# Stochastic Algorithms for Nonconvex Optimization and Reinforcement Learning

Mathukumalli Vidyasagar
Indian Institute of Technology Hyderabad
m.vidyasagar@iith.ac.in

November 15, 2025

# Preface

This is a *draft* manuscript. To quote Oliver Goldsmith, "There are a hundred faults in this Thing and a hundred things might be said to prove them beauties." I hope that, as time passes, the faults will decrease while the "beauties" will increase. In the meantime, "caveat emptor" is the watchword for the reader.

These notes are updated regularly. Please check the date of last update to ensure that you have the latest version, which is available at

https://people.iith.ac.in/m_vidyasagar/RL-Notes.pdf

Feedback of all kinds would be gratefully received at m.vidyasagar@iith.ac.in

These notes address some aspects of two somewhat disparate disciplines, namely: Nonconvex Optimization, and Reinforcement Learning (RL). Each of these disciplines ha a long history and s a vast literature. Thus the choice of topics covered in this book is dictated by the fact that the solution is obtained by *stochastic algorithms*. Within optimization, the techniques presented here can be used to minimize not only *convex* objective functions, but also some classes of *nonconvex* functions. Within Reinforcement Learning (RL), we discuss all of the standard topics such as value computation, Temporal Difference learning, and $Q$-Learning. However, within this subareas, the focus is on problems that can be solved using stochastic algorithms.

The topic of optimization dates back a few centuries, but the analysis was mostly confined to finding "closed-form" solutions. The main constraint was the unavailability of tools to carry out numerical computations at a large scale. The subject really picked up steam in the 1960s with the advent of digital computation, when the emphasis shifted to *iterative* methods that did not even attempt to find the solution "in closed form." Rather, the emphasis was on constructing a sequence of approximate solutions that converged to the true solution. Advances in computing (both in terms of increasing capability and decreasing cost) made the scientific community aspire to solve ever larger problems. In this setting, it is noticeably easier to deal with *convex* optimization problems than with nonconvex problems. However, the present-day widespread use of deep neural networks has led to greater emphasis on *nonconvex* optimization.

Stochastic algorithms are natural when the information about the problem to be solved is *uncertain*, or prone to measurement errors. In optimization problems, if the measurements of the objective function at each iteration, and/or its gradient, are subject to measurement, then it is imperative to use stochastic algorithms that are guaranteed to converge even in the presence of such uncertainties. Even without uncertain measurements, some problems become more tractable when some element of randomness is introduced into the algorithm. The framework presented in this paper is rich enough to handle such randomized algorithms as well, though those are not the main focus.

Reinforcement Learning (RL) is one of the most active areas of research in AI (Artificial Intellligence), or Artificial General Intelligence (AGI), and Machine Learning (ML). One can think of AI or AGI as a desire to enable computers to mimic various aspects of human intelligence, and ML as a set of tools and/or algorithms to achieve AI/AGI. Thus AI (AGI) is the *destination*, and ML is the *path* to that destination.

As mentioned above, the literature in both areas is vast. Therefore the aim of these notes is to provide a treatment of *some aspects* of nonconvex optimization as well as RL. The unifying theme in the treatment of various problems is a well-established technque known as Stochastic Approximation (SA). Stochastic Approximation was introduced in 1951 as a method for solving equations of the form $\mathbf{f}(\boldsymbol{\theta}^*) = \mathbf{0}$ when only noisy measurements of $\mathbf{f}(\cdot)$ are available. Since then the theory has expanded substantially. One of the objectives of these notes is to show how SA can be used to prove the convergence of iterative algorithms in

nonconvex optimization, and in RL. While the use of SA to analyze RL algorithms is well-established, going back to the late 1980s, the method has begun to be applied to nonconvex optimization relatively recently.

A useful feature of SA is that it is relatively easy to prove that various algorithms converge *almost surely*. This is in contrast to other proof techniques that lead to weaker conclusions such as convergence in probability, or in expectation. Since iterative stochastic algorithms generate *one sample path* of a stochastic process, it is worthwhile to know that almost all sample paths converge to the desired limit.

These notes are an outcome of having taught this material three times. Previously, I had offered courses at UC Berkeley (remotely) during the Fall semester of 2020, and at IIT Hyderabad during the First Semester of 2022-23. On both occasions, the focus of the course was solely on RL. I taught the course for the third time during the First Semester of 2024-25, this time with as added focus on optimization.

Over time, the nature of the notes underwent some changes. I had originally envisaged a *textbook* that covered most of the widely studied ideas in RL, even if the material did not contain any original research. However, subsequently I decided to narrow the scope to material that either represented original research, or streamlining and/or unification of existing proofs. Thus, in their current version, the notes are much more like a *research monograph* than a textbook. In the process, I trust that the breadth of coverage has not been unduly sacrificed, and that most of the relevant topics in RL are still included.

These notes are organized as follows:

Given that the notes are written in the style of a research monograph, very few "exercises" are included. Instead, the reader is advised to work out all the proofs in detail. This is the best way to master the content, and not via accepting various theorems at face value.

# Contents

**Preface** **i**

**1 Introducton** **1**
1.1 Introduction to Optimization . . . 1
1.1.1 Introduction to Optimization . . . 1
1.1.2 Classes of Functions . . . 2
1.1.3 Some Popular Algorithms . . . 4
1.1.4 Sources of Stochasticity . . . 5
1.2 Introduction to Reinforcement Learning . . . 7
1.2.1 Introduction to Reinforcement Learning . . . 7
1.2.2 Some Examples of Reinforcement Learning . . . 11
1.3 About These Notes . . . 16

**2 Convergence of Stochastic Processes** **17**
2.1 Random Variables and Stochastic Processes . . . 17
2.1.1 Random Variables . . . 18
2.1.2 Joint and Conditional Probabilities, Independence . . . 23
2.1.3 Conditional Expectations . . . 26
2.2 Markov processes . . . 29
2.2.1 Markov Processes: Basic Properties . . . 29
2.2.2 Stopping Times and Hitting Probabilities . . . 35
2.2.3 Maximum Likelihood Estimation of Markov Processes . . . 38
2.3 Some Convergence Theorems . . . 41
2.3.1 Introduction to Martingales . . . 41
2.3.2 Some Convergence Theorems . . . 47

**3 Stochastic Approximation: Algorithms and Convergence** **55**
3.1 An Overview of Stochastic Approximation . . . 55
3.2 Convergence of Synchronous Stochastic Approximation . . . 58
3.2.1 Convergence Theorems for SA via Lyapunov Theory . . . 58
3.2.2 Some Applications . . . 63
3.2.3 Existence of Suitable Lyapunov Functions . . . 66
3.3 Block Asynchronous Stochastic Approximation . . . 70
3.3.1 Problem Formulation . . . 71
3.3.2 Intermittent Updating: Convergence and Rates . . . 72
3.3.3 Boundedness of Iterations . . . 78
3.3.4 Convergence of Iterations with Rates . . . 85
3.4 Variants of Standard Stochastic Approximation . . . 88
3.4.1 Averaged Stochastic Approximation . . . 88

3.4.2 Two Time Scale Stochastic Approximation . . . 89
3.4.3 Finite-Time Stochastic Approximation . . . 89
3.4.4 Markovian Stochastic Approximation . . . 89

**4 Applications to Optimization 93**
4.1 Some Invex Functions . . . 93
4.2 Review of Some Standard Algorithms . . . 96
4.2.1 Stochastic Gradient Descent . . . 96
4.2.2 Momentum-Based Methods . . . 101
4.3 Stochastic Gradient Descent . . . 103
4.4 A Unified Theory for Momentum-Based Methods . . . 111
4.4.1 A Unified Momentum-Based Algorithm . . . 111
4.4.2 Literature Review . . . 112
4.4.3 Statements of Main Theorems . . . 116
4.4.4 Proofs of the Main Results . . . 119
4.4.5 Nonviability of an Earlier Iterative Scheme . . . 128
4.5 Stochastic Algorithms with Block Updating . . . 131
4.5.1 Various Block Updating Schemes . . . 131
4.5.2 Convergence of SGD with Block Updating . . . 132
4.5.3 Convergence of the Unified Momentum Algorithms . . . 134

**5 Markov Decision Processes 135**
5.1 Markov Reward Processes . . . 135
5.1.1 Discounted Reward Processes . . . 136
5.1.2 Average Reward Markov Processes . . . 138
5.2 Markov Decision Processes . . . 141
5.2.1 Markov Decision Processes: Problem Set-Up . . . 141
5.2.2 Markov Decision Processes: Analysis . . . 144

**6 Reinforcement Learning 155**
6.1 Value Determination Using Temporal Differences . . . 155
6.1.1 $TD(\lambda)$-Learning Without Function Approximation . . . 156
6.2 $TD(\lambda)$-Learning With Function Approximation . . . 157
6.2.1 Discounted Reward Processes . . . 157
6.2.2 Average Reward Processes . . . 157
6.3 Simultaneous Value and Policy Approximation . . . 157
6.3.1 Two Time-Scale Stochastic Approximation: Reprise . . . 157
6.3.2 Average Reward Processes: Reprise . . . 157
6.3.3 Policy Gradient Theorem . . . 157
6.3.4 Actor-Critic Methods . . . 158
6.4 $Q$-Learning . . . 158
6.5 Zap $Q$-Learning . . . 158
6.5.1 Stochastic Newton-Raphson Approximation . . . 158
6.5.2 Zap $Q$-Learning . . . 158

**7 Background Material 159**
7.1 Contraction Mapping Theorem . . . 159
7.2 Some Elements of Lyapunov Stability Theory . . . 160

# Chapter 1

# Introducton

## 1.1 Introduction to Optimization

In this chapter, we give a brief overview of the type of optimization problems studied in this book. Further details can be found in subsequent chapters, specifically Chapter 4.

### 1.1.1 Introduction to Optimization

Suppose $J : \mathbb{R}^d \to \mathbb{R}$ is some function; we will add more assumptions on $J(\cdot)$ as we go along. The core problem of optimization is to find one or more vectors $\boldsymbol{\theta}^* \in \mathbb{R}^d$ that *minimize* $J(\cdot)$. It is clear that, by replacing $J(\cdot)$ by $-J(\cdot)$, the problem of *maximizing* a function can be readily reformulated as one of minimizing its negative. Hence in this book we shall study only problems of minimization. It is customary to refer to $J(\cdot)$ as the **ojbective function**. With this convention, we next distinguish between two different problems.

- Unconstrained vs. constrained minimization
- Global vs. local minimization

Let us begin with the first item. In *unconstrained* minimization, we study a problem of the form

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}),$$

whereas in *constrained* minimization, we study a problem of the form

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \text{ s.t. } \boldsymbol{\theta} \in S,$$

where $S \subseteq \mathbb{R}^d$ is a specified region of $\mathbb{R}^d$, usually referred to as the "feasible" region. Clearly, if $S = \mathbb{R}^d$, there is no difference between the two. In this book, we restrict our attention to unconstrained minimization problems, even though many of the techniques presented here can be made to apply to constrained minimization with suitable modifications. Next, a vector $\boldsymbol{\theta}^*$ is a **local minimizer** of $J(\cdot)$ if there exists a neighborhood $S$ of $\boldsymbol{\theta}^*$ such that

$$J(\boldsymbol{\theta}^*) \leq J(\boldsymbol{\theta}) \; \forall \boldsymbol{\theta} \in S, \tag{1.1.1}$$

while $\boldsymbol{\theta}^*$ is a **unique local minimizer** of $J(\cdot)$ if

$$J(\boldsymbol{\theta}^*) < J(\boldsymbol{\theta}) \; \forall \boldsymbol{\theta} \in S \setminus \{\boldsymbol{\theta}^*\}. \tag{1.1.2}$$

A vector $\boldsymbol{\theta}^*$ is said to be a **global minimizer** of $J(\cdot)$ if (1.1.1) holds with $S$ replaced by $\mathbb{R}^d$, while $\boldsymbol{\theta}^*$ is said to be a **unique global minimizer** of $J(\cdot)$ if (1.1.2) holds with $S$ replaced by $\mathbb{R}^d$. In optimization problems,

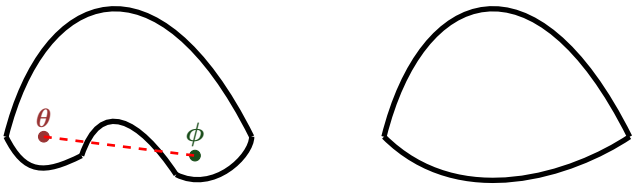



Figure 1.1: Examples of Convex and Nonconvex Sets

clearly everyone would like to find *global* minimizers, but often one has to settle for *local* minimizers. To the extent possible, in this book we strive to find global minimizers.

Before proceeding further, we clarify our usage of the terms "minimum" and "minimizer," something about which not every author is careful. If $\boldsymbol{\theta}^*$ is a satisfies (1.1.1), then we refer to $\boldsymbol{\theta}^*$ as the **minimizer**, and to $J(\boldsymbol{\theta}^*)$ as the **minimum**. Thus, the minimizer is the argument, while the minimum is the function value at the minimizer.

### 1.1.2 Classes of Functions

In order to make the problem of function minimization more tractable, it is necessary to introduce more "structure" into the problem, that is, to make some assumptions about $J(\cdot)$. Throughout the book, it is assumed that $J(\cdot)$ is $\mathcal{C}^1$, and that the gradient $\nabla J(\cdot)$ is globally $L$-Lipschitz continuous. This means that

$$\|\nabla J(\boldsymbol{\theta}) - \nabla J(\boldsymbol{\phi})\|_2 \leq L\|\boldsymbol{\theta} - \boldsymbol{\phi}\|_2, \ \forall \boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^d. \tag{1.1.3}$$

Thus, we do *not* study "non-smooth" objective functions such as $J(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$, nor functions of the form $J(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^4$.

With these assumptions, there are several classes of functions that are studied in this book. We introduce convex functions here, but as the title indicates, the focus of the book is on *nonconvex* objective functions. The reader is referred to Section 4.1 for several classes of nonconvex functions studied in the book.

We begin with the notion of a convex set, and then move to the notion of a convex function.

If $\boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^d$, and $\lambda \in [0, 1]$, then the vector $\lambda\boldsymbol{\theta} + (1 - \lambda)\boldsymbol{\phi}$ is called a **convex combination** of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. If $\lambda \in (0, 1)$ and $\boldsymbol{\theta} \neq \boldsymbol{\phi}$, then the vector $\lambda\boldsymbol{\theta} + (1 - \lambda)\boldsymbol{\phi}$ is called a **strict convex combination** of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Some authors also call this a "nontrivial" convex combination.

**Definition 1.1.** A subset $S \subseteq \mathbb{R}^d$ is said to be a **convex set** if

$$\lambda\boldsymbol{\theta} + (1 - \lambda)\boldsymbol{\phi} \in S \ \forall \lambda \in [0, 1], \ \forall \boldsymbol{\theta}, \boldsymbol{\phi} \in S. \tag{1.1.4}$$

Thus a set $S$ is convex if every convex combination of two elements of $S$ once again belongs to $S$. In two dimensions we can visualize a convex set very simply. If $\boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^2$, then the set $\{\lambda\boldsymbol{\theta} + (1 - \lambda)\boldsymbol{\phi} : \lambda \in [0, 1]\}$ is the line segment joining the two vectors $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Thus a set $S \subseteq \mathbb{R}^2$ is convex if and only if the line segment joining any two points in the set $S$ once again belongs to the set $S$. Therefore in Figure 1.1, the set on the left is not convex, because the line segment connecting $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ does not lie entirely in $S$; in contrast, the set on the right is convex. A similar interpretation also applies in higher dimensions, except that the "line" has to be imagined and cannot be drawn on a page.

Definition 1.1 is stated for a convex combination of *two* vectors, but can be easily extended to a convex combination of any finite number of vectors. Suppose $S \subseteq \mathbb{R}^d$ and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k \in S$. Then a vector of the form

$$\boldsymbol{\phi} = \sum_{i=1}^{k} \lambda_i \boldsymbol{\theta}_i, \lambda_i \geq 0 \ \forall i, \sum_{i=1}^{k} \lambda_i = 1$$

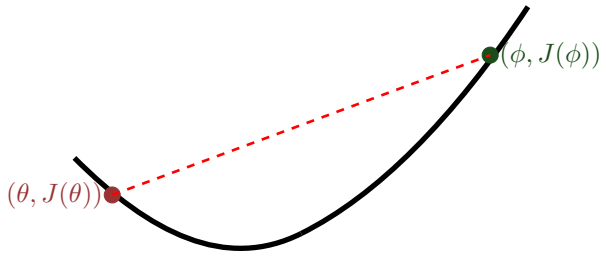



Figure 1.2: Graph Below Chord Interpretation of a Convex Function

is called a convex combination of the vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$. It is easy to show, by recursively applying Definition 1.1, that if $S \subseteq \mathbb{R}^d$ is a convex set then every convex combination of any finite number of vectors in $S$ again belongs to $S$.

**Example 1.1.** The $n$-dimensional simplex $\mathbb{S}_n$, which can be identified with the set of probability distributions on a finite alphabet of cardinality $n$, is a convex set. Thus if $\boldsymbol{\mu}, \boldsymbol{\nu}$ are $n$-dimensional probability distributions, then so is the convex combination $\lambda\boldsymbol{\mu} + (1-\lambda)\boldsymbol{\nu}$ for every $\lambda$ in $[0, 1]$.

**Definition 1.2.** Suppose $S \subseteq \mathbb{R}^d$ is a convex set and $J : S \to \mathbb{R}$. We say that the function $J$ is **convex** if

$$J[\lambda\boldsymbol{\theta} + (1-\lambda)\boldsymbol{\phi}] \leq \lambda J(\boldsymbol{\theta}) + (1-\lambda)J(\boldsymbol{\phi}), \ \forall \lambda \in [0,1], \ \forall \boldsymbol{\theta}, \boldsymbol{\phi} \in S. \tag{1.1.5}$$

We say that the function $J$ is **strictly convex** if

$$J[\lambda\boldsymbol{\theta} + (1-\lambda)\boldsymbol{\phi}] < \lambda J(\boldsymbol{\theta}) + (1-\lambda)J(\boldsymbol{\phi}), \ \forall \lambda \in (0,1), \ \forall \boldsymbol{\theta}, \boldsymbol{\phi} \in S, \boldsymbol{\theta} \neq \boldsymbol{\phi}. \tag{1.1.6}$$

Equations (1.1.5) and (1.1.6) are stated for a convex combination of *two* vectors $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. But we can make repeated use of these equations and prove the following facts. If $J$ is a convex function mapping a convex set $S$ into $\mathbb{R}$, and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k \in S$, then

$$J\left(\sum_{i=1}^{k} \lambda_i \boldsymbol{\theta}_i\right) \leq \sum_{i=1}^{k} \lambda_i J(\boldsymbol{\theta}_i), \ \text{whenever } [\lambda_1 \ldots \lambda_k] =: \boldsymbol{\lambda} \in \mathbb{S}_k.$$

The above definitions are all algebraic. But in the case where $S$ is an interval $[a, b]$ in the real line (finite or infinite), the various inequalities can be given a simple pictorial interpretation. Suppose we plot the graph of the function $J$. This consists of all pairs $(\theta, J(\theta)))$ as $\theta$ varies over the interval $[a, b]$. Suppose $(\theta, J(\theta))$ and $(\phi, J(\phi))$ are two points on the graph. Then the straight line joining these two points is called the "chord" of the graph. We can assume that the two points are distinct, because otherwise the inequalities (1.1.5) and (1.1.6) become trivial.) Equation (1.1.5) states that for any two points $\theta, \phi \in [a, b]$, the chord joining the two points $(\theta, J(\theta))$ and $(\phi, J(\phi))$ lies *above* the graph of the function $(z, J(z))$ whenever $z$ lies between $\theta$ and $\phi$. Equation (1.1.6) says that, not only does the chord joining the two points $(\theta, J(\theta))$ and $(\phi, J(\phi))$ lie *above* the graph of the function $(z, J(z))$ whenever $z$ lies between $x$ and $y$, but in fact the chord does not even touch the graph, except at the two end points $(\theta, J(\theta))$ and $(\phi, J(\phi))$.

It can be shown that, for all practical purposes, a convex function has to be continuous; see [125, Theorem 10.1]. But if a function is not merely continuous but also differentiable, then it is possible to give alternate characterizations of convexity that is more useful.

**Lemma 1.1.** *Suppose that $J : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable everywhere. If $J$ is convex, then*

$$J(\boldsymbol{\theta} + \boldsymbol{\phi}) \geq J(\boldsymbol{\theta}) + \langle \nabla J(\boldsymbol{\theta}), \boldsymbol{\phi} \rangle, \ \forall \boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^d. \tag{1.1.7}$$

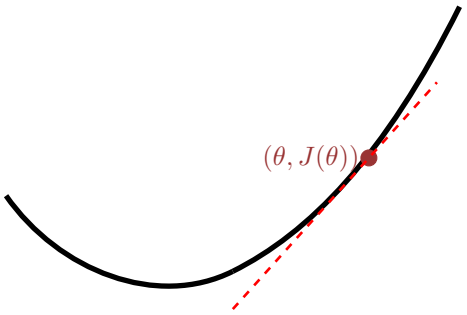



Figure 1.3: The Graph Above Tangent Property of a Convex Function

*If $J$ is strictly convex, then*

$$J(\boldsymbol{\theta}+\boldsymbol{\phi}) > J(\boldsymbol{\theta}) + \langle \nabla J(\boldsymbol{\theta}), \boldsymbol{\phi} \rangle, \ \forall \boldsymbol{\theta}, \ \forall \boldsymbol{\phi} \neq \mathbf{0} \in \mathbb{R}^d. \tag{1.1.8}$$

For a proof of Lemma 1.1, see [125, Theorem 25.1].

Now we give interpretations of the various inequalities above in the case where $d = 1$, so that $J : \mathbb{R} \to \mathbb{R}$. Suppose $J$ is continuously differentiable on some interval $(a, b)$. Then for every $\theta \in (a, b)$, the function $\phi \mapsto J(\theta) + J'(\theta)(\phi - \theta)$ is the tangent to the graph of $f$ at the point $(\theta, J(\theta))$. Thus (1.1.7) says that for a convex function, the tangent lies below the graph. This is to be contrasted with (1.1.5), which says that the chord lies above the graph. Equation (1.1.8) says that if the function is strictly convex, then not only does the tangent lie below the graph, but the tangent touches the graph only at the single point $(\theta, J(\theta))$. Figure 1.3 depicts the "graph above the tangent" property of a convex function, which is to be contrasted with the "graph below the chord" property depicted in Figure 1.2.

The above discussion allows us to introduce another relevant concept.

**Definition 1.3.** A $\mathcal{C}^1$ function $J : \mathbb{R}^d \to \mathbb{R}$ is said to be **$R$-strongly convex** if

$$J(\boldsymbol{\theta}+\boldsymbol{\phi}) \geq J(\boldsymbol{\theta}) + \langle \nabla J(\boldsymbol{\theta}), \boldsymbol{\phi} \rangle + \frac{R}{2}\|\boldsymbol{\phi}\|_2^2, \ \forall \boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^d. \tag{1.1.9}$$

The above concept is taken from [109, section 2.1.3], which also contains several consequences of strong-convexity.

If the function is in fact *twice* continuously differentiable, then we can give yet another set of characterizations of the various forms of convexity.

**Lemma 1.2.** *Suppose $S$ is an open convex subset of $\mathbb{R}^d$, and that $J : S \to \mathbb{R}$ is twice continuously differentiable on $S$. Denote $Q := \nabla^2 J(\cdot) : S \to \mathbb{R}^{d \times d}$. Then*

1. *$J$ is convex if and only if $Q(\boldsymbol{\theta})$ is positive semidefinite for all $\boldsymbol{\theta} \in \mathcal{S}$.*
2. *$J$ is strictly convex if $Q(\boldsymbol{\theta})$ is positive definite for all $\boldsymbol{\theta} \in \mathcal{S}$.*

For a proof of this result, see [125, Theorem 4.5].

### 1.1.3 Some Popular Algorithms

In this subsection, we discuss various popular approaches for minimizing a $\mathcal{C}^1$ objective function $J(\cdot)$. Again the contents of this subsection can be thought of as an overview, with more details being found in Chapter 4.

Historically, the oldest method for finding a minimizer of a $\mathcal{C}^1$ function $J(\cdot)$ is the **Steepest Descent** method, which goes back a few centuries. It can be described as follows:

1. Start with an initial guess $\boldsymbol{\theta}_0 \in \mathbb{R}^d$.
2. At step $t$, compute the gradient $\nabla J(\boldsymbol{\theta}_t)$, and choose a "step size" $\alpha_t$.
3. Update $\boldsymbol{\theta}_t$ via
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t \nabla J(\boldsymbol{\theta}_t). \tag{1.1.10}$$
4. Repeat

Note that the minus sign in front of $\alpha_t$ arises because the aim is to *minimize* the ojbective function. The recipe for choosing the step size $\alpha_t$ (which is also called the "learning rate" in the Machine Learning community) has varied over time. Initially, $\alpha_t$ was chosen via a "one-dimensional search," to minimize $J(bth_t - \alpha \nabla J(\boldsymbol{\theta}_t)$ as a function of $\alpha$. However, the current practice is to choose a predetermined "schedule" of step sizes. See Chapter 4 for details.

