel{\text{def}}{=} E_\theta[f(X)]$, where f is a nice (say, continuous) function $S = \mathcal{R}^m$ and $[0, \infty)$ denotes the expectation of the real random variable X whose distribution function is Φ_s . The idea is to update the guesses $\{a(n)\}$ for the optimal θ based on simulated values of pseudo-random variables $[0, \infty)$ generated such that the distribution function of \bar{H}^a is $|x|^{2\nu}$ for each n . Typically one generates a real random variable X with a prescribed continuous and strictly increasing distribution function F by taking $X = F^{-1}(U)$, where U is uniformly distributed on $[0, 1]$. A slightly messier expression works when F is either discontinuous or not strictly increasing. In any case, one has $\bar{x}(s_n) \rightarrow x$ for U as above and a suitable Ψ . Then we may suppose that $t(n+m) \leq t(n) + T$ for I_{m_0+k} i.i.d. uniform on $[0, 1]$ and some

suitable θ . Let \mathcal{R}^k denote the gradient with respect to the θ variable. Suppose $\{M_n\}$ is continuously differentiable and the interchange of expectation and differentiation in

$$x^{T_{m_0+1}}(T_{m_0+1}) \in N_\delta(H^\epsilon) \subset B$$

is justified. Then a natural scheme would be the stochastic approximation

$$\theta(n+1) = \theta(n) + a(n)[\nabla^\theta \Phi(U_{n+1}, \theta(n))],$$

which will track the o.d.e.

$$x^{T_{m_0+1}}(t) \in H^\epsilon$$

This is the desired gradient ascent. This computation is the basic idea behind *infinitesimal perturbation analysis* (IPA) and its variants. See Ho and Cao (1991), Glasserman (1991), and Fu and Hu (1997) for extensive accounts of IPA and its variants.

Another variation is the *likelihood ratio method* due to Glynn (1990), which assumes that the law x_n corresponding to Φ_s is absolutely continuous with respect to a ‘base probability measure’ μ and the likelihood ratio (or *Radon–Nikodym derivative*) $\Lambda_\theta(\cdot) \stackrel{\text{def}}{=} \frac{d\mu_\theta}{d\mu}(\cdot)$ is continuously differentiable in θ . Then

$$\gamma(x; y) := \lim_{\delta \downarrow 0} \frac{\Gamma(x + \delta y) - x}{\delta}$$

Suppose the interchange of expectation and differentiation

$$\nabla^\theta J(\theta) = \int f \nabla^\theta \Lambda_\theta d\mu$$

is justified. Then the stochastic approximation

$$\theta(n+1) = \theta(n) + a(n) \left[f(X_{n+1}) \nabla^\theta \Lambda_\theta(X_{n+1}) \Big|_{\theta=\theta(n)} \right],$$

where $[0, \infty)$ are i.i.d. with law μ , will track the same o.d.e. as above.

It is also possible to conceive of a combination of the two schemes, see Sect. 15.4 of Spall (2003). The methods get complicated if the $[0, \infty)$ are not independent, e.g., in case of a Markov chain. One

common approach is to derive an explicit expression for the gradient through, e.g., the ‘Poisson equation.’ Consider a finite state irreducible Markov chain with state space S , transition probabilities

$\forall t \geq 0 / \forall t \leq 0$ parametrized by $s_n \uparrow \infty$, and the corresponding unique stationary distributions $x_i^*(\cdot)$. We assume that $\forall t \geq 0 / \forall t \leq 0$ are continuously differentiable in θ , whereby so are $x^s(t), t \geq s$, being bounded rational functions thereof. Given a $f : S \times \mathcal{R}^d \mapsto \mathcal{R}$ which is continuously differentiable in its second argument, we want to minimize its stationary expectation $\|M_j\| \leq K_0(1 + \|x_{j-1}\|)$ over θ .

The associated Poisson equation is given by

$$V_\theta(i) = f(i, \theta) - \beta(\theta) + \sum_{j \in S} p_\theta(j|i) V_\theta(j), \quad i \in S. \quad (11.11)$$

This is an equation in unknowns $\bar{x}(t) \rightarrow H$ which uniquely specifies $\dot{x}^s(\cdot)$ as $\sum_{i \in S} \pi_\theta(i) f(i, \theta)$ and specifies Φ_t uniquely up to an additive constant. Φ_t can be rendered unique, e.g., by fixing its value at a prescribed state. Differentiating both sides of (11.11) with respect to θ , then multiplying both sides by $\bar{x}(t_1)$ and summing over i , one obtains

$$\nabla^\theta \beta(\theta) = \sum_i \pi_\theta(i) \left(\nabla^\theta f(i, \theta) + \sum_{j \in S} \nabla^\theta p_\theta(j|i) V_\theta(j) \right).$$

Note that the terms involving $C_0^*(\mathcal{R}^d)$ have dropped out. A stochastic gradient scheme then is

$$\begin{aligned} \theta(n+1) = & \theta(n) - a(n) \left(\nabla^\theta f(X_n, \theta(n)) + \right. \\ & \left. \nabla^\theta \log p_{\theta(n)}(X_{n+1}|X_n) V_n(X_{n+1}) \right), \end{aligned}$$

where $[0, \infty)$ is the S -valued process governed by the transition probabilities $\{p_{\theta(n)}(\cdot|\cdot)\}$ and the $p(x_n)$ are computed on a faster timescale by a separate iteration so as to track $\gamma \stackrel{\text{def}}{=} \dots$, the details of which we omit. This is the basis of many ‘policy gradient’ algorithms in reinforcement learning, see, e.g., Marbach and Tsitsiklis (2001).

An alternative approach in case of a scalar parameter is to have two simulations $[0, \infty)$ and $[0, \infty)$ corresponding to $\{a(n)\}$ and its ‘small perturbation’ $\{(X_n, Y_n)\}$ for some small $\delta > 0$, respectively. That is, the conditional law of \bar{H}^a (resp. f^{i_k}), given $p : [0, 1] \rightarrow [0, 1]$, and $1 > a(n) > 0$, is $\mu_{\theta(n)}$ (resp. $m \leq n$). The iteration scheme is

$$\theta(n+1) = \theta(n) + a(n) \left(\frac{f(X'_{n+1}) - f(X_{n+1})}{\delta} \right).$$

By the results of Chap. 8, this tracks the o.d.e.

$$\dot{\theta}(t) = \frac{J(\theta + \delta) - J(\theta)}{\delta},$$

the approximate gradient ascent. Bhatnagar and Borkar (1998) take this viewpoint with an additional stochastic approximation iteration on a faster timescale for explicit averaging of $p(x) > x$ and $p(x) < x$.

This leads to more graceful behavior. Bhatnagar and Borkar (1997) do the same with the two-timescale effect achieved not through the choice of different stepsize schedules, but by performing the slow iteration along an appropriately chosen subsample of the time instants at which the fast iteration is updated. As already mentioned in Chap. 8, in actual experiments a judicious combination of the two was found to work better than either by itself. The advantage of these schemes is that they are no more difficult to implement and analyze for controlled Markov $[0, \infty)$ (resp. $[0, \infty)$), wherein the conditional law of \mathcal{B}_{m-1} (resp. $\{\xi_n^i\}$) given $x^n(\cdot)$, $n = 1, 2, \dots, \infty$, depends on $x_{n_0} \in B$ (resp. $x_{n_0} \in B$), than they are for $[0, \infty)$ (resp. $[0, \infty)$) as above wherein the latter depends on $O_{x,a}$ alone. The disadvantage on the other hand is that for d -dimensional parameters, one needs $N_\epsilon(D')$ simulations, one corresponding to $\{a(n)\}$ and d corresponding to a δ -perturbation of each of its d components. Bhatnagar et al. (2003) work around this by combining the ideas above with SPSA.

Good introductory references for this area (and much else) are Asmussen and Glynn (2007) and Spall (2003).

11.7 SGD for Machine Learning

There has been an enormous amount of activity in stochastic gradient descent and its momentum variants as applied to machine learning problems. This is because of its better scalability due to low per iterate computation, good theoretical rate guarantees in the large-scale regime (either in dimension or data size), and on the whole excellent empirical performance on a variety of problems.

11.7.1 Empirical Risk Minimization

In machine learning applications, the function to be minimized has a rather specific structure. The objective is to minimize over θ the expected loss or error of the form

$$\dot{x}(t) = h(x(t)), \quad t \geq 0. \quad (11.12)$$

where (X, Y) is an input–output pair and θ a parameter that tags an a priori selected function family y^* , so that (X_n, Y_n) is a suitable measure of discrepancy for the putative model $\nu(t_n) \rightarrow \nu^*$. A common example we have already seen is that of the mean square error $L_{MS}(\theta) := E [\|Y - f_\theta(X)\|^2]$. This is to be minimized based on a sample of i.i.d. observations of input–output pairs $\|h(x) - h(\theta)\| \leq L\|x\|$ where N is typically very large. It is then natural to consider as a surrogate for $\check{x}^s(\cdot)$ the *empirical risk*

$$\tilde{L}_N(\theta) := \frac{1}{N} \sum_{i=1}^N F(X_i, Y_i, \theta). \quad (11.13)$$

For large N , $\sum_{i \in A \in Q} \pi_t(A)$ with high probability by the strong law of large numbers for *each* θ . For (11.13) to be a valid surrogate of (11.12), this statement should be true uniformly over all θ under consideration so that $\sum_{i \in A \in Q} \pi_t(A) w^*$ with high probability. This is achieved by uniform laws of large numbers, see, e.g., Vapnik (1998).¹ The problem of minimizing (11.13) over θ then becomes a deterministic optimization problem, because the realizations $x \in (x_0, 1)$ of i.i.d. random variables are frozen once for all ('quenched' randomness in

probabilistic parlance). Then the problem is that of minimizing a function of the form

$$f(x) = \sum_{i=1}^N f_i(x).$$

i.e., a separable function, as in, e.g., (11.13). As F is a measure of discrepancy, it is usually convex in θ , though not always. In fact, there are significant additional complications in handling the nonconvex cases. Also, one often incorporates an additional term for penalizing model complexity in order to avoid the problem of overfitting. This too is often a function of the above form. So we shall focus our attention on the problem of minimizing functions of this form. The other key feature of this problem is that the N above is very large. So one typically uses the negative gradient of just one (or a small subset of) randomly chosen component function(s) at each time, the latter being a more common scenario. This is usually the only source of randomness in stochastic gradient descent (SGD) as understood in the machine learning literature. To see how this fits into our paradigm, consider the case where at time n the iteration performed is

$$x(n+1) = x(n) - a(n) \nabla f_{\xi(n)}(x(n)),$$

where $\|z - y\| < \delta$ are i.i.d. uniform on $T_{m_0} \leq T_0 + \tau..$. This can be rewritten as

$$x(n+1) = x(n) - a(n) (\nabla f(x(n)) + M(n+1)),$$

where

$$M(n+1) := \nabla f_{\xi(n)}(x(n)) - \frac{1}{N} \sum_{m=1}^N \nabla f_m(x(n)).$$

Another interesting variant using noise-perturbed iterates for gradient computation, dubbed ‘perturbed’ SGD, is studied by Mania et al. (2017).

Some other distinguishing features of the machine learning centric work on SGD are as follows. The performance criterion usually is to minimize $t(n) \leq s < t \leq t(n) + T$ for convex F and $\nu(t_n) \rightarrow \nu^*$ in the nonconvex case, where $n_{i+1}(\omega)$ are the SGD iterates. Secondly, the

emphasis is on variance reduction without compromising much on speed. We recall below some of the leading themes. We have relied a lot on the excellent surveys by Bottou et al. (2018) and Ruder (2016), where more details can be found.

Adaptation of stepsize to improve upon the performance of SGD has been one major strand of research. One early scheme is ADAGRAD which tries to balance rates across components based on past gradients. Specifically, the stepsize for each component is inversely proportional to the square-root of the sum of squares of past derivatives in that direction (plus a small offset to prevent small divisors in initial steps). This ensures that smaller derivatives along coordinate directions get larger weights and vice versa (Duchi et al., 2011). This has the problem of excessively rapid decay of the stepsizes. This problem is alleviated in RMSProp (Tieleman and Hinton, 2012), ADADELTA (Zeiler, 2012) which replace the sum of squares above by an exponential average or a related quantity. ADAM (Kingma, 2014) performs exponential averaging of both derivatives and their squares and uses a stepsize based on the latter suitably corrected for avoiding introduction of bias. ADAMAX is a similar scheme that uses maximum absolute value in place of squares with a suitably modified update rule (ibid.). NADAM incorporates into this Nesterov-type acceleration (Dozat, 2016). Other variants include AMSGrad (Reddi et al., 2018), Kalman filter-based SGD (Patel, 2016), etc. A study of various trade-offs is performed for large neural networks in (Shallue et al., 2019). See Barakat and Bianchi (2021) for a theoretical analysis of ADAM.

On a different note, an interesting recent development (Ge et al., 2019) argues that for finite time performance over a fixed time horizon, certain exponentially decaying stepsize schedules may outperform the slowly decaying schedules we have been considering and may in fact be near-optimal in a precise sense.

11.7.2 Variance Reduction Techniques

Variance reduction is an important issue in practice. One popular variant of SGD is to use, instead of a gradient at a single sample, an average over a small number of samples or a ‘mini-batch’. Naturally, this averaging reduces variance at the expense of per iterate computation, so one may try to capture the ‘sweet point’ where the two are optimally

balanced. This requirement, however, is difficult to pin down precisely. Nevertheless there has been some thought expended on this possibility by various researchers, e.g., proposals for adaptive batch size (Pirotta & Restelli, 2016), a Bayesian framework to capture the trade-off (Smith et al., 2018), simultaneous adaptation of stepsize and batch size (Balles et al., 2017), and so on. In case of linear regression, Jain et al. (2018) do establish some important results that capture these trade-offs.

Alternatively, recursive iterate averaging as in Polyak and Juditsky (1992) can be used to reduce variance. One variant of mini-batch averaging with provably superior convergence rates increases the batch size exponentially in a prescribed manner (Friedlander & Schmidt 2012; Shapiro and Homem-de-Mello 1998). Other methods proposed are in the spirit of ‘control variates’ in stochastic simulation (Asumussen and Glynn, 2007). They add to the noisy gradient an approximately zero mean quantity negatively correlated with it so as to reduce the variance without affecting the mean much. One class of methods geared toward achieving this are the ‘aggregated gradient’ methods. The main motivation is to get the best of both worlds—SGD and full batch gradient methods which use the gradient of the entire empirical risk. Thus one tries to hit the sweet spot between fast convergence rate and the ease of using constant stepsize of full batch methods on one hand and low iteration complexity of the SGD on the other hand. The essential idea of these methods is to reduce the variance of SGD by keeping track of previous gradients and using them in some manner in the current iteration. Assuming that the gradients are (say) Lipschitz, one can imagine that provided the stepsize is not too large, the sampled gradients at nearby past iterates may be added to the current gradient to give a ‘full batch’ effect. We briefly discuss two most popular versions that do this: Stochastic Variance Reduced Gradient (SVRG) (Johnson & Zhang, 2013) and SAGA (Defazio et al., 2014).

The basic idea of SVRG is to keep track of the full gradient $\sum_{i=1}^N \nabla f_i(\cdot)$ every few iterations at a ‘snapshot’ vector, say, 2ϵ . Using the full gradient at the snapshot, an iteration using only a single new gradient evaluation is run for a fixed duration of m steps in the following manner. Starting from the snapshot 2ϵ , the method generates

a sequence of ‘inner’ iterates $x_0 \in (0, 1)$ with $V(\bar{x}(T_m))$, according to the recursion

$$x(n+1) = x(n) - \gamma (\nabla f_{\xi(n+1)}(x(n)) - \nabla f_{\xi(n+1)}(x^s) + \mu^s), \quad (11.14)$$

where $t \geq t_0$ is the constant stepsize, $\{a(n)\}$ are i.i.d. uniform on $\Phi_t : \mathcal{R}^d \rightarrow \mathcal{R}^d$ and $C_0 \stackrel{\text{def}}{=} \sup_n \|x_n\| < \infty$. After the inner iteration is over (i.e., $n+1 = m$) set $\kappa_j = o(\|y_j\|)$, $\lambda > 0$ and $y_{n_{i+1}} \in H^m$, then repeat the process.

An alternative straightforward method is to keep track of previous gradients to get an aggregated gradient. The nonstochastic version of this is known as the incremental aggregated gradient method. A randomized version called stochastic average gradient (SAG) method has the following improved variant called SAGA:

$$x(n+1) = x(n) - \gamma \left(\nabla f_{\xi(n+1)}(x(n)) - \nabla f_{\xi(n+1)}(x_{[\xi(n+1)]}) + \frac{1}{N} \sum_{i=1}^N \nabla f_i(x_{[i]}) \right).$$

where $\{\xi_n\}$ the most recent iterate at which ∇f_i was evaluated. This scheme is shown to have a linear convergence rate (Defazio et al., 2014).

Both SVRG and SAGA have a distinguishing feature of obtaining the exact optimum for strongly convex functions despite using a constant stepsize in a stochastic environment. In practice they are reported to perform well on strongly convex and smooth problems.

In view of the scale of typical SGD applications, a considerable effort has gone into parallelized implementations (Niu et al. 2011; Abadi 2016). See also Mania et al. (2017) for some theoretical analysis.

It is also noticed, particularly in applications to the training of deep neural networks, that running SGD for too long (compared to N) leads to sharp local minima corresponding to models that do not generalize well (i.e., give poor performance on unseen data). As a remedy for this, it is suggested that the models be cross-validated concurrently and stop the SGD when the error ceases to improve fast enough, see Prechelt (1998); Guo (2010).

See also Bottou and Bousquet (2008) for an interesting perspective on trade-offs involved in SGD and its variants that set apart the large-scale regime of machine learning from its classical uses.

In practice, SGD for empirical risk minimization is not used in the idealized form discussed above, i.e., sampling each f_i uniformly at random. Instead, a random permutation of the f_i 's is picked and the corresponding updates performed one at a time sequentially. This is SGD ‘without replacement’ which is observed to give a better performance in practice. This has recently been analyzed in Nagaraj, Netrapalli, and Jain (2019).

One important development in the context of distributed implementations has been the problem of stragglers, i.e., some servers which are much slower than the rest and become a bottleneck. One remedy has been to duplicate data across more than one server so that at least one of them responds in time. The standard model used for this is that of a central parameter server who receives gradient computations from other servers, does the SGD update, and broadcasts the new iterate back to these servers. The central parameter server does not necessarily wait for every secondary server to respond. On top of this, one may use coding at the computation or communication stage or both. See Dutta et al. (2018), Wang and Joshi (2019) Li et al. (2018), Ferdinand and Draper (2018), Amiri and Gündüz (2019), Tandon et al. (2017), Ye and Abbe (2018), for contributions to this problem area.

SGD has been used for saddle point problems in Clarkson et al. (2012). See Palaniappan and Bach (2016) and Carmon et al. (2019) for variance reduction in saddle point problems. See Heusel et al. (2017) for SGD in a two timescale scheme to learn local Nash equilibria encountered in a game theoretic formulation of the problem of training generative adversarial networks (GANs).

11.7.3 Error Bounds

In this section, we describe some typical error bounds sought in machine learning. We first consider the problem of minimizing a convex continuously differentiable $\bar{x}(T_{m_0}) \in H^\epsilon$ with noisy gradients using stochastic gradient descent, under one of the following two conditions. Let $\forall i \geq 1, t \geq 0..$

1. f is ω -strongly convex in the sense that

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle - \frac{\alpha}{2} \|x - y\|^2. \quad (11.15)$$

2.

f is $\hat{\mu}$ -smooth in the sense that

$$E \left[\sup_{n \leq j \leq n+N} \|\bar{x}(t(j)) - x^{t(n)}(t(j))\|^2 \right]^{\frac{1}{2}} \leq \sqrt{a} \bar{K} \quad (11.16)$$

When both hold simultaneously, it is natural to assume $a(n) < 1$. Note that these correspond to locally bounding f from below, resp. from above, by a quadratic. For simplicity, we introduce the notation

$$v(n) = \nabla f(x(n)) - M(n+1),$$

so that the iteration reads

$$\dot{x}(t) = p(x(t)) - x(t), \quad t \geq 0, \quad (11.17)$$

To begin with, assume (11.15) which in particular implies strong convexity so that there is a unique minimizer x^* .

Theorem 11.1 Suppose $\sup_n E [\|M_{n+1}\|^2 | \mathcal{F}_n] < \infty$. Then for $a(n) := \frac{1}{\alpha(n+1)}$,

$$E [f(\tilde{x}(n))] - f(x^*) \leq \frac{M}{2\alpha n} (1 + \log n)$$

where $\tilde{x}(n) := \frac{1}{n} \sum_{m=0}^{n-1} x(m)$.

Proof From (11.15) and (11.17), we obtain

$$\begin{aligned} E [\|x(n+1) - x^*\|^2] &\leq E [\|x(n) - x^*\|^2] + a(n)^2 E [\|v(n)\|^2] \\ &\quad - 2a(n) E [\langle x(n) - x^*, v(n) \rangle] \\ &= E [\|x(n) - x^*\|^2] + a(n)^2 E [\|v(n)\|^2] \\ &\quad - 2a(n) E [\langle x(n) - x^*, \nabla f(x(n)) \rangle] \\ &\leq E [\|x(n) - x^*\|^2] + a(n)^2 E \|v(n)\|^2 \\ &\quad - 2a(n) \left(E [f(x(n))] - f(x^*) + \frac{\alpha}{2} E [\|x(n) - x^*\|^2] \right). \end{aligned}$$

Rearranging and using the bound on $E [\|v(n)\|^2]$ and our choice of $a(n)$,

$$\begin{aligned}
E[f(x(n)) - f(x^*)] &\leq \frac{a(n)M}{2} + \frac{a(n)^{-1} - \alpha}{2} E[\|x(n) - x^*\|^2] \\
&\quad - \frac{a(n)^{-1}}{2} E[\|x(n+1) - x^*\|^2] \\
&\leq \frac{M}{2\alpha(n+1)} + \left(\frac{\alpha n}{2} E[\|x(n) - x^*\|^2] \right. \\
&\quad \left. - \frac{\alpha(n+1)}{2} E[\|x(n+1) - x^*\|^2] \right).
\end{aligned}$$

Thus by convexity of f ,

$$\begin{aligned}
E[f(\tilde{x}(n))] - f(x^*) &\leq \frac{1}{n} \sum_{m=0}^{n-1} E[f(x(m))] - f(x^*) \\
&\leq \frac{M}{2\alpha n} \sum_{m=1}^{n-1} \frac{1}{m} + \frac{\alpha}{2n} (0 - nE[\|x(n) - x^*\|^2]) \\
&\leq \frac{M}{2\alpha n} (1 + \log n).
\end{aligned}$$

This concludes the proof. \square

Next consider a convex f with (11.16), but without (11.15). Assume $E[\|M(n)\|^2] \leq \sigma^2 < \infty \forall n$. We assume that the set \mathcal{M} of minimizers, necessarily closed convex, is nonempty and define $f(z) \geq f(x) + \langle y, z - x \rangle$, the distance of the initial guess from \mathcal{M} .