During the 1960s, the Steepest Descent method was supplemented by **momentum-based** methods, wherein the current search direction $\nabla J(\boldsymbol{\theta}_t)$ in (1.1.10) is changed to some function of the current guess $\boldsymbol{\theta}_t$ and also the preceding guess $\boldsymbol{\theta}_{t-1}$. We discuss perhaps the two most popular momentum-based algorithms, namely the Heavy Ball method, and Nesterov's Accelerated Gradient method.

The Heavy Ball (HB) method was first introduced in [113]. The update rule in the Heavy Ball method is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \mu(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) - \alpha_t \nabla J(\boldsymbol{\theta}_t), \tag{1.1.11}$$

where $\alpha_t$ is the step size, $\mu$ is known as the "momentum parameter." The Nesterov Accelerated Gradient (NSG) algorithm was introduced in [107], and can be stated as follows (following [143, Eqs. (3)–(4)]):

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \mu_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) - \alpha_t \nabla J(\boldsymbol{\theta}_t + \mu_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1})). \tag{1.1.12}$$

The main difference between the HB method and NAG is that in HB, the search direction to which the step size is applied is $\nabla J(\boldsymbol{\theta}_t)$, whereas in NAG, it is $\nabla J(\boldsymbol{\theta}_t + \mu_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}))$.

### 1.1.4 Sources of Stochasticity

It can be seen that all of the algorithms discussed in the previous subsection are *deterministic*. The same was true of practically all the algorithms of that era. However, the title of this book is *stochastic* algorithms. So wherefrom does the stochasticity arise?

In recent years, the dimension $d$ of the optimization problems has increased enormously. In the design of contemporary neural networks or Large Language Models (LLMs), values of $d$ up to $10^{12}$ are not uncommon. Thus, even if the learner has the "ability" to compute the gradient $\nabla J(\boldsymbol{\theta}_t)$ (which belongs to $\mathbb{R}^d$) "exactly," the learner often uses an *approximate* gradient $\mathbf{h}_{t+1}$ whose computation is far less resource-intensive compared to computing $\nabla J(\boldsymbol{\theta}_t)$ exactly. Often, the approximate gradient $\mathbf{h}_{t+1}$ is also *random*. Thus, in contrast with deterministic algorithms where the iterations lead to a sequence of *deterministic* vectors $\{\boldsymbol{\theta}_t\}$, in this situation the output of the algorithms are a sequence of *random* vectors $\{\boldsymbol{\theta}_t\}$, that is, a **stochastic process**. Clearly, the tools required to analyze the behavior of a stochastic process are more involved than those needed to analyze the behavior of a sequence of vectors. Developing and presenting such tools is one the main objectives of this book. The same methodology can also be used to analyze the situation where there are unavoidable "measurement errors" in computing $\nabla J(\boldsymbol{\theta}_t)$.

As an illustration of these ideas, we study the problem of training a multi-layer neural network. Each neural network **architecture** can be thought of as a map $\mathbf{f} : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}^l$, where $d$ is the number of weights or adjustable parameters, $n$ is the number of inputs, and $l$ is the number of outputs. For ease of presentation, we show in Figure 1.4 a simple neural network with just a handful of neural elements and a few hidden layers. In reality, today's neural networks have millions of neurons if not more, and tens of billions of "weights" if not more. For each choice of weight vector $\boldsymbol{\theta} \in \mathbb{R}^d$, the neural network leads to an "input-output" map that associates an output $\mathbf{f}(\boldsymbol{\theta}, \mathbf{x})$. The "training" of a neural network takes place as

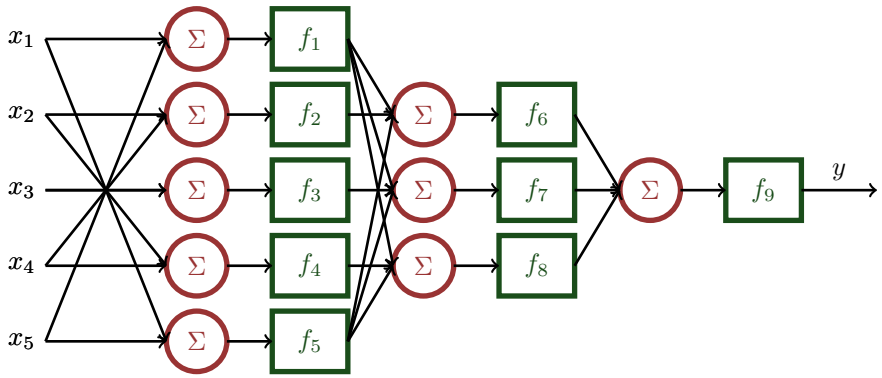



Figure 1.4: A simple multi-layer neural network

follows: The learner is given a large collection of "labelled" input-output pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$. The objective is to minimize the function

$$J(\boldsymbol{\theta}) := \frac{1}{m} L(\mathbf{y}_i, \mathbf{f}(\boldsymbol{\theta}, \mathbf{x}_i)), \tag{1.1.13}$$

where $L : \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}_+$ is known as the "loss function," and measures the difference between the actual output of the network $\mathbf{f}(\boldsymbol{\theta}, \mathbf{x}_i)$, and the desired output $\mathbf{y}_i$. The most commonly used loss function is the least-squares error

$$L(\mathbf{y}, \mathbf{z}) := \|\mathbf{y} - \mathbf{z}\|_2^2. \tag{1.1.14}$$

As discussed in the preceding subsection, almost all methods for minimizing $J(\cdot)$ require the computation of $\nabla J(\boldsymbol{\theta}_t)$, which clearly equals

$$\nabla J(\boldsymbol{\theta}_t) = \frac{1}{m} \nabla_{\boldsymbol{\theta}} L(\mathbf{y}_i, \mathbf{f}(\boldsymbol{\theta}, \mathbf{x}_i)). \tag{1.1.15}$$

Computing $\nabla J(\boldsymbol{\theta}_t)$ in this manner is known as the **batch approach**; see [26] for a discussion of such terms, as well as an excellent survey of optimization methods for large-scale ML problem. In principle, the above quantity is easy to compute, but for the fact that it requires $m$ different individual gradient computations. When $m$ is small, this approach is feasible. But in contemporary ML problems it is not uncommon for $m$, the number of training samples, to be in the billions or even trillions. To reduce the computational effort, a commonly used approach is called **mini-batch**. In this approach, an integer $k \ll m$ is selected. At each iteration, $k$ different indices $i_1, \cdots, i_k$ are chosen from the set $[m] := \{1, \cdots, m\}$, independently, and *with replacement*. (This makes the statistical analysis easier.) Then the approximate gradient $\mathbf{h}_{t+1}$ is defined as

$$\mathbf{h}_{t+1} := \frac{m}{k} \sum_{j=1}^{k} \nabla_{\boldsymbol{\theta}} L(y_{i_j}, \mathbf{f}(\boldsymbol{\theta}, \mathbf{x}_{i_j})). \tag{1.1.16}$$

Clearly $\mathbf{h}_{t+1}$ is a random vector. Since we are choosing the samples independently and with replacement, it is easy to see that the *expected value* of $\mathbf{h}_{t+1}$ equals the true gradient $\nabla J(\boldsymbol{\theta}_t)$. This is the reason for the term $m/k$. This is the motivation for referring to $\mathbf{h}_{t+1}$ as a **stochastic gradient**. In Chapter 4 we will present quite general conditions for the convergence of the **Stochastic Gradient Descent (SGD)** algorithm, where the true gradient $\nabla J(\boldsymbol{\theta}_t)$ is replaced by a random approximation to it denoted by $\mathbf{h}_{t+1}$. We also analyze momentum-based methods when a stochastic gradient is used.

## 1.2 Introduction to Reinforcement Learning

### 1.2.1 Introduction to Reinforcement Learning

As with many phrases in common usage, there is no precise definition of what constitutes "reinforcement learning," often abbreviated to just RL. In the present set of notes, this phrase is used to refer to decision-making with uncertain models, *and in addition, current decisions alter the future behavior of the system.* One consequence of this alteration is that, if the same decision is taken at a future time, the outcome might not be the same. This additional feature, namely that current decisions alter the dynamics of the system under study, usually though not always by altering the surrounding environment, is what distinguishes RL from "mere" decision-making under uncertainty.

Figure 1.5 rather arbitrarily divides decision-making problems into four quadrants. Examples from each quadrant can be given.

- Many if not most decision-making problems fall into the lower-left quadrant of "good model, no alteration." For example, a well-studied control system such as a fighter aircraft has an excellent model thanks to aerodynamical modelling and/or wind tunnel tests. To be specific, the dynamical model of a fighter aircraft depends on the so-called "flight condition," consisting of the altitude and velocity (measured as its Mach number). While the dependence of the dynamical model on the flight condition is nonlinear and somewhat complex, usually sufficient modelling studies are carried out, both before the aircraft is flown and afterwards, that the dynamical model can be assumed to be "known." In turn this permits the control system designers to formulate an optimal (or some other form of) control problem, which can be solved.

- Controlling a chemical reactor would be an example from the lower-right quadrant. As a traditional control system, it can be assumed that the dynamical model of such a reactor does not change as a consequence of the control strategy adopted. However, due to the complexity of a reactor, it is difficult to obtain a very accurate model, in contrast with a fighter aircraft for example. In such a case, one can adopt one of two approaches. The first, which is a traditional approach in the theory of control systems, is to use a nominal model of the system and to treat the deviations from the nominal model as uncertainties in the model. The second, which would move the problem from the lower right to the upper right quadrant, is to attempt to "learn" the unknown dynamical model by probing its response to various inputs. This approach is suggested in [145, Example 3.1]. A similar statement can be made about robots, where the geometry determines the *form* of the dynamical equations describing it, but not the *parameters* in the equations; see for example [139]. In this case too, it is possible to "learn" the dynamics through experimentation. In practice, such an approach is far slower than the traditional control systems approach of using a nominal model and designing a "robust" controller. However, "learning control" is a popular area in the world of machine learning.

- A classic example of a problem belonging to the upper-left corner is a Markov Decision Process (MDP). This topic is studied in Chapter 5 and it forms the backbone of one approach to RL. In an MDP, there is a state space $\mathcal{X}$, and an action space $\mathcal{U}$. While it is possible for the sets to be infinite, in this book we avoid a lot of technicalities by assuming that both sets are finite. Also, in realistic MDPs, the size of the action space $\mathcal{U}$ is very small. Often it is just two! However, though the state space $\mathcal{X}$ can be finite, its cardinality $|\mathcal{X}|$ can be enormous, as shown in some of the examples later in this chapter. In an MDP, at each time instant the learner (also referred to as actor or agent) decides on the action to be taken at that time. In turn the action affects the probabilities of the future evolution of the system. Board games without an element of randomness would also belong to the upper-left quadrant, at least in principle. Games such as tic-tac-toe belong here, because the rules of the game are clear, and the number of possible games is manageable. *In principle*, games such as chess which are "deterministic" (i.e., there is no throwing of dice as in Backgammon for example) would also belong here. Chess is a two-person game in which, for each board position, it is possible to assign the likelihood of the three

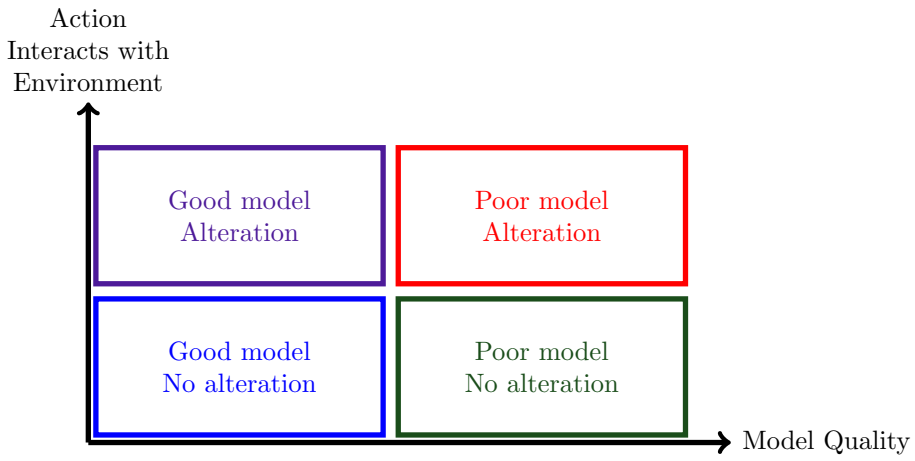



Figure 1.5: The four quadrants of decision-making under uncertainty

possible outcomes: White wins, Black wins, or it is a draw. However, due to the enormous number of possibilities, it is often not possible to *determine* these likelihoods precisely. It is pointed out explicitly in [133] that, merely because we cannot explicitly compute this likelihood function, that does not mean that the likelihood does not exist! However, as a practical matter, it is not a bad idea to treat this likelihood function as being unknown, and to *infer* it on the basis of experiment / experience. Thus, as with chemical reactors, it is not uncommon to move chess-playing from the lower-right corner to the upper-right corner.

- The upper-right quadrant is the focus of these notes. Any problems where the actions taken by the learner alter the environment, in ways that are not known to the learner, are referred to as "reinforcement learning" (RL). Despite the lack of knowledge about the consequences, the learner has no option but to keep trying out various actions in order to "explore" the environment in which the unknown system is operating. As time goes on, some amount of knowledge is gained, and it is therefore possible, at least in principle, to "exploit" the knowledge to improve decision making. The trade-off between exploration and exploitation is a standard topic in RL. A canonical example is MDPs *where the underlying parameters are not known*, and these occupy a major part of these notes. As mentioned above, often complex problems from the lower-right quadrant (such as chemical reactors), or the upper-left quadrant (such as Chess), are also treated as RL problems.

Now we will give a general description of the problem. In a RL problem, there is a state space $\mathcal{X}$ and another action space $\mathcal{U}$. At each time $t$, the learner (also known as the actor or the agent) measures the state $X_t \in \mathcal{X}$. Based on this measurement, the learner chooses an action $U_t$ from a menu of "actions," which is denoted by $\mathcal{U}$, and receives a reward $R(X_t, U_t)$. The rule by which the current action $U_t$ is chosen as a function of the current state $X_t$ is known as a **policy**. The idea is to find the best policy. Figure 1.6 depicts the situation.

While it is possible for the state space $\mathcal{X}$ and the range of possible actions $\mathcal{U}$ to be infinite, in these notes we simplify our lives by restricting $\mathcal{U}$ to be a finite set. In the same way, it is possible to treat "time" as a continuum, but again we simplify life by treating $t$ as a discrete variable assuming values in the set of natural numbers $\mathbb{N} = \{0, 1, \cdots\}$. Thus RL requires the agent to take a set of sequential decisions from a finite menu, at discrete instants of time. When the agent chooses an action $U_t \in \mathcal{U}$, two things happen.

1. The agent receives a "reward" $R_t$. The reward could either be deterministic, or random, and both possibilities are permitted in these notes. The reward could be a negative number, suggesting a

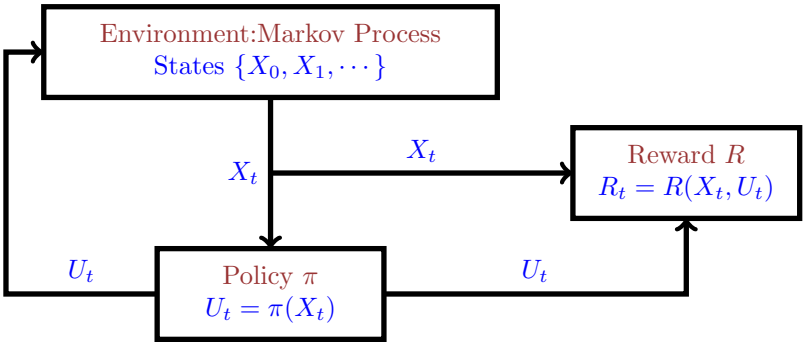



Figure 1.6: Depiction of a Reinforcement Learning Problem

penalty instead of a reward, but the phrase "reward" is standard phraseology. In case the reward is random, it is assumed that the reward lies in a bounded interval in $\mathbb{R}$ which is known *a priori*, in which case the reward can be translated to belong to an interval $[0, M]$. The same transformation can of course be applied if the reward is deterministic. Note that some authors speak of a "cost" which is to be minimized, rather than a reward which is to be maximized. The modifications required to tackle this situation are obvious and we will not comment upon this further. The reward depends not just on the action chosen $U_t$, but also the state $X_t$ of the environment at time $t$. There can be two sources of uncertainty in the reward. In a Markov Decision Problem (MDP), the reward could be a random function of $X_t$ and $U_t$, but with a known probability distribution. In an RL problem, even the probability distribution of the reward is not necessarily known. However, for technical reasons, it is assumed that the upper bound $M$ on the reward is known.

2. The action $U_t$ affects the dynamics of the system. A consequence is that the same action taken at a different time need not lead to the same reward, because in the meantime the "state" of the environment may have changed.

Over the years, the RL research community has given some "structure" to the above rather vague and general description. Specifically:

1. The environment is taken as a Markov process (see Section 2.2) with the state space $\mathcal{X}$, in which the state transition matrix depends on the action taken. So there are $|\mathcal{U}|$ state transition matrices, one for each possible action.

2. If $X_t$ denotes the state of the Markov process at time $t$ and $U_t$ is the action taken at time $t$, then the reward $R$ is taken to be a function $R(X_t, U_t)$. This formalism explains why the same action $U_t \in \mathcal{U}$ taken at a different time may lead to a different reward, because the state $X_t$ may have changed. It is also possible for $R$ to be a "random" function of $X_t$ and $U_t$, so that $X_t, U_t$ only specify the probability distribution of $R(X_t, U_t)$. In such a case, even if the same state-action pair $(X_t, U_t)$ were to occur at a different time, the resulting reward need not the same.

3. Yet another variation is that the reward $R(X_t, U_t)$ (whether random or deterministic) is paid at the *next* time instant $t + 1$. This is the case in some books, notably [148, 145]. In other words, if the Markov process is in state $X_t$ and the action $U_t$ is applied, the reward is $R_{t+1} = R(X_t, U_t)$. This allows those authors to consider the situation where the "next state" $X_{t+1}$ and "next reward" $R_{t+1}$ can share a joint probability distribution, which depends on $X_t$ and $U_t$. Some other authors assume that the reward is *immediate*, so that $R_t = R(X_t, U_t)$. This is the convention adopted in these notes.

4. There are two distinct types of Markov Decision Processes that are widely studied, namely: Discounted reward processes and average reward processes. Each of them has rather a distinct behavior from the

other. In discounted reward processes, there is a "discount factor" $\gamma \in (0,1)$ that is applied to future rewards. The objective is to maximize the sum of the future rewards, where the reward at time $t$ is discounted by the factor $\gamma^t$. Because this future discounted reward is itself random, we maximize the *expected value* of this random variable. In the average reward process, the objective is to minimize the expected value of the *average* of future rewards over time. Because there is no discounting of future rewards, a reward paid at any time contributes just as much to the average as a reward paid at any other time.

5. In the simplest version of the problem, the $|\mathcal{U}|$ state transition matrices, one for each possible action, are assumed to be known, as is the reward function. In the case where the reward is a random function of $X_t$ and $U_t$, it is assumed that the probability distribution of $R(X_t, U_t)$ is known. It is also assumed that the state $X_t$ of the Markov process can be observed by the agent, and can be used to decide the action $U_t$. A key concept in RL is that of a "policy" $\pi$ which is a map from the state space $\mathcal{X}$ of the Markov process to the set of actions $\mathcal{U}$. The objective here is to choose the optimal policy, which maximizes the expected value of the discounted future reward over all possible policies. This version of the problem is usually known as a **Markov Decision Process (MDP)**.[1] It is usually viewed as a precursor to RL. In "proper" RL, neither the Markovian dynamics nor the reward are assumed to be known, and must be learned on the fly so to speak. However, knowing the solution approaches to the MDP is very useful in solving RL problems. It should be pointed out that some authors also use the phrase RL to the problem of finding the optimal policy in an MDP where the parameters of the problem are completely known.

A dominant theme in RL is the trade-off between "exploration" and "exploitation." By definition, the agent in an RL problem is operating in an unknown environment. However, after sometime a reasonably good model of the environment is available, and a set of actions that is reasonably "rewarding" is also identified. Should the agent then persist with this set of actions, or occasionally attempt something new, just on the off-chance that there is a better set of actions available? Let us take a concrete example. A successful chess player would have evolved, over the years, a set of strategies that work well for him/her. Should the player persist with the time-proven strategies (exploitation) until someone starts beating him/her, or occasionally try something completely different just to see what happens (exploration)? The answer is not clear, and is likely to vary from one domain to another. To illustrate the domain dependence of the solution, suppose a person moves to a new town and wishes to find the best coffee shop. Then it is probably sufficient to try each nearby coffee shop just once (or just a few times), because most coffee shops have standardized protocols for preparing coffee, so that the quality is not likely to vary very much from one visit to the next. Therefore a person can stick to the coffee shop that is most appealing after a few visits, and there is very little incentive for further "exploration," only "exploitation." In contrast, it can be assumed that the course of a chess match between two players at the highest level almost invariably leads to a previously unexplored set of positions. Thus persisting with a stock strategy would invariably lead to suboptimal results, and there must be greater emphasis on exploration than in the coffee shop example.

There are a couple of methods for quantifying the trade-off between exploration and exploitation. We begin with the observation that almost any "sensible" learning algorithm would converge to a nearly optimal policy within a finite number of time steps. Here are two ways to measure how good the algorithm is:

1. Given an accuracy $\epsilon$, one can measure how many time steps are required for the policy to be within $\epsilon$ of the optimal policy.[2] The faster a policy becomes $\epsilon$-suboptimal, the better it is. Implicit in this characterization is the assumption that a policy is *not* penalized for how badly it performs before it achieves $\epsilon$-suboptimality – just the time it takes to achieve $\epsilon$-suboptimality status.

[1] There is a variant where the state $X_t$ cannot be observed directly; instead one observes an output $Y_t$ which is either a deterministic or a random function of $X_t$. This problem is known as a Partial Observed Markov Decision Process (POMDP). This problem is not discussed at all in these notes.

[2] This idea is made precise in subsequent chapters.

2. The other measure is to see what the reward would have been, had the learner somehow magically implemented the optimal policy right at the outset, and compare it against the actually achieved performance. This quantity is called the "regret" and is defined precisely later on. The difference between minimizing the regret and minimizing the time for achieving $\epsilon$-optimality is that in the latter, the performance of the algorithm before achieving $\epsilon$-optimality is not penalized, whereas it is counted as a part of the regret.

Clearly, the two criteria are not the same. A learning strategy that converges relatively quicky, but performs poorly along the way would be rated highly under the first criterion, and poorly under the second criterion.

Within the broad area of Machine Learning (ML) or Artificial Intelligence (AI), RL stands quite distinctly apart from other popular areas such as supervised learning (which is what many people mean when they talk about ML), and unsupervised learning. In supervised learning, the main goal is **generalization**. Thus the learner is shown an amount of labelled "training data." The labels could be binary, in which case the problem is called binary classification. Or the label set could be some finite set, in which case the problem is called multi-class classification. Finally, the label set could be a continuum, like $[0, 1]$ of the set of real numbers, in which case the problem is called regression. After the training phase, the learner is then shown "testing data" for which the correct labels are known to the evaluator, and the learner is asked to predict these correct labels. The extent to which the learner is able to predict the correct labels serves as a measure of the quality of the learning algorithm. For instance, detecting whether a credit card transaction is legitimate or fraudulent, or a growth represents a malignant cancer tumor or just a benign growth, are examples of supervised learning problems. A well-known recent example is the ImageNet database [65], created as a part of the LSVRC (Large Scale Visual Recognition Challenge). It consists of roughly 14 million images that are hand-curated. The full set, or some subset thereof, is presented to some supervised learning algorithm, whose parameters are then adjusted to achieve good performance on the training inputs. While there are several mathematical formalisms of this class of problems, the so-called PAC (Probably Approximately Correct) learning formulation is among the more popular approaches. Deep neural networks are an example of solving supervised learning problems using the PAC formalism.

At the other end of the spectrum lies unsupervised learning. In this problem, the learner is simply given a set of data, without any labels of any sort. The task of the learner is to collect the data into various "clusters" as they are known in the world of statistics. Once the training data is clustered, the learner is given a set of testing data. Each element of the testing data is then assigned to the cluster to which it most naturally belong. One way of stating the clustering problem is via the $K$-means algorithm. In this algorithm, the clusters are chosen in such a way that. the elements of each cluster are closer to the centroid of the cluster to which it belongs, than to the centroid of all other clusters. Figure 1.7 illustrates the outcome of one such clustering. It can be seen that there are five clusters, whose centroids are denoted by stars. In general, solving the $K$-means problem *exactly* is NP-hard. Hence various approximations are used. Unsupervised learning is not discussed further in these notes.