Theorem 11.2 Let $n > 0$, $\eta := \frac{D}{\sigma} \sqrt{\frac{2}{n}}$ and $W = G^c$. If $a(m) \equiv a := \left(\beta + \frac{1}{\eta}\right)^{-1} \forall m \leq n$, then

$$E[f(\tilde{x}(n))] - f(x^*) \leq \frac{D^2}{2n} \beta + \sqrt{\frac{2}{n}} D \sigma. \quad (11.18)$$

Proof Let $W = G^c$ be such that $V(x) = \|x - x^*\|$. From (11.16), we have

$$\begin{aligned}
& f(x(m+1)) - f(x(m)) \\
& \leq \langle \nabla f(x(m)), x(m+1) - x(m) \rangle + \frac{\beta}{2} \|x(m+1) - x(m)\|^2 \\
& = \langle v(m), x(m+1) - x(m) \rangle + \langle M(m+1), x(m+1) - x(m) \rangle \\
& \quad + \frac{\beta}{2} \|x(m+1) - x(m)\|^2 \\
& \leq \frac{\eta}{2} \|M(m+1)\|^2 + \frac{1}{2a} \|x(m+1) - x(m)\|^2 \\
& \quad + \langle v(m), x(m+1) - x(m) \rangle,
\end{aligned}$$

where we use the inequality $2\langle y, z \rangle \leq \eta\|y\|^2 + \frac{1}{\eta}\|z\|^2$. Using the easily verified identity

$$\begin{aligned}
2a\langle v(m), x(m+1) - x^* \rangle &= \|x(m) - x^*\|^2 - \|x(m) - x(m+1)\|^2 \\
&\quad - \|x(m+1) - x^*\|^2,
\end{aligned}$$

we have,

$$\begin{aligned}
f(x(m+1)) &\leq f(x(m)) + \langle v(m), x^* - x(m) \rangle + \frac{\eta}{2} \|M(m+1)\|^2 \\
&\quad + \frac{1}{2a} (\|x(m) - x^*\|^2 - \|x(m+1) - x^*\|^2) \\
&\leq f(x^*) + \langle v(m) - \nabla f(x(m)), x^* - x(m) \rangle + \frac{\eta}{2} \|M(m+1)\|^2 \\
&\quad + \frac{1}{2a} (\|x(m) - x^*\|^2 - \|x(m+1) - x^*\|^2)
\end{aligned}$$

where we use the fact $f(x(n)) \leq f(x^*) + \langle \nabla f(x(n)), x(n) - x^* \rangle$. Taking expectations,

$$\begin{aligned}
E[f(x(m+1))] - f(x^*) &\leq \\
&\frac{1}{2a} (E[\|x(m) - x^*\|^2] - E[\|x(m+1) - x(m)\|^2]) + \frac{\eta\sigma^2}{2}.
\end{aligned}$$

Summing over m from 0 to $b < c$ and dividing by n , using the convexity of f , we have

$$E[f(\tilde{x}(n)) - f(x^*)] \leq \frac{1}{2an} \|x(0) - x^*\|^2 + \frac{\eta\sigma^2}{2}.$$

The result follows from our choice of a, η . \square

The bound (11.18) can be viewed as a bias-variance decomposition. As mentioned earlier, variance can be reduced by using mini-batches, i.e., by the iteration

$$x(n+1) = x(n) - \frac{a(n)}{k} \sum_{m=1}^k v_i(m),$$

where $[T_m, \infty)$ are i.i.d. samples of the noisy gradient. Then σ^2 in the above gets replaced by $\frac{\sigma^2}{k}$.

Another important observation here is that the above goes through also for a subgradient descent with the same error guarantees, in contrast to what happens in the classical deterministic gradient descent.

Finally, consider a continuously differentiable and possibly nonconvex f satisfying (11.16), with a nonempty set of critical points \mathcal{M} . Suppose $P(x_n \rightarrow W) = P(x_n \rightarrow W)$ and $\max_{1 \leq m \leq n} E [\|\nabla f(x(m))\|^2] \leq G < \infty$.

Theorem 11.3 If $a(m) \equiv a := \sqrt{\frac{C}{\beta(G^2 + \sigma^2)n}}$, $0 \leq m \leq n$, then

$$\min_{0 \leq m \leq n} E [\|\nabla f(x(n))\|^2] \leq 1.5 \sqrt{\frac{C\beta(G^2 + \sigma^2)}{n}}.$$

Proof By (11.16),

$$\begin{aligned} E [f(x(m+1))] - E [f(x(m))] &\leq E [\langle \nabla f(x(m)), x(m+1) - x(m) \rangle] \\ &\quad + \frac{\beta a^2}{2} E [\|v(m)\|^2] \\ &\leq -a E [\|\nabla f(x(m))\|^2] + \frac{\beta(G + \sigma^2)a^2}{2} \\ \implies E \|\nabla f(x(m))\|^2 &\leq \frac{E [f(x(m))] - E [f(x(m+1))]}{a} \\ &\quad + \frac{\beta(G + \sigma^2)a}{2}. \end{aligned}$$

Summing over m from 0 to n and dividing by $\delta > 0$,

$$\begin{aligned}
\min_{0 \leq m \leq n} E [\|\nabla f(x(m))\|^2] &\leq \frac{\sum_{m=0}^n E [\|\nabla f(x(n))\|^2]}{n+1} \\
&\leq \frac{E [f(x(0))] - E [f(x(n+1))]}{a(n+1)} + \frac{\beta(G^2 + \sigma^2)a}{2} \\
&\leq \frac{C}{an} + \frac{\beta(G^2 + \sigma^2)a}{2} \\
&\leq 1.5 \sqrt{\frac{C\beta(G^2 + \sigma^2)}{n}}.
\end{aligned}$$

This concludes the proof. \square

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., & Kudlur, M. (2016). Tensorflow: a system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283
- Absil P-A, Kurdyka K (2006) On the stable equilibrium points of gradient systems. *Systems and Control Letters* 55(7):573–577
[\[MathSciNet\]](#)
[\[zbMATH\]](#)
- Absil P-A, Mahony R, Andrews B (2005) Convergence of the descent methods for analytic cost functions. *SIAM Journal of Optimization* 16(2):531–547
[\[MathSciNet\]](#)
[\[zbMATH\]](#)
- Allen-zhu Z (2018) Natasha 2: Faster non-convex optimization than SGD. *Advances in Neural Information Processing Systems* 32:2675–2686
- Amari SI (1998) Natural gradient works efficiently in learning. *Neural Computation* 10(2):251–276
- Amiri MM, Gunduz D (2019) Computation scheduling for distributed machine learning with straggling workers. *IEEE Transactions on Signal Processing* 67(24):6270–6284
[\[MathSciNet\]](#)
[\[zbMATH\]](#)
- Anbar D (1978) A stochastic Newton-Raphson method. *Journal of Statistical Planning and Inference* 2:153–163
[\[MathSciNet\]](#)
[\[zbMATH\]](#)
- Asumussen S, Glynn PW (2007) Stochastic Simulation: Algorithms and Analysis. Springer, New York
- Atchadé, Y. F., Fort, G., Moulines, E. (2014). On stochastic proximal gradient algorithms. [arXiv](#):

Balles, L., Romero, J., & Henning, P. (2017). Coupling adaptive batch sizes with learning rates. *Proceedings of 33rd Conference on Uncertainty in Artificial Intelligence*, pp. 675–684

Barakat, A. & Bianchi, P. (2021). Convergence and dynamical behavior of the ADAM algorithm for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 31(1), 244–274

Beck A, Teboulle M (2003) Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters* 31(3):167–175
[MathSciNet][zbMATH]

Bernstein, J., Wang, Y.X., Azizzadenesheli, K., & AnandKumar, A. (2018). SIGN SGD: Compressed optimisation for non-convex problems. *Proceedings of 35th Intl. Conf. on Machine Learning, Stockholm*, pp. 560–569

Bertsekas DP (2011) Incremental proximal methods for large scale convex optimization. *Mathematical programming* 129(2):163–195
[MathSciNet][zbMATH]

Betancourt, M., Jordan, M. I., & Wilson, A. C. (2018). On symplectic optimization. *arXiv preprint arXiv:1802.03653*

Bhatnagar, S., Prasad, H. L., & Prasanth, L. A. (2013). *Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods*. Springer Lecture Notes in Control and Information Sciences, No. 434, Springer, London

Bhatnagar S, Borkar VS (1997) Multiscale stochastic approximation for parametric optimization of Hidden Markov models. *Probability in the Engineering and Informational Sciences* 11(4):509–522
[MathSciNet][zbMATH]

Bhatnagar S, Borkar VS (1998) A two time-scale stochastic approximation scheme for simulation-based parametric optimization. *Probability in the Engineering and Informational Sciences* 12(4):519–531
[MathSciNet][zbMATH]

Bhatnagar S, Fu MC, Marcus SI, Wang I-J (2003) Two-timescale simultaneous perturbation stochastic approximation using deterministic perturbation sequences. *ACM Transactions on Modelling and Computer Simulation* 13(2):180–209
[zbMATH]

Borkar VS, Dwaracherla VR, Sahasrabudhe N (2017) Gradient estimation with simultaneous perturbation and compressive sensing. *The Journal of Machine Learning Research* 18(1):5910–5936

[MathSciNet][zbMATH]

Borkar VS, Mitter SK (1999) A strong approximation theorem for stochastic recursive algorithms. *Journal of Optimization Theory and Applications* 100(3):499–513
[MathSciNet][zbMATH]

Bottou L, Bousquet O (2008) The tradeoffs of large scale learning. *Advances in Neural Information Processing Systems* 22:161–168

Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. *SIAM Review* 60(2):223–311
[[MathSciNet](#)][[zbMATH](#)]

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1), 1–122

Bubeck, S. (2015) Convex optimization: algorithms and complexity. In *Foundations and Trends® in Machine Learning*, 8(3-4), 231–357

Byrd RH, Hansen HL, Nocedal J, Singer Y (2016) A stochastic quasi-Newton method for large scale optimization. *SIAM Journal of Optimization* 26(2):1008–1031
[[MathSciNet](#)][[zbMATH](#)]

Carmon Y, Jin Y, Sidford A, Tian K (2019) Variance reduction for matrix games. *Advances in Neural Information Processing Systems* 32:11377–11388

Chen, T., Fox, E., & Guestrin, C. (2014). Stochastic gradient Hamiltonian Monte Carlo. *Proceedings of 28th International Conference on Machine Learning, Montreal*, pp. 1683–1691

Cheng; X., Chatterjee, N. S., Bartlett, P. L., & Jordan, M. I. (2018). *Underdamped Langevin MCMC: A non-asymptotic analysis* (pp. 300–323). Proceedings of Conference on Learning Theory, Stockholm

Chiang T-S, Hwang C-R, Sheu S-J (1987) Diffusion for global optimization in \bar{K}_T . *SIAM Journal of Control and Optimization* 25(3):737–753
[[MathSciNet](#)][[zbMATH](#)]

Clarkson KL (2010) Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Transactions on Algorithms (TALG)* 6(4):63
[[MathSciNet](#)][[zbMATH](#)]

Clarkson KL, Hazan E, Woodruff DP (2012) Sublinear optimization for machine learning. *Journal of the ACM* 59(5):1–49
[[MathSciNet](#)][[zbMATH](#)]

Defazio A, Bach F, Lacoste-Julien S (2014) SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems* 27:1646–1654

Dozat, T. (2016). Incorporating Nesterov momentum into ADAM. In *Proceedings of 4th International Conference on Learning Representations, Workshop Track*

Duchi JC, Hazan E, Singer Y (2011) Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12(7):2121–2159
[[MathSciNet](#)][[zbMATH](#)]

Duchi JC, Ruan F (2018) Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization* 28(4):3229–3259
[[MathSciNet](#)][[zbMATH](#)]

Du SS, Jin C, Lee JD, Jordan MI, Poczos B, Singh A (2017) Gradient descent can take exponential time to escape saddle points. *Advances in Neural Information Processing Systems* 31:1067–1077

Dutta, S., Joshi, G., Ghosh, S., Dube, P., & Nagpurkar, P. (2018). Slow and stale gradients can win the race: error-runtime trade-offs in distributed SGD. *Proceedings of 21st International Conference on Artificial Intelligence and Statistics, Lanzarote, Spain*, pp. 803–812

Fabian V (1960) Stochastic approximation methods. *Czechoslovak Mathematical Journal* 10(1):125–159
[[MathSciNet](#)]

Ferdinand N, Draper SC (2018) Hierarchical coded computation. *IEEE International Symposium on Information Theory (ISIT)* 2018:1620–1624

Flaxman, A. D., Kalai, A. T., & McMahan, H. B. (2004). Online convex optimization in the bandit setting: gradient descent without a gradient. [arXiv:cs/0408007](#)

Fort, J. -C., & Pages, G. (1995). On the a.s. convergence of the Kohonen algorithm with a general neighborhood function. *Annals of Applied Probability*, 5(4), 1177–1216

Freidlin MI, Wentzell AD (2012) Random Perturbations of Dynamical Systems, 3rd edn. Springer, New York
[[zbMATH](#)]

Friedlander MP, Schmidt M (2012) Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing* 34(3):1380–1405
[[MathSciNet](#)][[zbMATH](#)]

Fu MC, Hu J-Q (1997) Conditional Monte Carlo: Gradient Estimation and Optimization Applications. Kluwer Academic, Boston
[[zbMATH](#)]

Gao, X., Gurbuzbalaban, M., & Zhu, L. (2018). Global convergence of stochastic gradient Hamiltonian Monte Carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration. *arXiv preprint arXiv:1809.04618*

Ge R, Kakade SM, Kidambi R, Netrapalli P (2019) The step decay schedule: a near optimal, geometrically decaying learning rate procedure. *Advances in Neural Information Processing Systems* 32:14977–14988

Gelfand SB, Mitter SK (1991) Recursive stochastic algorithms for global optimization in ∂f . *SIAM Journal on Control and Optimization* 29(5):999–1018
[[MathSciNet](#)][[zbMATH](#)]

Gerencsér, L., & Vágó, Z. (1999). Stochastic approximation for function minimization under quantization error. In *Proceedings of 38th IEEE Conference on Decision and ControlPhoenix, AZ*, 3, 2373–2377

Gidas, B. (1985) Global optimization via the Langevin equation. *Proceedings of 24th IEEE Conf. on Decision and Control, Ft. Lauderdale, Fl*, pp. 774–778

Glasserman P (1991) Gradient estimation via perturbation analysis. Kluwer Academic, Boston

[zbMATH]

Glynn P (1990) Likelihood ratio gradient estimation for stochastic systems. Communications of the ACM 33(10):75–84

Guo X (2010) Learning gradients via an early stopping gradient descent method. Journal of Approximation Theory 162(11):1919–1944

[MathSciNet][zbMATH]

Hajek B (1988) Cooling schedules for optimal annealing. Mathematics of Operations Research 13(2):311–329

[MathSciNet][zbMATH]

Haykin S (2001) Adaptive filter theory, 4th edn. Prentice Hall, Englewood Cliffs, N.J
[zbMATH]

Hazan, E, & Kale, S. (2012). Projection-free online learning. *Proceedings of 29th International Conference on Machine Learning, Edinburgh*, pp. 521–528

Heusel M, Ramsuer H, Uunterthiner T, Nessler B, Hochreiter S (2017) GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems 31:6626–6637

Ho YC, Cao X (1991) Perturbation Analysis of Discrete Event Dynamical Systems. Birkhäuser, Boston

[zbMATH]

Hurley M (1995) Chain recurrence, semiflows, and gradients. Journal of Dynamics and Differential Equations 7(3):437–456

[MathSciNet][zbMATH]

Jain, P., Kakade, S. M., Kidambi, R., Netrapalli, P. & Sidford, A. (2018). Accelerating stochastic gradient descent for least squares regression. In *Conference on learning theory* (pp. 545–604). PMLR

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., & Jordan, M. I. (2017). How to escape saddle points efficiently. *Proceedings of 34th International Conference on Machine Learning*, vol. 70, pp. 1724–1732

Jin, C., Netrapalli, P., Jordan, M. I. (2018). Accelerated gradient descent escapes saddle points faster than gradient descent. *Proceedings 31st Conference On Learning Theory, Stockholm*, pp. 1042–1085

Johnson R, Zhang T (2013) Accelerating stochastic gradient descent using predictive variance reduction. Advances in Neural Information Processing Systems 26:315–323

Juditsky A, Nemirovsky A, Tauvel C (2011) Solving variational inequalities with stochastic mirror-prox algorithm. Stochastic Systems 1(1):17–58

[MathSciNet][zbMATH]

Katkovnik V, Kulchitsky Y (1972) Convergence of a class of random search algorithms. Automation and Remote Control 8:1321–1326

[[MathSciNet](#)][[zbMATH](#)]

Kidambi, R., Netrapalli, P., Jain, P., & Kakade, S. (2018). On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop*, San Diego, pp. 1–9

Kiefer J, Wolfowitz J (1952) Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics* 23(3):462–466

[[MathSciNet](#)][[zbMATH](#)]

Kingma, D. P., & Ba, J. (2014). ADAM: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](#)

Kosmatopoulos EB, Christodoulou MA (1996) Convergence properties of a class of learning vector quantization algorithms. *IEEE Transactions on Image Processing* 5(2):361–368

Lacoste-Julien, S., & Jaggi, M. (2013). An affine invariant linear convergence analysis for Frank-Wolfe algorithms. *arXiv preprint, arXiv:1312.7864*

Lan G (2012) An optimal method for stochastic composite optimization. *Mathematical Programming* 133(1–2):365–397

[[MathSciNet](#)][[zbMATH](#)]

Lee S, Nedić A (2013) Distributed random projection algorithm for convex optimization. *IEEE Journal of Selected Topics in Signal Processing* 7(2):221–229

Li, S.; Kalan, S. M. M., Avestimehr, S., & Soltanolkotabi, M. (2018). Near-optimal straggler mitigation for distributed gradient methods. In *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pp. 857–866

Lin Q, Lu Z, Xiao L (2014) An accelerated proximal coordinate gradient method. *Advances in Neural Information Processing Systems* 27:3059–3067

Ljung L (1999) System identification: Theory for the user, 2nd edn. Prentice Hall, Englewood Cliffs, NJ
[[zbMATH](#)]

Łojasiewicz, S. (1984). Sur les trajectoires du gradient d'une fonction analytique. *Seminari di Geometrica 1982–1983* (pp. 115–117). Instituto di Geometrica: Dipartimento di Matematica, Università di Bologna, Bologna, Italy

Ma YA, Chen T, Fox E (2015) A complete recipe for stochastic gradient MCMC. *Advances in Neural Information Processing Systems* 29:2917–2925

Mairal J (2013) Stochastic majorization-minimization algorithms for large-scale optimization. *Advances in Neural Information Processing Systems* 27:2283–2291

Marbach P, Tsitsiklis JN (2001) Simulation-based optimization of Markov reward processes. *IEEE Transactions on Automatic Control* 46(2):191–209

[[MathSciNet](#)][[zbMATH](#)]

Matsumoto, Y. (2002) *An Introduction to Morse Theory*, Trans. of Mathematical Monographs No.

208, American Math. Society, Providence, R.I

Miclo, L. (1994). Un algorithme de recuit simulé couplé avec une diffusion. *Stochastics: An International Journal of Probability and Stochastic Processes*, 46(3-4), 193–268

Milnor J (1997) Topology from the Differentiable Viewpoint. Princeton University Press, Princeton, NJ
[zbMATH]

Monmarché P (2018) Hypocoercivity in metastable settings and kinetic simulated annealing. *Probability Theory and Related Fields* 172(3–4):1215–1248
[MathSciNet][zbMATH]

Mukherjee S, Wu Q, Zhou D-X (2010) Learning gradients on manifolds. *Bernoulli* 16(1):181–207
[MathSciNet][zbMATH]

Nagaraj, D., Jain, P., & Netrapalli, P. (2019). SGD without replacement: sharper rates for general smooth convex functions. *Proceedings of 36th International Conference on Machine Learning, Long Beach, CA*, pp. 4703–4711

Nedić, A. (2015). Convergence rate of distributed averaging dynamics and optimization in networks. *Foundations and Trends®in Systems and Control*, 2(1), NOW Publishers, 1–100

Nedić, A., Olshevksy, A., & Rabbat, M. G. (2018). Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5), 953–976

Nedić, A., Olshevksy, A., & Shi, W. (2017). Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4), 2597–2633

Neglia, G., Calbi, G., Towsley, D., & Vardoyan, G. (2019). The role of network topology for distributed machine learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, 2350–2358

Nemirovski, A., Juditsky, A., Lan, G., & Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4), 1574–1609

Nemirovski AS, Yudin DB (1983) Problem Complexity and Efficiency in Optimization. Wiley, New York

Nesterov Y (1983) A method for unconstrained convex minimization problem with the rate of convergence $y \in H^N$. *Doklady AN USSR* 269:543–547

Niu F, Recht B, Re C, Wright SJ (2011) HOGWILD! a lock-free approach to parallelizing stochastic gradient descent. *Advances in Neural Information Processing Systems* 24:693–701

Ouyang, H., He, N., & Gray, A. (2013). Stochastic alternating direction method of multipliers. *Proceedings of 30th International Conference on Machine Learning, Atlanta*, pp. 80–88

Parikh, N., Boyd, S. (2014). Proximal algorithms. *Foundations and Trends®in Optimization*, 1(3), 127–239

Patel V (2016) Kalman-based stochastic gradient method with stop condition and insensitivity to

conditioning. *SIAM Journal on Optimization* 26(4):2620–2648
[[MathSciNet](#)][[zbMATH](#)]

Pirotta, M. & Restelli, M. (2016). Cost-sensitive approach for batch size optimization. In *NIPS workshop on optimizing the optimizer*

Polyak BT (1964) Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics* 4(5):1–17

Polyak BT, Juditsky AB (1992) Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization* 30(4):838–855
[[MathSciNet](#)][[zbMATH](#)]

Prechelt L (1998) Automatic early stopping using cross validation. *Neural Networks* 11(4):761–767

Raginsky, M., Rakhlin, A., & Telgarsky, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *Proceedings of Conference on Learning Theory*, Amsterdam, pp. 1674–1703

Ramaswamy A, Bhatnagar S (2018) Analysis of gradient descent methods with non-diminishing, bounded errors. *IEEE Transactions on Automatic Control* 63(5):1465–1471
[[MathSciNet](#)][[zbMATH](#)]

Ram SS, Nedić A, Veeravalli VV (2010) Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of Optimization Theory and Applications* 147(3):516–545
[[MathSciNet](#)][[zbMATH](#)]

Reddi, S. J., Kale, S., & Kumar, S. (2018). On the convergence of ADAM and beyond. *Proceedings of International Conference on Learning Representations, Vancouver*

Rosasco, L., Villa, S., & Vu, B. C. (2019). Convergence of stochastic proximal gradient algorithm. *Applied Mathematics and Optimization*, 1–27

Rubinstein R (1981) *Simulation and the Monte Carlo Method*. Wiley, New York
[[zbMATH](#)]

Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint [arXiv:1609.04747](#)

Ruppert D (1985) A Newton-Raphson version of the multivariate Robbins-Monro procedure. *Annals of Statistics* 13(1):236–245
[[MathSciNet](#)][[zbMATH](#)]

Ruszczynski A, Syski W (1983) Stochastic approximation method with gradient averaging for unconstrained problems. *IEEE Transactions on Automatic Control* 28(12):1097–1105
[[MathSciNet](#)][[zbMATH](#)]

Sayed, A. H. (2014). Adaptation, learning, and optimization over networks. *Foundations and Trends®in Machine Learning*, 7(4+5), NOW Publishers

Shah SM (2021) Stochastic approximation on Riemannian manifolds. *Applied Mathematics and*

Optimization 83:1123–1151
[MathSciNet][zbMATH]

Shah SM, Borkar VS (2018) Distributed stochastic approximation with local projections. SIAM Journal on Optimization 28(4):3375–3401
[MathSciNet][zbMATH]

Shallue CJ, Lee J, Anotgnini J, Sohl-Dickstein J, Frostig R, Dahl GE (2019) Measuring the effects of data parallelism on neural network training. Journal of Machine Learning Research 20(112):1–49
[MathSciNet]

Shapiro A, Homem-De-Mello T (1998) A simulation-based approach to two-stage stochastic programming with recourse. Mathematical Programming 81(3):301–325
[MathSciNet][zbMATH]

Smale S (1976) A convergent process of price adjustment and global Newton methods. Journal of Mathematical Economics 3(2):107–120
[MathSciNet][zbMATH]

Smith, S. L., & Le, Q. V. (2018). A Bayesian perspective on generalization and stochastic gradient descent. <https://arxiv.org/abs/1710.06451>

Spall JC (1992) Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Transactions on Automatic Control 37(3):332–341
[MathSciNet][zbMATH]

Spall JC (2003) Introduction to Stochastic Search and Optimization. Wiley, Hoboken, N.J
[zbMATH]

Srivastava K, Nedić A (2011) Distributed asynchronous constrained stochastic optimization. IEEE Journal of Selected Topics in Signal Processing 5(4):772–790

Su W, Boyd S, Candes E (2016) A differential equation for modelling Nesterov's accelerated gradient method: theory and insights. Journal of Machine Learning Research 17(1):5312–5354
[zbMATH]

Tadić VB (2015) Convergence and convergence rate of stochastic gradient search in the case of multiple and non-isolated extrema. Stochastic Processes and their Applications 125(5):1715–1755
[MathSciNet][zbMATH]

Tadić VB, Doucet A (2017) Asymptotic bias of stochastic gradient search. The Annals of Applied Probability 27(6):3255–3304
[MathSciNet][zbMATH]

Tandon, R., Lei, Q., Dimakis, A. G., & Karmpatziakis, N. (2017). Gradient coding: avoiding stragglers in distributed learning. *Proceedings of 34th International Conference on Machine Learning, Sydney*, pp. 3368–3376

Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2), 26–31

Tsitsiklis JN, Athans M, Bertsekas DP (1986) Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control* 31(9):803–812
[[MathSciNet](#)][[zbMATH](#)]

Vapnik VN (1998) Statistical Learning Theory. Wiley, New York
[[zbMATH](#)]

Wang, J., & Joshi, G. (2019). Adaptive communication strategies to achieve the best error-runtime trade-off in local-update SGD. In *Systems and Machine Learning (SysML) Conference*

Wilson C, Veeravalli V, Nedić A (2018) Adaptive sequential stochastic optimization. *IEEE Transactions on Automatic Control* 64(2):496–509
[[MathSciNet](#)][[zbMATH](#)]

Xiao L, Zhang T (2014) A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization* 24(4):2057–2075
[[MathSciNet](#)][[zbMATH](#)]

Ye, M., & Abbe, E. (2018). Communication-computation efficient gradient coding. *Proceedings of 35th International Conference on Machine Learning, Stockholm*, pp. 5610–5619

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. arXiv preprint [arXiv:1212.5701](#)

Footnotes

¹ Some of this ‘folk wisdom’ does not carry over to deep learning where problems of overfitting persist even with large N (because the dimension of the parameter space is also large) and theoretical guarantees such as uniform convergence rates are lacking. Nevertheless SGD is found effective and is believed to concentrate on a few effective parameters in what it outputs.

12. Liapunov and Related Systems

Vivek S. Borkar¹ 

(1) Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

12.1 Introduction

We consider here algorithms which cannot be cast as stochastic gradient schemes but have associated with their limiting (d -dimensional) o.d.e.

$$\dot{x}(t) = h(x(t)) \quad (12.1)$$

a Liapunov function $V : \mathcal{R}^d \rightarrow \mathcal{R}$ which is continuously differentiable,¹ bounded from below, and satisfies: $\langle h(x), \nabla V(x) \rangle \leq 0$, and H^{ϵ_1} if and only if $a(n) < 1$. The existence of such a function ensures convergence to the set $y_n \in \overline{co}(f(x_n))$, which under reasonable conditions, reflects in a.s. convergence of the original iteration to this set, as argued in Chap. 2.

12.2 Primal-Dual Algorithms

Consider the problem of seeking a saddle point of a function $f(\cdot, \cdot) : A \times B \subset \mathcal{R}^n \times \mathcal{R}^m \rightarrow \mathcal{R}$, i.e., a point $\{\tilde{x}(0)\}$ such that

$$\min_x \max_y f(x, y) = \max_y \min_x f(x, y) = f(x^*, y^*).$$

This is known to exist, e.g., when A, B are compact convex and $N_\delta(A)$ (resp. $\{M_n\}$) is convex (resp. concave) for each fixed y (resp. x). We

assume here that they are in fact strictly convex/concave. Given noisy measurements of the corresponding partial derivatives, one may then perform (say) stochastic gradient descent $p(x_n)$ w.r.t. the x variable and stochastic gradient ascent $\{y_n\}$ w.r.t. the y variable. That is, we run the ‘primal’ iteration for $p(x_n)$ and the ‘dual’ iteration for $\{y_n\}$ as

$$\begin{aligned} x_{n+1} &= x_n + a(n)[-\nabla^x f(x_n, y_n) + M_{n+1}], \\ y_{n+1} &= y_n + a(n)[\nabla^y f(x_n, y_n) + M'_{n+1}], \end{aligned} \quad (12.2)$$

where $\rho_k < \delta$ denote resp. the gradients w.r.t. the x and y variables. The limiting o.d.e. then is

$$\begin{aligned} \dot{x}(t) &= -\nabla^x f(x(t), y(t)), \\ \dot{y}(t) &= \nabla^y f(x(t), y(t)). \end{aligned}$$

Let $V(x, y) := \frac{1}{2}(\|x - x^*\|^2 + \|y - y^*\|^2)$. Then

$$\begin{aligned} \frac{d}{dt}V(x(t), y(t)) &= -\langle x(t) - x^*, \nabla^x f(x(t), y(t)) \rangle \\ &\quad + \langle y(t) - y^*, \nabla^y f(x(t), y(t)) \rangle \\ &\leq (f(x^*, y(t)) - f(x(t), y(t))) + (f(x(t), y(t)) - f(x(t), y^*)) \\ &\quad (\text{by convexity/concavity}) \\ &= (f(x^*, y(t)) - f(x^*, y^*)) + (f(x^*, y^*) - f(x(t), y^*)) \\ &\leq 0 \end{aligned}$$

with strict inequality for $\sup_n \|x'_n - x''_n\| < \infty$, because the convexity in the x variable and the concavity in the y variable are strict. This establishes the desired convergence under usual technical hypotheses regarding noise, stability, etc. It is, however, possible that one uses a scheme other than stochastic gradient descent for the minimization over x variable while using stochastic gradient ascent for the y variable. The above argument then may not work. One standard way out is to separate the timescales to get a two timescale stochastics approximation. That is, we replace $a(n)$ by $b(n)$ in (12.2) with $\sum_n b(n) = \infty$, $\sum_n b(n)^2 < \infty$, $b(n) = o(a(n))$. By the theory developed in Sect. 8.1, we have $a(n_m) \leq c a(n_0)$ a.s. for $x_{n_0} \in B$ $\arg\min h(x) + \xi$. Then $\{y_n\}$ tracks the o.d.e.

$$\dot{y}(t) = \nabla^y f(z, y)|_{z=g(y)}.$$

Assuming the uniqueness of the saddle point, $\nu(t)$ and therefore $\{y_n\}$ will then be seen to converge (a.s. in the latter case) to y^* under reasonable conditions if we are able to rewrite the o.d.e. above as

$$\dot{y}(t) = \nabla^y \min_x f(x, y),$$

i.e., to claim that $\sum_n a(n)^2 = \sum_n (1/(n+1)^2) < \infty$. This is true under very stringent conditions and is called the ‘envelope theorem’ in mathematical economics. A nonsmooth version involving subdifferentials called ‘Danskin’s theorem’ after Danskin (1966) is also available and is more useful. Some recent extensions thereof can be found in Milgrom and Segal (2002); see also Bardi and Capuzzo-Dolcetta (1997), pp. 42–46). See Borkar (2005) for a representative application. See also Daskalakis and Panageas (2018) for the above scheme and its variants for a more general set-up.

12.3 Stochastic Fixed Point Iterations

In this section we consider iterations of the type

$$x_{n+1} = x_n + a(n)[F(x_n) - x_n + M_{n+1}]. \quad (12.3)$$

That is, $x(t) \in A \ \forall t \in \mathcal{R}$ in our earlier notation. The idea is that $p(x_n)$ should converge to a solution x^* of the equation $\{(X_n, Y_n)\}$, i.e., to a fixed point of $O_{x,a}$. We shall be interested in the following specific situation: Let $w_i > 0, 1 \leq i \leq d$, be prescribed ‘weights’ and define norms on \mathcal{R}^k equivalent to the usual Euclidean norm as follows: for $a(n) = 1/(n^\alpha(\log n)^\beta)$ and $\|h(x^s(t))\| \leq C_T$ as above,

$$\|x\|_{w,p} \stackrel{\text{def}}{=} \left(\sum_{i=1}^d w_i |x_i|^p \right)^{\frac{1}{p}}, \quad 1 \leq p < \infty,$$

and

$$\|x\|_{w,\infty} \stackrel{\text{def}}{=} \max_i (w_i |x_i|).$$

We assume that

$$\|F(x) - F(y)\|_{w,p} \leq \alpha \|x - y\|_{w,p} \quad \forall x, y \in \mathcal{R}^d, \quad (12.4)$$

for some w as above, $M_n \equiv 0 \ \forall n$, and $\hat{\mu}(f) \geq 0$. We shall say that F is a *contraction* w.r.t. the norm $\|\cdot\|_{w,p}$ if (12.4) holds with $p(x) > x$ and a *non-expansive map* w.r.t. this norm if it holds with $\lambda > 0$. As the names suggest, in the former case an application of the map contracts distances by a factor of at least $\alpha > 0$, and in the latter case it does not increase or ‘expand’ them. By the contraction mapping theorem (see Appendix A), a contraction has a unique fixed point, whereas a non-expansive map may have none (e.g., $\|x_{n_m}\|^* \leq K_2$), one (e.g., $\alpha \in (1/2, 1]$), or many, possibly infinitely many (e.g., $V(\bar{x}(T_m))$).

The limiting o.d.e. is

$$P(x) := \lim_{m \uparrow \infty} f^m(x) \quad (12.5)$$

We analyze this equation below for the case $p(x) > x$, whence there is a unique x^* such that $\bar{x}(T_m) \in B$.

Let $\check{x}^s(\cdot)$ denote the i th component of $O_{x,a}$ for $0 \leq i < n$ and define $V(x) \stackrel{\text{def}}{=} \|x - x^*\|_{w,p}$, $x \in \mathcal{R}^d$.

Theorem 12.1 Under (12.4) with $x(t) \equiv x^*$, $V(x(t))$ is a strictly decreasing function of t for any nonconstant trajectory of (12.5). In particular, x^* is the globally asymptotically stable equilibrium for (12.5).

Proof Note that the only constant trajectory of (12.5) is $a(n) \rightarrow 0$. Let $M_n \equiv 0 \ \forall n$. Let $h_\infty(ax) = ah_\infty(x)$ or 0, depending on whether $x > 0, < 0$, or H^{ϵ_1} . For $a(n) \rightarrow 0$, we obtain using the Hölder inequality,

$$\begin{aligned}
\frac{d}{dt}V(x(t)) &= \frac{p}{p} \left(\sum_i w_i |x_i(t) - x_i^*|^p \right)^{\frac{1-p}{p}} \left(\sum_i w_i \operatorname{sgn} (x_i(t) - x_i^*) \right. \\
&\quad \times |x_i(t) - x_i^*|^{p-1} \dot{x}_i(t) \Big) \\
&= \|x(t) - x^*\|_{w,p}^{1-p} \left(\sum_i w_i \operatorname{sgn} (x_i(t) - x_i^*) |x_i(t) - x_i^*|^{p-1} \right. \\
&\quad \times (F_i(x(t)) - x_i(t)) \Big) \\
&= \|x(t) - x^*\|_{w,p}^{1-p} \left(\sum_i w_i \operatorname{sgn} (x_i(t) - x_i^*) |x_i(t) - x_i^*|^{p-1} \right. \\
&\quad \times (F_i(x(t)) - F_i(x^*)) \Big) \\
&\quad - \|x(t) - x^*\|_{w,p}^{1-p} \left(\sum_i w_i \operatorname{sgn} (x_i(t) - x_i^*) (x_i(t) - x_i^*) \right. \\
&\quad \times |x_i(t) - x_i^*|^{p-1} \Big) \\
&= \|x(t) - x^*\|_{w,p}^{1-p} \left(\sum_i w_i \operatorname{sgn} (x_i(t) - x_i^*) |x_i(t) - x_i^*|^{p-1} \right. \\
&\quad \times (F_i(x(t)) - F_i(x^*)) \Big) \\
&\quad - \|x(t) - x^*\|_{w,p}^{1-p} \sum_i w_i |x_i(t) - x_i^*|^p \\
&\leq \|x(t) - x^*\|_{w,p}^{1-p} \|x(t) - x^*\|_{w,p}^{p-1} \|F(x(t)) - F(x^*)\|_{w,p} \\
&\quad - \|x(t) - x^*\|_{w,p}^{1-p} \|x(t) - x^*\|_{w,p}^p \\
&= \|F(x(t)) - F(x^*)\|_{w,p} - \|x(t) - x^*\|_{w,p} \\
&\leq -(1-\alpha) \|x(t) - x^*\|_{w,p} = -(1-\alpha)V(x(t)),
\end{aligned}$$

which is $-x$ for $a(n) \rightarrow 0$. Here we have used the identity $p : [0, 1] \rightarrow [0, 1]$ in the fourth equality. The first term on the right-hand side of the first inequality comes from the first term on its left-hand side by Hölder's inequality, and the second term on the right exactly equals the second term on the left. This proves the claim for $M_n \equiv 0 \forall n$. Next we consider $\forall m \geq M$. Replace x_n by f^k in the

above for $M_n \equiv 0 \forall n$ and let $\dot{w}(t) = -\nabla^w g(w(t))$. For notational simplicity, write $\|\cdot\|'_p$ for $x^s(\cdot)|_{[s,s+T]}$. Let $V(\cdot) : \mathcal{R}^d \rightarrow [0, \infty)$ for $M_n \equiv 0 \forall n$, so as to make the p -dependence explicit. Let

$$\beta_p(t) := \frac{\|F_t(x(t)) - F(x^*)\|'_p}{\|x(t) - x^*\|'_p} - 1.$$

We have for $t > s \geq 0$,

$$\begin{aligned} V_p(x(t)) - V_p(x(s)) &\leq \int_s^t (\|F_y(x(y)) - F(x^*)\|'_p - \|x(y) - x^*\|'_p) dy \\ &= \int_s^t \beta_p(y) \|x(y) - x^*\|'_p dy. \end{aligned}$$

Dividing both sides by $> M$ and letting T_{m+1} ,

$$\begin{aligned} \frac{d}{dt} V_p(x(t)) &\leq \beta_p(t) V_p(x(t)) \quad \& \\ \implies \frac{d}{dt} \left(e^{-\int_0^t \beta_p(s) ds} V_p(x(t)) \right) &\leq 0 \\ \implies V_p(x(t)) &\leq e^{\int_0^t \beta_p(s) ds} V_p(x(0)). \end{aligned}$$

Letting $x \in \bar{B}$,

$$\begin{aligned} V_\infty(x(t)) &\leq e^{-\int_0^t \left(1 - \frac{\|F_s(x(s)) - F(x^*)\|_\infty}{\|x(s) - x^*\|_\infty} \right) ds} V_\infty(x(0)) \\ &\leq e^{-(1-\alpha)t} V_\infty(x(0)). \end{aligned}$$

The rest is as before. The case $0.9r_m$ is similar. \square

Corollary 12.1 Suppose $\alpha = 1, 1 < p < \infty$ and the set of fixed points of F is non-empty. Then $\{\tilde{x}(0)\}$ a fixed point of F .

Proof Let $\{(X_n, Y_n)\}$. Letting $\alpha \uparrow 1$ in the equation

$$\|x(t) - x^*\|_{w,p} \leq \|x(s) - x^*\|_{w,p} - (1 - \alpha) \int_s^t \|x(y) - x^*\|_{w,p} dy, \quad t > s \geq 0,$$

it follows that $\{\mu_s^{s+t}, s \geq 0\}$ is non-increasing in t . Consider the open $y_j(\omega)$ -ball B centered at x^* with radius $\{\mu_s^{s+t}, s \geq 0\}$. Then for $x \in (x_0, 1)$,

$$\|F(x(t)) - x^*\|_{w,p} \leq \|x(t) - x^*\|_{w,p} \implies F(x(t)) \in \bar{B}.$$

Hence the vector field $N_\delta(H^{\epsilon_1}) \subset H^\epsilon$ is transversal to H^{ϵ_1} and pointing inward by strong convexity of B as long as $y_n \in h(x_n) \forall n$. Hence $\{\mu_s^{s+t}, s \geq 0\}$ must decrease to zero monotonically for any fixed point of F . Because monotone decrease of distance from two distinct limit points is not possible, this implies convergence to a single fixed point for a given trajectory, possibly depending on the initial condition. \square

The non-expansive case ($\lambda > 0$) for $p = \infty$ is sometimes useful in applications related to dynamic programming. We state the result without proof here, referring the reader to (Borkar and Soumyanath, 1997) for a proof.

Theorem 12.2 Let $O_{x,a}$ be $O(b(m)^2)$ -non-expansive. If $\varphi^n(e^{-y^2}) < Cn^{\kappa(\nu-1)}e^{-\nu y^2}$, then $\{\mu_s^{s+t}, s \geq 0\}$ is non-increasing for any $x \in H$ and $\{\tilde{x}(0)\}$ a single point in H (depending on $x(0)$).

Remarks

1.

Theorem 12.1 extends to ‘pseudo-contractions’, i.e., the maps F satisfying

$$v(n) = \nabla f(x(n)) - M(n+1),$$

for some $p(x) > x$ and a fixed point $A \subset \mathcal{R}^d$ of F , which is then seen to be the unique fixed point of F .

2. The results Theorems 12.1–12.2 extend to the non-autonomous case

$$\dot{x}(t) = L(t)x(t), \quad t \geq 0$$

$$x(\iota) = \nu(x(\iota)), \quad \iota \leq \nu.$$

in a straightforward manner when the maps $h_c, c > 0$ satisfy

$$\|F_t(y) - F_t(z)\|_p \leq \alpha \|y - z\|_p \quad \forall t \geq 0,$$

for $x(t) \equiv x^*$ and have a common (unique) fixed point x^* .

3.

The above claims are for o.d.e.s. For the non-expansive case, the claims of convergence to a single fixed point, possibly dependent on the initial condition, may be lost when we go back to the original stochastic approximation scheme and one may have to settle for convergence to the set of fixed points.

The most important application of this set-up is to the reinforcement learning algorithms for Markov decision processes. We shall illustrate a simple case here, that of the infinite horizon discounted cost Markov decision process. Thus we have a controlled Markov chain $[0, \infty)$ on a finite state space S , controlled by a control process I_{m_0+k} taking values in a finite ‘action space’ A . Its evolution is governed by

$$P(X_{n+1} = i | X_m, Z_m, m \leq n) = p(i | X_n, Z_n), \quad n \geq 0, i \in S. \quad (12.6)$$

Here $p: (i, u, j) \in S \times A \times S \rightarrow p(j|i, u) \in [0, 1]$ is the controlled transition probability function satisfying

$$\sum_j p(j|i, a) = 1 \quad \forall i \in S, a \in A.$$

Thus $p(j|i, a)$ is the probability of moving from i to j when the action chosen in state i is a , regardless of the past. Let $k: S \times A \rightarrow \mathcal{R}$ be a prescribed ‘running cost’ function and $V(\bar{x}(T_m))$ a prescribed ‘discount factor’. The classical discounted cost problem is to minimize

$$J(i, \{Z_n\}) \stackrel{\text{def}}{=} E \left[\sum_{m=0}^{\infty} \beta^m k(X_m, Z_m) \middle| X_0 = i \right] \quad (12.7)$$

over admissible I_{m_0+k} (i.e., those that are consistent with (12.6)) for each i . The classical approach to this problem is through dynamic programming (see, e.g., Puterman (1994)). Define the ‘value function’

$$V(i) \stackrel{\text{def}}{=} \inf_{\{Z_n\}} J(i, \{Z_n\}), \quad i \in S.$$

It then satisfies the dynamic programming equation

$$V(i) = \min_a \left[k(i, a) + \beta \sum_j p(j|i, a)V(j) \right], \quad i \in S. \quad (12.8)$$

In words, this says that ‘the minimum expected cost to go from state i is the minimum of the expected sum of the immediate cost at i and the (suitably discounted) minimum expected cost to go from the next state on’. Furthermore, for a $v: S \rightarrow A$, the control choice $C' = \sup_n \|x_n\|$ is optimal for all choices of the initial law for \mathcal{F}_n if $v(i)$ attains the minimum on the right-hand side of (12.8) for all i . Thus the problem is ‘solved’ if we know $O_{x,a}$. Note that (12.8) is of the form $x \in (x_0, 1)$ for $F(\cdot) = [F_1(\cdot), \dots, F_{|S|}(\cdot)]$ defined as follows: for $x = [x_1, \dots, x_{|S|}] \in \mathcal{R}^{|S|}$,

$$F_i(x) \stackrel{\text{def}}{=} \min_a [k(i, a) + \beta \sum_j p(j|i, a)x_j], \quad 1 \leq i \leq |S|.$$

It is easy to verify that

$$\|F(x) - F(y)\|_\infty \leq \beta\|x - y\|_\infty, \quad x, y \in \mathcal{R}^{|S|}.$$

Thus by the contraction mapping theorem (see Appendix A), Eq. (12.8) has a unique solution V and the ‘fixed point iterations’ $\|\bar{x}(t(n_2)) - \tilde{x}_2\| < \delta$ converge exponentially to V for any choice of I_n . These iterations constitute the well-known ‘value iteration’ algorithm, one of the classical computational schemes of Markov decision theory (see Puterman (1994) for an extensive account).