One can explain the difference between supervised learning and Reinforcement Learning as follows (though other explanations are also possible). In supervised learning, the learner gets immediate and (mostly) accurate feedback about the correctness of the label assigned to the testing data. However, in RL, the feedback to the learner is long-term, and statistical in nature.

### 1.2.2 Some Examples of Reinforcement Learning

In this section we briefly discuss a few motivating problems that can serve as illustrations of reinforcement learning. We will return to a couple of these problems again in future chapters.

There are several examples of reinforcement learning available in the literature. The books [119, 145] contain several examples, while the book [27] is primary devoted to examples of RL in a variety of areas, including healthcare, transportation, finance etc. Perhaps the most "famous" application of RL is a general-purpose algorithm that can be taught to play a variety of games, including Chess, Shogi and Go [37, 136]. Robot control, including path-planning in the presence of (possibly unknown) obstacles is another popular

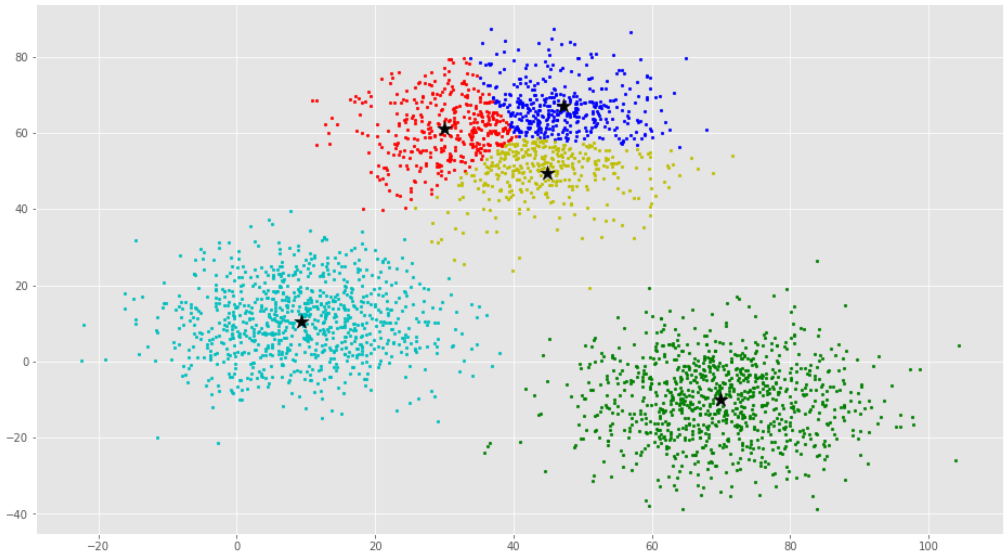



Figure 1.7: Typical output of a clustering algorithm

application. Some RL texts and papers study the problem of balancing a stick on a moving cart, which is known in control theory as the "inverse pendulum" problem. This might not be a good application of RL, because the system can be modelled very precisely, which in turn leads to very efficient control laws. However, by viewing this well-studied contro-theoretic problem as a problem in RL, the RL research community has developed several new and interesting learning paradigms. Another application is that of deciding an optimal strategy for the game of Blackjack, sometimes also called Twenty One. We will study this example, either in its full form or in a simplified form, in detail at appropriate places in these notes.

**Multi-Arm Bandit Problems**

This problem is a generalization of the "slot machine" in gambling casinos around the world, whereby the player pulls a lever and receives a random payoff. In order to pull the lever, the player has to insert some money, and the expected value of the payoff is less than the amount to be inserted; that is how the casino makes money. However, in our model, we ignore the fact that a player has to pay to play, and focus strictly on the payout part of it.

Suppose a player is facing $m$ slot machines, or "bandits," each of which has random payout. Specifically, let $X_i$ denote the random payout of the $i$-th bandit. Then $X_i$ has an unknown expected (mean value) payout, as well as an unknown probability distribution around this mean value. To avoid unnecessary technicalities, it is assumed that all returns are nonnegative, and that there is a fixed known upper bound $M$ on the payout of each machine, which can be taken as 1 without any loss of generality. Therefore the return of each arm has a probability distribution $\phi_i$ is supported on the set $[0, 1]$. Define

$$\mu_i = \int_0^1 x\phi_i(x)dx$$

to be the mean or expected value of $X_i$. Of course, the player does not know either $\mu_i$ or $\phi_i(\cdot)$. But the player is able to "pull the arm" of each bandit and see what happens. This generates (we assume) statistically independent samples $x_{i1}, \cdots, x_{im}$ of the random variable $X_i$. Based on the outcome of these experiments, the player is able to make *some estimate* of $\mu_i$ for each bandit $i$. These estimates can be used to determine future strategies.

Note that if the quantities $\mu_1, \cdots, \mu_m$ are known, then the problem is simple: The player should always

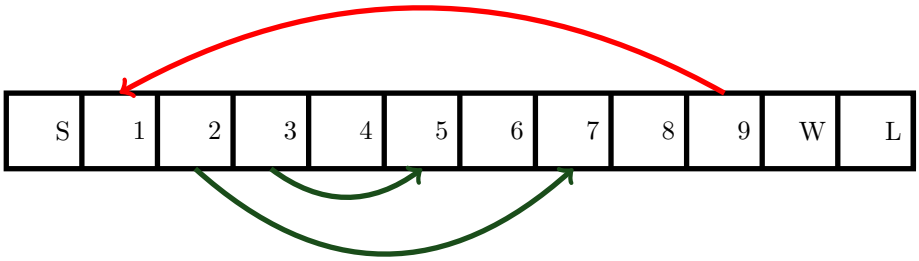



Figure 1.8: Toy Snakes and Ladders Game

play the machine that has the highest expected payout. But the challenge is to determine which machine this is, on the basis of experimentation. As stated above, there are many reasonable algorithms that will asymptotically (as the number of trials increases towards infinity) determine the arm(s) with the best return(s). Therefore one way to assess the performance of an algorithm is its "regret," that is, the return achieved over the course of learning, subtracted from the optimal return of always choosing the arm with the highest return. Bandit theory is a very well-developed branch of Reinforced Learning, which is somewhat orthogonal to Markov Decision Problems. So that topic, while central to RL, is not discussed further in this book.

**Snakes and Ladders**

We all know the ancient snakes and ladders game, where the objective is for a player to pass from the start to the end while avoiding the snakes and taking advantage of the ladders. We will modify the game slightly by adding the possibility of losing if the player overshoots the last square. A toy version of the game is shown below (it is also studied in Section 2.2).

The rules of the game are as follows:

- Initial state is S.
- A four-sided, fair die is thrown at each stage.
- Player advances as many squares as the outcome of the throw, followed by the impact of the snake or ladder, if any.
- Player must land exactly on $W$ to win.
- If implementing a move causes the player to hit or to cross $L$, then the player loses. Landing exactly on $L$ also loses.
- Hitting the square $W$ leads to a reward of 5 and hitting the square $L$ leads to a reward of $-5$. The reward in every other square is 0.

At each stage of the game, the player has two choices: to roll the die and take a chance on the outcome, or not to roll it. We can ask: What is the best strategy for a player as a function of the square currently being occupied? Clearly, it depends on whether the expected return from playing exceeds the expected return from not playing.

**Blackjack**

Blackjack is a popular game in gambling casinos around the world. The player plays against the "house."[3] The player and the house draw cards in alternation. The objective is to draw cards such that the total of

[3]Actually, it is possible to have more than one player plus the "house." However, to simplify the problem, we study only the case of one player against the "house."

| $(P, H)$ | $R$ |
|---|---|
| $P < H$ | -2 |
| $P = H$ | 1 |
| $P > H, P \neq W$ | 2 |
| $(P, H) = (W, *)$ | 5 |
| $(P, H) = (L, *)$ | -5 |

Table 1.1: Reward Table for Simplified Blackjack Game

the cards is as close to 21 as possible without exceeding it. That is why sometimes Blackjack is also called "Twenty-One." The formulation of Blackjack as a problem in RL is discussed in [145, Example 5.1]. At each time instant, the player has ony two possible actions: To ask for one more card, or not. These are known as "hit" and "stick" respectively. So the set of possible actions $\mathcal{U}$ has cardinality two. If the player draws a card, the outcome is obviously random. Either way, the house also draws a card whose outcome is random. It is shown in [145, Example 5.1] that the process can be modelled by a Markov process with 200 states, so that $|\mathcal{X}| = 200$. However, tracing out all possible future evolutions of the game, starting from the current state, is nearly impossible, and simulations are the only way to analyze the problem.

We now present a simplified version of Blackjack. Obviously, drawing a card leads to the player's total increasing by anywhere from 1 to 11.[4] So if the player's current total is 10 or less, the player cannot possibly lose by drawing, and may get closer to winning. So the optimal strategy from such a position is not in doubt. With that in mind, we replace the drawing of a card by the rolling of a fair four-sided die, with all four outcomes being equally probable. It does not matter what the "target" total is, because if the target total is $T$, then so long as the player's total is $T - 4$ or less, the player should roll the die. With this in mind, we can think of the player's states as $\{0, 1, 2, 3, W, L\}$, with $W$ and $L$ denoting Win and Lose respectivey. If the player's current total plus the outcome of the die *exactly equals* 4, the player wins, and if the total exceeds 4, the player loses. But there is an added complication, which is the total of the "House." Let us assume that the House policy is to "stick" whenever it gets within 3 of the designated total. Hence it can be assumed that the House total is in $\{1, 2, 3\}$. Now the object of the game is not merely to get as close to $W$ without going over, but also to beat the House total. Hence the reward for this game can be specified as shown in Table 1.1. With this reward structure, at each position, the player has the option of rolling the die, or not. It turns out that this game is more complex than just the player playing snakes and ladders. We will analyze this game also in later chapters.

**Backgammon**

Backgammon is a board game played by two players on a board with (essentially) 24 positions, with each player throwing two six-sided dice at each turn. Figure 1.9 shows a typical board position. The game combines chance (random outcome of throwing the dice) and strategy (what a player does based on the outcome of the dice).

Unlike in Blackjack, the range of possible actions available to a player at each turn is quite large. This game is well-suited to a technique called "temporal difference" or TD-learning, which is studied in Section 4.2. Tesauro has published several articles on how to program a computer to play backgammon, including [154, 155, 156]. See [145, Section 16.1] for a detailed description of the rules of backgammon and the TD implementation of Tesauro.

**AlphaGo and AlphaZero**

It would not be a exaggeration to say that a great deal of the public attention to artificial intelligence arises from the success of two programs, namely AlphaGo and AlphaZero. In 2016, a UK-based company called Deep Mind (since acquired by Google) created a program called AlphaGo to play Go, a board game played on a grid of $19 \times 19$ places. In a five-game match held in Seoul, Korea between the 9th and 15th of March, AlphaGo played against Lee Sedol, who was an eighteen-time world champion, though he was

[4] An Ace can be counted as either 1 or 11 as per the player's choice.

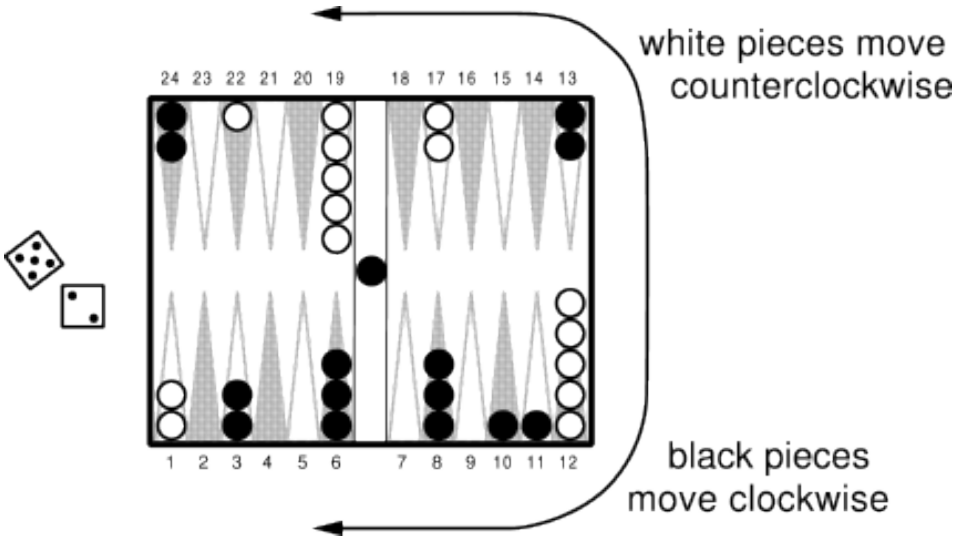



Figure 1.9: A typical board position in backgammon

not world champion at that time. AlphaGo won four out of the five games. It was the first instance of a computer defeating a ranking Go player. A year later, in 2017, AlphaGo defeated the top-ranked player Ke Jie. In a series of three matches played between 23rd and 27th May, AlphaGo won all three matches.

Twenty years earlier IBM had developed the Deep Blue platform to play chess. Obviously, over such a long period of time, there would be massive improvements in computing hardware. Indeed, AlphaGo ran on a collection of Tensor Processing Units (TPUs), which are specially designed to carry out the type of computations required by AlphaGo (as opposed to general-purpose CPUs, or Central Processing Units).

Even at that time, Deep Mind had in its possession a more advanced program called AlphaZero, but did not deploy it against Ke Jie. AlphaZero could be programmed to play chess, Go and shogi (Japanese chess). AlphaZero defeated AlphaGo while playing Go, defeated Stockfish (a popular chess-playing program), and Elmo (a popular program to play shogi). However, in the eyes of many, the real interest in AlphaZero arose from the manner in which is trained itself. Recall that the Deep Blue platform developed by IBM relied on human inputs, and a search technique, in order to analyze board positions and determine its next move. In contrast, AlphaZero used an entirely different approach, whereby it improved itself through "self-play", through a mathematical method known as Monte Carlo tree search (MCTS) algorithm. Thus *the same program* is able to "teach itself" to play different games. A popular description how AlphaZero goes about its self-appointed task can be found in [37]. Those interested in the mathematical details can find them in [136].

One of the intriguing philosophical aspects of AlphaZero is the fact that, as its name implies, AlphaZero starts from zero, that is, *without any prior knowledge*. Its superior performance compared to other programs that make use of prior knowledge has been interpreted by some AI researchers to claim that "prior knowledge" is not necessary to achieve top performance. To understand why this is interesting, let us consider the same question, but changing "chess" to "cooking." Suppose you wish to become a master chef. Should you first learn under someone who is already a master chef, and experiment on your own only *after* you have achieved some level of proficiency? Or is it better for you to undertake trial and error right from Day One? Most of us would instinctively answer that learning from a master (i.e., tapping domain knowledge) would be better. One of the intriguing aspects of the success of AlphaZero is that, when it comes to a computer learning to play chess, domain knowledge *apparently* does not confer any advantage. However, at the moment the role of prior domain knowledge in AI is still a topic for further research. It is not clear whether the success of AlphaZero is a one-off phenomenon, or a manifestation of a more universally applicable principle.

## 1.3 About These Notes

This section will be written last.

## Notes and References

The problem of *deterministic* methods of optimization has been studied for decades, and picked up momentum during the 1960s and the 1970s. Historically, these algorithms assumed that various measurements (such as function evaluations or gradient evaluations) were *noise-free*. Some of these deterministic algorithms in turn inspired various stochastic algorithms that are widely used, many of which are studied in this book. An excellent book on deterministic optimization algorithms is [47]. In [128], the author briefly discusses the relationship between the stochastic approximation algorithm and deterministic algorithms, specifically the Newton-Raphson algorithm, when only noisy measurements are available. This connection is made precise in the material on Zap $Q$-learning in Chapter 6.

There are a great many books devoted to both convex analysis and convex optimization. Among the many excellent books on convex analysis, two noteworthy ones are [125, 62]. These books discuss convex functions in general, though [62] also has some discussion of convex optimization. Among the bext books dedicated to convex optimization is [28].

We will discuss the formulation of Reinforcement Learning (RL) in Chapters 5 and 6. For the present, the reader is directed to the following *representative sample* of papers that discuss the practical applications of RL: [157, 1, 105, 67, 164, 78, 54].

# Chapter 2

# Convergence of Stochastic Processes

As mentioned in Section 1.1.4, when attempting to solve very high-dimensional optimization problems, it is often desirable to introduce some randomness, to avoid computing gradient vectors of large dimensions. One consequence of this is that, in contrast with deterministic optimization algorithms that produce a *sequence of vectors*, stochastic algorithms produce a *stochastic process*. A similar statement applies to Reinforcement Learning problems. The central problem studied in this book is the convergence of various stochastic processes that arise in nonconvex optimization and Reinforcement Learning. Therefore, in this chapter we present some "universal" (that is, widely applicable) theorems that can be used to establish the convergence of stochastic processes. The actual applications of these convergence theorems to specific situations are deferred to subsequent chapters. Specifically, applications to nonconvex optimization are studied in Chapter 4, and applications to RL are studied from Chapter 5 onwards.

Note that the contents of the chapter are a mixture of "standard" material and "advanced" material. Secifically, the material contained in Sections 2.1.1 and 2.1.2 is quite basic, and can be found in several texts. Nevertheless, even a knowledgeable reader may wish to browse these sections in order to become familiar with the notation used in this book. However, the contents of Section 2.1.3 are at a more advanced level. Good references for this material are [44, 173]. Similarly, the material in Section 2.2.1 is quite standard. However, the material in Sections 2.2.2 and 2.2.3 is not so standard. Some of it can be found in [111], but much of it is stated here for the first time, so far as the author is aware. Finally, while the material in Section 2.3.1 is standard, some of the material in Section 2.3.2 is new and presented in book form for the first time (though it is contained some publications by the author).

A typical theorem in this domain gives *sufficient conditions* for convergence. Thus, if the hypotheses of the theorem hold, then convergence is guaranteed. However, convergence might take place even when the hypotheses of the theorem do not hold. Constantly expanding the realm of applicability of convergence theorems is an on-going and vital activity.

## 2.1 Random Variables and Stochastic Processes

In this first section of the chapter, we introduce various topics related to measure, probability, and random variables. The contents of the first two subsections are fairly elementary, and the treatment is fairly cursory. It is suggested that readers who are encountering these topics for the first time should supplement this material by the references cited here. The concept of the conditional expectation of a random variable with respect to a $\sigma$-algebra is not elementary. It is introduced in this subsection to facilitate a precise definition of a Markov process in Secton 2.2.

Probability theory is a well-dveloped subject, and there is no dearth of excellent texts. Thus the suggested reading list is limited to slightly more advanced texts,, wherein the topics of conditional expection (in this section) and martingales (Section 2.3) are covered in depth.

Axiomatic probability theory can be said to have been started by Kolmogorov, and his very brief monograph [79] gives a good motivation for the subject. While the specific contents of this book have been superceded by later books, [79] is a valuable resource for shedding light on the origins of probability theory, and its evolution during its early years. For a thorough treatment of topics from measure theory, the reader can consult [10]. A very good overview of probability is found in the book (with same name) [29]. Topics such as conditional expectation and martingales are briefly discussed in [20, 29], and a more detailed treatment can be found in [173, 16, 44]. In particular, [44] has both a large number of examples as well as a lot of exercises.

### 2.1.1 Random Variables

**Definition 2.1.** Suppose $\Omega$ is a set and that $\mathcal{F}$ is a collection of subsets of $X$. Then $\mathcal{F}$ is said to be a **$\sigma$-algebra**[1] if $\mathcal{F}$ satisfies the following axioms:

(S1). $\Omega \in \mathcal{F}$.

(S2). If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, where $A^c$ denotes the complement of $A$ in $\Omega$.[2]

(S3). If $\{A_i\}_{i\geq 1}$ is any countable sequence of sets belonging to $\mathcal{F}$, then

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}. \tag{2.1.1}$$

The pair $(\Omega, \mathcal{F})$ is called a **measurable space**.

**Definition 2.2.** Suppose $(\Omega, \mathcal{F})$ is a measurable space. A function $P : \mathcal{F} \to [0, 1]$ is called a **probability measure** if it satisfies the following axioms:

P1. $P(\Omega) = 1$.

P2. $P$ is *countably additive*; that is: Whenever $\{A_i\}$ are pairwise disjoint sets from $\mathcal{F}$, we have that

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \tag{2.1.2}$$

The triple $(\Omega, \mathcal{F}, P)$ is called a **probability space**.

Note that if $\Omega$ is a finite or countable set, it is customary to take $\mathcal{F}$ to be the "power set" of $\Omega$, that is, the collection of all subsets of $\Omega$, often denoted by $2^{\Omega}$. Suppose $\{p_i\}$ is a sequence of nonnegative numbers, of the same cardinality as $\Omega$, and that $\sum_i p_i = 1$. Let us enumerate the elements of $\Omega$ in some fashion as $\{\omega_1, \cdots, \}$, and assign a nonnegative weight $p_i$ to each element $\omega_i \in \Omega$. Now suppose $A \subseteq \Omega$, and define

$$P(A) = \sum_{\omega_i \in A} p_i = \sum_{\omega_i \in \Omega} I_{\{\omega_i \in A\}}, \tag{2.1.3}$$

where $I_{\{\omega_i \in A\}}$ is the indicator function that equals $\omega_1$ if $i \in A$ and 0 if $\omega_i \notin A$. Then it is easy to verify that $(\Omega, 2^{\Omega}, P)$ is a probability space. However, if $\Omega$ is an uncountable set, e.g., the real numbers, then the above approach of assigning weights to individual elements does not work.

**Definition 2.3.** Suppose $(\Omega, \mathcal{F})$ and $(\mathcal{X}, \mathcal{G})$ are measurable spaces. Then a map $f : \Omega \to \mathcal{X}$ is said to be **measurable** if $f^{-1}(S) \in \mathcal{F}$ for all $S \in \mathcal{G}$.

[1] The term $\sigma$-field is more popular, but this terminology is preferred here.

[2] Note that (S1) and (S2) together imply that $\emptyset \in \mathcal{F}$.

Thus a map from $\Omega$ into $\mathcal{X}$ is measurable if the preimage of every set in $\mathcal{G}$ under $f$ belongs to $\mathcal{F}$.

**Definition 2.4.** Suppose $(\Omega, \mathcal{F}, P)$ is a probability space, and $(\mathcal{X}, \mathcal{G})$ is a measurable space. A function $X : \Omega \to \mathcal{X}$ is said to be a **random variable** if it is measurable, that is,

$$X^{-1}(S) \in \mathcal{F}, \ \forall S \in \mathcal{G}. \tag{2.1.4}$$

In such a case, for each set $S \in \mathcal{G}$, the quantity

$$P(X^{-1}(S)) =: P_X(S)$$

is called the **probability that** $X \in S$.

In the above definition, $(\mathcal{X}, \mathcal{G})$ is called the "event space," and the sets belonging to the $\sigma$-algebra $\mathcal{G}$ are called "events," because each such set has a probability associated with it via (2.1.4). The triple $(\Omega, \mathcal{F}, P)$ is called the "sample space."

**Example 2.1.** Suppose we wish to capture the notion of a two-sided coin that comes up $H$ for heads 60% of the time, and $T$ for tails 40% of the time. In such a case, the event space (the set of possible outcomes) is just $\mathcal{X} = \{H, T\}$. Because the set $\mathcal{X}$ is finite, the corresponding $\sigma$-algebra $\mathcal{G}$ can be just $2^{\mathcal{X}} = \{\emptyset, \{H\}, \{T\}, \mathcal{X}\}$. The sample space $(\Omega, \mathcal{F}, P)$ can be anything, as can the map $f : \Omega \to \mathcal{X}$, provided only that two conditions hold: First,

$$f^{-1}(\{H\}) = \{\omega \in \Omega : f(\omega) = H\} \in \mathcal{F}, f^{-1}(\{T\}) = \{\omega \in \Omega : f(\omega) = T\} \in \mathcal{F}.$$

(Actually, either one of the conditions would imply the other.) Second,

$$P(f^{-1}(\{H\}) = 0.6, P(f^{-1}(\{T\}) = 0.4.$$

□

**Definition 2.5.** Suppose $X$ is a random variable defined on the sample space $(\Omega, \mathcal{F}, P)$ taking values in $(\mathcal{X}, \mathcal{G})$. Then the **$\sigma$-algebra generated by** $X$ is defined as the smallest $\sigma$-algebra contained in $\mathcal{F}$ with respect to which $X$ is measurable, and is denoted by $\sigma(X)$.