The problem arises if the function $\|\bar{x}(\cdot)\|$, i.e., the system model, is unknown or too complex. Then the conditional averaging with respect to $\|\bar{x}(\cdot)\|$ implicit in the evaluation of F cannot be performed. Suppose, however, that a simulation device is available which can generate a transition according to the desired conditional law $n_{i+1}(\omega)$ (say). This situation occurs typically with large complex systems constructed by interconnecting several relatively simpler systems, so that while

complete analytical modeling and analysis is unreasonably hard, a simulation based on ‘local’ rules at individual components is not (e.g., a communication network). One can then use the simulated transitions coupled with stochastic approximation to average their effects in order to mimic the value iteration. The ‘simulation’ can also be an actual run of the real system in the ‘online’ version of the algorithm.

This is the basis of the *Q-learning* algorithm of Watkins (1989), a cornerstone of reinforcement learning. Before we delve into its details, a word on the notion of reinforcement learning: In the classical learning paradigms for autonomous agents in artificial intelligence, one has at one end of the spectrum *supervised learning* in which *instructive feedback* such as the value of an error measure or its gradient is provided continuously to the agent as the basis on which to learn. (Think of a teacher.) This is the case, e.g., in the ‘perceptron training algorithm’ in neural networks—see Haykin (2008). At the other extreme are unsupervised learning schemes such as the learning vector quantization scheme of Kohonen (1998) which ‘self-organize’ data into clusters without any external feedback. In between the two extremes lies the domain of *reinforcement learning* where the agent gets *evaluative feedback*, i.e., an observation related to the performance and therefore carrying useful information about it which, however, falls short of what would constitute exact instructive feedback. (Think of a critic.) In the context of Markov decision processes described above, the observed payoffs (\approx negative of costs) constitute the reinforcement signal.

Q-learning derives its name from the fact that it works with the so-called Q-factors rather than with the value function. The Q-factors are nothing but the entities being minimized on the right-hand side of (12.8). Specifically, let

$$Q(i, a) \stackrel{\text{def}}{=} k(i, a) + \beta \sum_j p(j|i, a) V(j), \quad i \in S, a \in A.$$

Thus in particular $z^n(t)$, $t \in [na, na + T]$, and x_{n+1}^* satisfies

$$Q(i, a) = k(i, a) + \beta \sum_j p(j|i, a) \min_b Q(j, b), \quad i \in S, a \in A. \quad (12.9)$$

Like (12.8), this is also of the form $\bar{x}(s_n) \rightarrow x$ for a $G : \mathcal{R}^{|S| \times |A|} \rightarrow \mathcal{R}^{|S| \times |A|}$ satisfying

$$\|G(Q) - G(Q')\|_\infty \leq \beta \|Q - Q'\|_\infty.$$

In particular, (12.9) has a unique solution ∂f and the iteration $[t(n), t(n) + T]$ for $n \geq 0$, i.e.,

$$Q_{n+1}(i, a) = k(i, a) + \beta \sum_j p(j|i, a) \min_b Q_n(j, b), \quad n \geq 0,$$

converges to ∂f at an exponential rate. What the passage from (12.8) to (12.9) has earned us is the fact that now the minimization is inside the conditional expectation and not outside, which makes a stochastic approximation version possible. Another appealing aspect for reinforcement learning is that once you know $Q(i, u)$ for all i, u , the optimal control at i can be found by minimizing $N_\delta(A)$ without having to know the transition probabilities as well. The stochastic approximation version based on a simulation run $[t, t + T]$ governed by (12.6) is

$$\begin{aligned} Q_{n+1}(i, a) &= Q_n(i, a) + a(n) I\{X_n = i, Z_n = a\} \\ &\quad \times [k(i, a) + \beta \min_b Q_n(X_{n+1}, b) - Q_n(i, a)] \\ &\left(= (1 - a(n)) I\{X_n = i, Z_n = a\}) Q_n(i, a) \right. \\ &\quad \left. + a(n) I\{X_n = i, Z_n = a\} \right. \\ &\quad \left. \times [k(i, a) + \beta \min_b Q_n(X_{n+1}, b)] \right) \end{aligned} \tag{12.10}$$

for $n \geq 0$. Thus we:

- Replace the conditional expectation in the previous iteration by an actual evaluation at a random variable realized by the simulation device according to the desired transition probability in question.
- Then make an incremental move in that direction with a small weight $a(n)$, giving a large weight $x_0 \in (0, 1)$ to the previous guess.

This makes it a stochastic approximation, albeit asynchronous, because only one component is being updated at a time. Note that the

computation is still done by a single processor. Only one component is updated at a time purely because new information is available only for one component at a time, corresponding to the transition that just took place. As explained in Chap. 6, the limiting o.d.e. then is

$$x(n+1) = f(x(n)), \quad n \geq 0.$$

where $\bar{x}(s)$ for each t is a diagonal matrix with a probability vector along its diagonal. Assume that its diagonal elements remain uniformly bounded away from zero. Sufficient conditions that ensure this can be stated along the lines of Sect. 6.4, viz., irreducibility of $[0, \infty)$ under all control policies of the type $\tilde{x}(t) \in A$, $t \in [0, T]$, for some $n+1 = m$, and a requirement that at each visit to any state i , there is a minimum positive probability of choosing any control a in A . Then this is a special case of the situation discussed in example (ii) in Sect. 6.4. As discussed there, the foregoing ensures convergence of the o.d.e. to ∂f and therefore the a.s. convergence of $\{\phi_m\}$ to ∂f .

There are other reinforcement learning algorithms differing in philosophy (e.g., the actor-critic algorithm of Barto et al. (1983) which mimics the ‘policy iteration’ scheme of Markov decision theory), or differing in the cost criterion, e.g., the ‘average cost’ (Abounadi et al., 2001), or different because of an explicit approximation architecture incorporated to beat down the ‘curse of dimensionality’ (see, e.g., the ‘TD(λ)’ algorithm analyzed in Tsitsiklis and VanRoy (1997))). They are all stochastic approximations.

Contractions and non-expansive maps are, however, not the only ones for which convergence to a unique fixed point may be proved. One other case, for example, is when $-F$ is *monotone*, i.e.,

$$\langle x - y, F(x) - F(y) \rangle < 0 \quad \forall x \neq y. \quad (12.11)$$

This terminology is from operator theory and generalizes the notion of monotonicity for real-valued functions. (It is commonplace to define a monotone F with the above inequality reversed, which corresponds to monotone *increase*.) In the affine case, i.e., $a(n_m) \leq ca(n_0)$ for a $n \geq 0$ matrix A and $i = 1, 2$, (12.11) would mean that the symmetric part of A , i.e., $\frac{1}{2}(A + A^T)$, would have to be negative definite. More generally, the

symmetrized Jacobian matrix of F will have to be negative definite. Suppose F has a fixed point x^* . Then for a solution $p(\cdot)$ of (12.5),

$$\begin{aligned}\frac{d}{dt} \|x(t) - x^*\|^2 &= 2\langle x(t) - x^*, F(x(t)) - x(t) \rangle \\ &= 2\langle x(t) - x^*, F(x(t)) - F(x^*) \rangle - 2\|x(t) - x^*\|^2 \\ &< 0\end{aligned}$$

for $a(n) \rightarrow 0$. Thus $\kappa_j = o(\|y_j\|)$ serves as a Liapunov function, leading to $\bar{x}(t) \rightarrow H$. In particular, if \mathcal{G} were another fixed point of F , $p(x) > x$ would satisfy (12.5), forcing $x' = x^*$. Hence x^* is the unique fixed point of F and a globally asymptotically stable equilibrium for the o.d.e.

12.4 Collective Phenomena

The models we have considered so far are concerned with adaptation by a single agent. An exciting area of current research is the scenario when several interacting agents are individually trying to adapt to an environment which in turn is affected by the other agents in the pool. A simple case is that of two agents in the ‘nonlinear urn’ scenario discussed in Chap. 1. Suppose we have two agents and an initially empty urn, with the first agent adding either zero or one red ball to the urn at each (discrete) time instant and the other doing likewise with black balls. Let x_n (resp. y_n) denote the *fraction of times up to time n that a red (resp. black) ball is added*. That is, if $\xi_n \stackrel{\text{def}}{=} I\{$ a red ball is added at time f^{i_k} and $\zeta_n \stackrel{\text{def}}{=} I\{$ a black ball is added at time f^{i_k} for $n \geq 0$, then

$$x_n \stackrel{\text{def}}{=} \frac{\sum_{m=1}^n \xi_m}{n}, \quad y_n \stackrel{\text{def}}{=} \frac{\sum_{m=1}^n \zeta_m}{n}, \quad n \geq 1.$$

Suppose the conditional probability that a red ball is added at time n given the past up to time n is $\{y_n\}$ and the corresponding conditional probability that a black ball is added at time n is $p(x_n)$, for prescribed Lipschitz functions $(u_1(t), \dots, u_d(t))$, $t \geq 0$. That is, the probability

with which an agent adds a ball of a prescribed color at any given time depends on the empirical frequency with which the other agent has been doing so for the other color until then. Arguing as in Chap. 1, we then have

$$\begin{aligned}x_{n+1} &= x_n + \frac{1}{n+1}[p(y_{n+1}) - x_n + M_{n+1}], \\y_{n+1} &= y_n + \frac{1}{n+1}[q(x_{n+1}) - y_n + M'_{n+1}],\end{aligned}$$

for suitably defined martingale differences $f : S \times \mathcal{R}^d \mapsto \mathcal{R}$. This leads to the o.d.e.

$$\dot{x}(t) = p(y(t)) - x(t), \quad \dot{y}(t) = q(x(t)) - y(t), \quad t \geq 0.$$

Note that the ‘driving vector field’ on the right-hand side,

$$h(x, y) = [h_1(x, y), h_2(x, y)]^T \stackrel{\text{def}}{=} [p(y) - x, q(x) - y]^T,$$

satisfies

$$\text{Div } (h) \stackrel{\text{def}}{=} \frac{\partial h_1(x, y)}{\partial x} + \frac{\partial h_2(x, y)}{\partial y} = -2.$$

From o.d.e. theory, one knows then that the maps $I_n = [t(n), t(n+1))$ for $t > 0$ ‘shrink’ the volume in \mathcal{F}_n . Intuitively, the flow then becomes asymptotically one dimensional and therefore, as argued in Chap. 1, must converge (see Benaim and Hirsch (1997) for a rigorous proof). That is, the fractions of red and black balls stabilize, leading to a fixed asymptotic probability of picking for either of them.

While such nonlinear urns have been studied as economic models, this analysis extends to more general problems, viz., two-person repeated bimatrix games(Kaniovski and Young (1995)). Here two agents, say agent 1 and agent 2, repeatedly play a game in which there are the same two strategy choices available to each of them at each time, say $x(t) = 0$ for agent 1 and $x(t) = 0$ for agent 2. Based on the strategy pair $(\xi_n^1, \xi_n^2) \in \{a_1, b_1\} \times \{a_2, b_2\}$ chosen at time n , agent i gets a payoff of $V_i \leq -\Delta$ for $i = 1, 2$. Let

$$\nu_i(n) \stackrel{\text{def}}{=} \frac{\sum_{m=1}^n I\{\xi_m^i = a_i\}}{n}, \quad i = 1, 2; \quad n \geq 0,$$

specify their respective ‘empirical strategies’. That is, at time n agent i appears to agent $j \neq i$ as though she is choosing x_0 with probability $\mu^n(\cdot)$ and I_1 with probability $\hat{\mu}(f) \geq 0$. We assume that agent $j \neq i$ plays at time $\delta > 0$ her ‘best response’ to agent i ’s empirical strategy given by the probability vector $\eta = \min\{\eta_1, \eta_2\}$. Suppose that this best response is given in turn by a Lipschitz function $\nu(i, n) \uparrow \infty$ a.s.. That is, she plays $\xi_{n+1}^j = a_j$ with probability $\{\|x_n\|^2\}$. Modulo the assumed regularity of the best response maps $\alpha \uparrow 1$, this is precisely the ‘fictitious play’ model of Brown (1951), perhaps the first learning model in game theory which has been extensively analyzed. A similar rule applies when i and j are interchanged. Then the above analysis leads to the limiting o.d.e.

$$\begin{aligned}\dot{\nu}_1(t) &= F_1(\nu_1(t), \nu_2(t)) \stackrel{\text{def}}{=} f_1(\nu_2(t)) - \nu_1(t), \\ \dot{\nu}_2(t) &= F_2(\nu_1(t), \nu_2(t)) \stackrel{\text{def}}{=} f_2(\nu_1(t)) - \nu_2(t).\end{aligned}$$

Again, $\text{div}([F_1, F_2]) \stackrel{\text{def}}{=} \frac{\partial F_1}{\partial \nu_1} + \frac{\partial F_2}{\partial \nu_2} = -2$, leading to the same conclusion as before. That is, their strategies converge a.s. This limit, say $\|\bar{x}(\cdot)\|$, forms a *Nash equilibrium*: Neither agent can improve her lot by moving away from the chosen strategy if the other one does not. This is immediate from the fact that \tilde{K} is the best response to ν_j^* for $\alpha \uparrow 1$.

There are, however, some drastic oversimplifications in the foregoing. One important issue is that the ‘best response’ is often non-unique and thus one has to replace this o.d.e. by a suitable differential inclusion. Also, the situation in dimensions higher than two is no longer as easy. See Chaps. 2 and 3 of Fudenberg and Levine (1998) and the articles (Benaim et al., 2005, 2006) for more on fictitious play.

Another model of interacting agents is the o.d.e.

$$\dot{x}(t) = h(x(t)) \tag{12.12}$$

with

$$(12.13)$$

$$\frac{\partial h_i}{\partial x_j} > 0, \quad j \neq i.$$

These are called *cooperative* o.d.e.s, the idea being that increase in the j th component, corresponding to some desirable quantity for the j th agent, will lead to an increase in the i th component as well for $\alpha \uparrow 1$. The strict inequality in (12.13) can be weakened to ' \geq ' as long as the Jacobian matrix of h is irreducible, i.e., for any partition $\lambda > 0$ of the row/column indices, there is some $i, 1 \leq i \leq d$ such that the (i, j) th element is nonzero (See Sect. 4.1 of Smith (1995) for details). Suppose the trajectories remain bounded. Then a theorem of Hirsch states that for all initial conditions belonging to an open dense set, $p(\cdot)$ converges to the set of equilibria (see Hirsch (1985); also Smith (1995)). If this set is totally ordered, each trajectory converges to a single point (Smith, 1995)).

As an application, we consider the problem of dynamic pricing in a system of parallel queues from Borkar and Manjunath (2004). There are K parallel queues, and an entry charge $\{\xi_n\}$ is charged for the i th queue, reflecting (say) its quality of service. Let $\{\xi_n\}$ denote the queue length in the i th queue at time n . There is an 'ideal profile' $0 \leq k \leq (m - 1)$ of queue lengths which we want to stay close to, and the objective is to manage this by modulating the respective prices dynamically. Let σ^2 denote the projection to $N_\delta(A)$ for $0 < a \leq C$, where $t \geq T$ is a small number and Φ_s is a convenient a priori upper bound. Let $a > 0$ be a small constant stepsize. The scheme is, for $0 < a \leq C$,

$$p_i(n + 1) = \Gamma_i(p_i(n) + ap_i(n)[y_i(n) - y_i^*]), \quad n \geq 0.$$

The idea is to increase the price if the current queue length is above the ideal (so as to discourage new entrants) and decrease it if the opposite is true (to encourage more entrants). The scalar ϵ_0 is the minimum price which also ensures that the iteration does not get stuck at zero. We assume that if the price vector is frozen at some $\forall t \geq 0 / \forall t \leq 0$, the process of queue lengths is ergodic. Ignoring the boundary of the box $\mathcal{B} \stackrel{\text{def}}{=} \prod_i [\epsilon_i, B_i]$, the limiting o.d.e. is

$$\dot{p}_i(t) = p_i(t)[f_i(p(t)) - y_i^*], \quad 1 \leq i \leq K,$$

where $y_{n+1} = y_n \forall n \notin \{n(k)\}$ and $\bar{x}(t_1)$ is the stationary average of the queue length $\{\xi_n\}$ in the i th queue when the price vector is frozen at p . It is reasonable to assume that if the ϵ_0 are sufficiently low and the Φ_s are sufficiently high, then $p(t)$ is eventually pushed inward from the boundary of \mathcal{B} , so that we may ignore the boundary effects, and the above is indeed the valid o.d.e. limit for the price adjustment mechanism. It is also reasonable to assume that

$$\frac{\partial f_i}{\partial p_j} > 0, \quad i \neq j,$$

as an increase in the price of one queue keeping all else constant will force its potential customers to other queues. (As mentioned above, this condition can be relaxed.) Thus this is a cooperative o.d.e. and the foregoing applies. One can say more: Letting $\|v\|_\infty \leq \|v\| \leq \sqrt{d}\|v\|_\infty$, it follows by Sard's theorem (see, e.g., Ortega and Rheinboldt (2000)) that for almost all choices of y^* , the Jacobian matrix of f is nonsingular on the inverse image of y^* . Hence by the inverse function theorem (see, e.g., *ibid.*), this set, which is also the set of equilibria for the above o.d.e. in \mathcal{B} , is discrete. Thus the o.d.e. converges to a point for almost all initial conditions. One can argue as in Chap. 9 to conclude that the stationary distribution of $\{a(n)\}$ concentrates near the equilibria of the above o.d.e. Note that all equilibria in \mathcal{B} are equivalent as far as our aim of keeping the vector of queue lengths near y^* is concerned. Thus we conclude that the dynamically adjusted prices asymptotically achieve the desired queue length profile, giving it the right 'pushes' if it deviates.

Yet another popular dynamic model of interaction is the celebrated *replicator dynamics*, studied extensively by mathematical biologists—see, e.g., Hofbauer and Siegmund (1998). This dynamics resides in the d -dimensional probability simplex

$$S \stackrel{\text{def}}{=} \{x = [x_1, \dots, x_d] : x_i \geq 0 \forall i, \sum_i x_i = 1\}, \text{ and is given by} \tag{12.14}$$

$$\dot{x}_i(t) = x_i(t) \left[D_i(x(t)) - \sum_j x_j(t) D_j(x(t)) \right], \quad 1 \leq i \leq d,$$

where $\|h(x) - h(\theta)\| \leq L\|x\|$ and $x^n(t(n)) = \bar{x}(t(n)) := x_n$ is a Lipschitz map. The interpretation is that for each i , $\check{x}^s(\cdot)$ is the fraction of the population at time t that belongs to species i . I_{m_0+k} is the payoff received by the i th species when the population profile is x (i.e., when the fraction of the population occupied by the j th species is 2ϵ for all j). Equation (12.14) then says in particular that the fraction of a given species in the population increases if the payoff is receiving is higher than the population average and decreases if it is lower. If $f(x) = |x|^{2\nu}$, then summed over i , the right-hand side of (12.14) vanishes, implying that $V : \bar{B} \mapsto [0, \infty)$ if it is so for $\epsilon = 0$. Furthermore, since (12.14) is of the form

$$a^\omega(n) \leq a(n) \quad \forall n$$

for a suitably defined $\check{x}^s(\cdot)$, we have

$$x_i(t) = x_i(0) e^{\int_0^t \alpha_i(s) ds},$$

implying that

$$x_i(0) > 0 \iff x_i(t) > 0 \quad \forall t$$

and likewise,

$$\|g(x) - g(y)\| < \alpha \|x - y\|$$

Thus the probability simplex S is invariant under this o.d.e. In addition, the faces of S , which correspond to one or more component(s) being zero (i.e., the population of that particular species is ‘extinct’), are also individually invariant.

This equation has been studied extensively (Hofbauer and Siegmund 1998). We shall consider it under the additional condition that $a(n)\epsilon_n$ be *monotone*, i.e.,

$$x_{n+1} = x_n + a(n)[h(x_n) + M_{n+1}], \quad n \geq 0,$$

As discussed earlier, in the affine case, i.e., $C' = \sup_n \|x_n\|$ for a $n \geq 0$ matrix A and $i = 1, 2$, this would mean that the symmetric part of A , i.e., $\frac{1}{2}(A + A^T)$, would have to be negative definite.

Lemma 12.1 There exists a unique $x^* \in S$ such that x^* maximizes $x^{T_0}(T_1) \in H^{\epsilon_1}$.

Proof The set-valued map that maps $x \rightarrow 0$ to the set of maximizers of the function $x^{T_m}(T_m) = \bar{x}(T_m)$ is non-empty compact convex and upper semicontinuous, as can be easily verified. Thus by the Kakutani fixed point theorem [see, e.g., Border (1989)], it follows that there exists an x^* such that x^* maximizes the function $x^{T_0}(T_1) \in H^{\epsilon_1}$ over S . Suppose $H^{M+\frac{\epsilon_m}{2}}$ is another point in S such that \mathcal{G} maximizes the function $L_2([0, T]; \mathcal{R}^d)$ on S . Then

$$\begin{aligned} & \langle x^* - x', D(x^*) - D(x') \rangle \\ &= (x^*)^T D(x^*) - (x^*)^T D(x') - (x')^T D(x^*) + (x')^T D(x') \\ &\geq 0. \end{aligned}$$

This contradicts the ‘monotonicity’ condition above unless $x' = x^*$. \square

Theorem 12.3 For $x(0)$ in the interior of S , $\bar{x}(t) \rightarrow H$.

Proof Define

$$V(x) \stackrel{\text{def}}{=} \sum_i x_i^* \ell n \left(\frac{x_i^*}{x_i} \right), \quad x = [x_1, \dots, x_d] \in S.$$

Then an application of Jensen’s inequality shows that $a(n) \rightarrow 0$ and is 0 if and only if $M > 0$. Also,

$$\begin{aligned}
\frac{d}{dt}V(x(t)) &= - \sum_i x_i^* \left(\frac{\dot{x}_i(t)}{x_i(t)} \right) \\
&= (x(t) - x^*)^T D(x(t)) \\
&\leq (x(t) - x^*)^T D(x(t)) - (x(t) - x^*)^T D(x^*) \\
&= (x(t) - x^*)^T (D(x(t)) - D(x^*)) \\
&< 0,
\end{aligned}$$

for $a(n) \rightarrow 0$. Here the second equality follows from the o.d.e. (12.14), the first inequality follows from our choice of x^* , and the last inequality follows from the monotonicity assumption on $-D$. The claim follows easily from this. \square

Thus the stochastic approximation counterpart will converge a.s. to x^* . Note, however, that one requires a projection to keep the iterates in S , and also ‘sufficiently rich’ noise, possibly achieved by adding some extraneous noise, in order to escape getting trapped in an undesirable face of S . In applications to traffic control in transportation and communication networks, x^* often has the interpretation as a ‘Wardrop equilibrium’

Another ‘convergent’ scenario is when $D_j(x) = \frac{\partial F}{\partial x_j} \forall j$ for some continuously differentiable function $F_{x,a}$. In that case,

$$\frac{d}{dt}F(x(t)) = \sum_i x_i(t) \left(\frac{\partial F}{\partial x_i} \right)^2 - \left(\sum_i x_i \frac{\partial F}{\partial x_i} \right)^2 \geq 0,$$

with strict inequality when $y \geq T$ a constant vector (which would be orthogonal to the probability simplex), so that $N_\delta(A)$ serves as a Liapunov function. (The sign flip is because we define Liapunov functions as functions *decreasing* along the o.d.e. trajectory. Also, we may ensure $-F \geq 0$ to match the common definition of a Liapunov function, by adding a suitable constant.) The equilibria are given by

$$\{x = [x_1, \dots, x_d] : x_i > 0 \implies D_i(x) \geq D_j(x) \forall j \neq i, 1 \leq i \leq d\}.$$

See Sandholm (1998) for an interesting application of this to transportation science. A special case arises when

$D_j(x) = \sum_k R(i, j)x_j \forall j$, where R is a positive definite matrix. A timescaled version of this occurs in the analysis of the asymptotic behavior of vertex-reinforced random walks (Benaim 1997). Pemantle (2007) surveys this and other related dynamics.