**Example 2.2.** Consider again the random variable studied in Example 2.1. Thus $\mathcal{X} = \{H, T\}$ and $\mathcal{G} = 2^{\mathcal{X}} = \{\emptyset, \{H\}, \{T\}, \mathcal{X}\}$. Now suppose $X$ is a measurable map from some $(\Omega, \mathcal{F}, P)$ into $(\mathcal{X}, \mathcal{G})$. Then all possible preimages of sets in $\mathcal{G}$ are:

$$\emptyset = X^{-1}(\emptyset), \Omega = X^{-1}(\mathcal{X}), A := X^{-1}(\{H\}), B := X^{-1}(\{T\}) = A^c = \Omega \setminus A.$$

Thus the *smallest possible* $\sigma$-algebra on $\Omega$ with respect to which $X$ is measurable consists of $\{\emptyset, A, A^c, \Omega\}$. Therefore this is the $\sigma$-algebra of $\Omega$ generated by $X$. Any other sets in $\mathcal{F}$ are basically superfluous. We can carry this argument further and simply take the sample space $\Omega$ to be the same as the event space $\mathcal{X}$, and $X$ to be the identity operator on $(\Omega, \mathcal{F}, P)$ into $(\Omega, \mathcal{F})$. Thus $\Omega = \mathcal{X} = \{H, T\}$, and $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$. Further, we can define $P(\{H\}) = 0.6$, $P(\{T\}) = 0.4$. This is sometimes called the **canonical representation** of the random variable $X$. Usually we can do this whenever the event space is finite or countable. □

Suppose $\mathcal{S}$ is a collection of sets, each of which belongs to $\mathcal{F}$. Then $\sigma(\mathcal{S})$ denotes the smallest $\sigma$-algebra containing all sets in the collection $\mathcal{S}$. Previously we had defined $\sigma(X)$, the $\sigma$-algebra generated by a random variable $X$. The two usages are consistent, in the following sense. Suppose $X$ is a random variable on $(\Omega, \mathcal{F}, P)$ mapping $\Omega$ into $(\mathcal{X}, \mathcal{G})$, and let $\mathcal{S}$ consist of all preimages in $\Omega$ of sets in $\mathcal{G}$. Then $\sigma(X)$ and $\sigma(\mathcal{S})$ are the same.

Originally, the phrase "random variable" was used only for the case where the event space $\mathcal{X} = \mathbb{R}$, and the $\sigma$-algebra is the so-called **Borel $\sigma$-algebra** denoted by $\mathcal{B}$, which is defined as the smallest $\sigma$-algebra of subsets

of $\mathbb{R}$ that contains all closed subsets of $\mathbb{R}$. Random quantities such as the outcomes of coin-toss experiments were called something else (depending on the author). Subsequently, the phrase "random variable" came to be used for *any* situation where the outcome is uncertain, as defined above. Much of Reinforcement Learning (RL) has to with Markov Decision Processes (MDPs), which are introduced in Chapter 5. In the context of MDPs, often the Markov process evolves over a finite set, and the action space is also finite. So a lot of the heavy machinery above is not needed to describe the evolution of an MDP. However, in RL, the parameters of the MDP need to be estimated, including the reward–and these are real-valued quantities. Similarly, in optimization problems, the objective is to find a vector $\boldsymbol{\theta}^* \in \mathbb{R}^d$ that minimizes some objective function $J : \mathbb{R}^d \to \mathbb{R}$. This is attempted by generating a sequence $\{\boldsymbol{\theta}_t\}$, which, one hopes, will converge to $\boldsymbol{\theta}^*$. The process of determining $\boldsymbol{\theta}_{t+1}$ from the past history $\boldsymbol{\theta}_0^t$ is the "algorithm." When the algorithm is statistical in nature, the resulting sequence of iterations $\{\boldsymbol{\theta}_t\}$ is a stochastic process assuming values in $\mathbb{R}^d$. So in this book, we introduce the axiomatic foundation to deal with random variables that assume values in a continuum such as $\mathbb{R}$. When we do that, the event set $\mathcal{X}$ equals $\mathbb{R}$ or some subset thereof, and $\mathcal{F}$ equals the Borel $\sigma$-algebra.

Next we introduce the concept of the "expected value" of a real-valued random variable that assumes only finitely many values. Suppose $X$ is a real-valued random variable assuming values in some finite set $\mathcal{X} = \{x_1, \cdots, x_n\}$, and that $f : \mathcal{X} \to \mathbb{R}$. Then we can think of $f(X)$ as a *real-valued* random variable. Moreover, if $X = x_i$ with probability $p_i$, then $f(X) = f(x_i)$ with probability $p_i$. Note that the values $\{f(x_1), \cdots, f(x_n)\}$ need not all be distinct. We define the **expected value** of $f(X)$ as

$$E[f(X), \mathbf{p}] := \sum_{i=1}^{n} f(x_i) p_i. \tag{2.1.5}$$

Note that the above definition is valid even if the values $\{f(x_1), \cdots, f(x_n)\}$ are not all distinct. Moreover, while $X$ can be an abstract random variable, $f(X)$ has to be real-valued; otherwise we cannot talk about its expected value. In particular, if $X$ is tself a *real-valued* random variable assuming finitely many real values $\{x_1, \cdots, x_n\} \subseteq \mathbb{R}$, then its expected value can be defined as

$$E[X, \mathbf{p}] := \sum_{i=1}^{n} x_i p_i \tag{2.1.6}$$

if $X$ has only finitely many values. However, if $X$ takes values in a continuum, then the summation has to be replaced by an integral. Hence we digress to give a very brief introduction to real-valued random variables that are not restricted to assuming values in a finite set. Here we skirt over many technical issues. For a proper treatment of the concepts below, the reader is referred to [29, 10].

We begin by introducing the cumulative distribution function, often abbreviated to just cdf. Good references for this and related topics are [29, Section 2.5], [16, Section 14] and [43, Section 9.1]. Suppose $X$ is a real-valued random variable, mapping some sample space $(\Omega, \mathcal{F}, P)$ into the event space $(\mathbb{R}, \mathcal{B})$. Since every semi-infinite interval $(-\infty, a]$ belongs to $\mathcal{B}$, the preimage $X^{-1}(-\infty, a] \in \mathcal{F}$. Therefore the probability

$$P(X^{-1}((-\infty, a])) =: \Pr\{X \leq a\}$$

is well-defined for each $a \in \mathbb{R}$.

**Definition 2.6.** Suppose $X$ is a real-valued random variable, which maps some probability space $(\Omega, \mathcal{F}, P)$ into the event space $(\mathbb{R}, \mathcal{B})$. Then the function $\Phi_X : \mathbb{R} \to [0, 1]$ defined by

$$\Phi_X(a) := \Pr\{X \leq a\} \tag{2.1.7}$$

is called the **cumulative distribution function (cdf)** of $X$.

Some properties of the cdf are given next. Most of these are easy to verify.

- If $a < b$ then $\Phi_X(a) \leq \Phi_X(b)$. So the cdf is *nondecreasing.*
- $\Phi_X(a) \to 0$ as $a \to -\infty$ and $\Phi_X(a) \to 1$ as $a \to +\infty$.
- $\Phi_X(\cdot)$ is continuous from the right, and has limits from the left. Thus

$$\lim_{a \to b^+} \Phi_X(a) = \Phi_X(b), \tag{2.1.8}$$

$$\lim_{a \to b^-} \Phi_X(a) \text{ exists and is } \leq \Phi_X(b), \tag{2.1.9}$$

- If $\Phi_X(\cdot)$ is not continuous at $b$, but has a "jump", that is

$$\lim_{a \to b^-} \Phi_X(a) < \lim_{a \to b^+} \Phi_X(a),$$

  then

$$\Pr\{X = b\} = \lim_{a \to b^+} \Phi_X(a) - \lim_{a \to b^-} \Phi_X(a).$$

- The set of points at which the cdf is not continuous is either finite or countable.
- If the cdf is continuous at $b$, then $\Pr\{X = b\} = 0$.

In case there is a function $\phi_X : \mathbb{R} \to \mathbb{R}_+$ such that

$$\Phi_X(a) = \int_{-\infty}^{a} \phi_X(u)\, du, \tag{2.1.10}$$

Where the integration can be with respect to the Lebesgue measure, then $\phi(\cdot)$ is called the **probability density function** (pdf) of $X$.

For a r.v. with density, the quantity (if it exists)

$$\mu(X) := \int_{-\infty}^{\infty} u\phi_X(u)\, du$$

is called the **mean** of $X$. Similarly, the quantity (if it exists)

$$V(X) := \int_{-\infty}^{\infty} [u - \mu(X)]^2 \phi_X(u)\, du$$

is called the **variance** of $X$, and $\sigma(X) := \sqrt{V(X)}$ is called the **standard deviation** of $X$. It can be shown that if $V(X) < \infty$, then $\mu(X) < \infty$; see below. If $X$ does not have a density, then the integrals above can be interpreted as Riemann-Stiltjes integrals with respect to the cdf.

Next we introduce the concept of an $L_p$ real-valued random variable $X$. Suppose $X$ is a measurable map from $(\Omega, \mathcal{F}, P)$ to $(\mathbb{R}, \mathcal{B})$, i.e., $X$ is a real-valued random variable. We can attempt to integrate the function $X(\omega), \omega \in \Omega$ using the measure $P$. The concept of integration with respect to an arbitrary measure is somewhat advanced, and we skip lightly over the details. As mentioned above, [10] is an extremely good reference for such topics.

**Definition 2.7.** For $1 \leq p < \infty$, we define $L_p(\Omega, \mathcal{F}, P)$ as the set of functions whose $p$-th powers are absolutely integrable, or

$$L_p(\Omega, \mathcal{F}, P) := \{f : \Omega \to \mathbb{R} \text{ s.t. } \int_\Omega |f(\omega)|^p\, P(d\omega) < \infty\}. \tag{2.1.11}$$

The $L_p$-norm of a function $f \in L_p(\Omega, \mathcal{F}, P)$ is defined as

$$\|f\|_p := \left[\int_\Omega |f(\omega)|^p P(d\omega)\right]^{1/p}. \tag{2.1.12}$$

If $p = \infty$, we define $L_\infty(\Omega, \mathcal{F}, P)$ to be the set of functions that are **essentially bounded**, that is, bounded except on a set of measure zero, and define the corresponding norm as the "essential supremum" of $f(\cdot)$, that is

$$\|f\|_\infty = \inf\{c : P\{|f(\omega)| \geq c\} = 0\}. \tag{2.1.13}$$

In the above formulas, the integration is with respect to the probability measure $P$ on the space $(\Omega, \mathcal{F})$. If $X$ has a density $\phi_X(\cdot)$, then we can take $(\Omega, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$, and $P(d\omega)$ in the above equations can be replaced by $\phi_X(\omega)d\omega$.

Next we introduce two very useful inequalities. The numbers $p, q \in [1, \infty]$ are said to be **conjugate indices** if

$$\frac{1}{p} + \frac{1}{q} = 1.$$

In particular, if $p \in (1, \infty)$, then $q = p/(p-1)$. If $p = 1$, then $q = \infty$ and vice versa. If $p = 2$, then $q = 2$.

**Theorem 2.1.** *(**Hölder's inequality**) If $f \in L_p(\Omega, \mathcal{F}, P)$ and $g \in L_q(\Omega, \mathcal{F}, P)$ where $p$ and $q$ are conjugate indices, then the product $fg \in L_1(\Omega, \mathcal{F}, P)$, and*

$$\int_\Omega |f(\omega)g(\omega)|P(d\omega) \leq \left[\int_\Omega |f(\omega)|^p P(d\omega)\right]^{1/p} \cdot \left[\int_\Omega |g(\omega)|^q P(d\omega)\right]^{1/q}, \tag{2.1.14}$$

*or more compactly,*

$$\|fg\|_1 \leq \|f\|_p \cdot \|g\|_q. \tag{2.1.15}$$

Using Hölder's Inequality and the fact that $P(\Omega) = 1$, it can be shown that

$$L_q(\Omega, \mathcal{F}, P) \subseteq L_p(\Omega, \mathcal{F}, P) \text{ whenever } q > p,$$

or equivalently,

$$q > p, f \in L_q(\Omega, \mathcal{F}, P) \implies f \in L_p(\Omega, \mathcal{F}, P).$$

In particular, if a real random variable $X$ is square-integrable, it is also absolutely integrable. Thus if $X$ has finite variance, it also has a finite mean or expected value.

In particular, choosing $p = q = 2$ in Hölder's inequality leads to **Schwarz' inequality**:

**Theorem 2.2.** *if $f, g \in L_2(\Omega, P)$, then $fg \in L_1(\Omega, P)$, and*

$$\int_\Omega |f(\omega)g(\omega)|P(d\omega) \leq \left[\int_\Omega |f(\omega)|^2 P(d\omega)\right]^{1/2} \cdot \left[\int_\Omega |g(\omega)|^2 P(d\omega)\right]^{1/2}, \tag{2.1.16}$$

*or, more compactly*

$$\|fg\|_1 \leq \|f\|_2 \cdot \|g\|_2. \tag{2.1.17}$$

For future use, we introduce definitions of what it means for a sequence of real-valued random variables to converge. Three commonly used notions of convergence are convergence probability, almost sure convergence, and convergence in the mean. All are defined here. A good reference for this material is [29, Section 2.8]; see also [10, Section 21].

**Definition 2.8.** Suppose $\{X_n\}_{n\geq 0}$ is a sequence of real-valued random variables, and $X^*$ is a real-valued random variable, on a common probability space $(\Omega, \mathcal{F}, P)$. Then the sequence $\{X_n\}_{n\geq 0}$ is said to converge to $X^*$ **in probability** if

$$P(\{\omega \in \Omega : |X_n(\omega) - X^*(\omega)| > \epsilon\}) \to 0 \text{ as } n \to \infty, \ \forall \epsilon > 0. \tag{2.1.18}$$

The sequence $\{X_n\}_{n\geq 0}$ is said to converge to $X^*$ **almost surely** (or almost everywhere) if

$$P(\{\omega \in \Omega : X_n(\omega) \to X^*(\omega) \text{ as } n \to \infty\}) = 1. \tag{2.1.19}$$

Now suppose that $X_n, X^* \in L_1(\Omega, P)$. Then the sequence $\{X_n\}_{n\geq 0}$ is said to converge to $X^*$ **in the mean** if

$$\|X_n - X^*\|_1 \to 0 \text{ as } n \to \infty. \tag{2.1.20}$$

More generally, for any $p \in (1, \infty)$, "convergence in the $p$-th mean" can be defined in the space $L_p(\Omega, P)$, as $\|X_n - X^*\|_p \to 0$ as $n \to \infty$. However, this terminology is rarely used.

The extension of Definition 2.8 to random variables assuming values in a vector space $\mathbb{R}^d$ is obvious and is left to the reader.

The relationship between the various types of convergence is as follows: Again, see [29, Section 2.8], [10, Section 21].

**Theorem 2.3.** *Suppose $\{X_n\}, X^*$ are random variables defined on some probability space $(\Omega, \mathcal{F}, P)$. Suppose $X_n \to X^*$ in probability as $n \to \infty$. Then every subsequence of $\{X_n\}$ contains a subsequence that converges almost surely to $X^*$.*

**Theorem 2.4.** *Suppose $\{X_n\}, X^* \in L_1(\Omega, P)$. Then*

1. *$X_n \to X^*$ a.s. implies that $X_n \to X^*$ in probability.*
2. *$X_n \to X^*$ in the mean implies that $X_n \to X^*$ in probability*
3. *Suppose there is a nonnegative random variable $Z \in L_1(\Omega, P)$ such that $|X_n| \leq Z$ a.e., and suppose that $X_n \to X^*$ a.s.. Then $X_n \to X^*$ in the mean.*

These statements also apply to $\mathbb{R}^d$-valued random variables.

### 2.1.2 Joint and Conditional Probabilities, Independence

Until now we have discussed what might be called "individual" random variables. Now we discuss the concept of joint random variables, and the associated notion of joint probability. The definition below is for two joint variables, but it is obvious that a similar definition can be made for any finite number of joint random variables. In turn this reads to the concept of conditional probability.

**Definition 2.9.** Suppose $(\mathcal{X}, \mathcal{G})$ and $(\mathcal{Y}, \mathcal{H})$ are measurable spaces. Then the **product** of these two spaces is $(\mathcal{X} \times \mathcal{Y}, \mathcal{G} \otimes \mathcal{H})$ where $\mathcal{X} \times \mathcal{Y}$ is the usual Cartesian product of $\mathcal{X}$ and $\mathcal{Y}$, and $\mathcal{G} \otimes \mathcal{H}$ is the smallest $\sigma$-algebra of subsets of $\mathcal{X} \times \mathcal{Y}$ that contains all products of the form $G \times H, G \in \mathcal{G}, H \in \mathcal{H}$.

Note that $\mathcal{G} \otimes \mathcal{H}$ is called the "product" $\sigma$-algebra, which equals $\sigma(\mathcal{G} \times \mathcal{H})$, where $\sigma(\mathcal{S})$ denotes the smallest $\sigma$-algebra containing all sets in the collection $\mathcal{S}$, and $\mathcal{G} \times \mathcal{H}$ consists of all products $G \times H$ for $G \in \mathcal{G}, H \in \mathcal{H}$. The use of the tensor product symbol $\otimes$ to denote the product $s$-algebra is not entirely standard.

Suppose $(\Omega, \mathcal{F}, P)$ is a probability space, and that $(\mathcal{X}, \mathcal{G})$ and $(\mathcal{Y}, \mathcal{H})$ are measurable spaces. Let $(\mathcal{X} \times \mathcal{Y}, \mathcal{G} \otimes \mathcal{H})$ denote their product. Suppose further that $Z : \Omega \to \mathcal{X} \times \mathcal{Y}$ is measurable and thus a random variable taking values in $\mathcal{X} \times \mathcal{Y}$. Let $P_Z$ denote the probability measure of the random variable $Z$. Express $Z$ as $(X, Y)$ where $X, Y$ are the components of $Z$, so that $X : \Omega \to \mathcal{X}, Y : \Omega \to \mathcal{Y}$. Then it can be shown

that $\mathcal{X}$ and $\mathcal{Y}$ are themselves measurable and are thus random variables in their own right. The probability measures associated with these two random variables are as follows:

$$P_X(S) := P_Z(S \times \mathcal{Y}),\ \forall S \in \mathcal{G}, \quad P_Y(T) := P_Z(\mathcal{X} \times T),\ \forall T \in \mathcal{H}. \tag{2.1.21}$$

We refer to $Z = (X, Y)$ as a joint random variable with **joint probability measure** $P_Z$, and to $P_X$ and $P_Y$ as the **marginal probability measures** (or just marginal probabilities) of $P_Z$ for $X$ and $Y$ respectively.

A common application of marginal probabilities arises when both $\mathcal{X}$ and $\mathcal{Y}$ are finite sets. In this case $X, Y, Z$ are random variables assuming values in finite sets $\mathcal{X}, \mathcal{Y}, \mathcal{X} \times \mathcal{Y}$ respectively. Suppose to be specific that $\mathcal{X} = \{x_1, \cdots, x_n\}$ and $\mathcal{Y} = \{y_1, \cdots, y_m\}$. Then it is convenient to represent the joint probability distribution of $Z = (X, Y)$ as an $n \times m$ matrix $\Theta$, where

$$\theta_{ij} = \Pr\{Z = (x_i, y_j)\} = \Pr\{X = x_i \& Y = y_j\}.$$

Let us denote the marginal probabilities as

$$\phi_i = \Pr\{X = x_i\}, \psi_j = \Pr\{Y = y_j\}.$$

Then it is easy to infer that

$$\phi_i = \sum_{j=1}^{m} \theta_{ij}, \quad \psi_j = \sum_{i=1}^{n} \theta_{ij},$$

or in vector notation

$$\boldsymbol{\phi}^\top = \Theta \mathbf{1}_n, \quad \boldsymbol{\psi} = \mathbf{1}_n^\top \Theta,$$

where $\mathbf{1}_k$ denotes a column vector of $k$ ones. Note that we follow the convention that a probability distribution is a row vector.

**Example 2.3.** We illustrate the concept of marginal probability using a simple example where the two sets $\mathcal{X}$ and $\mathcal{Y}$ are finite. Suppose $\mathcal{X} = \{x_1, x_2, x_3\}$ and $\mathcal{Y} = \{y_1, y_2, y_3, y_4\}$. Suppose $Z = (X, Y)$ is a random variable on the product set $\mathcal{X} \times \mathcal{Y}$, with the probability distribution $\Theta$ given by

$$\Theta = \begin{bmatrix} 0.0200 & 0.0400 & 0.0300 & 0.0100 \\ 0.1100 & 0.1700 & 0.1300 & 0.0900 \\ 0.0700 & 0.1200 & 0.0800 & 0.1300 \end{bmatrix},$$

where the rows represent the values of $X$ and the columns represent the values of $Y$. Thus $\Pr\{Z = (x_2, y_3)\} = 0.13$, and so on. To define the marginal probability $P_X$ of the random variable $X$. we simply sum over all possible values of $Y$, or sum each row. Since we view probability distributions as row vectors, we see that

$$\mathbf{p}_X = \boldsymbol{\phi} = (\Theta \mathbf{1}_4)^\top = [\ 0.1000 \quad 0.5000 \quad 0.4000\ ].$$

Similarly the marginal probability $P_Y$ is obtained as

$$\mathbf{p}_Y = \boldsymbol{\psi} = \mathbf{1}_3^\top \Theta = [\ 0.2000 \quad 0.3300 \quad 0.2400 \quad 0.2300\ ].$$

□

**Definition 2.10.** Suppose $(\Omega, \mathcal{F}, P)$ is a probability space. Suppose $(\mathcal{X}, \mathcal{G})$ and $(\mathcal{Y}, \mathcal{H})$ are measurable spaces, and let $Z = (X, Z) : \Omega \to \mathcal{X} \times \mathcal{Y}$ be a joint random variable. Finally, suppose $S \in \mathcal{G}, T \in \mathcal{H}$ are events involving $X$ and $Y$ respectively. Then the **conditional probability** $\Pr\{X \in S | Y \in T\}$ is defined as

$$\Pr\{X \in S | Y \in T\} = \frac{\Pr\{Z = (X, Y) \in S \times T\}}{\Pr\{Y \in T\}} = \frac{P_Z(S \times T)}{P_Y(T)}. \tag{2.1.22}$$

In the definition of the conditional probability (2.1.22), it is assumed that $P_Y(T) > 0$. If $P_Y(T) = 0$, by convention $\Pr\{X \in S | Y \in T\}$ is taken as $P_X(S)$.

Let us fix a set $T \in \mathcal{H}$, and define the function $P_{\{X|Y\in T\}} : \mathcal{G} \to [0,1]$ by

$$P_{\{X|Y\in T\}}(S) := \Pr\{X \in S | Y \in T\}. \tag{2.1.23}$$

Then it is easy to verify that $P_{\{X|Y\in T\}}$ is a probability measure on $(\mathcal{X}, \mathcal{G})$. We can think of it as the probability measure conditioned on the event that $Y \in T$. If $P_Y(T) = 0$, then $P_{\{X|Y\in T\}} = P_X$.

**Example 2.4.** Let us return to Example 2.3. First, define the event $T$ as $Y = y_2$. Then the corresponding conditional probability distribution on $X$ is given by

$$\mathbf{p}_{\{X|Y\in T\}} = (1/0.33)[\ 0.04 \quad 0.17 \quad 0.12\ ].$$

Now let us define $T = \{y_1, y_3\}$. Then $P_Y(T) = 0.2 + 0.24 = 0.44$, and

$$\begin{aligned} P_{\{X|Y\in T\}} &= \frac{1}{0.20 + 0.24}[\ 0.02 + 0.03 \quad 0.11 + 0.13 \quad 0.07 + 0.08\ ] \\ &= \frac{1}{0.44}[\ 0.05 \quad 0.24 \quad 0.15\ ]. \end{aligned}$$

Observe that $P_{\{X|Y\in T\}}$ is a convex combination of $P_{\{X|Y=y_1\}}$ and $P_{\{X|Y=y_3\}}$, namely

$$P_{\{X|Y\in T\}} = \frac{0.20}{0.44} \times \frac{1}{0.20}[\ 0.02 \quad 0.11 \quad 0.07\ ] + \frac{0.24}{0.44} \times \frac{1}{0.24}[\ 0.03 \quad 0.13 \quad 0.08\ ].$$

This property holds in general. □

After defining the concept of a conditional probability, it is straight-forward to define the conditional expected value. If $f : \mathcal{X} \to \mathbb{R}$ and $T \subseteq \mathcal{Y}$ is some event involving $Y$, then $E[f(\mathcal{X})|Y \in T]$ is just $E[f(X), P_{\{X|Y\in T\}}]$.

**Example 2.5.** Let us continue Example 2.4. Suppose $f : \mathcal{X} \to \mathbb{R}$ is defined by

$$[\ f(x_1) \quad f(x_2) \quad f(x_3)\ ] = [\ 2 \quad -7 \quad 4\ ].$$

Let $T = \{y_1, y_3\}$. Then it is already known that

$$P_{\{X|Y\in T\}} = \frac{1}{0.44}[\ 0.05 \quad 0.24 \quad 0.15\ ].$$

Therefore

$$E[f(\mathcal{X})|Y \in T] = \frac{1}{0.44}(0.1 - 1.68 + 0.6) = -\frac{0.98}{0.44}.$$

Similar computations can be carried out for other choices of the event $T$. In particular, since $T = \{y_1, y_3\}$, it follows that $E[f(X)|Y \in T]$ is a convex combination of $E[f(X)|Y = y_1$ and $E[f(X)|Y = y_3]$. □

All of the above definitions can be extended to more than two random variables.