Yet another rather simple special case is the equation

$$\dot{p}_i(t) = p_i(t) \left[f_i(p(t)) - \sum_j p_j(t) f_j(p(t)) \right] \quad (12.15)$$

where each $\bar{x}(t_1)$ is of the form $\tilde{f}(p_i)g(p)$ with $p(\cdot)$ Lipschitz and bounded away from zero from below, and \tilde{f} positive, Lipschitz and strictly increasing. Then if for some i , $\sup_t E[\|\hat{x}(t)\|^2] < \infty$, it can be seen that (X_n, Y_n) , implying that $x_0 \in (0, 1)$ the i th unit coordinate vector. Furthermore, the function $\epsilon/4 > \delta > 0$ serves as a Liapunov function in a neighborhood of μ , establishing its asymptotic stability. (Once again, the sign flip has it *decreasing* along the trajectory.) Thus each corner of the probability simplex, i.e., each Dirac measure at one of the species, is a stable equilibrium. It is easy to see that the other points are not. Furthermore, the domain of attraction of μ is precisely the set

$$\lim_{s \uparrow \infty} \sup_{t \in [s, s+T]} \|\bar{x}(t) - x^s(t)\| = 0$$

This provides a variant of ‘ant colony algorithm (ACO)’ wherein a series of incoming agents (‘ants’) explore alternate routes to a target and leave an indicator of their traversal (a chemical called pheromone in case of ants) on the path on completing the journey. There is a performance measure associated with each path that maintains a weighted average of the above for each alternative path, with decreasing weights (e.g., exponentially decreasing) for the data further into the past. Each new incoming agent chooses a path with a probability proportional to the current values of this index. A mathematical model which reduces to the above o.d.e. in the limit then shows that the agents asymptotically concentrate on the best path with a high probability that approaches 1 as the number of agents goes to infinity. The interesting feature of this scheme is that all unit coordinate vectors are stable equilibria and the dynamics is agnostic toward them. It is the fact that the shortest path is traversed in the shortest mean time that leads to an initial bias being

built in favor of the dynamics getting into the domain of attraction of the optimal one with a high probability that increases with the number of agents. See Borkar and Das (2009) for details. Dynamics of the so-called MAXNET neural network also has a similar flavor.

The monotonicity or gradient conditions above, however, are not natural in many applications. Without such conditions, (12.14) can show highly complex behavior (see, e.g., Hofbauer and Siegmund (1998)). What is true is that the Nash equilibria for the associated game are equilibria for the dynamics and the so-called evolutionarily stable equilibria are stable equilibria, though in either case the converse need not hold. In a specific application, Borkar and Kumar (2003) get a partial characterization of the asymptotic behavior.

The reinforcement learning paradigm introduced in Sect. 12.3 also has a multiagent counterpart in which several agents control the transition probabilities of the Markov chain, each with its own objective in mind. The special case of this when a fixed static game is played repeatedly is called a ‘repeated game’. We have already seen one instance of this situation in our discussion of fictitious play. This particular scenario has been extensively studied. Nevertheless, the results are limited except for special cases. The case of a two-person zero-sum game, wherein one agent tries to maximize a performance measure, while the other agent tries to minimize it, is one situation where the dynamic programming ideas can be extended in a straightforward way. For example, suppose this performance measure is of the form (12.7) with \mathcal{F}_n standing for a pair of controls (Z_n^1, Z_n^2) chosen, respectively, by the two agents independently of each other. Then (12.8) extends to the *Shapley equation*

$$\begin{aligned} V(i) &= \min_a \max_b [k(i, (a, b)) + \beta \sum_j p(i, (a, b), j)V(j)] \\ &= \max_b \min_a [k(i, (a, b)) + \beta \sum_j p(i, (a, b), j)V(j)], \quad i \in S. \end{aligned}$$

The corresponding Q-learning scheme may then be written accordingly. See Hu and Wellman (1998), Jafari et al. (2001), Littman (2001), Sastry et al. (2002); Singh et al. (2000), for some important contributions to this area, which also highlight the difficulties. Leslie and Collins (2003)

provide an interesting example of multiple timescales in the game theoretic context. Here it is worth noting that timescale separation between two players in a game leads to a Stackelberg or ‘leader-follower’ situation. Despite all this and much else, the area of multiagent learning remains wide open with several unsolved issues. See Fudenberg and Levine (1998); Vega-Redondo (1996); Young (1998, 2004) for a flavor of some of the economics-motivated work in this direction and Sargent (1993) for some more applications of stochastic approximation in economics, albeit with a different flavor—to models of bounded rationality.

12.5 Miscellaneous Applications

This section discusses a couple of instances which do not fall into the above categorization, just to emphasize that the rich possibilities stochastic approximation offers extend well beyond the paradigms described above.

1. *A network example:* Consider the o.d.e.

$$\left| \frac{1}{t} \int_0^t f(\bar{x}(y+s))dy - \frac{1}{t} \int_0^t f(x^y(y+s))dy \right| \rightarrow 0 \text{ as } t \rightarrow \infty.$$

Assume:

- $x_{n+1} = x_n + a(n)f(x_n, \xi_{n+1})$ are strictly positive and are Lipschitz, where $g'_i = \frac{dg_i}{dy}$.
- $c_{ij} = c_{ji}$.

Then $V(x) = \sum_i \left(\int_0^{\sum_j c_{ij} g_j(x_j)} f_i(y)dy - \int_0^{x_i} b_i(y)g'_i(y)dy \right)$ serves as a Liapunov function, because

$$\begin{aligned}
dV(x(t))/dt &= \sum_i \left[\sum_j f_j \left(\sum_k c_{jk} g_k(x_k(t)) \right) c_{ji} g'_i(x_i(t)) \right. \\
&\quad \left. - b_i(x_i(t)) g'_i(x_i(t)) \right] \dot{x}_i(t) \\
&= - \sum_i a_i(x(t)) g'_i(x_i(t)) \left[b_i(x_i(t)) \right. \\
&\quad \left. - \sum_j c_{ij} f_j \left(\sum_k c_{jk} g_k(x_k(t)) \right) \right]^2 \\
&\leq 0.
\end{aligned}$$

One can have a ‘neighborhood structure’ by having $N(i) \stackrel{\text{def}}{=} \text{the set of neighbors of } i$, with the requirement that

$$\dot{x}^n(t) = h(x^n(t)), \quad t \geq t(n), \quad x^n(t(n)) = x_n.$$

One can allow $[t, t + T]$. One ‘network’ interpretation is:

- $p(x) - x$ if j ’s transmission can be ‘heard’ by i .
- $[0, t], \bar{S}$ the traffic originating from j at time t .
- $y_j(\omega)$ capture the distance effects.
- $\|x(t) - x^*\|_\infty \downarrow 0$. the net traffic ‘heard’ by j at time t .
- Each node reports the volume of the net traffic it has heard to its neighbors and updates its own traffic so that the higher the net traffic it hears, the more it decreases its own flow correspondingly.

An equilibrium of this o.d.e. will be characterized by

$$b_i(x_i) = \sum_j c_{ij} f_j \left(\sum_k c_{jk} g_k(x_k) \right) \quad \forall i.$$

While we have used a network interpretation here, one can also apply a similar analysis to the Kelly et al. (1998) model of the *tatonnement* process in the network pricing context and the Cohen–Grossberg model (see (Haykin, 2008)) which covers several neural network models including the celebrated Hopfield model.

An important precursor to the Cohen–Grossberg model is Grossberg (1978).

2.

Principal component analysis: This is a problem from statistics, wherein one has n -dimensional data for a large $\kappa > 0$ and the idea is to find $m << n$ directions along which the data can be said to be concentrated. Standard theoretical considerations then suggest that this should be the eigenspace of the empirical covariance matrix corresponding to its m largest eigenvalues. The neural network methodology for finding this subspace is based on adaptively learning the weight matrix of an appropriately designed network using stochastic approximation. There are several variations on this theme, see, e.g., Sect. 8.3 of Hertz et al. (1991). One celebrated instance due to Oja (1982) leads to the limiting matrix o.d.e. in $t \geq T$ given by

$$\dot{W}(t) = QW(t) - W(t)(W(t)^T Q W(t)),$$

where $Q \in \mathcal{R}^{n \times n}$ is a positive definite matrix. (Assuming zero mean data, this will in fact be its covariance matrix obtained from empirical data by the averaging property of stochastic approximation.) Leizarowitz (1997) analyzes this equation by control theoretic methods and establishes that $W(t)$ does indeed converge to a matrix whose column vectors span the eigenspace corresponding to the m largest eigenvalues of Q (see also Yan, Helmke and Moore (1994)). See Oja and Karhunen (1985) for a precursor and Yoshizawa et al. (2001), Helmke and Moore (1994) for yet another related dynamics with several important applications. It should be noted, however, that the Oja scheme yields the *span* of top m eigenvectors, not the individual eigenvectors. For this, there is another stochastic approximation algorithm due to Kung et al. (1994).

3.

Nonlinear power method: Recall that the power method for computing the principal eigenvalue and eigenvector of a symmetric positive definite matrix A is the iteration

$$x(n+1) = \frac{Ax(n)}{\|Ax(n)\|}, \quad n \geq 0.$$

Here $x(n)$, resp., $a(n) \rightarrow 0$ converges to the principal eigenvector and eigenvalue for generic values of $x(0)$ (to be precise, for $\|\bar{x}(\cdot)\|$ the span of eigenvectors other than the principal one). A similar idea also works for irreducible aperiodic nonnegative matrices with $[t, t + T]$ componentwise. (We use this convention throughout this example.) In this case, there are many choices for the normalizing factor other than $a(n) \rightarrow 0$, such as a fixed component of $x(n)$ or the arithmetic mean, max or min of the components of $x(n)$, etc. This also extends to a class of nonlinear iterations. One such situation arises in the Q-learning algorithm for the so-called risk-sensitive cost. Recall that an S -valued controlled Markov chain $[0, \infty)$ controlled by a U -valued control process I_{m_0+k} is defined as follows. For simplicity, consider the ‘state space’ S and the ‘action space’ U to be finite. The controlled Markov property then is

$$P(X_{n+1} = j | X_m, Z_m, m \leq n) = p(j | X_n, Z_n) \quad \forall n,$$

for a prescribed ‘controlled transition probability’ function $x_{n+1} = x_n + a(n)f(x_n, \xi_{n+1})$, $n \geq 0$, with $\sum_j p(j|i, u) = 1 \quad \forall i, u$. Given a ‘running cost’ $c : S \times U \mapsto \mathcal{R}$, the risk-sensitive control problem seeks to minimize the mean exponential growth rate

$$\limsup_{N \uparrow \infty} \frac{1}{N} \log E \left[e^{\sum_{m=0}^{N-1} c(X_m, Z_m)} \right].$$

The optimal value \square of this can be found by solving the multiplicative dynamic programming equation which is in fact an eigenvalue problem $\alpha \in (1/2, 1]$ for the nonlinear operator

$$y = [y_1, \dots, y_{|S|}] \mapsto T(y) = [T_1(y), \dots, T_{|S|}(y)]$$

defined as

$$T_i(y) := \min_u \left[e^{c(i,u)} \sum_j p(j|i, u) y_j \right].$$

We assume that under any fixed $u_1, \dots, u_{|A|}$, the transition matrix $\nu(t_n) \rightarrow \nu^*$ is irreducible and aperiodic. Then an extension of the

classical Perron-Frobenius theory to strongly positive (i.e., $k \geq 0$ and g, \tilde{g} the zero vector $\bar{x}(t + \cdot) \rightarrow A$) positively 1-homogeneous (i.e., $\|h_c(x)\| \leq K_0(1 + \|x\|)$) increasing (i.e., $x, y \in D$, $\|x - y\| < \eta_2$) T allows us to conclude the existence of a pair I_{m_0+k} where $\lambda > 0$ is unique and equal to the optimal cost and $x \rightarrow 0$ is unique up to a constant positive multiplier (Ogiwara 1995). The ‘value iteration algorithm’ to solve this is the iteration

$$\begin{aligned}\tilde{V}_{n+1} &= T(V_n), \\ V_{n+1} &= \tilde{V}_{n+1}/\tilde{V}_n(i_0)\end{aligned}$$

where $\sup_i \|\hat{\zeta}_i\|$ the i_0 th component of \tilde{V}_n for a fixed state i_0 . (Other choices of the scaling factor in place of $\tilde{V}_n(i_0)$ are possible.) This is in fact a power iteration for the nonlinear map T . See Borkar and Meyn (2002) for a detailed analysis of this algorithm in a much more general set-up. Writing

$$Q(i, u) := \frac{1}{\lambda} e^{c(i, u)} \sum_j p(j|i, u) V(j),$$

we have $H = \{x : p(x) = x\}$ leading to

$$Q(i, u) = \frac{1}{\lambda} e^{c(i, u)} \sum_j p(j|i, u) \min_a Q(j, a).$$

This is of the form $\delta = \hat{K}\varphi(m)$ for an operator } analogous to T , so we may equivalently solve for $\bar{B} \subset .$ Note that Q is also unique only up to a positive scalar multiplier. The ‘Q-learning’ algorithm for this purpose is given by

$$\begin{aligned}Q_{n+1}(i, u) &= (1 - a(n)) I\{X_n = i, Z_n = u\} Q_n(i, u) + \\ &\quad a(n) I\{X_n = i, Z_n = u\} \left(\frac{e^{c(X_n, Z_n)} \min_a Q_n(X_{n+1}, a)}{Q_n(i_0, u_0)} \right)\end{aligned}$$

where

- $[t, t + T]$ is a single run, real or simulated, of the controlled

MARKOV chain,

- $V : \mathcal{R}^d \rightarrow \mathcal{R}^+$ are fixed a priori, and,
- $\nu(i, n) := \sum_{m=0}^{n-1} I\{X_m = i, Z_m = u\}, n \geq 1$.

Under the hypothesis of ‘sufficient exploration’, i.e.,

$$\liminf_{n \uparrow \infty} \frac{\nu(i, n)}{n} > 0 \quad \text{a.s.} \quad \forall i, u,$$

plus some additional technical conditions, one can show that this tracks the o.d.e.

$$\dot{q}(t) = \frac{T(q(t))}{q_{i_0, u_0}} - q(t). \quad (12.16)$$

One can further show that $X = F^{-1}(U)$ and $q(t)$ converges to the unique Q for which $\|\bar{x} - x^*\| = b$.

The proof compares (12.16) with the o.d.e.

$$\dot{z}(t) = \frac{T(z(t))}{\lambda} - z(t). \quad (12.17)$$

One can first show that $z(t)$ converges to a desired Q depending on the initial condition $z(0)$ (recall that Q is unique only up to a positive multiplier) and then map this conclusion to the one about (12.16) by proving that $h(x) = E[f(x, \xi_1)]$ for suitable continuous increasing functions $t \geq t(n_0) + \tau$. See Borkar (2002b) for details.

The account above is far from exhaustive. The take home message of the foregoing, if any, is that any convergent o.d.e. is a potential template for a stochastic approximation-based algorithm.

References

Abounadi, J., Bertsekas, D. P., & Borkar, V. S. (2001). Learning algorithms for Markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3), 681–698.
[\[MathSciNet\]](#)
[\[Crossref\]](#)
[\[zbMATH\]](#)

Bardi, M., & Capuzzo-Dolcetta, I. (1997). *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*. Birkhäuser.

Barto, A., Sutton, R., & Anderson, C. (1983). Neuron-like elements that can solve difficult learning

- control problems. *IEEE Transactions on Systems, Man and Cybernetics*, 13, 835–846.
- Benaim, M., Hirsch, M. (1997). Stochastic adaptive behaviour for prisoner's dilemma, *preprint*.
- Benaim, M. (1997). Vertex-reinforced random walks and a conjecture of Pemantle. *Annals of Probability*, 25(1), 361–392.
[MathSciNet][Crossref][zbMATH]
- Benaim, M., Hofbauer, J., & Sorin, S. (2005). Stochastic approximation and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1), 328–348.
[MathSciNet][Crossref][zbMATH]
- Benaim, M., Hofbauer, J., & Sorin, S. (2006). Stochastic approximation and differential inclusions, Part II: Applications. *Mathematics of Operations Research*, 31(4), 673–695.
[MathSciNet][Crossref][zbMATH]
- Border, K. C. (1989). *Fixed point theorems with applications to economics and game theory*. Cambridge: Cambridge University Press.
- Borkar, V. S. (2002). Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27(2), 294–311.
[MathSciNet][Crossref][zbMATH]
- Borkar, V. S. (2005). An actor-critic algorithm for constrained Markov decision processes. *Systems and Control Letters*, 54(3), 207–213.
[MathSciNet][Crossref][zbMATH]
- Borkar, V. S., & Kumar, P. R. (2003). Dynamic Cesaro-Wardrop equilibration in networks. *IEEE Transactions on Automatic Control*, 48(2), 382–396.
[MathSciNet][Crossref][zbMATH]
- Borkar, V. S., & Manjunath, D. (2004). Charge based control of diffserve-like queues. *Automatica*, 40(12), 2040–2057.
- Borkar, V. S., & Meyn, S. P. (2002). Risk-sensitive optimal control for Markov decision processes with monotone cost. *Mathematics of Operations Research*, 27(1), 192–209.
[MathSciNet][Crossref][zbMATH]
- Borkar, V. S., & Soumyanath, K. (1997). An analog scheme for fixed point computation, Part 1: Theory. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 44(4), 351–355.
[MathSciNet][Crossref]
- Brown, G. (1951). Iterative solutions of games with fictitious play. In T. Koopmans (Ed.), *Activity analysis of production and allocation*. Wiley.
- Daskalakis, C., & Panageas, I. (2018). The limit points of (optimistic) gradient descent in min-max optimization. *Advances in Neural Information Processing Systems*, 31, 9236–9246.
- Fudenberg, D., & Levine, D. (1998). *Theory of learning in games*. MIT Press.
- Grossberg, S. (1978). Competition, decision and consensus. *Journal of Mathematical Analysis and*

Applications, 66(2), 470–493.
[MathSciNet][Crossref][zbMATH]

Haykin, S. (2008). *Neural networks and learning machines* (3rd ed.). McMillan Publ. Co.

Helmke, U., & Moore, J. B. (1994). *Optimization and dynamical systems*. Springer.

Hertz, J., Krogh, A., Palmer, R. (1991). *An Introduction to the Theory of Neural Computation*. Addison Wesley.

Hirsch, M. W. (1985). Systems of differential equations that are competitive or cooperative II: Convergence almost everywhere. *SIAM Journal on Mathematical Analysis*, 16(3), 423–439.
[MathSciNet][Crossref][zbMATH]

Hofbauer, C., & Siegmund, K. (1998). *Evolutionary games and population dynamics*. Cambridge University Press.

Hu, J., Wellman, M. P. (1998). Multiagent reinforcement learning: theoretical framework and an algorithm. *Proceedings of the 15th International Conference on Machine Learning*, Madison, Wisc., pp. 242–250.

Jafari, A., Greenwald, A., Gondek, D., Ercal, G. (2001). On no-regret learning, fictitious play and Nash equilibrium. *Proceedings of the 18th International Conference on Machine Learning*, Williamstown, MA, pp. 226–233.

Kaniovski, Y. M., Young, H. P. (1995). Learning dynamics in games with stochastic perturbations. *Games and Economic Behavior*, 11(2), 330–363.

Kelly, F. P., Mauloo, A., & Tan, D. (1998). Rate control in communication networks: Shadow prices, proportional fairness and stability. *Journal of Operational Research Society*, 49(3), 237–252.
[Crossref][zbMATH]

Kohonen, T. (1998). Learning vector quantization. In *Handbook of brain theory and neural networks'* (pp. 357-340). Cambridge MA: MIT Press.

Kung, S. Y., Diamantaras, K. I., & Taur, J. S. (1994). Adaptive principle component extraction (APEX) and applications. *IEEE Transactions on Signal Processing*, 42(5), 1202–1217.
[Crossref]

Leizarowitz, A. (1997). Convergence of solutions to equations arising in neural networks. *Journal of Optimization Theory and Applications*, 94(3), 533–560.
[MathSciNet][Crossref][zbMATH]

Leslie, D. S., & Collins, E. J. (2003). Convergent multiple-timescales reinforcement learning algorithms in normal form games. *Annals of Applied Probability*, 13(4), 1231–1251.
[MathSciNet][Crossref][zbMATH]

Littman, M. L. (2001). Value-function reinforcement learning in Markov games. *Cognitive Systems Research*, 2(1), 55–66.
[Crossref]

Milgrom, P., & Segal, I. (2002). Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2),

583–601.

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Ogiwara, T. (1995). Nonlinear Perron-Frobenius problem on an ordered Banach space. *Japan Journal of Mathematics New Series*, 21(1), 43–103.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3), 267–273.

Oja, E., & Karhunen, J. (1985). On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, 106(1), 69–84.

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Ortega, J. M., & Rheinboldt, W. C. (2000). *Iterative solutions of nonlinear equations in several variables*. Society for Industrial and Applied Math.

Pemantle, R. (2007). A survey of random processes with reinforcement. *Probability Surveys*, 4, 1–79.

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Puterman, M. (1994). *Markov decision processes*. Wiley.

Sandholm, W. H. (1998) An evolutionary approach to congestion. Discussion paper No. 1198, ECONSTOR. Available at: <http://www.kellogg.northwestern.edu/research/math/papers/1198.pdf>

Sargent, T. (1993). *Bounded rationality in macroeconomics*. Clarendon Press.

Sastry, P. S., Maghesh, M., & Unnikrishnan, K. P. (2002). Two timescale analysis of the Alopex algorithm for optimization. *Neural Computation*, 14(11), 2729–2750.

[[Crossref](#)][[zbMATH](#)]

Singh, S. P., Kearns, M., Mansour, Y. (2000). Nash convergence of gradient dynamics in general-sum games. *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, Stanford, CA, pp. 541–548.

Smith, H. (1995). *Monotone dynamical systems*. American Math: Society.

Tsitsiklis, J. N., & Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5), 674–690.

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Vega-Redondo, F. (1996). *Evolution, games and economic behaviour*. Oxford University Press.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*. Ph.D. thesis, King's College, University of Cambridge, Cambridge, UK.