Next we briefly discuss the concept of independence. Kolmogorov, who laid down the foundations of probability theory, remarks on [79, p. 8] (in English translation) that

> Historically, the independence of experiments and random variables represents the very mathematical concept that has given the theory of probability its peculiar stamp.

This statement, together with the text that precedes it, can be paraphrased as: Without the concept of independence, there is essentially no difference between measure theory and probability theory. Thus the concept of independence is fundamental (and unique) to probability theory.

**Definition 2.11.** Suppose $(\Omega, \mathcal{F}, P)$ is a probability space. Then two events $S, T \in \mathcal{F}$ are said to be **independent** if

$$P(S \cap T) = P(S)P(T).$$

Suppose now that $\mathcal{F}_1, \mathcal{F}_2$ are sub-$\sigma$-algebras of $\mathcal{F}$. Then $\mathcal{F}_1$ and $\mathcal{F}_2$ are said to be **independent** if

$$P(S \cap T) = P(S)P(T),\ \forall S \in \mathcal{F}_1, T \in \mathcal{F}_2. \tag{2.1.24}$$

Two random variables $X_1, X_2$ defined on $(\Omega, \mathcal{F}, P)$ are said to be **independent** if the corresponding $\sigma$-algebras $\sigma(X_1), \sigma(X_2)$ are independent. Further, suppose $Z = (X, Y)$ is a joint random variable defined on a product space $(\mathcal{X} \times \mathcal{Y}, \mathcal{G} \otimes \mathcal{H})$. Then $X$ and $Y$ are said to be **independent random variables** if

$$P_Z(S \times T) = P_X(S) \times P_Y(T),\ \forall S \in \mathcal{G}, T \in \mathcal{H}. \tag{2.1.25}$$

The extension of the above definition to any finite number of events, or $\sigma$-algebras, or random variables, is quite obvious. For more details, see [29, Section 3.1] or [173, Chapter 4].

### 2.1.3 Conditional Expectations

The concept of conditional *probability* discussed in the preceding subsection can be applied to even "abstract" random variables, that is, random variables assuming values in some abstract set. In contrast, concepts such as expected value (both unconditional and conditional) are meant to be used with *real-valued* random variables. The ideas extend readily to vector-valued random variables by applying them componentwise. The objective of this subsection is to introduce another concept known as the "conditional expectation" of a random variable with respect to a $\sigma$-algebra. While the conditional *expected value* is a real number, the conditional *expectation* is a random variable. There is a close relationship between these two concepts, as will be brought both through the theory as well as an example. The discussion below requires an understanding of *integration* with respect to a probability measure. We do not go into too many details regarding the abstract concept of integration with respect to a measure, because that would be rather tangential to the main discussion. Instead we refer interested reader to [10] for details.

Throughout this subsection, we deal with *real-valued* random variables. Thus, when we say that $X$ is a real random variable on $(\Omega, \mathcal{F}, P)$, we mean that $X$ is a measurable map from $(\Omega, \mathcal{F}, P)$ to $(\mathbb{R}, \mathcal{B})$.

In the discussion below, we often deal with two random variables $X$ and $X'$ that differ only on a set of measure zero, that is,

$$P\{\omega : X(\omega) \neq X'(\omega)\} = 0.$$

In such a case, we write $X = X'$ a.e., or $X = X'$ a.s.

Next, we define the concept of the conditional expectation of a random variable with respect to a $\sigma$-algebra. The precise (and rather abstract) definition is given first, followed by some properties of the conditional expectation. Then some concrete examples are given. Suppose $X, Y$ are random variables, and as before, let $\sigma(Y)$ denote the $\sigma$-algebra generated by $Y$. Then one can think of the conditional expectation $E(X|\sigma(Y))$ as a natural generalization of the conditional expected value of $X$, as $Y$ ranges over all its possible values.

**Definition 2.12.** (See [29, Definition 4.16], [44, Section 4.1] or [173, Section 9.2].) Suppose $(\Omega, \mathcal{F}, P)$ is a probability space, and that $X$ is a real random variable belonging to $L_1(\Omega, \mathcal{F}, P)$. Suppose that $\mathcal{G} \subseteq \mathcal{F}$ is another $\sigma$-algebra on $\Omega$. Then the **conditional expectation** of $X$ with respect to $\mathcal{G}$, denoted by $E(X|\mathcal{G})$, is any random variable $Y$ such that (i) $Y$ is measurable with respect to $(\Omega, \mathcal{G})$, and (ii)

$$\int_D X(\omega)P(d\omega) = \int_D Y(\omega)P(d\omega),\ \forall D \in \mathcal{G} \tag{2.1.26}$$

Note that $E(X|\mathcal{G})$ is a $(\Omega, \mathcal{G})$-measurable approximation to $X$ such that, when restricted to sets in $\mathcal{G}$, $E(X|\mathcal{G})$ is functionally equivalent to $X$, as stated in (2.1.26). Note that (2.1.26) can also be expressed as

$$\int_\Omega X(\omega) I_D(\omega) P(d\omega) = \int_\Omega Y(\omega) I_D(\omega) P(d\omega), \ \forall D \in \mathcal{G}$$

where $I_D(\cdot)$ is the indicator function of the set $D$.

To make the discussion below easier to follow, we employ the notation $Y \in \mathcal{M}(\mathcal{G})$ to indicate that $Y$ maps $\Omega$ into $\mathbb{R}$, and is measurable with respect to $(\Omega, \mathcal{G})$ and $(\mathbb{R}, \mathcal{B})$. Also, since in these notes we deal with both the conditional expectation (which is a random variable) and the conditional expected value (which is a real number), we use parenthesis to denote the conditional expectation and square brackets to denote the expected value, conditional or otherwise.

In the above definition, it is not clear that such a conditional expectation exists. Any $Y$ that satisfies (2.1.26) is called a "version" in [173, 44]. The next theorem summarizes, without proof, some key properties of the conditional expectation. These details can be found in [173, Chapter 9] and/or [44, Section 4.1].

**Theorem 2.5.** *Suppose $X \in L_1(\Omega, \mathcal{F}, P)$ and that $\mathcal{G} \subseteq \mathcal{F}$ is another $\sigma$-algebra on $\Omega$. Then*

1. *(Existence) There is at least one $Y \in \mathcal{M}(\mathcal{G})$ such that (2.1.26) holds.*
2. *(Uniqueness) If $Y, Y' \in \mathcal{M}(\mathcal{G})$ both satisfy (2.1.26), then $Y(\omega) = Y'(\omega)$ a.s..*
3. *(Expected Value Preservation) Every conditional expectation $Y = E(X|\mathcal{G})$ belongs to $L_1(\Omega, \mathcal{F}, P)$. Moreover $X$ and $Y = E(X|\mathcal{G})$ have the same expected value. Thus*

$$E[Y, P] = E[X, P], \ \text{or} \ \int_\Omega Y(\omega) P(d\omega) = \int_\Omega X(\omega) P(d\omega). \tag{2.1.27}$$

4. *(Self-Replication) If $X \in \mathcal{M}(\mathcal{G})$, then $E(X|\mathcal{G}) = X$ a.s..*
5. *(Iterated Conditioning) If $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$ are $\sigma$-algebras, then*

$$E(E(X|\mathcal{G})|\mathcal{H}) = E(X|\mathcal{H}). \tag{2.1.28}$$

6. *(Idempotency) If $p, q$ are conjugate indices, and $Z \in L_q(\Omega, \mathcal{G}, P)$, $X \in L_p(\Omega, \mathcal{F}, P)$, then*

$$E((ZX)|\mathcal{G}) = ZE(X|\mathcal{G}) \ \text{a.s.}. \tag{2.1.29}$$

7. *(Linearity) If $X_1, X_2 \in L_1(\Omega, \mathcal{F}, P)$ and $a_1, a_2 \in \mathbb{R}$, then*

$$E((a_1 X_1 + a_2 X_2)|\mathcal{G}) = a_1 E(X_1|\mathcal{G}) + a_2 E(X_2|\mathcal{G}) \ \text{a.s.}. \tag{2.1.30}$$

8. *(Nonnegativity) If $X(\omega) \geq 0$ a.s., then $E(X|\mathcal{G})(\omega) \geq 0$ a.s..*
9. *(Projection Property) If $X \in L_2(\Omega, \mathcal{F}, P)$ (and not just $L_1(\Omega, \mathcal{F}, P)$), then*

$$E(X|\mathcal{G}) = \underset{Y \in L_2(\Omega, \mathcal{G}, P)}{\arg\min} \|Y - X\|_2^2 \ \text{a.s.}. \tag{2.1.31}$$

Now we interpret some of the statements in the theorem. The obvious ones are not discussed. Item 3 states that the expected value of the conditional expectation is the same as the expected value of the original random variable. Item 5 states that if we were to first take the conditional expectation of $X$ with respect to $\mathcal{G}$, and then take the conditional expectation of the resulting random variable with respect to a smaller $\sigma$-algebra $\mathcal{H}$, then the answer would be the same as if we had directly taken the conditional expectation of $X$ with respect to $\mathcal{H}$. Note that this property is called the "tower property" on [173, p. 88]. Item 6 states that

$X \in L_p(\Omega, \mathcal{F}, P)$ is multiplied by a $Z \in L_q(\Omega, \mathcal{G}, P)$ where $p$ and $q$ are conjugate indices (so that the product $ZX$ belongs to $L_1(\Omega, \mathcal{F}, P)$), then the term $Z$ can be moved outside the conditional expectation operation. Items 6 and 7, taken together, imply the following: If $Z_1, Z_2 \in L_q(\Omega, \mathcal{G}, P)$ and $X_1, X_2 \in L_p(\Omega, \Omega, P)$, where $p$ and $q$ are conjugate indices, then

$$E((Z_1X_1 + Z_2X_2)|\mathcal{G}) = Z_1E(X_1|\mathcal{G}) + Z_2E(X_2|\mathcal{G}).$$

A ready consequence of Items 7 and 8 is that, if $X_1 \geq X_2$ almost surely, then $E(X_1|G) \geq E(X_2|G)$ almost surely. Finally, Item 9 states that if $X$ belongs to $L_2(\Omega, \mathcal{F}, P)$, which is a subspace of $L_1(\Omega, \mathcal{F}, P)$ and an inner product space, then its conditional expection $E(X|\mathcal{G})$ belongs to the subspace $L_2(\Omega, \mathcal{G}, P)$ of $L_2(\Omega, \mathcal{F}, P)$, and can be computed as the closest element in $L_2(\Omega, \mathcal{G}, P)$ to $X$ using the projection theorem.

In the above discussion, the $\sigma$-algebra $\mathcal{G}$ can be any sub-algebra of $\mathcal{F}$. In some applications, the following situation arises: Suppose $Z \in \mathcal{M}(\mathcal{F})$, and $\mathcal{G} = \sigma(Z)$, the $\sigma$-algebra generated by $Z$. In such a case, we can also use the alternate notation $E(X|Z)$ to denote $E(X|\mathcal{G}) = E(X|\sigma(Z))$. This notation proves to be convenient in analyzing problems in RL.

**Example 2.6.** In this example, we illustrate the concept of a conditional expectation in a very simple case, namely, that of a random variable assuming only finitely many values. Suppose $\mathcal{X} = \{x_1, \cdots, x_n\}$ and $\mathcal{Y} = \{y_1, \cdots, y_m\}$ are finite sets, and that $Z = (X, Y)$ is a joint random variable assuming values in $\mathcal{X} \times \mathcal{Y}$. Let $\Theta \in [0, 1]^{n \times m}$ denote the joint probability distribution of $Z$ written out as a matrix, and let $\boldsymbol{\phi}, \boldsymbol{\psi}$ denote the marginal probability distributions of $X$ and $Y$ respectively, written out as row vectors. Finally, suppose $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a given function. Then $f(Z)$ is a real-valued random variable assuming values in some finite set.

Because both $X$ and $Y$ are finite-valued, we can use the canonical representation, and choose $\Omega = \mathcal{X} \times \mathcal{Y}$, $\mathcal{F} = 2^\Omega$, and $P = \Theta$. Now suppose we define $\mathcal{G}$ to be the $\sigma$-algebra generated by $Y$ alone. Thus $\mathcal{G} = \{\emptyset, \mathcal{X}\} \otimes 2^\mathcal{Y}$. Again, because $f(X, Y)$ assumes only finitely many values over a finite set, it is a bounded random variable. Therefore $E(f|\mathcal{G})$ is the best approximation to $f(X, Y)$ using a function of $Y$ alone. From Item 9 of Theorem 2.5, this conditional expectation can be determined using projections.

It can be assumed without loss of generality that every component of $\boldsymbol{\psi}$ is positive. If $\psi_j = 0$ for some $j$, then $\theta_{ij} = 0$ for all $i$; therefore the value $y_j$ can be omitted from the set $\mathcal{Y}$. Therefore the ratio

$$\frac{\theta_{ij}}{\psi_j} = \Pr\{X = x_i | Y = y_j\}$$

is well-defined, though it could be zero.

In order to determine $E(f|\mathcal{G})$, we should find a function $g : \mathcal{Y} \to \mathbb{R}$ such that the error $E[(f - g)^2, \Theta]$ is minimized. Let $g_1, \cdots, g_m$ denote the values of $g(\cdot)$, and define the objective function

$$J = \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{n} (g_j - f_{ij})^2 \theta_{ij}.$$

Note that $J$ is the sum of $m$ terms, where the $j$-th term depends only on $g_j$. Then the objective is to choose the constants $g_1, \cdots, g_m$ so as to minimize $J$. This happens when

$$0 = \frac{\partial J}{\partial g_j} = \sum_{i=1}^{n} (g_j - f_{ij}) \theta_{ij} i, \ \forall j \in [m].$$

This expression can be rewritten as

$$0 = g_j \sum_{i=1}^{n} \theta_{ij} - \sum_{i=1}^{n} f_{ij} \theta_{ij} = g_j \psi_j - \sum_{i=1}^{n} f_{ij} \theta_{ij},$$

or

$$g_j = \sum_{i=1}^{n} f_{ij} \frac{\theta_{ij}}{\psi_j} = E[f(X,Y)|Y = y_j].$$

This formula explains the terminology "conditional expectation." For each outcome $Y = y_j$, the conditional expectation $E(f|\mathcal{G}) = g_j$ equals the expected value of $f$ conditioned on the event that $Y = v_j$. However, since $Y$ is itself random, so is the conditional expected value $E[X|Y = y_j]$. This is precisely the conditional expectation $E(f|\mathcal{G})$. Since $Y = y_j$ with probability $\psi_j$, the conditional expectation $E(f|\mathcal{G})$ equals $E[f|Y = y_j]$ with probability $\psi_j$. The same expression also shows that

$$E[E(f|\mathcal{G}), \boldsymbol{\psi}] = \sum_{j=1}^{m} g_j \psi_j = \sum_{j=1}^{m} \sum_{i=1}^{n} f_{ij} \theta_{ij} = E[f, \Theta].$$

□

Though simple in appearance, the derivation in Example 2.6 will be used repeatedly in applications to Reinforcement Learning, when the underlying MDP evolves over finite state and action spaces.

## 2.2 Markov processes

In this section, we introduce the concept of a Markov process that assumes values in a *finite* set $\mathcal{X} = \{x_1, \cdots, x_n\}$, where the elements $x_i$ could represent abstract symbols, and the "time index" of the process is the set of natural numbers. It is possible to define Markov processes where both the state space and the time index set are a continuum; but such generality is not needed in this book.

### 2.2.1 Markov Processes: Basic Properties

Suppose $\mathcal{X}$ is a set of finite cardinality, say $\mathcal{X} = \{x_1, \cdots, x_n\}$, and suppose that $\{X_t\}_{t \geq 0}$ is a stochastic process assuming values in $\mathcal{X}$, that is, $\{X_t\}_{t \geq 0}$ is a sequence of random variables assuming values in $\mathcal{X}$. Let the symbol $X_0^t$ denote the (finite) collection of random variables $(X_0, \cdots, X_t)$.

**Definition 2.13.** The process $\{X_t\}_{t \geq 0}$ is said to **possess the Markov property**, or to be a **Markov process**, if

$$E(X_{t+1}|X_0^t) = E(X_{t+1}|X_t), \ \forall t \geq 0. \tag{2.2.1}$$

The above abstract definition states simply that the conditional expectation of the "state" $X_{t+1}$ conditioned on the entire past $X_0^t$ is the same as the conditional expectation given only the most recent "state" $X_t$. This abstract definition can be "operationalized" as follows: For every $y \in \mathcal{X}$ and every $\mathbf{u}_o^t \in \mathcal{X}^{t+1}$, it is true that

$$\Pr\{X_{t+1} = y|X_0^t = \mathbf{u}_0^t\} = \Pr\{X_{t+1} = y|X_t = u_t\}. \tag{2.2.2}$$

In other words, the conditional probability of the state $X_{t+1}$ depends only on the most recent value of $X_t$; adding information about the past values of $X_\tau$ for $\tau < t$ does not change the conditional probability. One can also say that $X_{t+1}$ is independent of $X_0^{t-1}$ given $X_t$. This property is sometimes paraphrased as "the future is conditionally independent of the past, given the present."

A Markov process over a finite set $\mathcal{X}$ is completely characterized by the probability distribution $\boldsymbol{\phi}_0$ of the initial state $X_0$, and its sequence of **state transition matrices** $A^{(t)} \in [0,1]^{n \times n}$, where

$$a_{ij}^t := \Pr\{X_{t+1} = x_j|X_t = x_i\}, \ \forall x_i, x_j \in \mathcal{X}.$$

Thus in $a_{ij}^t$, $i$ denotes the current state and $j$ the future state. The reader is cautioned that some authors interchange the roles of $i$ and $j$ in the above definition. If the transition probability does not depend on $t$, then the Markov process is said to be **stationary**; otherwise it is said to be **nonstationary**. We do

not deal with nonstationary Markov processes in these notes. A stationary Markov process is completely characterized by its state transition matrix $A \in [0,1]^{n\times n}$, and the probability distribution $\boldsymbol{\phi}_0$ of its initial state.

Note that $a_{ij} \in [0,1]$ for all $i, j$. Also, at any time $t+1$, it must be the case that $X_{t+1} \in \mathcal{X}$, no matter what $X_t$ is. Therefore, the sum of each row of $A$ equals one, i.e.,

$$\sum_{j=1}^{n} a_{ij} = 1, i = 1, \dots n. \tag{2.2.3}$$

The above equation can be expressed compactly as

$$A\mathbf{1}_n = \mathbf{1}_n, \tag{2.2.4}$$

where $\mathbf{1}_n$ denotes the column vector consisting of $n$ ones. For future purposes, let us refer to a matrix $A \in [0,1]^{n\times n}$ that satisfies (2.2.3) as a **row-stochastic matrix**, and denote by $\mathbb{S}_{n\times n}$ the set of all row-stochastic matrices of dimension $n \times n$.

The matrix $A$ is often called the "one-step" transition matrix, because row $i$ of $A$ gives the probablity distribution of $X_{t+1}$ if $X_t = x_i$. So we can ask: What is the $k$-step transition matrix? In other words, what is the probability distribution of $X_{t+k}$ if $X_t = x_i$? It is not difficult to show that this conditional probability is just the $i$-th row of $A^k$. Thus the $k$-step transition matrix is just $A^k$. Therefore, if $X_0$ has the probability distribution $\boldsymbol{\phi}$ (denoted by $X_0 \sim \boldsymbol{\phi}$), then $X_t \sim \boldsymbol{\phi}A^t$. Note that the probability distributions are viewed as *row* vectors.

**Example 2.7.** A familiar example of a Markov process is the "snakes and ladders" game. Take for example the board shown in Figure 2.1. Suppose the player throws a four-sided die with each of the outcomes $(1, 2, 3, 4)$ being equally probable. Then the resulting sequence of positions $\{X_t\}_{t\geq 0}$ is a stochastic process. Suppose for example that the player is on square 60. Then with probability of $1/4$, the position at time $t+1$ will be 61, 19 (snake on 62), 81 (ladder on 63) and 60 (snake on 64). Note that what happens next after a player has reached square 60 (or any other square) *does not depend on how the player reached that square.* That is why the sequence of positions is a Markov process. Moreover, the states corresponding to any square that has either a snake or a ladder can be deleted from the state space. Thus the true state space is not $\{1, \cdots, 100\}$ but some subset thereof. In this case, there are eight snakes and eight ladders, so the state space consists of 84 elements, namely $\{1, 2, 3, 5, 6, \cdots\}$. The element 4 is missing because it is the starting point of a ladder. Thus, in row corresponding to square 60 of the $84 \times 84$ state transition matrix, there are elements of $1/4$ in columns $19, 60, 61, 81$ and zeros in the remaining 80 columns. In the same manner, the entire $84 \times 84$ state transition matrix can be determined.

Let us suppose that the snakes and ladders game always starts with the player being in square 1. Thus $X_0$ is not random, but is deterministic, and the "probability distribution" of $X_0$, viewed as a row vector, has a 1 in column 1 and zeros elsewhere. If we multiply this row vector by $A^k$ for any integer $k$, we get the probability distribution of the player's position after $k$ moves.

□

An application of the Gerschgorin circle theorem [63, Theorem 6.1.1] shows that, whenever $A$ is row-stochastic, the spectral radius $\rho(A) \leq 1$. Moreover, the relationship (2.2.3) shows that $\lambda = 1$ is an eigenvalue of $A$ with column eigenvector $\mathbf{1}_n$, so that in fact $\rho(A) = 1$. Thus one can ask: What does the row eigenvector corresponding to $\lambda = 1$ look like? If there is a *nonnegative* row eigenvector $\boldsymbol{\mu} \in \mathbb{R}^n_+$, then it can be scaled so that $\boldsymbol{\mu}\mathbf{1}_n = 1$. Such a $\boldsymbol{\mu}$ is called a **stationary distribution** of the Markov process, because if $X_t$ has the probability distribution $\boldsymbol{\mu}$, then so does $X_{t+1}$. More generally, if $X_0$ has the probability distribution $\boldsymbol{\mu}$, then so does $X_t$ for all $t \geq 0$.

**Theorem 2.6.** *(See [11, Theorem 3.2, p. 8].) Every row-stochastic matrix $A$ has a nonnegative row eigenvector corresponding to the eigenvalue $\lambda = 1$.*

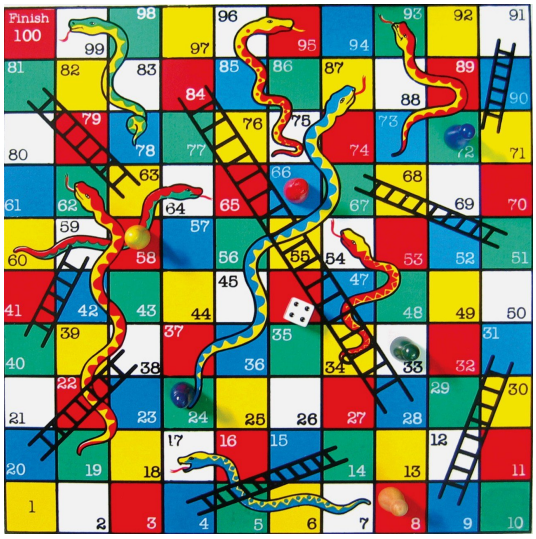



Figure 2.1: Snakes and Ladders Game

Note that Theorem 2.6 is a very weak statement. It states only that *there exists* a stationary distribution; nothing is said about whether this is unique or not. However, by making some assumptions about $A$, it is possible to derive stronger conclusions. The ideas discussed in the remainder of this subsection are discussed in far greater detail in [11, Chapter 6], [131] and [167, Chapter 3].

**Definition 2.14.** A row-stochastic matrix $A$ is said to be **irreducible** if it is not possible to partition the permute the rows and columns symmetrically (via a permutation matrix $\Pi$) such that

$$\Pi^{-1} A \Pi = \begin{bmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{bmatrix}.$$

Thus a row-stochastic matrix is irreducible if it is not possible to turn it into a block-triangular matrix through symmetric row and column permutations. The notion of irreducibility plays a crucial role in the theory of Markov processes. So it is worthwhile to give an alternate characterization of irreducibility.

**Lemma 2.1.** *A row-stochastic matrix $A$ is irreducible if and only if, for any pair of states $y_s, y_f \in \mathcal{X}$, there exists a sequence of states $y_1, \cdots y_l \in \mathcal{X}$ such that, with $y_0 = y_s$ and $y_{l+1} = y_f$, we have that*

$$a_{y_k y_{k+1}} > 0, k = 0, \ldots, l.$$

Thus the matrix $A$ is irreducible if and only if, for every pair of states $y_s$ and $y_f$, there is a path from $y_s$ to $y_f$ such that every step in the path has a positive probability. In such a case we can say that $y_f$ is reachable from $y_s$.

**Example 2.8.** The Markov process corresponding to the Snakes and Ladders game of Example 2.7 is *not irreducible*. To illustrate just a few combinations, there is no path from 3 to 2, nor from 6 to 5. (However, there *is* a path from 8 to 7 by travelling from 8 to 17 which has a "snake" leading back to 7.)

There are several equivalent characterizations of irreducibility, and for nonnegative matrices in general, not necessarily satisfying (2.2.3). In fact, the discussion in the references [11, Chater 6], [131] and [167, Chapter 3] deal with nonnegative matrices in general, and are not restricted to stochastic matrices alone. One such characterization of irreducibility is given next.

**Theorem 2.7.** *(See [167, Corollary 3.8].) A row-stochastic matrix $A$ is irreducible if and only if*

$$M_{n-1} := \sum_{l=0}^{n-1} A^l > 0, \tag{2.2.5}$$

*where $A^0 = I$ and the inequality is componentwise.*

So we can start with $M_0 = I$ and define recursively $M_{l+1} = I + AM_l$. Then $M_l$ is the partial sum up to the $A^l$ term in (2.2.5). If $M_l > 0$ for any $l$, then (2.2.5) is satisfied, because higher powers of $A$ are nonnegative; hence $A$ is irreducible. If we get up to $M_{n-1}$ and this matrix is not strictly positive, then $A$ is not irreducible.

**Theorem 2.8.** *(See [167, Theorem 3.25].) Suppose $A$ is an irreducible row-stochastic matrix. Then*

1. *$\lambda = 1$ is a simple eigenvalue of $A$.*
2. *The corresponding row eigenvector of $A$ has all positive elements.*
3. *Thus $A$ has a unique stationary distribution, whose elements are all positive.*
4. *There is an integer $p$, called the **period** of $A$, such that the spectrum of $A$ is invariant under rotation by $\exp(\mathbf{i}2\pi/p)$.*
5. *In particular, $\exp(\mathbf{i}2l\pi/p)$, $l = 0, \cdots, p-1$ are all eigenvalues of $A$.*

Now we introduce a concept that is stronger than irreducibility.

**Definition 2.15.** A row-stochastic matrix $A$ is said to be **primitive** if there exists an integer $l$ such that $A^l > 0$.

To connect the two notions of irreducibility and primitivity, we introduce another important concept called the period of an irreducible Markov process. An aperiodic Markov process is one whose period equals one.

Suppose $A$ corresponds to an irreducible Markov process. Then, by Lemma 2.1, there is a path between every pair of states. Now let $x_i$ be any state. Then there is always at least one "cycle," that is, a path starting at $x_i$ back to itself. To see this, pick any other state $x_j \neq x_i$. Then by irreducibility, there exists a path from $x_i$ to $x_j$, and also a path from $x_j$ to $x_i$. Taken together, they form a cycle from $x_i$ back to itself. There can of course be multiple cycles from a state back to itself.

**Definition 2.16.** Fix a state $x_i$. The **period of the state** $x_i$ is defined as the greatest common divisor (g.c.d.) of the lengths of all cycles from $x_i$ back to itself. As shown in [167, Theorem 3.12], every state in an irreducible Markov process has the same period, which is defined to be the **period** of the process. A Markov process is said to be **aperiodic** if its period equals one.

**Theorem 2.9.** *If a row-stochastic matrix $A$ is irreducible and aperiodic, then $\lambda = 1$ is the only eigenvalue of $A$ with magnitude one.*

**Theorem 2.10.** *(See [167, Theorem 3.15].) A row-stochastic matrix $A$ is primitive if and only if it is irreducible and aperiodic.*

**Example 2.9.** Suppose

$$A_1 = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \end{bmatrix}, A_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Then $A_1$ is primitive, while $A_2$ is irreducible but not primitive; it has a period $p = 3$. □

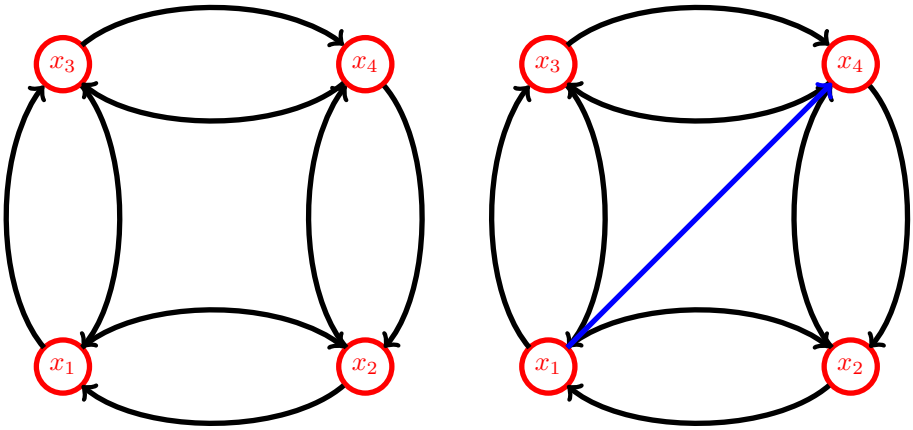



Figure 2.2: Irreducible vs. aperiodic Markov processs

**Example 2.10.** In this example, we illustrate the concept of aperiodicity using just the graphical features of a Markov process, without having to specify the transition probabilities.

In Figure 2.2, we deliberately *do not specify* the values of the transition probabilities; rather, the presence of an arrow indicates that the corresponding transition probability is positive, while the absence of an arrow indicates that the corresponding transition probability is zero. The objective is to highlight that the properties of a Markov process depend only on the pattern of the values (zero or nonzero), and not on the actual values.

From the diagram, it is clear that the Markov process consists of two "cyclical" processes, one with the transitions $x_1 \to x_3 \to x_4 \to x_2 \to x_1$, and other one being $x_1 \to x_2 \to x_4 \to x_3 \to x_1$. Therefore there is a path from every vertex to every other vertex, so that the Markov process is irreducible. Moreover, there are cycles of length 4 as well as of length 2, but no cycles whose length is an odd number. So the period of the Markov process is two.

Now suppose we add just one transition, say from $x_1$ to $x_4$. Then there are two cycles from $x_1$ to itself: one of length four, namely $x_1 \to x_3 \to x_4 \to x_2 \to x_1$, and another of length three, namely $x_1 \to x_4 \to x_2 \to x_1$, Since 3 and 4 have no common divisors (other than 1), the process is now aperiodic, and thus primitive. □

In some situations, the following result is useful.

**Theorem 2.11.** *(See [167, Lemma 4.12].) Suppose $A$ is an irreducible row-stochastic matrix, and let $\boldsymbol{\mu}$ denote the corresponding stationary distribution. Then*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} A^t = \mathbf{1}_n \boldsymbol{\mu}. \tag{2.2.6}$$

Therefore, the average of $I, A, \cdots, A^{T-1}$ approaches the rank one matrix $\mathbf{1}_n \boldsymbol{\mu}$. Recall that, if $\boldsymbol{\phi}$ is *any* probability distribution on $\mathcal{X}$, and the Markov process is started off with the initial distribution $\boldsymbol{\phi}$, then the distribution of the state $X_t$ is $\boldsymbol{\phi} A^t$. Note that, because $\boldsymbol{\phi}$ is a probability distribution, we have that $\boldsymbol{\phi}\mathbf{1}_n = 1$. Therefore (2.2.6) implies that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \boldsymbol{\phi} A^t = \boldsymbol{\phi}\mathbf{1}_n \boldsymbol{\mu} = \boldsymbol{\mu}, \ \forall \boldsymbol{\phi}. \tag{2.2.7}$$

The above relationship holds for *every* $\boldsymbol{\phi}$ and forms the basis for the so-called **Markov chain Monte Carlo (MCMC)** algorithm. Suppose $\{X_t\}_{t\geq 0}$ is a Markov process evolving over the state space $\mathcal{X}$, with an irredcible state transition matrix $A$ and stationary distribution $\boldsymbol{\mu}$. Suppose further that $f : \mathcal{X} \to \mathbb{R}$ is a real-valued function defined on the state space $\mathcal{X}$. We wish to compute the expected value of the random variable $f(X_t)$ with respect to the stationary distribution $\boldsymbol{\mu}$, namely

$$E[f(X), \boldsymbol{\mu}] = \sum_{x_i \in X} f(x_i)\mu_i. \tag{2.2.8}$$

While we may know $A$, often we may not know $\boldsymbol{\mu}$ or may not wish to spend the effort to compute it due to the high dimension of $A$. In such a case, we start off the Markov process with an arbitrary initial probability distribution $\boldsymbol{\phi}$, let it run for some time $t_0$, and then compute the quantity

$$\hat{f}_T = \frac{1}{T} \sum_{t=t_0+1}^{t_0+T} f(X_t). \tag{2.2.9}$$

Because this quantity is based on the observed state $X_t$ which is random, $\hat{f}_T$ is also random. However, the expected value of $\hat{f}_T$ is precisely $E[f(X), \boldsymbol{\mu}]$. Moreover, its sample-path average $\hat{f}_T$ converges to $E[f(X), \boldsymbol{\mu}]$ as $T \to \infty$, and is a good approximation for the expected value for finite $T$.

The next result is analogous to Theorem 2.11 for primitive matrices.

**Theorem 2.12.** *(See [167, Corollary 4.13].) Suppose $A$ is a primitive row-stochastic matrix, and let $\boldsymbol{\mu}$ denote the corresponding stationary distribution. Then*

$$A^l \to \mathbf{1}_n^\top \boldsymbol{\mu} \text{ as } l \to \infty. \tag{2.2.10}$$

Now we prove a couple of useful lemmas about irreducible and primitive matrices respectively. These are useful when we study so-called Markov Decision Processes.

**Theorem 2.13.** *Suppose $A$ is a nontrivial convex combination of row stochastic matrices $A_1, \cdots, A_k$, and that at least one $A_i$ is irreducible. Then $A$ is irreducible.*

*Proof.* Without loss of generality, renumber the matrices such that $A_1$ is irreducible. Write

$$A = \sum_{i=1}^{k} \gamma_i A_i,$$

and note that $\gamma_1 > 0$ while $A_1$ is irreducible. Then

$$A^l \geq \gamma_1^l A_1^l \ \forall l,$$

where the inequality holds componentwise, because all other "cross-product" terms in the expansion of $A^l$ are nonnegative matrices. Because $A_1$ is irreducible, it follows from Theorem 2.7 that

$$\sum_{l=0}^{n-1} A_1^l > 0,$$

where again the inequality is componentwise. Combining this with the above inequality shows that

$$\sum_{l=0}^{n-1} A^l \geq \sum_{l=0}^{n-1} \gamma_1^l A_1^l \geq \gamma_1^{n-1} \sum_{l=0}^{n-1} A_1^l > 0.$$

Therefore $A$ is irreducible. □

**Corollary 2.1.** *The set of irreducible matrices is convex.*

**Theorem 2.14.** *Suppose $A$ is a nontrivial convex combination of row stochastic matrices $A_1, \cdots, A_k$, and that at least one $A_i$ is primitive. Then $A$ is primitive.*

The proof is similar to that of Theorem 2.12, except that Theorem 2.7 is replaced by Definition 2.15.

**Corollary 2.2.** *The set of primitive matrices is convex.*

### 2.2.2 Stopping Times and Hitting Probabilities

The contents of this subsection are very useful when we study reinforcement learning using "episodes." These results are presented, but without using vector notation, in [111, Section 1.3] and [167, Section 4.2.2]. The derivations given here are cleaner.

**Definition 2.17.** A state $x_i \in \mathcal{X}$ is said to be an **absorbing state** if $X_t = x_i$ implies that $X_{t+1} = x_i$, or equivalently, that $X_\tau = x_i$ for all $\tau \geq t$. Another equivalent defintion is that row $i$ of the state transtion matrix $A$ consists of a 1 in column $i$ and zeros elsewhere. More generally, a subset $\mathcal{S} \subseteq \mathcal{X}$ is said to be a **set of absorbing states** if $X_t \in \mathcal{S} \implies X_\tau \in \mathcal{S}$ for all $\tau > t$.

Now we illustrate the concepts of absorbing states, and of absorbing sets. For convenience, we change notation slightly. Assume that the state space $\mathcal{X}$ of a Markov process can be partitioned as $\mathcal{T} \cup \mathcal{S}$, where $\mathcal{T}$ denotes the set of "transient" states, and $\mathcal{S}$ is an absorbing set. Suppose further that $\mathcal{T} = \{x_1, \cdots, x_m\}$, and $\mathcal{S} = \{a_1, \ldots, a_s\}$. The logic behind the phrases "transient" and "absorbing" is brought in [131, 167]. It is a ready consequence of Definition 2.17 that the state transition matrix $M$ of the Markov process has the form (note the change in notation):

$$M = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix}, \tag{2.2.11}$$

where $C \in \mathbb{S}_{s \times s}$ is a row stochastic matrix in itself, and the matrix $B$ has at least one nonzero element. Note too that the set $S$ can be absorbing, even if no individual state in $S$ is absorbing. For example, suppose $C$ is a permutation matrix over $s$ indices. However, if $C = I_s$, the identity matrix, then not only is the set $S$ absorbing, but every individual state in $S$ is absorbing. In this case the matrix $M$ looks like

$$M = \begin{bmatrix} A & B \\ 0 & I_s \end{bmatrix}. \tag{2.2.12}$$

An illustration of an absorbing state is provided by the snakes and ladders game. If the player's position hits 100, then the game is over. So 100 is an absorbing state. In other games like Blackjack, there are *two* absorbing states, namely $W$ and $L$ (for win and lose). In the Markov process literature, any sample path $X_0^l$ such that $X_l$ is an absorbing state is called an **episode**.

It can be shown that if the state $X_t$ of the Markov process enters the absorbing set $S$ with probability one as $t \to \infty$, then $B \neq 0$, that is, $B$ contains at least one nonzero element, and further, $\rho(A) < 1$. See specifically Items 3 and 6 of [167, Theorem 4.7]. More details can be found in [167, Section 4.2.2]. (Note that notation in [167] is different.) For the purposes of RL, it is useful to go beyond these facts, and to compute the probability distribution of the time at which the state trajectory enters $S$. In turn this gives the average number of time steps needed to reach the absorbing set. In case there are multiple absorbing states, it is also possible to compute the probability of hitting an individual absorbing state $a_i$ within the overall absorbing set $\mathcal{S}$. To be specific, define $\theta_{iS}$ to be the *first time* that a sample path $\{X_0^\infty\}$ hits the set $S$, starting at $X_0 = x_i$. Further, if $M$ is of the form (2.2.12) so that each set in $S$ is absorbing, define $\theta_{ik}$ to be the first time that a sample path $\{X_0^\infty\}$ hits the absorbing state $a_k$, starting at $X_0 = x_i$. Note that both $\theta_{iS}$ and $\theta_{ik}$ are integer-valued random variables. Then we have the following result:

**Theorem 2.15.** *With the above notation, we have that*

$$\Pr\{\theta_{iS} = l\} = \mathbf{e}_i^\top A^{l-1} B \mathbf{1}_s \ \forall l \geq 1, \tag{2.2.13}$$

*where $\mathbf{e}_i$ denotes the $i$-th elementary column vector with a 1 in row $i$ and zeros elsewhere. If $M$ has the form (2.2.12), the for each $k \in [s]$, we have*

$$\Pr\{\theta_{ik} = l\} = \mathbf{e}_i^\top A^{l-1} \mathbf{b}_k \ \forall l \geq 1 \tag{2.2.14}$$

*where $\mathbf{b}_k$ denotes the $k$-th column of $B$. The probability that a sample path $X_0^\infty$ with $X_0 = x_i$ terminates in the absorbing state $a_k$ is given by*

$$p_{ik} = \mathbf{e}_i^\top (I - A)^{-1} \mathbf{b}_k. \tag{2.2.15}$$

*Moreover,*

$$\sum_{k=1}^{s} p_{ik} = 1, \ \forall i \in [m].$$

*The vector of probabilities that a sample path $X_0^\infty$ terminates in the absorbing state $a_k$ is given by*

$$\mathbf{p}_k = (I - A)^{-1} \mathbf{b}_k. \tag{2.2.16}$$

*For each transient initial state $x_i \in \mathcal{T}$, define the average hitting time to reach the absorbing set $S$ starting from the initial state $x_i$ to be the expected value of $\theta_{iS}$, that is*

$$\bar{\theta}_{iS} = \sum_{l=1}^{\infty} l \Pr\{\theta_{iS} = l\},$$

*and the vector of average hitting times as $\bar{\boldsymbol{\theta}}_S \in \mathbb{R}^m$. Then*

$$\bar{\boldsymbol{\theta}}_S = (I - A)^{-1} \mathbf{1}_n. \tag{2.2.17}$$

*Proof.* We begin by deriving the expressions for the probability distributions. Note that $\theta_{iS} = l$ if and only if (i) the states $X_1, \cdots, X_{l-1}$ belong to $\mathcal{T}$, and (ii) $X_l \in \mathcal{S}$. For each pair of indices $i, j \in [m]$ and each integer $l$, the value $(A^l)_{ij}$ is the probability that, starting in state $x_i$ at time $t = 0$, the state $X_l$ at time $l$ equals $x_j$, *while staying within the set $\mathcal{T}$*. Thus the probability that $\theta_{iS} = l$ is given by

$$\Pr\{\theta_{iS} = l\} = \sum_{j=1}^{m} (A^{l-1})_{ij} (B\mathbf{1}_s)_j = \mathbf{e}_i^\top A^{l-1} B \mathbf{1}_s.$$

Here the summation is over all states $j \in \mathcal{T}$. This is (2.2.13). If $S$ consists of individual absorbing states, and we wish to determine the probability distribution that $X_l = a_k$ given that $X_0 = x_i$, then we simply replace $B\mathbf{1}_s$ by the corresponding $k$-th column of $B$. This is (2.2.14). Equation (2.2.15) is obtained by observing that, since $\rho(A) < 1$, we have that

$$\sum_{l=1}^{\infty} A^{l-1} = (I - A)^{-1}.$$

Therefore the probability that a trajectory starting at $x_i$ terminates in state $a_k$ is given by

$$\sum_{l=1}^{\infty} \mathbf{e}_i^\top A^{l-1} \mathbf{b}_k = \mathbf{e}_i \left[ \sum_{l=1}^{\infty} A^{l-1} \right] \mathbf{b}_k = \mathbf{e}_i (I - A)^{-1} \mathbf{b}_k.$$

This is (2.2.15). Stacking these probabilities as $i$ varies over $[m]$ gives (2.2.16).

Next we deal with the hitting times. Define the vector $\mathbf{b} = B\mathbf{1}_s$, and consider the modified Markov process with the state transition matrix

$$M = \begin{bmatrix} A & \mathbf{b} \\ 0 & 1 \end{bmatrix}.$$

In effect, we have aggregated the set of absorbing states into one "virtual state." From the standpoint of computing $\bar{\boldsymbol{\theta}}$, this is permissible, because once the trajectory hits the set $S$, or the virtual "last state" in the modified formulation, the time counter stops. To prove (2.2.17), suppose the Markov process starts in state $x_i$. Then there are two possibilities: First, with probability $b_i$, the trajectory hits the last virtual state. In this case the counter stops, and we can say that the hitting time is 1. Second, with probability $a_{ij}$ for each $j$, the trajectory hits the state $x_j$. In this case, the hitting time is now $1 + \bar{\theta}_j$. Therefore we have

$$\bar{\theta}_i = b_i + \sum_{j=1}^{n} a_{ij}(1 + \bar{\theta}_j).$$

Observe however that

$$b_i = 1 - \sum_{j=1}^{n} a_{ij}.$$

Substituting in the previous equation gives

$$\bar{\theta}_i = 1 + \sum_{j=1}^{n} a_{ij}\bar{\theta}_j,$$

or in matrix form

$$(I - A)\bar{\boldsymbol{\theta}} = \mathbf{1}_m.$$

Clearly this is equivalent to (2.2.17). □

**Example 2.11.** Consider the "toy" snakes and ladders game with two extra states, called $W$ and $L$ for win and lose respectively. The rules of the game are as follows:

- Initial state is S.
- A four-sided, fair die is thrown at each stage.
- Player must land exactly on $W$ to win and exactly on $L$ to lose.
- If implementing a move causes crossing of $W$ and $L$, then the move is not implemented.

There are twelve possible states in all: S, 1, ... , 9 , $W$, $L$. However, 2, 3, 9 can be omitted, leaving nine states, namely S, 1, 4, 5, 6, 7, 8, $W$, $L$. At each step, there are at most four possible outcomes. For example, from the state S, the four outcomes are 1, 7, 5, 4. From state 6, the four outcomes are 7, 8, 1, and W. From state 7, the four outcomes are 8, 1, W, 7. From state 8, there four possible outcomes are 1, $W$, $L$ and 8 with probability 1/4 each, because if the die comes up with 4, then the move cannot be implemented. It is time-consuming but straight-forward to compute the state transition matrix as

| | $S$ | 1 | 4 | 5 | 6 | 7 | 8 | $W$ | $L$ |
|---|---|---|---|---|---|---|---|---|---|
| $S$ | 0 | 0.25 | 0.25 | 0.25 | 0 | 0.25 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0.25 | 0.50 | 0 | 0.25 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0.25 | 0.25 | 0.25 | 0.25 | 0 | 0 |
| 5 | 0 | 0.25 | 0 | 0 | 0.25 | 0.25 | 0.25 | 0 | 0 |
| 6 | 0 | 0.25 | 0 | 0 | 0 | 0.25 | 0.25 | 0.25 | 0 |
| 7 | 0 | 0.25 | 0 | 0 | 0 | 0 | 0.25 | 0.25 | 0.25 |
| 8 | 0 | 0.25 | 0 | 0 | 0 | 0 | 0.25 | 0.25 | 0.25 |
| $W$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| $L$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

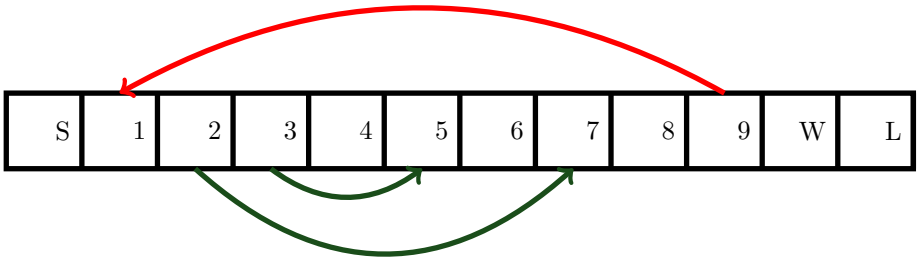



Figure 2.3: Toy Snakes and Ladders Game

The average duration of a game, which is the expected time before hitting one of the two absorbing states $W$ or $L$, is given by (2.2.17), and is

$$\boldsymbol{\theta} = \begin{bmatrix} 5.5738 \\ 5.4426 \\ 4.7869 \\ 4.9180 \\ 3.9344 \\ 3.1475 \\ 3.1475 \end{bmatrix}.$$

To compute the probabilities of reaching the absorbing states $W$ or $L$ from any nonabsorbing state, define $A$ to be the $7 \times 7$ submatrix on the top left, and $B$ to be the $7 \times 2$ submatrix on the top right. Then the probabilities of hitting $W$ and $L$ are given by (2.2.16), and are given by

$$[P_W \;\; P_L] = (I - A)^{-1} B = \begin{bmatrix} 0.5433 & 0.4567 \\ 0.5457 & 0.4543 \\ 0.5574 & 0.4426 \\ 0.5550 & 0.4450 \\ 0.6440 & 0.3560 \\ 0.5152 & 0.4848 \\ 0.5152 & 0.4848 \end{bmatrix}.$$

Not surprisingly, the two columns add up to one in each row, showing that, irrespective of the starting state, the sample path with surely hit either $W$ or $L$. Also not surprisngly, the probability of hitting $W$ is maximum in state 6, because it is possible to win in one throw of the die, but impossible to lose in one throw.

□

### 2.2.3 Maximum Likelihood Estimation of Markov Processes

Suppose $\{X_t\}_{t \geq 0}$ is a Markov process evolving over a finite state space (or alphabet) $\mathcal{X} = \{x_1, \ldots, x_n\}$, with an unknown state transition matrix $A$. We are able to observe a sample path $y_0^l := \{y_0, y_1, \ldots, y_l\}$ of the process, where each $y_i \in \mathcal{X}$. From this observation, we wish to determine *the most likely* state transition matrix $\hat{A}$, that is, the matrix $\hat{A}$ that maximizes the likelihood of the observed sample path. As it turns out, the solution is very simple.