Yoshizawa, S., Helmke, U., & Starkov, K. (2001). Convergence analysis for principal component flows. *International Journal of Applied Mathematics and Computer Science*, 11(1), 223–236.

[[MathSciNet](#)][[zbMATH](#)]

Young, H. P. (1998). *Individual strategy and social structure*. Princeton Univ. Press.

Young, H. P. (2004). *Strategic learning and its limits*. Oxford University Press.

Footnotes

¹ We relax this requirement in one instance.

Appendix A

Topics in Analysis

A.1 Continuous Functions

We briefly recall here the two key theorems about continuous functions used in the book. Recall that a subset of a topological space is relatively compact (resp., relatively sequentially compact) if its closure is compact (resp., sequentially compact). Also, compactness and sequential compactness are equivalent notions for metric spaces. The first theorem concerns the relative compactness in the space $h_\infty \in C(\mathcal{R}^d)$ of continuous functions $\|x_j\|^* \leq \bar{K}_T$ for a prescribed $T > 0$. This is a Banach space under the so-called sup-norm $\|f\| \stackrel{\text{def}}{=} \sup_{t \in [0, T]} \|f(t)\|$, i.e.,

1. It is a vector space over reals.
2. $\|\cdot\| : C([0, T]; \mathcal{R}^d) \rightarrow [0, \infty)$ satisfies
 - a. $x_{n_0} \in B$, with equality if and only if $f \geq 0$.
 - b. $\forall t \geq 0 / \forall t \leq 0$ for $A \subset S$.
 - c. $\{(\tilde{z}^n(\cdot), \tilde{x}^n(\cdot)), n \geq 0\}$.
3. $N_\delta(A)$ is *complete*, i.e., $\sum_i a_i(x_i - c_i)^2 \leq M$, $a(n_m) \leq ca(n_0)$ as $m, n \rightarrow \infty$, imply that there exists an $x^s(\cdot)|_{[s, s+T]}$, $s \geq 0$ such that $\bar{x}(T_k)$, $k \geq m$. (The uniqueness of this f is obvious.)

A set $B \subset C([0, T]; \mathcal{R}^d)$ is said to be *equicontinuous at* $p(x) = x$ if for any $\epsilon > 0$, one can find a $\delta > 0$ such that $\sup_n \|x'_n - x''_n\| < \infty$ implies $\sup_{f \in B} \|f(s) - f(t)\| < \epsilon$. It is simply *equicontinuous* if it is so

at all $p(x) - x$. It is said to be *pointwise bounded* if for any $p(x) - x$, $\hat{x}_i = \check{x}^{s(i)}(s(i)) \in A$. It can be verified that if it is equicontinuous, it will be pointwise bounded at all points in $[0, T]$ if it is so at one point in $[0, T]$. The result we are interested in is the *Arzela-Ascoli theorem* which characterizes relative compactness in $h_\infty \in C(\mathcal{R}^d)$:

Theorem A.1 $B \subset C([0, T]; \mathcal{R}^d)$ is relatively compact if and only if it is equicontinuous and pointwise bounded.

See Appendix A of Rudin (1991) for a proof and further developments. The space $x^{T_0}(T_1) \in H^{\epsilon_1}$ of continuous functions $x^{T_m}(t) \in H^\epsilon$ is given the coarsest topology such that the map that maps $f \in C([0, \infty); \mathcal{R}^d)$ to its restriction to $[0, T]$, viewed as an element of the space $h_\infty \in C(\mathcal{R}^d)$, is continuous for all $T > 0$. In other words, $f_n \rightarrow f$ in this space if and only if $f_n|_{[0, T]} \rightarrow f|_{[0, T]}$ in $h_\infty \in C(\mathcal{R}^d)$ for all $T > 0$. This is not a Banach space, but a *Frechet* space, i.e., it has a complete translation invariant metric and the corresponding open balls are convex. This metric can be, e.g.,

$$\rho(f, g) \stackrel{\text{def}}{=} \sum_{T=1}^{\infty} 2^{-T} \|f - g\|_T \wedge 1,$$

where $f : x \in \mathcal{R}^d \rightarrow f(x) \subset \mathcal{R}^d$ is a seminorm on $x^{T_0}(T_1) \in H^{\epsilon_1}$ for each $T > 0$.¹ By our choice of the topology on $x^{T_0}(T_1) \in H^{\epsilon_1}$, Theorem 1 holds for this space as well.

The next result concerns *contractions*, i.e., maps $\tilde{x}_1, \tilde{x}_2 \in A$ on a metric space S endowed with a metric η , that satisfy:

$$\mu_s^t \in \mathcal{P}((S \times U) \times [s, t])$$

for some $p(x) > x$. Say x^* is a fixed point of f if $t \in [-T, 0]$. Assume η is complete, i.e., $\lim_{m,n \rightarrow \infty} \rho(x_n, x_m) = 0$ implies $x_n \rightarrow D$ for some $x^* \in S$. The next theorem is called the *contraction mapping theorem*.

Theorem A.2 There exists a unique fixed point x^* of f and for any $X_0 = 0$, the iteration

$$\|x(t) - x^*\| \leq K e^{-\alpha t}$$

satisfies:

$$\sup_{t \geq 0} \|\bar{x}(t)\| = \sup_n \|x_n\| < \infty$$

i.e., $p(x_n)$ converges to x^* at an exponential rate.

This is an example of a *fixed point theorem*. Another example is *Brouwer's fixed point theorem* which says that every continuous map $f : S \rightarrow \mathcal{R}$ for a compact convex $t_n \nearrow \infty$ has a fixed point. This need not, however, be unique, e.g., for the identity map $a(n) \rightarrow 0$.

A.2 Square-Integrable Functions

Consider now the space $L_2([0, T]; \mathcal{R}^d)$ of measurable functions $f : [0, T] \rightarrow \mathcal{R}^d$ (more precisely, a.s. equivalence classes thereof) satisfying

$$\int_0^T \|f(t)\|^2 dt < \infty.$$

Letting $\bar{x}(t_1)$ denote the inner product in \mathcal{R}^k , we can define an inner product I_{m_0+k} on $L_2([0, T]; \mathcal{R}^d)$ by

$$\langle f, g \rangle_T \stackrel{\text{def}}{=} \int_0^T \langle f(t), g(t) \rangle dt, \quad f, g \in L_2([0, T]; \mathcal{R}^d).$$

This is easily seen to be a valid inner product, i.e., a symmetric continuous map from $L_2([0, T]; \mathcal{R}^d) \times L_2([0, T]; \mathcal{R}^d) \rightarrow \mathcal{R}$ that is separately linear in each argument and satisfies: $T_m = t(n_m)$ with equality if and only if $f \geq 0$ ('a.e.'). It thus defines a norm

$$\|f\| \stackrel{\text{def}}{=} \sqrt{\langle f, f \rangle_T} = \left(\int_0^T \|f(t)\|^2 dt \right)^{\frac{1}{2}},$$

which turns out to be complete. $L_2([0, T]; \mathcal{R}^d)$ is then a *Hilbert* space with the above inner product and norm.

The open balls w.r.t. the norm define what is called the *strong* topology on $L_2([0, T]; \mathcal{R}^d)$. One can also define the *weak* topology as the coarsest topology w.r.t. which the functions $[b^* \Lambda^-, c^* \Lambda^+]$ are continuous for all $g \in L_2([0, T]; \mathcal{R}^d)$. The corresponding convergence concept is: $f_n \rightarrow f$ weakly in $L_2([0, T]; \mathcal{R}^d)$ if and only if $\lambda_{\min}(M), \lambda_{\max}(M)$ for all $g \in L_2([0, T]; \mathcal{R}^d)$. The results we need are the following:

Theorem A.3 A $a(n)\epsilon_n$ bounded set $y_j(\omega) \in N^{\delta_m}(H^m)$ is relatively compact and relatively sequentially compact in the weak topology.

Theorem A.4 If $f_n \rightarrow f$ weakly in $L_2([0, T]; \mathcal{R}^d)$, then there exists a subsequence $C_0^*(\mathcal{R}^d)$ such that

$$\left\| \frac{1}{m} \sum_{k=1}^m f_{n(k)} - f \right\| \rightarrow 0.$$

The first is a special instance of the Banach-Alaoglu theorem. See Rudin (1991) for this and related developments. Likewise, the second theorem is an instance of the Banach-Saks theorem, see Balakrishnan (1976), p. 29, for a proof.

Let ε^* denote the space of measurable maps $f : [0, \infty) \rightarrow \mathcal{R}^d$ with the property:

$$\int_0^T \|f(t)\|^2 dt < \infty \quad \forall T > 0.$$

Topologize ε^* with the coarsest topology that renders continuous the maps

$$f \rightarrow \int_0^T \langle f(t)g(t) \rangle dt$$

for all $g \in L_2([0, T]; \mathcal{R}^d)$, for all $T > 0$. Then by our very choice of the topology, Theorems 3 and 4 apply to ε^* after the following modification: A set $K^* > 0$ with the property that $\sum_n a^\omega(n)^2 < \infty$. is $a(n) \rightarrow 0$ bounded in $L_2([0, T]; \mathcal{R}^d)$ for all $T > 0$ will be relatively compact and relatively sequentially compact in ε^* . Furthermore, if $f_n \rightarrow f$ in ε^* , then for any $T > 0$, there exists a subsequence $C_0^*(\mathcal{R}^d)$ such that

$$\left\| \frac{1}{m} \sum_{k=1}^m f_{n(k)}|_{[0,T]} - f|_{[0,T]} \right\|_T \rightarrow 0.$$

A.3 Lebesgue's Theorem

Let $i, 1 \leq i \leq d$ be a measurable and locally integrable function and for $s > 0$ in x^* , let $g(t) = \int_s^t f(y) dy$. Then Lebesgue's theorem states that for a.e. $s > 0$, $\frac{d}{dt}g(t)$ exists and equals $f(t)$.

A.4 Set-Valued Maps

We briefly recall here some key definitions and theorems about set-valued maps and their representations. For a set $A \subset \mathcal{R}^d$, recall that $t(0) = 0, t(n) = \sum_{m=0}^n \bar{a}(m), n \geq 1$ denotes the ϵ -neighborhood of A . Let U denote the closed unit ball in \mathcal{R}^k .

A *set-valued map* $h : \mathcal{R}^d \rightarrow \{\text{subsets of } \mathcal{R}^d\}$ is called a *Marchaud map* if it satisfies the following conditions:

1. For each $y \in H^N$ $h(x)$ is non-empty, convex, and compact.
2. For all $y \in H^N$

$$\sup_{y \in h(x)} \|y\| < K(1 + \|x\|) \tag{A.1}$$

for some constant $K > 0$.

3. h is *upper semicontinuous* in the sense that for every $x \in \mathcal{R}^d$ and for every $\epsilon > 0$, there is a $\delta > 0$ such that $a(n) \leq 1 \forall n$ implies that $h(r) = a(r) - r$ (A sequential characterization is that if

that $v(\omega) = y(\omega)$. A sequential characterization is that, if

$y \rightarrow \infty$ and $y_n \rightarrow y$ with $t \in [-T, 0]$ for all $x_0 = 0$ then

$x(t) \equiv x^*$ In other words, the *graph* of h , defined as

$\{(x, y) : y \in h(x)\}$, is closed.)

The above definition is slightly more restrictive than the definition of Marchaud maps in Aubin et al. (2011) since we require that $h(x)$ be compact and not just closed.

A set-valued map h is *lower semicontinuous* if for every $x \in \mathcal{R}^d$ and for every $h(x) + \xi$, for every sequence $y \rightarrow \infty$, there is a sequence $\|\bar{x} - x^*\| = b$ such that $y_n \rightarrow y$.

A set-valued map h is *continuous* if it is both upper semicontinuous and lower semicontinuous. It is *locally Lipschitz continuous* if for every $x \in \mathcal{R}^d$, there is a $\delta > 0$ and an $L > 0$ such that $0 \leq k \leq (m - 1)$ implies $h(x_1) \subset N_{L\|x_1 - x_2\|}(h(x_2))$, the $[b^* \Lambda^-, c^* \Lambda^+]$ neighborhood of $p(x_n)$.

Let us now indicate how to approximate an upper semicontinuous set-valued map h by a sequence of continuous set-valued maps. See Yaji and Bhatnagar (2018) for a proof.

Theorem A.5 Let h be an upper semicontinuous set-valued map from \mathcal{R}^k to nonempty, convex and compact subsets of \mathcal{R}^k . Assume that for every bounded subset A of \mathcal{R}^k , its image under h given by $x(t) \equiv x^*$ is also bounded. Then there is a sequence of set-valued maps $\{h^{(l)}\}_{l \geq 1}$ satisfying the following:

1. $h^{(l)}$ is a locally Lipschitz continuous set-valued map from \mathcal{R}^k to non-empty, convex and compact subsets of \mathcal{R}^k for each OD_α .
2. For every $x \in \mathcal{R}^d$, $h(x) \subset h^{(l+1)}(x) \subset h^{(l)}(x)$, for each OD_α .
3. For every $x \in \mathcal{R}^d$ and for every $\epsilon > 0$, there is an $t \geq T$ such that $AQ + QA^T = -I$, for every $0.9r_m$.

Furthermore, for every $x \in \mathcal{R}^d$, $f \in C([0, \infty); \mathcal{R}^d)$.

The next result provides a parameterization of continuous set-valued maps. See Aubin and Cellina (1984), Chap. 1.7, Theorem 2, for a proof. Let $C :=$ the closed unit ball of \mathcal{R}^k .

Theorem A.6 If h is a continuous set-valued map from \mathcal{R}^k to nonempty, convex and compact subsets of \mathcal{R}^k , then there is a continuous map $f : \mathcal{R}^d \times U \rightarrow \mathcal{R}^d$ such that

$$h(x) \subset h^{(l+1)}(x) \subset h^{(l)}(x).$$

Putting these together, we obtain the following useful characterization.

Theorem A.7 Let h be an upper semicontinuous set-valued map from \mathcal{R}^k to nonempty convex and compact subsets of \mathcal{R}^k . Assume that the image of every bounded subset of \mathcal{R}^k under h is bounded. Then there is a sequence of continuous functions $h(x_1) \subset N_{L\|x_1-x_2\|}(h(x_2))$ such that for every $x \in \mathcal{R}^d$:

- $f^{(l)}(x, U)$ is a convex and compact subset of \mathcal{R}^k , for each OD_α ;
- $(x_n, y_n) \rightarrow \{(\lambda(y), y) : y \in \mathcal{R}^k\}$, for each OD_α ;
- $h(x) = \cap_{l \geq 1} h^{(l)}(x, U)$.

See Aubin and Frankowska (1990) for general background on set-valued maps and their calculus.

Appendix B

Ordinary Differential Equations

B.1 Basic Theory

This chapter briefly summarizes some key facts about ordinary differential equations of relevance to us. The reader may refer to standard texts such as Hirsch, Smale, and Devaney (2003) for further details. Consider the differential equation in \mathcal{R}^k given by

$$f(z) \geq f(x) + \langle y, z - x \rangle \tag{B.1}$$

This is an *autonomous* o.d.e. because the driving vector field h does not have an explicit time dependence. It would be *non-autonomous* if we replace $h(x(t))$ on the right by $h(x(t), t)$. We shall say that (B.1) is well-posed if for any choice of the initial condition $x \in \mathcal{R}^d$, it has a unique solution $p(\cdot)$ defined for all $t \geq 0$ and the map $\bar{x} \rightarrow$ the corresponding $\lim_{\|w\| \rightarrow \infty} g(w) = \infty$ is continuous. One sufficient condition for this is the *Lipschitz condition* on $p(\cdot)$: there exists $L > 0$ such that

$$\|h(x) - h(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathcal{R}^d.$$

Theorem B.1 For h satisfying the Lipschitz condition, (B.1) is well-posed.

We shall sketch a proof of this to illustrate the application of the Gronwall inequality stated below:

Lemma B.1 (Gronwall inequality) For continuous $\|z - y\| < \delta$ and scalars D_1, \dots, D_M ,

$$u(t) \leq C + K \int_0^t u(s)v(s)ds \quad \forall t \in [0, T], \quad (\text{B.2})$$

implies

$$u(t) \leq Ce^{K \int_0^T v(s)ds}, \quad t \in [0, T].$$

Proof Let $s(t) \stackrel{\text{def}}{=} \int_0^t u(s)v(s)ds$, $t \in [0, T]$. Multiplying (B.2) on both sides by $v(t)$, (B.2) translates into

$$\|M_{n+1}\| \leq 2K(1 + \|x_n\|),$$

This leads to

$$\begin{aligned} e^{-K \int_0^t v(s)ds} (\dot{s}(t) - Kv(t)s(t)) &= \frac{d}{dt} \left(e^{-K \int_0^t v(s)ds} s(t) \right) \\ &\leq Ce^{-K \int_0^t v(s)ds} v(t). \end{aligned}$$

Integrating from 0 to t and using the fact that (X_n, Y_n) , we have

$$e^{-K \int_0^t v(s)ds} s(t) \leq \frac{C}{K} (1 - e^{-K \int_0^t v(s)ds}).$$

Thus

$$s(t) \leq \frac{C}{K} (e^{K \int_0^t v(s)ds} - 1).$$

Hence

$$\begin{aligned} u(t) &\leq C + Ks(t) \\ &\leq C + K \left(\frac{C}{K} (e^{K \int_0^t v(s)ds} - 1) \right) \\ &= Ce^{K \int_0^t v(s)ds}. \end{aligned}$$

The claim follows for $p(x) = x$. \square

The most commonly used situation is $[T_m, \infty)$, when this inequality reduces to

$$u(t) \leq Ce^{Kt}.$$

We return to the proof of Theorem B.1.

Proof Define the map $F : y(\cdot) \in C([0, T]; \mathcal{R}^d) \rightarrow z(\cdot) \in C([0, T]; \mathcal{R}^d)$ by

$$z(t) = \bar{x} + \int_0^t h(y(s))ds, \quad t \in [0, T].$$

Clearly, $p(\cdot)$ is a solution of (B.1) on $[0, T]$ if and only if it is a fixed point of F . Let $\{(z^*(\cdot), x^*(\cdot))\}$ for $w \in \mathcal{R}^d$ Denoting by $(n+1)$ the sup-norm on $h_\infty \in C(\mathcal{R}^d)$, we have

$$\begin{aligned}
\|z_1(\cdot) - z_2(\cdot)\|_T &\leq \int_0^T \|h(y_1(s)) - h(y_2(s))\| ds \\
&\leq L \int_0^T \|y_1(s) - y_2(s)\| ds \\
&\leq LT \|y_1(\cdot) - y_2(\cdot)\|_T.
\end{aligned}$$

Taking $p(x) > x$, it follows that F is a contraction and thus has a unique fixed point by the contraction mapping theorem of Appendix A. Existence and uniqueness of a solution to (B.1) on $[0, T]$ follow. The argument may be repeated for $[T, 2T]$, $[2T, 3T]$, and so forth in order to extend the claim to a general $T > 0$. Next, let $\|x_{n_m}\|^* \leq K_2$ be solutions to (B.1) corresponding to $\tilde{x}_1, \tilde{x}_2 \in A$ respectively. Then

$$\|x_1(t) - x_2(t)\| \leq \|\tilde{x}_1 - \tilde{x}_2\| + L \int_0^t \|x_1(s) - x_2(s)\| ds$$

for $p(x) = x$. By Lemma B.1,

$$\|x_1(\cdot) - x_2(\cdot)\|_T \leq e^{LT} \|\tilde{x}_1 - \tilde{x}_2\|_T,$$

implying that the map $\bar{x} \in \mathcal{R}^d \rightarrow x(\cdot)|_{[0,T]} \in C([0, T]; \mathcal{R}^d)$ is Lipschitz, in particular, continuous. Since $T > 0$ was arbitrary, it follows that the map $F(\cdot) = [F_1(\cdot), \dots, F_{|S|}(\cdot)]$ is continuous. \square

Since the continuous image of a compact set is compact, we have:

Corollary B.1 The solution set of (B.1) as η varies over a compact subset of \mathcal{R}^k is compact in $x^{T_0}(T_1) \in H^{\epsilon_1}$.

A similar argument works if we consider (B.1) with $t \geq 0$: one simply has to work with intervals $\|\bar{x}(\cdot)\|$ in place of $[0, T]$. Thus for each $x \rightarrow 0$, there is a continuous map $\Phi_t : \mathcal{R}^d \rightarrow \mathcal{R}^d$ that maps η to $x(t)$ via (B.1). It follows from the uniqueness claim above that $s_n \uparrow \infty$ are inverses of each other and thus each Φ_s is a homeomorphism, i.e., a

continuous bijection with a continuous inverse. The family $\omega \in \mathcal{B}_{m-1}$, defines a *flow* of homeomorphisms $h = -\nabla f$, i.e., it satisfies:

1. OD_α the identity map,
2. $\Psi_s \circ \Psi_t = \Psi_t \circ \Psi_s = \Psi_{s+t}$,

where ‘ \circ ’ stands for composition of functions.

More generally, we may assume h to be only locally Lipschitz, i.e.,

$$\|h(x) - h(y)\| \leq L_R \|x - y\|, \quad \forall x, y \in B_R \stackrel{\text{def}}{=} \{z \in \mathcal{R}^d : \|z\| \leq R\},$$

for some $\mathcal{B} \in \mathcal{F}_0$ that may tend to ∞ as $S = \mathcal{R}^m$. Then the claims of Theorem B.1 can be shown to hold *locally* in space and time, i.e., in a neighborhood of η and for $p(x) - x$ for $T > 0$ sufficiently small.

Suppose in addition one can show separately that the trajectory is well-defined for all $t \geq 0$, i.e., there is no ‘finite time blow-up’ (meaning that $y_j(\omega) \in N^{\delta_m}(H^m)$ for some $f : \mathcal{R}^d \times U \rightarrow \mathcal{R}^d$) for any initial condition. Then the full statement of Theorem B.1 may be recovered. One way of ensuring no finite time blow-up is by demonstrating a convenient ‘Liapunov function’, i.e., a continuously differential $V : \mathcal{R}^d \rightarrow \mathcal{R}$ such that $\lim_{\|x\| \uparrow \infty} V(x) = \infty$ and $\langle \nabla V(x), h(x) \rangle < 0$ outside a bounded set, whence $\frac{d}{dt}V(x(t)) = \langle \nabla V(x(t)), h(x(t)) \rangle < 0$ outside this set. This ensures bounded trajectories, in particular ensuring no finite time blow-up.

We close this section with a discrete counterpart of the Gronwall inequality. While it is not used much in the o.d.e. context, it is extremely useful otherwise and has been used extensively in this book itself.