**Theorem 2.16.** *Suppose we are given a sample path $y_0^l$, For each pair $(x_i, x_j) \in \mathcal{X}^2$, let $\nu_{ij}$ denote the number of times that the string $x_i x_j$ occurs (in that order) in the sample path $y_0^l$. Next, define*

$$\bar{\nu}_i := \sum_{j=1}^{n} \nu_{ij}. \tag{2.2.18}$$

*Then the maximum likelihood estimate of $A$ is given by*

$$\hat{a}_{ij} = \frac{\nu_{ij}}{\bar{\nu}_i}. \tag{2.2.19}$$

*Proof.* For a given sample path $y_0^l$, the likelihood that this sample path is generated by a Markov process with state transition matrix $A$ is given by

$$\begin{aligned} L(y_0^l|A) &= \Pr\{y_0\} \prod_{t=1}^{l} Pr\{X_t = y_t | X_{t-1} = y_{t-1}, A\} \\ &= \Pr\{y_0\} \prod_{t=1}^{l} a_{y_{t-1} y_t}. \end{aligned} \tag{2.2.20}$$

The formula becomes simpler if we take the logarithm of the above. Clearly, maximizing the log-likelihood of observing $y_0^l$ is equivalent to maximizing the likelihood of observing $y_0^l$. Thus

$$LL(y_0^l|A) = \log \Pr\{y_0\} + \sum_{t=1}^{l} \log a_{y_{t-1} y_t}. \tag{2.2.21}$$

A further simplification is possible. For each pair $(x_i, x_j) \in \mathcal{X}^2$, let $\nu_{ij}$ denote the number of times that the string $x_i x_j$ occurs (in that order) in the sample path $y_0^l$. Next, define $\bar{\nu}_i$ as in (2.2.18). Note that $\bar{\nu}_i$ is the number of times that the state $x_i$ occurs *in the sample path* $y_0^{l-1}$. The last symbol $y_l$ does not affect $\bar{\nu}_i$. It is easy to see that, instead of summing over strings $y_{t-1} y_t$, we can sum over strings $x_i x_j$. Thus $y_{t-1} y_t = x_i x_j$ precisely $\nu_{ij}$ times. Therefore

$$LL(y_0^l|A) = \log \Pr\{y_0\} + \sum_{i=1}^{n} \sum_{j=1}^{n} \nu_{ij} \log a_{ij}. \tag{2.2.22}$$

We can ignore the first term as it does not depend on $A$. Now, $A$ needs to satisfy the stochasticity constraint

$$\sum_{j=1}^{n} a_{ij} = 1, i = 1, \ldots, n. \tag{2.2.23}$$

So we want to maximize the right side of (2.2.22) (without the term $\log \Pr\{y_0\}$) subject to (2.2.23). For this purpose we form the Lagrangian

$$J = \sum_{i=1}^{n} \sum_{j=1}^{n} \nu_{ij} \log a_{ij} + \sum_{i=1}^{n} \lambda_i \left( 1 - \sum_{j=1}^{n} a_{ij} \right),$$

where $\lambda_1, \ldots, \lambda_n$ are the Lagrange multipliers. Next, observe that

$$\frac{\partial J}{\partial a_{ij}} = \frac{\nu_{ij}}{a_{ij}} - \lambda_i.$$

Setting the partial derivatives to zero gives

$$\lambda_i = \frac{\nu_{ij}}{a_{ij}}, \text{ or } a_{ij} = \frac{\nu_{ij}}{\lambda_i}.$$

The value of $\lambda_i$ can be determined from (2.2.23), which gives

$$\sum_{j=1}^{n} a_{ij} = \frac{1}{\lambda_i} \sum_{j=1}^{n} \nu_{ij} = \frac{\bar{\nu}_i}{\lambda_i} = 1 \implies \lambda_i = \bar{\nu}_i.$$

Therefore the maximum likelihood estimate for the state transition matrix of a Markov process, based on the sample path $y_0^l$, is given by (2.2.19). □

Here the caret over $a$ indicates that it is only an estimate of the true but unknown value $a_{ij}$. The only issue that remains to be settled is: What happens if a particular state $x_i$ does not appear at all in the sample path $y_0^l$? In this case, $\bar{\nu}_i = 0$, which perforce implies that $\nu_{ij} = 0$ for all $j$. Therefore (2.2.19) becomes indeterminate. The answer is that in this case, we can assign *any* vector in $\mathbb{S}_n$ as the $i$-th row of $\hat{A}$, and every such matrix is a maximum likelihood estimate of $A$.

**Example 2.12.** As a toy example to illustrate ML estimation, suppose $\mathcal{X} = \{0, 1\}$. Suppose we observe a sample path of length 21, so that $l = 20$, as follows:

$$y_0^{20} = 011011000110111010111.$$

Then

$$\nu_{00} = 2, \nu_{01} = 6, \nu_{10} = 5, \nu_{11} = 7, \bar{\nu}_0 = 8, \bar{\nu}_1 = 12.$$

Observe that the last element of 1 adds to $\nu_{11}$ but not to $\bar{\nu}_1$. Therefore the maximum likelihood estimate of the state transition matrix is

$$\hat{A} = \begin{bmatrix} 2/8 & 6/8 \\ 5/12 & 7/12 \end{bmatrix}.$$

Next we study a situation that arises frequently in applications, namely: Instead of having one long sample path, we have several sample paths, which are statistically independent of each other.

**Theorem 2.17.** *Suppose we are given $N$ different sample paths $(y_0^{l_k})^k, k \in [N]$, of a Markov process over $\mathcal{X} = \{x_1, \cdots, x_n\}$, with an unknown state transition matrix $A$. Suppose further that these sample paths are pairwise independent. Then the maximum likelihood estimate of $A$, denoted by $\hat{A}$ is obtained as follows: For each index $k \in [N]$, define the corresponding coefficients $\nu_{ij}^{(k)}, \bar{\nu}_i^{(k)}$ as in* (2.2.18) *and* (2.2.19) *respectively. Then the maximum likelihood estimate of $A$ is given by*

$$\hat{a}_{ij} = \frac{\sum_{k=1}^{N} \nu_{ij}^{(k)}}{\sum_{k=1}^{N} \bar{\nu}_i^{(k)}} \tag{2.2.24}$$

**Remark:** The expression (2.2.24) can be made clearer by fixing $k \in [N]$. Let us define

$$\hat{a}_{ij}^{(k)} = \frac{\nu_{ij}^{(k)}}{\bar{\nu}_i^{(k)}}, \tag{2.2.25}$$

which is the $ij$-th element of the maximum likelihood estimate based on only the $k$-th sample path. Then

$$\hat{a}_{ij} = \sum_{k=1}^{N} \frac{\bar{\nu}_i^{(k)}}{\sum_{k'=1}^{N} \bar{\nu}_i^{(k')}} \hat{a}_{ij}^{(k)}. \tag{2.2.26}$$

Thus each element of $\hat{A}$ is a *convex combination* of the corresponding $k$ maximum likelihood estimates, based on each of the $k$ sample paths. Moreover, the weights depend only on the row index $i$, but not on the column index $j$.

*Proof.* To reduce notational clutter, we will study the case of just two independent sample paths, and also change notation to $y_0^l$ and $z_0^m$. Then, for a given state transition matrix, the independence of the sample paths implies that

$$L(y_0^l \& z_0^m | A) = L(y_0|A) \cdot L(z_0^m|A).$$

Therefore

$$\begin{aligned} LL(y_0^l \& z_0^m|A) &= LL(y_0|A) + LL(z_0^m|A) \\ &= \log(y_0) + \sum_{t=1}^{l} \log(a_{y_{t-1},y_t}) + \log(z_0) + \sum_{t=1}^{m} \log(a_{z_{t-1},z_t}) \\ &= \log(y_0) + \log(z_0) + \sum_{i=1}^{n}\sum_{j=1}^{n} [\nu_{ij}^{(y)} + \nu_{ij}^{(z)}] \log a_{ij}. \end{aligned}$$

Now we can apply the same reasoning as in the proof of Theorem 2.16, and deduce that

$$\hat{a}_{ij} = \frac{\nu_{ij}^{(y)} + \nu_{ij}^{(z)}}{\bar{\nu}_i^{(y)} + \bar{\nu}_i^{(z)}}.$$

Finally, the above sum can be rewritten as

$$\begin{aligned} \hat{a}_{ij} &= \frac{\bar{\nu}_i^{(y)}}{\bar{\nu}_i^{(y)} + \bar{\nu}_i^{(z)}} \frac{\nu_{ij}^{(y)}}{\bar{\nu}_i^{(y)}} + \frac{\bar{\nu}_i^{(z)}}{\bar{\nu}_i^{(y)} + \bar{\nu}_i^{(z)}} \frac{\nu_{ij}^{(z)}}{\bar{\nu}_i^{(z)}} \\ &= \frac{\bar{\nu}_i^{(y)}}{\bar{\nu}_i^{(y)} + \bar{\nu}_i^{(z)}} \hat{a}_{ij}^{(y)} + \frac{\bar{\nu}_i^{(z)}}{\bar{\nu}_i^{(y)} + \bar{\nu}_i^{(z)}} \hat{a}_{ij}^{(z)}. \end{aligned}$$

Thus $\hat{a}_{ij}$ is a convex combination of $\hat{a}_{ij}^{(y)}$ and $\hat{a}_{ij}^{(z)}$. The reasoning can be readily extended to more than two independent sample paths. □

## 2.3 Some Convergence Theorems

In this section, we introduce the concepts of a martingale, supermartingale, and submartingale. Then we state and prove some convergence theorems that are based on these concepts.

### 2.3.1 Introduction to Martingales

Originally, martingales represented an abstract representation of a "fair game." In the context of optimization algorithms, martingales enter the picture to capture the notion that successive noise-corrupted measurements are unbiased. Therefore martingale difference sequences play a central role in analyzing the convergence of stochastic processes. In this subsection we briefly summarize some of the basic results. In turn, these lead to contemporary results on the convergence of the kind of stochastic processes arising in optimization and/or Reinforcement Learning,

Further details about this topic can be found in [173, 29, 20, 44]. In particular, [173, Part B] is a very good source of theorems and examples, while the corresponding exercises in [173, Part E] provide additional useful material. Similarly, [44, Chapter 4] has a wealth of material, including several examples and problems, that is relevant to the material below.

Before proceeding to a general discussion of martingales, let us recall the concepts of joint random variables, but with the twist that now we need to deal with *infinitely many* random variables, rather than just a finite number of them. Suppose we are interested in stochastic processes assuming values in the space $\mathbb{R}^d$ for fixed integer $d$. In this situation, the $\sigma$-algebra of subsets of $\mathbb{R}^d$ is the Borel $\sigma$-algebra, defined as the

**smallest $\sigma$-algebra that contains all open subsets of $\mathbb{R}^d$.** For convenience, this collection is denoted hereafter by $\mathcal{B}_d$. Note that we could also have used closed subsets instead of open subsets, and this would lead to the same $\sigma$-algebra. This is because a set is open if and only if its complement is closed, and a $\sigma$-algebra is closed under complementation.

One can think of a stochastic process $\{X_t\}$ evolving over the set $\mathbb{R}^d$ as a sequence of random variables, indexed by $t \geq 0$. Note that the sequence $\{X_t\}$ belongs to the product space $\prod_{t=0}^{\infty} \mathbb{R}^d$, which can also be denoted by $(\mathbb{R}^d)^{\mathbb{N}}$ where $\mathbb{N} = \{0, 1, \cdots\}$ denotes the set of natural numbers. Since our objective in these notes is to study the *convergence* of such a sequence as $t \to \infty$, we need to specify the associated $\sigma$-algebra. Specifically: What are the sample space and the event space, as defined in Definition 2.4. For this purpose, we need to define the infinite product measurable space $(\prod_{t=0}^{\infty} \mathbb{R}^d, \mathbb{S})$, where $\mathbb{S}$ is some suitable $\sigma$-algebra of subsets of $\prod_{t=0}^{\infty} \mathbb{R}^d$. Recall that in Definition 2.9, we have defined the product of *two* measurable spaces, which can be readily extended to any *finite* product. To extend this definition to an *infinite product* of measurable spaces, we proceed as follows.

**Definition 2.18.** A subset

$$S = \prod_{t=0}^{\infty} S_i$$

is called a **cylinder set** if (i) each $S_i \in \mathcal{B}_d$ for all $i$, and (ii) $S_i = \mathbb{R}^d$ for all but finitely many indices $i$. The smallest $\sigma$-algebra of subsets of $\prod_{t=0}^{\infty} \mathbb{R}^d$ that contains all cylinder sets is denoted by $\otimes_{t=0}^{\infty} \mathcal{B}_d$. The pair $(\prod_{t=0}^{\infty} \mathbb{R}^d, \otimes_{t=0}^{\infty} \mathcal{B}_d)$ is the product measurable space.

Thus, for an $\mathbb{R}^d$-valued stochastic process, the *event space* is the measurable space $(\prod_{t=0}^{\infty} \mathbb{R}^d, \otimes_{t=0}^{\infty} \mathcal{B}_d)$. We use the "canonical representation," whereby the underlying sample space is also $(\prod_{t=0}^{\infty} \mathbb{R}^d, \otimes_{t=0}^{\infty} \mathcal{B}_d)$. However, we still need to specify the *probability measure* $P$ on $(\prod_{t=0}^{\infty} \mathbb{R}^d, \otimes_{t=0}^{\infty} \mathcal{B}_d)$, which governs the behavior of the stochastic process.

We will use this definition and convention in various examples. However, in the iterests of completeness, we define filtrations and martingales in a general situation.

**Definition 2.19.** Suppose that $(\Omega, \mathcal{F}, P)$ is a probability space, as described in Section 2.1. A sequence of $\sigma$-algebras $\{\mathcal{F}_t\}_{t \geq 0}$ on $\Omega$ is called a **filtration** if

$$\mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \mathcal{F}, \ \forall t \geq 0. \tag{2.3.1}$$

Now suppose that $\{Z_t\}_{t \geq 0}$ is an $\mathbb{R}^d$-valued stochastic process on $(\Omega, \mathcal{F}, P)$. We say that $\{Z_t\}$ is **adapted** to the filtration $\{\mathcal{F}_t\}$, or that the pair $(\{Z_t\}, \{\mathcal{F}_t\})$ is adapted, if $Z_t$ is measurable with respect to $(\Omega, \mathcal{F}_t)$, (i.e., with $\mathcal{F}$ replaced by $\mathcal{F}_t$).

Clearly (2.3.1) implies that

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \mathcal{F}, \ \forall t \geq 0. \tag{2.3.2}$$

Since the underlying set $\Omega$ and probability measure $P$ are fixed, and the only thing varying is $\mathcal{F}_t$, we denote this by $Z_t \in \mathcal{M}(\mathcal{F}_t)$. In view of (2.3.1), we can make the following observations:

1. $Z_t \in \mathcal{M}(\mathcal{F}_\tau)$ whenever $\tau \geq t$.
2. Let $Z_0^t \in \mathbb{R}^{d(t+1)}$ denote $(Z_0, Z_1, \cdots, Z_t)$. Then $Z_0^t \in \mathcal{M}(\mathcal{F}_t)$.

If $\{Z_t\}_{t \geq 0}$ is an $\mathbb{R}^d$-valued stochastic process on $(\Omega, \mathcal{F}, P)$, then we can define the "natural filtration" by

$$\mathcal{F}_t = \sigma(Z_0^t),$$

where $\sigma(Z_0^t) \subseteq \mathcal{F}$ is the $\sigma$-algebra generated by $Z_0^t$. However, much of the discussion below applies even if we do not use the natural filtration, but use a larger filtration.

**Definition 2.20.** Suppose $\{\mathcal{F}_t\}$ is a filtration on $(\Omega, \mathcal{F})$, and that $\{Z_t\}_{t\geq 0}$ is an $\mathbb{R}^d$-valued stochastic process on $(\Omega, \mathcal{F}, P)$ wherein $Z_t \in L_1(\Omega, \mathcal{F}_t, P)$ for all $t$. (In other words, $\{(Z_t, \mathcal{F}_t)\}$ is adapted.) Then the pair $(\{Z_t\}, \{\mathcal{F}_t\})$ is said to be a **martingale** if

$$E(Z_{t+1}|\mathcal{F}_t) = Z_t, \text{ a.s., } \forall t \geq 0. \tag{2.3.3}$$

If (2.3.3) is replaced by

$$E(Z_{t+1}|\mathcal{F}_t) \leq Z_t, \text{ a.s., } \forall t \geq 0, \tag{2.3.4}$$

then $\{Z_t\}_{t\geq 0}$ is called a **supermartingale**, whereas if (2.3.6) is replaced by

$$E(Z_{t+1}|\mathcal{F}_t) \geq Z_t, \text{ a.s., } \forall t \geq 0, \tag{2.3.5}$$

then $\{Z_t\}_{t\geq 0}$ is called a **submartingale**.

If we use the natural filtration $\mathcal{F}_t = \sigma(Z_0^t)$, then (2.3.3) can be replaced by

$$E(Z_{t+1}|Z_0^t) = Z_t, \text{ a.s., } \forall t \geq 0. \tag{2.3.6}$$

Similar remarks apply to supermartingales and submartingales.

Several useful consequences of the definition are obtained by applying Theorem 2.5. If $\{Z_t\}$ is a martingale, then by the iterated conditioning property (Item 6 of Theorem 2.5), it follows that

$$E(Z_\tau|\mathcal{F}_t) = Z_t, \text{ a.s., } \forall \tau \geq t+1, \ \forall t \geq 0. \tag{2.3.7}$$

The equality is replaced by $\leq$ for a supermartingale, and by $\geq$ for a submartingale. Next, by the expected value preservation property (Item 3 of Theorem 2.5), it follows that[3]

$$E[Z_t, P] = E[Z_0, P], \ \forall t \geq 0. \tag{2.3.8}$$

It similarly follows that if $\{Z_t\}$ is a supermartingale, then

$$E[Z_t, P] \leq E[Z_0, P], \ \forall t \geq 0, \tag{2.3.9}$$

where as if $\{Z_t\}$ is a submartingale, then

$$E[Z_t, P] \geq E[Z_0, P], \ \forall t \geq 0. \tag{2.3.10}$$

Thus, in a supermartingale, $\{E[Z_t, P]\}$ is a nonincreasing sequence of real numbers, while in a submartingale, $\{E[Z_t, P]\}$ is a nondecreasing sequence of real numbers.

Next, let $\{\xi_t\}_{t\geq 0}$ be a stochastic process adaptated to a filtration $\{\mathcal{F}_t\}$, such that $E[|\xi_t|, P] < \infty$ for all $t$, and define

$$Z_t = \sum_{\tau=0}^{t} \xi_\tau. \tag{2.3.11}$$

Then it is obvious that $\{Z_t\}$ is also adapted to $\{\mathcal{F}_t\}$. The sequence $(\{\xi_t\}, \{\mathcal{F}_t\})$ is said to be a **martingale difference sequence** if $(\{Z_t\}, \{\mathcal{F}_t\})$ is a martingale. It is easy to show using (2.3.3) that, if $\{\xi_t\}$ is a martingale difference sequence, then

$$E(\xi_{t+1}|\mathcal{F}_t) = 0, \text{ a.s., } \forall t \geq 0. \tag{2.3.12}$$

If, in addition, $\xi_0 = 0$ almost surely, then it follows that $E[\xi_t, P] = 0$ for all $t \geq 1$. The picture is clearer if each $\xi_t$ belongs to $L_2(\Omega, \mathcal{F}_t)$. Then, by the projection property (Item 9) of Theorem 2.5, (2.3.12) is equivalent to the statement that $\xi_{t+1}$ is orthogonal to every element of $L_2(\Omega, \mathcal{F}_t)$.

[3]The reader is reminded that, wherever possible, we use parentheses for the conditional expectation, which is a random variable, and square brackets for the expected value, which is a real number.

**Example 2.13.** Suppose $\{\xi_t\}$ is a sequence of random variables such that $\xi_t \in L_1(\Omega, \mathcal{F}, P)$ for each $t$, and in addition,

$$E(\xi_{t+1}|\mathcal{F}_t) = 0 \text{ a.s. } \forall t.$$

Then $\{\xi_t\}$ is a martingale difference sequence, and the sequence $\{Z_t\}$ defined in (2.3.11)is a martingale. □

**Example 2.14.** In this example we study a coin-tossing game as an illustration of a martingale. The game starts at time $t = 0$ with the person having a prespecified amount of money, which can be taken as $X_0 = 0$ without loss of generality. At each time $t \geq 0$, a fair coin is tossed. If the coin turns up "heads," then the person receives 10 units of money, whereas if the coin turns "tails," then the person must pay up 10 units of money. To keep the notation consistent, let us suppose that the reward for the coin toss at time $t$ (positive or negative) is paid at time $t + 1$. Let $X_0 = 0$, and let $X_t, t \geq 1$ denote the payoff at time $t$ (corresponding to the coin toss at time $t - 1$). Define

$$Z_t = \sum_{\tau=0}^{t} X_t.$$

Note that we can also start the summation from time 1, because $X_0 = 0$.

To study this situation formally, we use the structure introduced in Definition 2.18. Thus the stochastic process $\{X_t\}$ evolves on the measurable space $(\mathbb{R}^{\mathbb{N}}, \otimes_{t=0}^{\infty}\mathcal{B})$. Let $X_0^t$ denote the tuple of random variables $(X_0, \cdot, X_t)$, and define $\mathcal{F}_t = \sigma(X_0^t)$. Then each $\mathcal{F}_t$ is a sub-$\sigma$-algebra of $\otimes_{t=0}^{\infty}\mathcal{B}$. Moreover, $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$, which means that $\{\mathcal{F}_t\}$ is a filtration. To complete the specification of the stochastic process, we need to define the probability law of the sequence $\mathcal{X}_0^{\infty}$. To keep the discussion simple, it is assumed that each $X_t$ is independent of $X_0^{t-1}$. Moreover, $X_t = \pm 10$ with equal probability.

Let us analyze the stochastic process $\{Z_t\}$. Observe that $Z_{t+1} = Z_t + X_{t+1}$, and that $E(X_{t+1}|\mathcal{F}_t) = 0$, because $X_{t+1}$ is independent of previous tosses, and because the coin is fair. Therefore

$$E(Z_{t+1}|\mathcal{F}_t) = E(Z_t|\mathcal{F}_t) + E(X_{t+1}|\mathcal{F}_t) = Z_t,$$

because $Z_t \in \mathcal{M}(\mathcal{F}_t)$, and $E(X_{t+1}|\mathcal{F}_t) = 0$. Thus $\{Z_t\}$ is a martingale.

Next, let us change the problem specification so that $E(X_{t+1}|\mathcal{F}_t)$ is no longer zero. This can be done in either of two ways, which are mathematically equivalent: First, the coin can remain fair, but the payoffs for Heads and Tails are not equal. Second, the payoffs can remain equal, but the coin is not fair (it has a bias in favour of Heads or Tails). If the coin is fair but the payoff for Heads is larger in magnitude than the penalty for Tails, then $E(X_{t+1}|\mathcal{F}_t) > 0$, and as a result

$$E(Z_{t+1}|\mathcal{F}_t) = E(Z_t|\mathcal{F}_t) + E(X_{t+1}|\mathcal{F}_t) > Z_t,$$

and the process is now a submartingale. Reversing the magnitudes so that $E(X_{t+1}|\mathcal{F}_t) < 0$ causes the process to be a supermartingale. □

**Example 2.15.** Now we continue the previous example by introducing the concept of "marginal utility" from economics. If an individual has a quantum of money $M$, its "utility" is defined as $U(M)$. Usually $U$ is taken as map from $\mathbb{R}_+$ to $\mathbb{R}_+$, but since the coin toss can result in both a loss as well as a profit, we take $U$ as a map from $\mathbb{R}$ to $\mathbb{R}$ such that $U(0) = 0$. Two key attributes of $U(\cdot)$ are:

1. $U(.)$ is strictly increasing. Thus

$$x_1 > x_2 \implies U(x_1) > U(x_2).$$

2. $U(\cdot)$ is strictly concave. Thus

$$x_1 \neq x_2, \lambda \in (0,1) \implies U[\lambda x_1 + (1-\lambda)x_2] > \lambda U(x_1) + (1-\lambda)U(x_2).$$

The second property is also referred to as "diminishing marginal utility" or something similar.

Now let us return to the coin-flipping game. Suppose the outcomes $H$ and $T$ has probabilities $p$ and $1-p$, and the payoffs are $a$ and $-b$ (the latter being a penalty). Suppose further that the coin toss is "fair" in the sense that the expected value of the payoff is zero, that is

$$pa - (1-p)b = 0.$$

Let $X_t$ denote the accumulated payoffs at time $t$, starting at $X_t = 0$. Now let us examine the utility $U(X_t)$. Both $\{X_t\}$ and $\{U_t\}$ are stochastic processes on $(\mathbb{R}^{\mathbb{N}}, \otimes_0^\infty \mathcal{B})$. We know that

$$X_{t+1} = \begin{cases} X_t + a & \text{w.p. } p, \\ X_t - b & \text{w.p. } 1-p. \end{cases}$$

Also

$$p(X_t + a) + (1-p)(X_t - b) = X_t.$$

Therefore $\{X_t\}$ is a martingale. This relationship also shows that $X_t$ is a convex combination of $X_t + a$ and $X_t - b$. Hence

$$E[U(X_{t+1}|X_t] = pU(X_t + a) + (1-p)U(X_t - b) < U(X_t).$$

Hence $\{U_t\}$ is a supermartingale. □

Now we present some important results related to martingales, which are useful in themselves, though they are not directly used in this book. The material is taken from [173, Chapter 12] and/or [44, Chapter 4] and is stated without proof. Citations from these sources are given for individual results stated below.