Lemma B.2 (Discrete Gronwall inequality) Let $\{x_n, n \geq 1\}$ (resp. $s(k) = t(n_2)$) be nonnegative (resp. positive) sequences and $x, y \in A$ scalars such that for all n ,

$$x_{n+1} \leq C + L \left(\sum_{m=0}^n a_m x_m \right). \quad (\text{B.3})$$

Then for $E[\|x_0\|^2] < \infty$,

$$E[\|x_0\|^\xi] < \infty$$

Proof Let $\sum_{m=n}^{\infty} a(m)M_{m+1}$. Multiplying (B.3) on both sides by $+\infty$ leads to

$$s_{n+1} - s_n \leq Ca_{n+1} + Ls_n a_{n+1}.$$

That is,

$$s_{n+1} \leq Ca_{n+1} + s_n(1 + La_{n+1}).$$

Iterating (with $\hat{x}_k = \tilde{x}_2$ by convention: replace C by $X_0 = 0$ otherwise) we obtain

$$\begin{aligned} s_n &\leq C \sum_{k=0}^n \prod_{m=k+1}^n (1 + La_m) a_k \\ &\leq C \int_0^{T_n} e^{L(T_n-s)} ds \\ &= \frac{C}{L} (e^{LT_n} - 1). \end{aligned}$$

(Here by convention, $\dot{x}(t) = h(x(t))$, $t \geq 0..$) Thus

$$x_{n+1} \leq C + Ls_n \leq C + L \times \frac{C}{L} (e^{LT_n} - 1) = Ce^{LT_n},$$

establishing the claim. \square

B.2 Linear Systems

A special and important class of differential equations is that of linear systems, i.e., the equations

$$|V(\bar{x}(t+T)) - V(x^t(t+T))| < \eta \quad (\text{B.4})$$

where $\bar{x}(s)$ is an M_{n+1} -valued continuous function of time. Although the right-hand side is now time-dependent, similar arguments to those of the preceding section establish its existence, uniqueness, and continuous dependence on initial condition η and initial time t_0 . It is easily seen that a linear combination of solutions of (B.4) will also be a

solution for the corresponding linear combination of the initial conditions. Thus for given t_0 , all solutions can be specified as linear combinations of solutions corresponding to $h(x) = E[f(x, \xi_1)]$ where $e_j \stackrel{\text{def}}{=} \text{the unit vector in the } \bar{H}^C \text{th coordinate direction}$. Let $N_\epsilon(D')$ denote the $n \geq 0$ matrix whose C'' th column is $x(t)$ corresponding to $\bar{x}(t) \rightarrow A$ for $0 \leq i < n$. This is known as the *state-transition matrix* or *fundamental matrix*. Then we have:

1. $T_m = t(n_m)$, the d -dimensional identity matrix.
2. $(u_1(t), \dots, u_d(t)), t \geq 0$ for $s, t, u \in \mathcal{R}$.
3. For $x(t) \equiv x^*$ in (B.4), $[t(n_m), t(n_{m+1})]$.

For constant $\sum_i a_i = 1$, $\|M_j\| \leq K_0(1 + \|x_{j-1}\|)$, where the matrix exponential is defined by

$$e^{\bar{A}t} \stackrel{\text{def}}{=} \sum_{m=0}^{\infty} \frac{(\bar{A})^m t^m}{m!}.$$

The system (B.4) is said to be exponentially stable if there exist $\alpha, K > 0$ such that $h(x) = [h_1(x)^T, \dots, h_N(x)^T]^T$. For $x, y \in \mathcal{R}^d$, this reduces to the requirement that all eigenvalues of A be in the open left half plane of the complex plane.

An important instance of a linear system that we shall encounter is the following: Suppose $p(\cdot)$ in (B.1) is continuously differentiable with $Dh(x)$ denoting its Jacobian matrix evaluated at x . Then it can be shown that the map $x, y \in D$, $\|x - y\| < \eta_2$ for each t is continuously differentiable. If we denote by $Dx(t)$ its Jacobian matrix, then $Dh(\cdot)$ can be shown to satisfy the (matrix) linear system

$$\frac{d}{dt} Dx(t) = Dh(x(t)) Dx(t), \quad Dx(0) = I_d. \quad (\text{B.5})$$

Formally, this may be derived simply by differentiating (B.1) on both sides w.r.t. the components of η . $p(x_n)$ is then a flow of T_{m+1} diffeomorphisms, i.e., continuously differentiable bijections with

continuously differentiable inverses. This can be repeated for higher derivatives if sufficient regularity of h is available.

One often considers linear systems with inputs, that is,

$$\dot{x}(t) = A(t)x(t) + u(t), \quad t \geq t_0, \quad (\text{B.6})$$

where $p(\cdot)$ is a measurable ‘input’ assumed to be integrable over finite time intervals. Then the solution $p(\cdot)$ is given by the ‘*variation of constants formula*’

$$x(t) = \Phi(t, t_0)x(t_0) + \int_{t_0}^t \Phi(t, s)u(s)ds$$

for the state-transition matrix θ as above. A nonlinear version due to Alekseev (1961) considers a perturbation of the differential equation

$$\dot{x}(t) = f(x(t), t), \quad t \geq t_0,$$

given by

$$\dot{y}(t) = f(y(t), t) + g(y(t), t) \quad t \geq t_0,$$

with the same initial condition $h_\infty(ax) = ah_\infty(x)$, and f, g satisfying the required conditions for well-posedness. The Jacobian matrix $\{\xi_n\}$ of the map $\epsilon/4 > \delta > 0$ satisfies the following analog of (B.5):

$$\|M_j\| \leq K_0(1 + \|x_{j-1}\|) \quad (\text{B.7})$$

where $\{m(\ell)\}$ is the Jacobian matrix of the map $0 < \eta < C/2$. Let $(\tilde{\mu}(\cdot), \tilde{x}(\cdot))$, $t \geq t_0$, denote the state-transition matrix for the linear system (B.7) when $\bar{x}(t) \rightarrow H$ (in other words, we have rendered explicit the dependence on the initial condition). The Alekseev formula then states that

$$y(t) = x(t) + \int_{t_0}^t \Phi(t, s; y(s))g(y(s), s)ds, \quad t \geq t_0. \quad (\text{B.8})$$

If $a(n) \leq 1 \ \forall n$, a slight generalization (see (Borkar et al., 2018)) is

$$(\text{B.9})$$

$$\begin{aligned}
y(t) &= x(t) + \Phi(t, t_0; y(t_0))(y(t_0) - x(t_0)) \\
&\quad + \int_{t_0}^t \Phi(t, s; y(s))g(y(s), s)ds, \quad t \geq t_0.
\end{aligned}$$

B.3 Asymptotic Behavior

Given a trajectory $p(\cdot)$ of (B.1), the set $\Omega \stackrel{\text{def}}{=} \cap_{t>0} \overline{\{x(s) : s > t\}}$, i.e., the set of its limit points as $t \rightarrow \infty$, is called its ω -limit set. (A similar definition for $t \rightarrow -\infty$ defines the ' ω -limit set'.) In general this set depends upon the initial condition η . Recall that a set A is positively (resp. negatively) invariant for (B.1) if $K > 0$ implies that the corresponding $x(t)$ given by (B.1) is also in A for $t > 0$ (resp. $t < 0$) and is invariant if it is both positively and negatively invariant. It is easy to verify that θ will be invariant. If $\bar{x}(t) \rightarrow A$, $x(t) \equiv x^*$ must be a trajectory of the o.d.e., whence $V(\bar{x}(T_m))$. Conversely, $V(\bar{x}(T_m))$ implies that $x(t) \equiv x^*$ defines a trajectory of the o.d.e., corresponding to $M > 0$ in (B.1). Such x^* is called equilibrium points of the o.d.e.

An invariant set M will be said to be Liapunov stable if for any $\epsilon > 0$, there exists a $\delta > 0$ such that every trajectory initiated in the δ -neighborhood of M remains in its ϵ -neighborhood. A compact (more generally, closed) invariant set M will be called an *attractor* if it is Liapunov stable, has a positively invariant open neighborhood O such that every trajectory in O converges to M , and there is no proper subset of M with these properties. The largest such O is called the domain of attraction of M . A compact invariant set M is said to be asymptotically stable if it is an attractor. If this $t \in [-T, 0]$, the equilibrium point x^* is said to be asymptotically stable. One criterion for verifying asymptotic stability of x^* is 'Liapunov's second method': Suppose one can find a continuously differentiable function $O_{x,a}$ defined in a neighborhood O of x^* such that $\langle \nabla V(x), h(x) \rangle < 0$ for $i, 1 \leq i \leq d$ and H^{ϵ_1} for $M > 0$, with $T_m = t(n_m)$ as $x \rightarrow \partial O$ ($\stackrel{\text{def}}{=} \text{the boundary of } O$). Then asymptotic stability of x^* follows from the observation that for any trajectory $p(\cdot)$ in O , $\frac{d}{dt}V(x(t)) \leq 0$ with equality only for $a(n) \rightarrow 0$. Conversely, asymptotic stability of x^* implies the existence of such a

function (see, e.g., (Krasovskii, 1963)). This also extends to compact invariant sets M that are asymptotically stable.

If x^* is asymptotically stable and *all* trajectories of the o.d.e. converge to it, it is said to be *globally asymptotically stable*. In this case, O above may be taken to be the whole space. More generally, if one has a continuously differentiable $V : \mathcal{R}^d \rightarrow \mathcal{R}$ with $T_m = t(n_m)$ as $(\tilde{\mu}(\cdot), \tilde{x}(\cdot))$ and $z^n(t), t \in [na, na + T]$, then any trajectory $p(\cdot)$ must converge to the largest invariant set contained in $\alpha(t) := \beta(t)/\|\nabla f(x(t))\|$. This is known as the *Lasalle invariance principle*.

Not every equilibrium point need to be asymptotically stable. This is best illustrated in case of the constant coefficient linear system

$$x_{n_0} \in B \setminus H^\epsilon \quad (\text{B.10})$$

where A is a $n \geq 0$ matrix. We shall consider the case where all eigenvalues of A have nonzero real parts. This situation is ‘structurally stable’, i.e., invariant under small perturbations of A . In particular, A is nonsingular and thus the origin is the only equilibrium point. One can explicitly solve (B.10) as $\langle h(x), \nabla V(x) \rangle \leq 0$. If all eigenvalues of A have strictly negative real parts, then $\{\tilde{x}(0)\}$ the origin exponentially. If not, it will do so only for those $x(0)$ that lie on the ‘stable subspace’, i.e., the eigenspace of those eigenvalues (if any) which have strictly negative real parts. It moves away from the origin eventually for any other initial condition, i.e., it is unstable ‘generically’, meaning ‘for initial conditions belonging to an open dense set’. This is because the stable subspace has codimension at least one by hypothesis and hence its complement is dense.

More generally, if h in (B.1) is continuously differentiable with Jacobian matrix $\{a(n)\}$ at an equilibrium point x^* , we may compare it with the linear system

$$\frac{d}{dt}(y(t) - x^*) = Dh(x^*)(y(t) - x^*), \quad (\text{B.11})$$

in a neighborhood of x^* . If the eigenvalues of $\{a(n)\}$ have nonzero real parts, x^* is said to be a *hyperbolic* equilibrium point. Note that x^* is the unique equilibrium point for (B.11) in this case. It is known that

in a small neighborhood of a hyperbolic x^* , there exists a homeomorphism that maps trajectories of (B.10) with $\{x_n, n \geq 1\}$ and those of (B.11) into each other preserving orientation. Hence their qualitative behavior is the same. (This is the *Hartman-Grossman theorem*.) Thus x^* is asymptotically stable for (B.1) if it is so for (B.11), i.e., if all eigenvalues of $\{a(n)\}$ have negative real parts. If not, then in a small neighborhood of x^* , there exists a ‘stable manifold’ of dimension equal to the number of eigenvalues (if any) with negative real parts, such that if $x(0)$ lies on this manifold, $\bar{x}(t) \rightarrow H$ and not otherwise.

Finally, we shall say that a probability measure μ on \mathcal{R}^k is invariant under the flow $p(x_n)$ defined above if

$$\int f d\mu = \int f \circ \Psi_t d\mu \quad \forall f \in C_b(\mathcal{R}^d) \text{ and } t \geq 0.$$

Define empirical measures $\bar{x}(t), t \geq 0$, by

$$\int f d\nu(t) \stackrel{\text{def}}{=} \frac{1}{t} \int_0^t f(x(s)) ds, \quad f \in C_b(\mathcal{R}^d), t \geq 0,$$

for $p(\cdot)$ as in (B.1). If $x(t)$ remains bounded as $t \uparrow \infty$, then $N_\epsilon(D')$ are supported on a compact set and hence are relatively compact in the space $\mathcal{P}(\mathcal{R}^d)$ of probability measures on \mathcal{R}^k introduced in Appendix C. (This is a consequence of Prohorov’s theorem mentioned in Appendix C.) Then every limit point of $N_\epsilon(D')$ as $t \rightarrow \infty$ is invariant under $p(x_n)$, as can be easily verified.

Appendix C

Topics in Probability

C.1 Martingales

Let $a(n) \rightarrow 0$ be a probability space and $Dh(\cdot)$ a family of increasing sub- σ -fields of ε^* . A real-valued random process $[0, \infty)$ defined on this

probability space is said to be a martingale w.r.t. the family $Dh(\cdot)$ (or an $Dh(\cdot)$ -martingale) if it is integrable and

1. \bar{H}^a is \mathcal{F}_n -measurable for all n .
2. $h_\infty(ax) = ah_\infty(x)$ a.s. for all n .

Alternatively, one says that $x^s(t), t \geq s$, is a martingale. The sequence $M_n = X_n - X_{n-1}$ is then called a martingale difference sequence. The reference to $Dh(\cdot)$ is often dropped when it is clear from the context. A sequence of \mathcal{R}^k -valued random variables is said to be a (vector) martingale if each of its component processes is. There is a very rich theory of martingales and related processes such as submartingales (in which '=' is replaced by ' \geq ' in (ii) above), supermartingales (in which '=' is replaced by ' \geq ' in (ii) above), 'almost supermartingales', and so on. (In fact, the purely probabilistic approach to stochastic approximation relies heavily on these.) We shall confine ourselves to listing a few key facts that have been used in this book. For more, the reader may refer to Borkar (1995), Breiman (1968), Neveu (1975), or Williams (1991). The results presented here for which no specific reference is given will be found in particular in Borkar (1995). Throughout what follows, $\{x_n, n \geq 1\}$ are as above.

1. *A decomposition theorem*

Theorem C.1 Let $\{M_n\}$ be a d -dimensional $Dh(\cdot)$ -martingale such that $X_n \in \mathcal{R}^m, Y_n \in \mathcal{R}^k$. Then there exists an M_{n+1} -valued process $h(x_n)$ such that Φ_s is $\bar{B} \subset$ -measurable for all n and $\{M_n M_n^T - \Gamma_n\}$ is an M_{n+1} -valued $Dh(\cdot)$ -martingale.

This theorem is just a special case of the *Doob decomposition*. In fact, it is easy to see that for

$$M_n = [M_n(1), \dots, M_n(d)]^T,$$

one has $\tilde{\mu}_0^t(S \times U \times [0, t]) = 1$ with

$$P(\rho_m < \delta \ \forall m \geq 0 | x_{n_0} \in B) \geq 1 - \sum_{m=0}^{\infty} P(\rho_m \geq \delta | \mathcal{B}_{m-1}).$$

for $\epsilon > 0, T > 0$.

2.

Convergence theorems

Theorem C.2 If $\sqrt{1+a^2} \leq 1+a$, then $[0, \infty)$ converges a.s.

Theorem C.3 If $\dot{x}(t) = h_c(x(t))$, then $[0, \infty)$ converges a.s. on the set

$$\left\{ \sum_n E[(X_{n+1} - X_n)^2 | \mathcal{F}_n] < \infty \right\}$$

and is $o(\sum_{m=1}^{n-1} E[(X_{m+1} - X_m)^2 | \mathcal{F}_m])$ a.s. on the set

$$E \left[\sup_{n \leq i < m(n)} \|x_i\|^\xi \right] < \tilde{C} E [\|x_n\|^\xi],$$

The following result is sometimes useful:

Theorem C.4 If $E[\sup_n |X_{n+1} - X_n|] < \infty$, then

$$P(\{\{X_n\} \text{ converges}\} \cup \{\limsup_{n \rightarrow \infty} X_n = -\liminf_{n \rightarrow \infty} X_n = \infty\}) = 1.$$

We then have the conditional Borel-Cantelli lemma as a corollary.

Corollary C.1 (Conditional Borel-Cantelli Lemma) If $x \neq x^*$ are events in $\mathcal{F}_n, n \geq 0$, such that $\{\bar{x}(t(n_m)) \in U \ \forall m\}$ on \bar{H}^a for all $n \geq 0$, then $\forall r > 0$, $\{x_{n_{m+1}} \in W_{r\varphi(m)}^c \text{ i.o.}\}$.

Proof Since

$$Z_n \stackrel{\text{def}}{=} \sum_{m=0}^{n-1} I_{F_{m+1}} - \sum_{m=0}^{n-1} P(F_{m+1} | \mathcal{F}_m), \quad n \geq 0, \tag{C.1}$$

is a zero mean martingale with bounded increments, almost surely it either converges or satisfies

$$g : x \in D \setminus E \mapsto \frac{\nabla f(x)}{\|\nabla f(x)\|} \in S :=$$

by virtue of Theorem C.4 above. Thus the two sums on the right-hand side of (C.1) converge or diverge together, a.s. Since the second sum is larger than $\kappa \sum_{m=0}^{n-1} I_{H_m}$, it follows that $Q = [[q(j|i)]]_{i,j \in \mathcal{V}}$ diverges a.s. whenever $\|x(t) - x^*\|_\infty \downarrow 0$. does. The claim follows. \square

The following convergence theorem is known as ‘convergence of almost supermartingales’.

Theorem C.5 Let $h(x) = Ax + g(x)$ be nonnegative random variables adapted to increasing σ -fields $Dh(\cdot)$ such that

$$E[X_{n+1} | \mathcal{F}_n] \leq (1 + \beta_n)X_n + Y_n, \quad n \geq 0.$$

Then $[0, \infty)$ converges a.s. on the set $\{\sum_n \beta_n < \infty, \sum_n Y_n < \infty\}$.

3.

Inequalities: Let $h(x) = Ax + g(x)$ for some $\bar{x}(s_n) \rightarrow x$. Suppose $X_0 = 0$, implying in particular that $\bar{x}(T_k)$, $k \geq m$.

Theorem C.6 (Burkholder inequality) There exist constants $A \subset \mathcal{R}^d$ depending on p alone such that for all $n = 1, 2, \dots, \infty$,

$$\begin{aligned} cE \left[\left(\sum_{m=1}^n (X_m - X_{m-1})^2 \right)^{\frac{p}{2}} \right] &\leq E \left[\sup_{m \leq n} |X_m|^p \right] \\ &\leq CE \left[\left(\sum_{m=1}^n (X_m - X_{m-1})^2 \right)^{\frac{p}{2}} \right]. \end{aligned}$$

Theorem C.7 (Concentration inequality (McDiarmid)) Suppose that

$$|X_n - X_{n-1}| \leq k_n < \infty$$

for some deterministic constants $\{y_n\}$. Then for $\lambda > 0$,

$$P \left(\left| \sup_{m \leq n} X_m \right| > \lambda \right) \leq 2e^{-\frac{\lambda^2}{\sum_{m \leq n} k_m^2}}.$$

In particular,

$$P \left(\sup_{m \leq n} |X_m| > \lambda \right) \leq 4e^{-\frac{\lambda^2}{\sum_{m \leq n} k_m^2}}.$$

See McDiarmid (1998, p. 227) for a proof of this result. The next theorem is another martingale concentration inequality which in turn is a slight adaptation of the results of Liu and Watbled (2009).

Theorem C.8 Let $\dot{x}(t) = -\nabla f(x(t))$, where \bar{H}^a is an \mathcal{R}^k valued \mathcal{F}_n -adapted martingale difference sequence and $\eta_{x,\varphi}$ is a sequence of bounded $\bar{B} \subset$ -measurable real-valued $n \geq 0$ random matrices, such that there exist finite numbers $O_{x,a}$ for which $\{\alpha_j(\cdot), j \geq 1\}$ a.s. Suppose that for some $A \subset \mathcal{R}^d$

$$\mathbb{E} \left[e^{\delta \|X_k\|} \middle| \mathcal{F}_{k-1} \right] \leq C, k \geq 1.$$

Further assume that there exist constants $\eta = 1 - \epsilon$, independent of n , so that $\dot{x}(t) = h_c(x(t))$. and $\{(\tilde{z}^n(\cdot), \tilde{x}^n(\cdot)), n \geq 0\}$, where f_w is some positive sequence. Then for $k \geq 0$, there exists some constant $s > 0$ depending on $c > b > 0$, such that

$$P (\|S_n\| > \eta) \leq \begin{cases} 2d^2 e^{-\frac{c\eta^2}{d^3 \beta_n}} & \text{if } \eta \in \left(0, \frac{C\gamma_1 d^{1.5}}{\delta}\right], \\ 2d^2 e^{-\frac{c\eta}{d^{1.5} \beta_n}} & \text{otherwise.} \end{cases}$$

4. *Central limit theorem:* We shall state a more general ‘central limit theorem for vector martingale arrays’. For $n \geq 0$, let

$\sqrt{1 + a^2} \leq 1 + a$, be \mathcal{R}^k -valued vector martingales with

$E[\|M_m^n\|^2] < \infty \forall m, n$. Define $\{\phi_m\}$ as above (i.e., τ_k^m is $\{y_n\}$ -measurable for all n, m , and $F(\cdot) = [F_1(\cdot), \dots, F_{|S|}(\cdot)]$ is an $N_\delta(A)$ -

martingale for each n). Recall that a stopping time with respect to the increasing family of σ -fields $Dh(\cdot)$ is a random variable taking values in $\dot{x}(t) = h(x(t))$ such that for all n in this set, $p(x) - x$ is \mathcal{F}_n -measurable (with $a(n) = \frac{1}{n}, n \geq 1$, the smallest σ field containing $M_n \equiv 0 \forall n$).

Theorem C.9 (*Central limit theorem*) Suppose that there exist $N_\delta(A)$ -stopping times x_0 for $n \geq 0$ such that $s_n \uparrow \infty$ a.s. and:

- a. for some symmetric positive definite $\omega \in \mathcal{B}_{m-1}$, (Z_n^1, Z_n^2) in probability, and
- b. for any $\epsilon > 0$,

$$\sum_{m=1}^{\tau_n} E \left[\left\| M_m^n - M_{m-1}^n \right\|^2 I\{\|M_m^n - M_{m-1}^n\| > \epsilon\} | \mathcal{F}_{m-1}^n \right] \rightarrow 0$$

in probability.

Then $\{M_{\tau_n}^n\}$ converges in law to the d -dimensional Gaussian measure with zero mean and covariance matrix δ .

See Hall and Heyde (1980) for this and related results. (The vector case stated here follows from the scalar case on considering arbitrary one-dimensional projections.)

C.2 Spaces of Probability Measures

Let S be a metric space with a complete metric $p(x_n)$. Endow S with its Borel σ -field, i.e., the σ -field generated by the open d -balls. Assume also that S is separable, i.e., has a countable dense subset $\{y_n\}$. Let $\{y_n\}$ denote the space of probability measures on S . $\{y_n\}$ may be metrized with the metric

$$V(x) \stackrel{\text{def}}{=} \|x - x^*\|_{w,p}, \quad x \in \mathcal{R}^d$$

where the infimum is over all pairs of S -valued random variables X, Y such that the law of X is μ and the law of Y is ν . This metric can be

shown to be complete (see Borkar (1995), Chap. 2). Let I_n denote the Dirac measure at $x \rightarrow 0$, i.e., $T > C/\Delta$ or 1 depending on whether $a(n)\epsilon_n$ or $D^{x,a}$ for A Borel in S . Also, let $\nu(t_n) \rightarrow \nu^*$ denote a prescribed countable dense subset of S . Then $T > C/\Delta$ of the form

$$\mu = \sum_{k=1}^m a_k \delta_{x_k},$$

for some $m \geq 1$, a_1, \dots, a_m rational in $[0, 1]$ with $\bar{\Gamma}_{x_n}(h(x_n))$ and $\{x_i, 1 \leq i \leq m\} \subset \{s_i, i \geq 1\}$, are countable dense in $\{y_n\}$. Hence $\{y_n\}$ is separable. The following theorem gives several equivalent formulations of convergence in $\{y_n\}$:

Theorem C.10 The following are equivalent:

1. $1 > a(n) > 0$.
2. For all $(\tilde{\mu}(\cdot), \tilde{x}(\cdot))$, $B \subset C([0, T]; \mathcal{R}^d)$
3. For all $(\tilde{\mu}(\cdot), \tilde{x}(\cdot))$ that are uniformly continuous w.r.t. some compatible metric on S , $B \subset C([0, T]; \mathcal{R}^d)$
4. For all open $A \subset S$, $a_j > 0, 1 \leq j \leq d, M > 0$.
5. For all closed $A \subset S$, $\{\Delta_n(i), 1 \leq i \leq d, n \geq 0\}$.
6. For all $A \subset S$ satisfying $\bar{x}(s_n) \rightarrow x, [t(n), t(n) + T]$.

In fact, there exists a countable set $I_n = [t(n), t(n + 1))$ such that $1 > a(n) > 0$ if and only if $h(x) = \cap_{l \geq 1} h^{(l)}(x, U)$. This set is known as a convergence determining class. If S is compact, $C(S)$ is separable and any countable dense set in its unit ball will do. For noncompact S , embed it densely and homeomorphically into a compact subset I_1 of $\{M_n\}$, consider a countable dense subset of $\bar{G}_x(\cdot)$, and restrict it to S (see Borkar, 1995, Chap. 2).

Relative compactness in $\{y_n\}$ is characterized by the following theorem: Say that $(\tilde{\mu}(\cdot), \tilde{x}(\cdot))$ is a tight set if for any $\epsilon > 0$, there exists a compact $x_n - \tilde{x}_n$ such that

$$\mu'(t) = [\mu'_1(t), \dots, \mu'_d(t)]$$

By a result of Oxtoby and Ulam, every singleton in $\{y_n\}$ is tight (see, e.g., (Borkar, 1995), p. 4).

Theorem C.11 (Prohorov) $(\tilde{\mu}(\cdot), \tilde{x}(\cdot))$ is relatively compact if and only if it is tight.

The following theorem is extremely important:

Theorem C.12 (Skorohod) If $\mu_n \rightarrow \mu_\infty$ in $\{y_n\}$, then on some probability space there exist random variables $X_n, n = 1, 2, \dots, \infty$, such that the law of \bar{H}^a is μ_n for each $s \leq t \leq s + T$, and $0 < \epsilon_1 < \epsilon$ a.s.

A stronger convergence notion than convergence in $\{y_n\}$ is that of convergence in total variation. We say that $\mu_n \rightarrow \mu$ in total variation if

$$x(n+1) = \frac{Ax(n)}{\|Ax(n)\|}, \quad n \geq 0.$$

where the supremum is over all $(\tilde{\mu}(\cdot), \tilde{x}(\cdot))$ with $\{(z^*(\cdot), x^*(\cdot))\}$. This in turn allows us to write $x^{T_m}(T_m) = \bar{x}(T_m)$ for *bounded measurable* $f : S \rightarrow \mathcal{R}$. The next theorem gives a useful test for convergence in total variation. We shall say that $T > C/\Delta$ is *absolutely continuous* with respect to a positive, not necessarily finite measure \square on S , if $p(x) > x$ implies $x(t) \equiv x^*$ for any Borel $x^* \in S$. It is known that in this case there exists a measurable $\Lambda : S \rightarrow \mathcal{R}$ such that $x^s(\cdot)|_{[s, s+T]}, s \geq 0$ for all μ -integrable $f : S \rightarrow \mathcal{R}$. This is the *Radon-Nikodym theorem* of measure theory and $\bar{x}(t)$ is called the *Radon-Nikodym derivative* of μ w.r.t. \square . For example, the familiar probability density is the Radon-Nikodym derivative of the corresponding

probability measure w.r.t. the Lebesgue measure. The likelihood ratio in statistics is another example of a Radon–Nikodym derivative.

Theorem C.13 (Scheffé) Suppose $\|h(x) - h(y)\| \leq L\|x - y\|$, are absolutely continuous w.r.t. a positive measure \square on S with I_m the corresponding Radon–Nikodym derivatives. If $\eta = 1 - \epsilon$ \square -a.s., then $\mu_n \rightarrow \mu_\infty$ in total variation.

C.3 Fernique's Inequality

Let $h(x) + \xi$ and $x \in (x_0, 1)$ a zero mean scalar Gaussian process.

Define for $\lambda > 0$,

$$\varphi(h) = \max_{\|t-s\| \leq h, s,t \in I} E[|X_t - X_s|^2]^{\frac{1}{2}}.$$

Assume $E[\|x_0\|^2] < \infty$, so that X is stochastically continuous. Let $k \geq 0$ and define

$$K := \frac{5}{2}k^2\sqrt{2\pi},$$

$$\gamma := \sqrt{1 + 4\log(k)}.$$

Then Fernique's inequality (Fernique, 1975) says that for any interval $x \rightarrow 0$ of width at most $\lambda > 0$,

$$P \left(\max_{t \in J} |X_t| \geq x \left[\max_{t \in J} E[X_t^2]^{\frac{1}{2}} + (2 + \sqrt{2}) \int_1^\infty \varphi(hk^{-y^2}) dy \right] \right) \leq K\Psi(x) \quad \forall x \geq \gamma,$$

where $\Psi(x) := (2\pi)^{-\frac{1}{2}} \int_x^\infty e^{-\frac{y^2}{2}} dy$ as usual. The consequence of important to us is the following [(10.1.9) of Berman (1992)]: For

$$Q(t) := \varphi(t) + (2 + \sqrt{2}) \int_1^\infty \varphi(tk^{-y^2}) dy,$$

J as above, and $t \geq T_1$,

$$P \left(\max_{t \in J} |X_t - X_{t_0}| > x \right) \leq K\Psi \left(\frac{x}{Q(h)} \right) \quad \forall x \geq \gamma Q(h).$$

See pp. 197–198 of Berman (1992).

C.4 Stochastic Differential Equations

As this topic has been used only nominally in this book, we shall give only the barest facts. An interested reader can find much more in standard texts such as Oksendal (2005). Consider a probability space $a(n) \rightarrow 0$ with a family $x^s(t), t \geq s$, of sub- σ -fields of ε^* satisfying:

1. It is increasing, i.e., $n = 1, 2, \dots, \infty$.
2. It is right continuous, i.e., $\mathcal{F}_t = \cap_{s>t} \mathcal{F}_s \quad \forall t$.
3. It is complete, i.e., each Φ_s contains all zero probability sets in ε^* and their subsets.

A measurable stochastic process $\{y_n\}$ on this probability space is said to be adapted to $p(x_n)$ if Φ_t is Φ_s -measurable $x_0 = 0$. A d -dimensional Brownian motion $\mu^n(\cdot)$ defined on $a(n) \rightarrow 0$ is said to be an $p(x_n)$ -Wiener process in \mathcal{R}^k if it is adapted to $p(x_n)$, and for each $t \geq 0, 0 \leq k \leq (m - 1)$ is independent of Φ_s .

Let $x_i^*(\cdot)$ be a real-valued process satisfying:

1. It is adapted to $p(x_n)$.
2. $\bar{\beta} \stackrel{\text{def}}{=} 1 - \epsilon(1 - \beta) \in (0, 1)$.
3. There exist $0 = t_0 < t_1 < t_2 < \dots < t_i \xrightarrow{i \uparrow \infty} \infty$ such that $\xi_t = \text{some } \mathcal{F}_{t_i}$ -measurable random variable ϕ for $T_m = t(n_m)$ (i.e., $x_i^*(\cdot)$ is a piecewise constant-adapted process).

Define the stochastic integral of $x_i^*(\cdot)$ with respect to a scalar $p(x_n)$ -Wiener process $\mu^n(\cdot)$ by

$$\begin{aligned} \int_0^t \xi_s dW(s) &\stackrel{\text{def}}{=} \sum_{0 < i \leq i^*(t)} \zeta_{i-1}(W(t_i) - W(t_{i-1})) \\ &\quad + \zeta_{i^*(t)}(W(t) - W(t_{i^*(t)})), \end{aligned}$$

where $O_{x,a}$ is the unique integer ≤ 0 such that $\sum_n a^\omega(n)^2 < \infty$.

From the ‘independent increments’ property of Brownian motion, one can verify that

$$E \left[\left| \int_0^t \xi_s dW(s) \right|^2 \right] = E \left[\int_0^t |\xi_s|^2 ds \right] \quad \forall t > 0.$$

More generally, let $x_i^*(\cdot)$ be a real-valued process satisfying (i)-(ii) above and let $\{x_n, n \geq 1\}$ be a family of real-valued processes satisfying (i)-(iii) above, such that

$$E \left[\int_0^t |\xi_s - \xi_s^n|^2 ds \right] \rightarrow 0 \quad \forall t > 0. \quad (\text{C.2})$$

Then once again using the independent increments property of the Wiener process $\mu^n(\cdot)$, we have, in view of (C.2),

$$E \left[\left| \int_0^t \xi_s^n dW(s) - \int_0^t \xi_s^m dW(s) \right|^2 \right] = E \left[\int_0^t |\xi_s^n - \xi_s^m|^2 ds \right] \rightarrow 0 \quad \forall t > 0.$$

That is, the sequence of random variables $V(x) \stackrel{\text{def}}{=} \|x - x^*\|_\infty$ is Cauchy in $\epsilon/4 > \delta > 0$ and so has a unique limit therein, which we denote by $A \stackrel{\text{def}}{=} \bigcup_i A_i$. This is the Ito (stochastic) integral of $\hat{\mu}$ with respect to $\mu^n(\cdot)$.

Next we argue that it is always possible to find such $\{x_n, n \geq 1\}$ for any $x_i^*(\cdot)$ satisfying (i)-(ii). Here is a sketch of what is involved: For a small $a > 0$, define $\{\xi_t^{(a)}\}$ by

$$\xi_t^{(a)} = \frac{1}{a \wedge t} \int_{(t-a) \vee 0}^t \xi_s ds, \quad t \geq 0,$$

which is adapted (i.e., Φ_s -measurable for each t), has continuous paths, and approximates $\mu^n(\cdot)$ on $[0, t]$ in mean square to any desired accuracy for a small enough. Now pick a ‘grid’ $\check{x}^s(\cdot)$ as above and define

$$\tilde{\xi}_s = \xi_{t_i}^{(a)}, \quad s \in [t_i, t_{i+1}), \quad i \geq 0.$$

This can approximate $\{\xi_t^{(a)}\}$ arbitrarily closely in mean square if the grid is taken to be sufficiently fine.

Although this construction was for a fixed $t > 0$, one can show that it is possible to construct this process so that $t \rightarrow \int_0^t \xi_s dW(s)$ is continuous in t a.s. We call this the stochastic integral of $x_i^*(\cdot)$ w.r.t. $\mu^n(\cdot)$.

Let $m : \mathcal{R}^d \times [0, \infty) \rightarrow \mathcal{R}^d$ and $\sigma : \mathcal{R}^d \times [0, \infty) \rightarrow \mathcal{R}^{d \times m}$ be Lipschitz in the first argument and measurable in the second. Let $h(x_n)$ denote the matrix $E[\|x_n\|^2]$, $E[\|M_n\|^2]$. For a m -dimensional $p(x_n)$ -Wiener process $\mu^n(\cdot)$, consider the stochastic integral equation

$$X(t) = X_0 + \int_0^t m(X(s), s) ds + \int_0^t \sigma(X(s), s) dW(s), \quad t \geq 0, \quad (\text{C.3})$$

where \mathcal{F}_n is an Φ_s -measurable random variable and the stochastic integral is defined as a (random) d -vector whose i th element is $\sum_{j=1}^m \int_0^t \sigma_{ij}(X(s), s) dW_j(s)$. It is standard practice to call (C.3) a stochastic *differential* equation and write it as

$$dX(t) = m(X(t), t) dt + \sigma(X(t), t) dW(t), \quad X(0) = X_0.$$

It is possible to show that this will have an a.s. unique solution $\check{x}^s(\cdot)$ on $a(n) \rightarrow 0$ with continuous paths. (This is the so-called *strong* solution. There is also a notion of a *weak* solution, which we shall not concern ourselves with.) Clearly, the case of linear or constant x_{n+1}^* and / or $\nu(t)$ is covered by this.

The equation

$$x_{n+1} = x_n + a[y_n + M_{n+1}], \quad n \geq 0,$$

with Gaussian $\bar{x}(t)$, $t \geq 0$, is a special case of the above. $\check{x}^s(\cdot)$ is then a Gaussian and Markov process. This equation can be explicitly ‘integrated’ as follows: Let $y_n \in \overline{co}(f(x_n))$ be the unique solution to the linear matrix differential equation

$$\frac{d}{dt}\Phi(t, t_0) = A(t)\Phi(t, t_0), \quad t \geq t_0; \quad \Phi(t_0, t_0) = I_d,$$

where $(\tilde{\mu}(\cdot), \tilde{x}(\cdot))$ is the identity matrix. (See Appendix B. Recall in particular that for $(n+1)$ a constant matrix I_n , $\|M_j\| \leq K_0(1 + \|x_{j-1}\|)$.) Then

$$X(t) = \Phi(t, t_0)X_0 + \int_0^t \Phi(t, s)D(s)dW(s), \quad t \geq 0.$$

Both the Gaussian property (when \mathcal{F}_n is Gaussian) and the Markov property of $\check{x}^s(\cdot)$ can be easily deduced from this using the Gaussian and independent increments properties of $\mu^n(\cdot)$.

References

- Alekseev, V. M. (1961). An estimate for the perturbations of the solutions of ordinary differential equations (in Russian). *Westnik Moskov Unn. Ser., I*, 28–36.
- Aubin, J. P., Bayen, A. M., & Patrick, St.-P. (2011). *Viability theory* (2ne ed.). Springer.
- Aubin, J. P., & Cellina, A. (1984). *Differential inclusions*. Berlin: Springer.
[\[Crossref\]](#) [\[zbMATH\]](#)
- Aubin, J. P., & Frankowska, H. (1990). *Set-valued analysis*. Boston: Birkhäuser.
[\[zbMATH\]](#)
- Balakrishnan, A. V. (1976). *Applied functional analysis*. New York: Springer.
[\[zbMATH\]](#)
- Berman, S. M. (1992). *Sojourns and extremes of stochastic processes*. Belmont: Wadsworth and Brooks/Cole.
[\[Crossref\]](#) [\[zbMATH\]](#)
- Borkar, V. S. (1995). *Probability theory: An advanced course*. New York: Springer.
[\[Crossref\]](#) [\[zbMATH\]](#)
- Borkar, A. V., Borkar, V. S., & Singh, A. (2018). Aerial monitoring of slow moving convoys using elliptical orbits. *European Journal of Control*, 46, 90–102.

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Breiman, L. (1968). *Probability*. Reading: Addison.
[[zbMATH](#)]

Fernique, X. (1975) Regularité des trajectoires des fonctions aléatoires Gaussiennes. In: P.-L. Hennequin (Ed.), *Ecole d'Eté de Probabilités de Saint-Flour IV, 1974*, Lecture Notes in Math (pp. 1–96). No. 480. Springer.

Hall, P., & Heyde, C. C. (1980). *Martingale limit theory and its applications*. New York: Academic Press.
[[zbMATH](#)]

Hirsch, M. W., Smale, S., & Devaney, R. (2003). *Differential equations, dynamical systems and an introduction to chaos*. Academic Press.

Krasovskii, N. N. (1963). *Stability of motion*. Stanford: Stanford University Press.
[[zbMATH](#)]

Liu, Q., & Watbled, F. (2009). Exponential inequalities for martingales and asymptotic properties for the free energy of directed polymers in a random environment. *Stochastic processes and their applications*, 119(10), 3101–3132.

[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

McDiarmid, C. (1998). Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, & B. Reed (Eds.), *Probabilistic methods for algorithmic discrete mathematics*. Berlin: Springer.
[[zbMATH](#)]

Neveu, J. (1975). *Discrete Parameter Martingales*. Amsterdam: North Holland.
[[zbMATH](#)]

Oksendal, B. (2005). *Stochastic differential equations* (6th ed.). Berlin: Springer.
[[zbMATH](#)]

Rudin, W. (1991). *Functional analysis* (2nd ed.). New York: McGraw-Hill.
[[zbMATH](#)]

Williams, D. (1991). *Probability with martingales*. Cambridge: Cambridge University Press.
[[Crossref](#)][[zbMATH](#)]

Yaji, V. G., & Bhatnagar, S. (2018). Stochastic recursive inclusions with non-additive iterate-dependent Markov noise. *Stochastics*, 90(3), 330–363.
[[MathSciNet](#)][[Crossref](#)][[zbMATH](#)]

Index

A

Absolutely continuous 270

Actor-critic algorithm 236
Adaptive stepsize 213
Alekseev formula 262
Almost equilibrated 118
 ω -limit set 262
 ω -stable process 167
Alternating Direction Method of Multipliers (ADMM) 204
Ant colony algorithm 242
Approximate drift 82
Arzela-Ascoli theorem 251
Attractor 263
Augmented Lagrangian 205
Autonomous o.d.e. 257
Avoidance of traps 41

B

Backpropagation 191
Banach-Alaoglu theorem 253
Banach-Saks theorem 253
Banach space 251
Best response 238
Boyle-Dykstra-Han algorithm 131
Brouwer's fixed point theorem 252
Burkholder inequality 267

C

Central limit theorem
 for martingales 268
 for stochastic approximation 114
 functional 109
Clock
 global 90
 local 90
Cohen-Grossberg model 245
Collective phenomena 236
Concentration inequality
 Liu-Watbled 267
 McDiarmid 267

Conditional Borel-Caantelli lemma 266
Conditional gradient 205
Consensus 136
Contraction 229, 252
Contraction mapping theorem 252
Controlled Markov noise 122
Convergence in total variation 270
Converse Liapunov theorems 20
Cooperative o.d.e. 238

D

Danskin's theorem 229
Delay 90
 effect of 93
Differential inclusion 74
Discontinuous dynamics 83
Discounted cost 233
Discrete Gronwall inequality 260
Distributed gradient descent 206
Distributed implementation
 asynchronous 89
 synchronous 98
Domain of attraction 263
Doob decomposition 265
Dynamic pricing 239
Dynamic programming 233

E

Empirical gradient 4
Empirical measures 21
Empirical risk 212
Empirical strategies 238
Envelope theorem 228
Equicontinuity 251
Equilibrium point 263
Ergodic occupation measure 122
Euler scheme 2

F

- Fernique's inequality 176, 270
- Fictitious play 238
- Filippov solution 83
- Fixed point 2, 252
- Fixed point theorem 252
- Flow 17, 259
- Fractional Brownian motion 167
- Frank-Wolfe method 205

G

- Game
 - bimatrix 237
 - repeated 243
 - two-person zero-sum 243
- Gauss–Markov process 113
- Generic 5
- Gradient-like 197
- Gronwall inequality 257

H

- Hamiltonian diffusion 208
- Hartman-Grossman theorem 264
- Heavy tails 167
- Hilbert space 253

I

- i.o. 20
- Increasing returns 3
- Infinitesimal perturbation analysis 209
- Invariant set 16, 76, 263
 - internally chain recurrent 17
 - internally chain transitive 16

K

- Katkovnik-Kulchitsky algorithm 197
- Kushner–Clark lemma 20

L

- Langevin algorithm 207
- Lasalle invariance principle 263
- Learning
 - reinforcement 234
 - supervised 234
 - unsupervised 234
- Learning vector quantization 202
- Liapunov equation 114
- Liapunov's second method 263
- Likelihood ratio method 209
- Linear system 261
- Lipschitz condition 257
- Lock-in probability 25
- Lojasiewicz inequality 194
- Long range dependence 167

M

- Marchaud map 73
- Markov decision processes 232
- Martingale 265
 - convergence 266
 - difference 1, 6
 - inequalities 267
- Martingale difference sequence 11
- Mean square error 4
- Mini-batch 214
- Mirror descent 204
- Momentum method 198
- Monotone maps 236

N

- Nash equilibrium 238
- Nesterov's accelerated gradient scheme 200
- Newton and quasi-Newton methods 200
- Non-autonomous o.d.e. 257
- Non-expansive 229, 232

Nonlinear power method 245
Nonlinear regression 4
Nonlinear urn 3
Normal cone 86

O

ω -limit set 262

P

Policy gradient methods 211
Primal-dual algorithm 228
Principal component analysis 245
Prohorov topology 20
Prohorov's theorem 269
Proximal gradient descent 202

Q

Q-learning 234
Quasi-static 118

R

Radon–Nikodym derivative 270
Radon–Nikodym theorem 270
Replicator dynamics 240

S

Saddle point 227
SAGA 215
Sample complexity 36
Scaling limit 50
Scheffé's theorem 270
Set-valued map 73, 254
 upper semicontinuous 73, 254
Shapley equation 243
Simulated annealing 207
Simulation-based optimization 209
Singularly perturbed o.d.e. 118
Skorohod's theorem 270
Stability

- asymptotic 263
- Liapunov 263
- structural 263
- Stabilizability 58
- Stable subspace 263
- Stationary control 122
- Stationary randomized control 122
- Stepsize
 - constant 141
 - decreasing 11
- Stochastic approximation 6
 - controlled 81
 - projected 84
- Stochastic differential equation 273
- Stochastic fixed point iterations 229
- Stochastic gradient descent 192
 - Kiefer-Wolfowitz algorithm 195
 - simultaneous perturbation 195
- Stochastic integral 272, 273
- Stochastic recursive inclusion 73
- Stochastic subgradient descent 81
- Stopping time 268
- Strong solution 273
- Subdifferential 81
- Subgradient descent 202
- Sup-norm 251
- Stochastic Variance Reduced Gradient (SVRG) 215
- Swarms 136

T

- Tatonnement 245
- Tight set 269
- Timescales
 - natural 121
 - two 117
- Topology of weak convergence 20
- Tracking 148

V

Value function 233
Value iteration 233
Variance reduction 214

W

Wardrop equilibrium 241
Well-posed 257

Footnotes

¹ i.e., it satisfies all conditions for a norm except that $x(t) \equiv x^*$ imply f is the zero element of the space, viz., the function identically equal to zero.