The first result we present is the Doob decomposition theorem. To state this theorem, we introduce a new concept. Suppose $\{\mathcal{F}_t\}_{t\geq 0}$ is a filtration, and $\{A_t\}_{t\geq 1}$ is a stochastic process that is adapted to $\mathcal{F}_t$, that is, $A_t \in \mathcal{M}(\mathcal{F}_t)$ for all $t \geq 1$. We say that $\{(A_t, \mathcal{F}_t)\}$ is **predictable** if $A_{t+1} \in \mathcal{M}(\mathcal{F}_t)$ for all $t \geq 0$. Note that there is no $A_0$ for a predictable process. Also, note that in [173], such processes are said to be "previsible." However, the phrase "predictable" is used in [44] and appears to be more commonly used. We say that a martingale $\{Z_t\}$ (adapted to $\mathcal{F}_t$) is **null at zero** if $Z_0 = 0$ a.s., and that a predictable process $\{A_t\}$ is **null at zero** (because there is no $A_0$).

**Theorem 2.18.** *(Doob decomposition theorem.) See [173, Theorem 12.11] or [44, Theorem 4.3.2].) Suppose $\{\mathcal{F}_t\}$ is a filtration and $\{Y_t\}$ is a stochastic process adapted to $\{\mathcal{F}_t\}$. Then $Y_t$ can be expressed as*

$$Y_t = Z_t + A_t + Y_0, \tag{2.3.13}$$

*where $\{Z_t\}_{t\geq 0}$ is a martingale null at zero, and $\{A_t\}$ is a predictable process null at zero. If $\{Z'_t\}$ and $\{A'_t\}$ also satisfy the above conditions, then*

$$P\{\omega : Z_t(\omega) = Z'_t(\omega) \& A_t(\omega) = A'_t(\omega),\ \forall t\} = 1. \tag{2.3.14}$$

*(In other words, the decomposition is essentially unique.) Moreover, $\{Y_t\}$ is a submartingale if and only if $\{A_t\}$ is an increasing process, that is*

$$P\{\omega : A_{t+1}(\omega) \geq A_t(\omega),\ \forall t\} = 1. \tag{2.3.15}$$

*Similarly, $\{Y_t\}$ is a supermartingale if and only if $\{A_t\}$ is a decreasing process, that is*

$$P\{\omega : A_{t+1}(\omega) \leq A_t(\omega),\ \forall t\} = 1. \tag{2.3.16}$$

*Proof.* Define

$$A_{t+1} = \sum_{\tau=0}^{t} E((Y_{\tau+1} - Y_\tau)|\mathcal{F}_\tau) = \sum_{\tau=0}^{t} [E(Y_{\tau+1}|\mathcal{F}_\tau) - Y_\tau]. \tag{2.3.17}$$

Then it is obvious that $A_{t+1} \in \mathcal{M}(\mathcal{F}_t)$; hence $\{A_t\}$ is a predictable process. Also, $A_t$ satisfies the recursion

$$A_{t+1} = E(Y_{\tau+1}|\mathcal{F}_\tau) - Y_\tau + A_t. \tag{2.3.18}$$

Now define a stochastic process $\{Z_t\}$ by $Z_0 = 0$, and

$$Z_{t+1} = Y_{t+1} - A_{t+1} - Y_0, \text{ or } Y_{t+1} = Z_{t+1} + A_{t+1} + Y_0, \ \forall t \geq 0. \tag{2.3.19}$$

It is now shown that $\{Z_t\}$ is a martingale, which would prove the first statement. Observe that

$$\begin{aligned} E(Z_{t+1}|\mathcal{F}_t) &= E(Y_{t+1}|\mathcal{F}_t) - A_{t+1} - Y_0 \\ &= E(Y_{t+1}|\mathcal{F}_t) - [E(Y_{\tau+1}|\mathcal{F}_t) - Y_t + A_t] + Y_0 \\ &= Y_t - A_t + Y_0 = Z_t, \end{aligned}$$

where in the first step we use the fact that $A_{t+1} \in \mathcal{M}(\mathcal{F}_t)$, and in the second step we use the (2.3.18).

To prove the uniqueness of the decomposition, we essentially reverse the above steps. Suppose

$$Y_{t+1} = Z'_{t+1} + A'_{t+1} + Y_0, \tag{2.3.20}$$

where $\{Z'_t\}$ is a martingale null at $t = 0$, and $\{A'_t\}$ is predictable. Then

$$\begin{aligned} E(Y_{t+1}|\mathcal{F}_t) &= E(Z'_{t+1}|\mathcal{F}_t) + A'_{t+1} + Y_0 \\ &= Z'_t + A'_{t+1} + Y_o \\ &= Z'_t + A'_t + Y_0 + (A'_{t+1} - A'_t) \\ &= Y_t + (A'_{t+1} - A'_t) \end{aligned}$$

Therefore

$$A'_{t+1} - A'_t = E(Y_{t+1}|\mathcal{F}_t) - Y_t, \text{ or } A'_{t+1} = \sum_{\tau=0}^{t} [E(Y_{\tau+1}|\mathcal{F}_\tau) - Y_\tau].$$

Since this is the same summation as in (2.3.17), it follows that $A'_t = A_t$ almost surely. Substituting this into (2.3.19) leads to

$$Z'_{t+1} = Y_{t+1} - A'_{t+1} - Y_0 = Y_{t+1} - A_{t+1} - Y_0 = Z_{t+1} \text{ a.s.}$$

This shows that the decomposition is unique modulo differing on a set of measure zero.

To prove the last part of the theorem, rewrite (2.3.18) as

$$A_{t+1} - A_t = \sum_{\tau=0}^{t} [E(Y_{\tau+1}|\mathcal{F}_\tau) - Y_\tau].$$

So $A_{t+1} \geq A_t$ for all $t$ if and only if $\{Y_t\}$ is a submartingale, and So $A_{t+1} \leq A_t$ for all $t$ if and only if $\{Y_t\}$ is a supermartingale. □

Next, suppose $Y_t = M_t^2$, where $\{M_t\}$ is a martingale in $L_2(\Omega, P)$ null at zero. Then it is easy to show using the conditional Jensen's inequality (not covered here) that $\{Y_t\}$ is a submartingale null at zero. Therefore the Doob decomposition of $Y_t = M_t^2$ is

$$M_t^2 = Z_t + A_t, \tag{2.3.21}$$

where $\{Z_t\}$ is a martingale and $\{A_t\}$ is an increasing predictable process, both null at zero. It is customary to refer to $\{A_t\}$ as the **quadratic variation process** and to denote it by $\langle M_t \rangle$. Note that

$$A_{t+1} - A_t = E((M_{t+1}^2 - M_t^2)|\mathcal{F}_t) = E((M_{t+1} - M_t)^2|\mathcal{F}_t). \tag{2.3.22}$$

Define $A_\infty(\omega) = \lim_{t\to\infty} A_t(\omega)$ for (almost all) $\omega \in \Omega$. Then we have the following:

**Theorem 2.19.** *(See [173, Theorem 12.13].) If $A_\infty(\cdot)$ is bounded almost everywhere as a function of $\omega$, then $\{M_t(\omega)\}$ converges almost everywhere at $t \to \infty$.*

Actually [173, Theorem 12.13] is more powerful and gives "almost necessary and sufficient" conditions for convergence. We have simply extracted what is needed for present purposes.

We will use several versions of the next theorem repeatedly when analyzing the convergence of various stochastic algorithms.

**Theorem 2.20.** *(See [44, Theorem 4.2.12].) If $\{Z_t\}$ is nonnegative (i.e., $Z_t \geq 0$ a.s.) supermartingale, then there exists a $\zeta \in L_1(\Omega, P)$ such that $Z_t \to \zeta$ almost surely, and $E[\zeta, P] \leq E[Z_0, P]$.*

Note what the theorem *does not* say: There is no guarantee that $Z_t$ converges to $\zeta$ *in the mean* as $t \to \infty$. However, if $Z_t \in L_p(\Omega, P)$ for some $p > 1$, we can make a stronger statement.

A slight variation of the above theorem is also useful.

**Corollary 2.3.** *Suppose $\{Z_t\}$ is a supermartingale, and that there exists a fixed constant $c$, independent of $\omega$, such that*

$$Z_t(\omega) \geq -c \text{ a.s.} \tag{2.3.23}$$

*Then there exists a $\zeta \in L_1(\Omega, P)$ such that $Z_t \to \zeta$ almost surely, and $E[\zeta, P] \leq E[Z_0, P]$.*

*Proof.* Observe that, since $c$ is a fixed constant independent of $\omega$, the process $\{Z_t - c\}$ is also a supermartingale. Moreover, this process is nonnegative. Now apply Theorem 2.20. □

**Theorem 2.21.** *Suppose $\{\mathbf{Z}_t\}$ is a martingale wherein $Z_t \in L_p(\Omega, P)$ for some $p > 1$, and suppose further that the martingale is bounded in $\|\cdot\|_p$, that is*

$$\sup_t E[Z_t^p, P] < \infty. \tag{2.3.24}$$

*Then there exists a $\zeta \in L_p(\Omega, P)$ such that $Z_t \to \zeta$ as $t \to \infty$, almost surely and in the $p$-th mean.*

The above theorem is false if $p = 1$. The convergence is almost sure but need not be in the mean. See [44, Example 4.2.13].

### 2.3.2 Some Convergence Theorems

Recall that, throughout, we are dealing with stochastic processes defined on some probability space $(\Omega, \mathcal{F}, P)$, even if we do not always display this probability space explicitly. Thus when we write, for example, $\{z_t\}$, we really mean $\{z_t(\omega)\}$. For the most part, it is not necessary to display this dependence on $\omega$. Wherever it is necessary, we display it. But the $\omega$ is implicitly present throughout. Also, when we say $z_t \geq 0$, we mean that $z_t(\omega) \geq 0$ for almost all $\omega$.

The theorems presented in this subsection are the basis of all the proofs of convergence, and estimates of the rate of convergence, presented in these notes.

Theorem 2.22 below, originally due to [124], can be said to be the "workhorse" in this area, in the sense that practically every convergence theorem in this book can be traced back to this theorem, in one way or another. It is referred to as the "Robbins-Siegmund Theorem," or the "almost supermartingale convergence theorem." We prefer the former name. This result refines an earlier argument from [52], but was discovered independently. The proof given below is pretty much the same as in the original paper. Another proof, based on "stopping times" (not discussed in this book) can be found in [9, Section 5.2.1]. Yet another proof can be found in the survey paper [48].

**Theorem 2.22.** ***(Robbins-Siegmund Theorem)*** *Suppose $\{z_t\}, \{f_t\}, \{g_t\}, \{h_t\}$ are nonnegative stochastic processes adapted to some filtration $\{\mathcal{F}_t\}$, that satisfy*

$$E(z_{t+1}|\mathcal{F}_t) \leq (1 + f_t)z_t + g_t - h_t \text{ a.s., } \forall t. \tag{2.3.25}$$

*Define the set* $\Omega_0 \subseteq \Omega$ *by*

$$\Omega_0 := \{\omega : \sum_{t=0}^{\infty} f_t(\omega) < \infty\} \cap \{\omega : \sum_{t=0}^{\infty} g_t(\omega) < \infty\}, \tag{2.3.26}$$

*Then, for all*[4] $\omega \in \Omega_0$*, we have that (i)* $\lim_{t\to\infty} z_t(\omega)$ *exists and is finite, and (ii)*

$$\sum_{t=0}^{\infty} h_t(\omega) < \infty, \ \forall \omega \in \Omega_0. \tag{2.3.27}$$

*In particular, if* $P(\Omega_0) = 1$*, then* $\{z_t\}$ *is bounded almost surely, in the sense that*

$$P\{\omega \in \Omega : \sup_t z_t(\omega) < \infty\} = 1, \tag{2.3.28}$$

*and*

$$\sum_{t=0}^{\infty} h_t(\omega) < \infty \ \textit{a.s.} \tag{2.3.29}$$

*Proof.* We begin with a simple observation. For each $\omega \in \Omega_0$,

$$\sum_{t=0}^{\infty} f_t(\omega) < \infty \implies \prod_{t=0}^{\infty}[1 + f_t(\omega)] < \infty.$$

Therefore, for each $\omega \in \Omega_0$,

$$\prod_{t=0}^{t}[1 + f_t(\omega)]^{-1} \downarrow \prod_{t=0}^{\infty}[1 + f_t(\omega)]^{-1} =: b(\omega) > 0, \tag{2.3.30}$$

where the limit $b(\omega)$ could depend on $\omega$. Now define new processes (note the difference in the upper limits of the products)

$$z'_t = \prod_{t=0}^{t-1}[1 + f_t(\omega)]^{-1} z_t, \ \ g'_t = \prod_{t=0}^{t}[1 + f_t(\omega)]^{-1} g_t, \ \ h'_t = \prod_{t=0}^{t}[1 + f_t(\omega)]^{-1} h_t,$$

and observe that

$$\prod_{t=0}^{t}[1 + f_t(\omega)]^{-1} \in \mathcal{M}(\mathcal{F}_t).$$

With these definitions, we can compute from (2.3.25) that

$$\begin{aligned} E(z'_{t+1}|\mathcal{F}_t) &= \prod_{t=0}^{t}[1 + f_t(\omega)]^{-1} E(z_{t+1}|\mathcal{F}_t) \\ &\leq \prod_{t=0}^{t-1}[1 + f_t(\omega)]^{-1} z_t + \prod_{t=0}^{t}[1 + f_t(\omega)]^{-1} g_t - \prod_{t=0}^{t}[1 + f_t(\omega)]^{-1} h_t \\ &= z'_t + g'_t - h'_t. \end{aligned} \tag{2.3.31}$$

Next, define

$$u_t = z'_t - \sum_{\tau=0}^{t-1}(g'_\tau - h'_t). \tag{2.3.32}$$

[4]Here and elsewhere, "for all" really means "for almost all."

Fix some arbitrary constant $a > 0$, and define

$$T(a,\omega) := \inf\{t : \sum_{\tau=0}^{t} g'_t > a\},$$

with the understanding that if $\sum_{\tau=0}^{\infty} g'_t \le a$, then $T(a,\omega) = \infty$. Suppose now that $t < T(a,\omega)$. Then

$$\begin{aligned} E(u_{t+1}|\mathcal{F}_t) &= E(z'_{t+1} - \sum_{\tau=0}^{t}(g'_\tau - h'_t)|\mathcal{F}_t) \\ &\le z'_t + g'_t - h'_t - \sum_{\tau=0}^{t}(g'_\tau - h'_t) \\ &= z'_t - \sum_{\tau=0}^{t-1}(g'_\tau - h'_t) = u_t. \end{aligned} \tag{2.3.33}$$

Let us use the notation $t \wedge \tau$ to denote $\min\{t,\tau\}$, and define a new stochastic process $\{v_t(\omega)\}$ by

$$v_t(\omega) = u_{T(a,\omega)\wedge(t+1)}(\omega).$$

Thus

$$v_t(\omega) = \begin{cases} u_t(\omega), & \text{if } t < T(a,\omega), \\ u_{T(a,\omega)}(\omega), & \text{if } t \ge T(a,\omega). \end{cases}$$

It is now shown that $\{v_t\}$ is a supermartingale. If $t \le T(a,\omega)$, (2.3.33) implies that

$$E(v_{t+1}|\mathcal{F}_t) \le v_t,$$

whereas if $t \ge T(a,\omega)$, then $t+1 \ge T(a,\omega)$. Hence

$$v_{t+1}(\omega) = u_{T(a,\omega)} = v_t(\omega).$$

So $\{v_t\}$ is a supermartingale. Now, because both $\{z'_t\}$ and $\{g'_t\}$ are nonnegative processes, it follows that

$$u_{T(a,\omega)\wedge t}(\omega) \ge -\sum_{\tau=0}^{T(a,\omega)\wedge t} g'_\tau(\omega) \ge -\sum_{\tau=0}^{\infty} g'_\tau(\omega) \ge -a.$$

Therefore it follows from Corollary 2.3 that

$$\lim_{t\to\infty} u_{T(a,\omega)\wedge t}(\omega)$$

exists and is finite for all $\omega$ such that

$$\sum_{\tau=0}^{\infty} g'_\tau(\omega) \le a. \tag{2.3.34}$$

Now define

$$\Omega_1 := \{\omega : \sum_{\tau=0}^{\infty} g'_\tau(\omega) < \infty\}.$$

However, for all $\omega \in \Omega_0$, it follows from (2.3.30) that

$$b(\omega)g_t(\omega) \le g'_t(\omega) \le g_t(\omega).$$

And since

$$\sum_{\tau=0}^{\infty} g_\tau(\omega) < \infty \; \forall \omega \in \Omega,$$

it follows that we can take $\Omega_1 = \Omega_0$. Since $a$ is arbitrary, it follows that for all $\omega \in \Omega_0$, (2.3.33) holds for sufficiently large $a$ (which could depend on $\omega$), so that $T(a, \omega) = \infty$, and $u_{T(a,\omega)\wedge t}(\omega)$ for all $t$. Hence $\lim_{t\to\infty} u_t(\omega)$ exists and is finite (almost surely) for all $\omega \in \Omega_0$. In turn this implies that $u_t(\omega)$ is bounded for all $\omega \in \Omega$.

Next we study the consequences of $u_t(\omega)$ being bounded (though the bound could depend on $\omega$). Rewrite (2.3.32) as

$$u_t = z'_t + \sum_{\tau=0}^{t-1} h'_\tau - \sum_{\tau=0}^{t-1} g_t.$$

For $\omega \in \Omega_0$, the last term is bounded from below, while $u_t(\omega)$ is bounded. Hence there is a bound $\bar{c}(\omega)$ such that

$$z'_t + \sum_{\tau=0}^{t-1} \psi'_\tau \leq \bar{c}(\omega), \; \forall t.$$

Since both terms are nonnegative, this in turn implies that

$$z'_t \leq \bar{c}(\omega), \sum_{t=0}^{\infty} \psi'_\tau \leq \bar{c}(\omega). \tag{2.3.35}$$

Also, since $u_t(\omega)$ converges as $t \to \infty$, and

$$\sum_{\tau=0}^{t-1} \psi'_\tau(\omega) \uparrow \sum_{\tau=0}^{\infty} \psi'_\tau(\omega),$$

it follows that $z'_t(\omega)$ has a limit as $t \to \infty$, and the limit is finite, for all $\omega \in \Omega_0$. To complete the proof, all that remains is to replace $z'_t, \psi'_t$ by $z_t, \psi_t$ respectively. But this is straight-forward, because

$$\prod_{\tau=0}^{t} [1 + f_\tau(\omega)]^{-1} \downarrow b(w) \text{ as } t \to \infty,$$

as shown in (2.3.30). □

The above proof is taken from [124]. The same theorem is also proved in a very terse form on [9, page 343]. Because that proof involves the use of "stopping times," a concept that is not needed elsewhere in this book, we choose to give the original proof.

Now we present two convergence theorems, which are extensions of Theorem 2.22. The first one allows us to infer the convergence of a stochastic process to zero, while the second one provides bounds on the *rate of convergence* to zero. Both theorems are taken from [70, 71].

Theorem 2.23 below builds upon Theorem 2.22 by providing sufficient conditions to ensure that $z_t \to 0$ as $t \to \infty$. It draws upon the concept of a function of Class $\mathcal{B}$, defined in Definition 7.6. For the convenience of the reader, the definition is repeated below.

**Definition 2.21.** A function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is said to **belong to Class $\mathcal{B}$** if $\phi(0) = 0$, and in addition, for arbitrary real numbers $0 < \epsilon \leq M$, it is true that

$$\inf_{\epsilon \leq r \leq M} \phi(r) > 0.$$

**Theorem 2.23.** *Suppose $\{z_t\}, \{f_t\}, \{g_t\}, \{h_t\}, \{\alpha_t\}$ are $[0, \infty)$-valued stochastic processes defined on some probability space $(\Omega, \mathcal{F}, P)$, and adapted to some filtration $\{\mathcal{F}_t\}$. Suppose further that*

$$E_t(z_{t+1}) \leq (1 + f_t)z_t + g_t - \alpha_t h_t \text{ a.s., } \forall t. \tag{2.3.36}$$

*Define*

$$\Omega_0 := \{\omega \in \Omega : \sum_{t=0}^{\infty} f_t(\omega) < \infty \text{ and } \sum_{t=0}^{\infty} g_t(\omega) < \infty\}, \tag{2.3.37}$$

$$\Omega_1 := \{\omega \in \Omega : \sum_{t=0}^{\infty} \alpha_t(\omega) = \infty\}. \tag{2.3.38}$$

*Then*

1. *Suppose that $P(\Omega_0) = 1$. Then the sequence $\{z_t\}$ is bounded almost surely, and there exists a random variable $W$ defined on $(\Omega, \mathcal{F}, P)$ such that $z_t(\omega) \to W(\omega)$ almost surely.*
2. *Suppose that, in addition to $P(\Omega_0) = 1$, it is also true that $P(\Omega_1) = 1$. Then*

$$\liminf_{t\to\infty} h_t(\omega) = 0 \ \forall \omega \in \Omega_0 \cap \Omega_1. \tag{2.3.39}$$

3. *Further, suppose there exists a function $\eta(\cdot)$ of Class $\mathcal{B}$ such that $h_t(\omega) \geq \eta(z_t(\omega))$ for all $\omega \in \Omega_0$. Then $z_t(\omega) \to 0$ as $t \to \infty$ for all $\omega \in \Omega_0$.*

*Proof.* By Theorem 2.22, there exists a random variable $W$ such that $z_t(\omega) \to W(\omega)$ as $t \to \infty$ for almost all $\omega \in \Omega_0$. This implies that $z_t$ is bounded almost surely. This is Item 1.

Next we prove item 2. Again from Lemma 2.22,

$$\sum_{t=0}^{\infty} \alpha_t(\omega) h_t(\omega) < \infty, \ \forall \omega \in \Omega_0.$$

Now, by definition

$$\sum_{t=0}^{\infty} \alpha_t(\omega) = \infty, \ \forall \omega \in \Omega_0 \cap \Omega_1.$$

Therefore (2.3.39) follows. To prove Item 3, suppose that, for some $\omega \in \Omega_0 \cap \Omega_1$, we have that $W(\omega) > 0$, say $W(\omega) =: 2\epsilon > 0$. Choose a time $T$ such that $z_t(\omega) \geq \epsilon$ for all $t \geq T$. Also, by Item 1,

$$M := \sup_{t \geq T} z_t(\omega) < \infty.$$

Since $z_t(\omega) \to 2\epsilon$ as $t \to \infty$, it is clear that $M \geq 2\epsilon$. Next, since $\eta(\cdot)$ belongs to Class $\mathcal{B}$, it follows that

$$c := \inf_{\epsilon \leq r \leq M} \eta(r) > 0.$$

So, for $t \geq T$, we have that

$$h_t(\omega) \geq \eta(z_t(\omega)) \geq c.$$

Now, if we discard all terms for $t < T$, we get

$$\sum_{t=T}^{\infty} \alpha_t(\omega) h_t(\omega) < \infty, \ \forall \omega \in \Omega_0, \sum_{t=T}^{\infty} \alpha_t(\omega) = \infty, h_t(\omega) \geq c > 0,$$

which is clearly a contradiction. Therefore the set of $\omega \in \Omega_0 \cap \Omega_1$ for which $W(\omega) > 0$ has zero measure within $\Omega_0 \cap \Omega_1$. In other words, $z_t(\omega) \to 0$ for (almost) all $\omega \in \Omega_0 \cap \Omega_1$. This is Item 3. □

Theorem 2.23 above shows only that $z_t$ converges to 0 almost surely on sample paths in $\Omega_0 \cap \Omega_1$. In these notes, we are interested not only in the convergence of various algorithms, but also on the *rate* of convergence. With this in mind, we now state and prove an extension of Theorem 2.23 that provides such an estimate on rates.

But before that, we need to define what "rate of convergence" means for a stochastic process converges almost surely. Unlike convergence in the mean and convergence in probability, which readily lend themselves to the concept of "rate," the concept of the rate is somewhat tricky in the case of almost-sure convergence. Suppose $\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*$ in the quadratic mean. Then we can study $E[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2^2, P]$, and say that $\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*$ at the rate $\lambda$ if $E[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2^2, P] = O(t^{-\lambda})$. Prior to the contents of Chapters 3 and 4 were discovered in recent years, the above notion of the rate was the most widely-studied form. Similarly, if $\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*$ in probability, we can define the quantity

$$q(t, \epsilon) := \Pr\{\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2 > \epsilon\}.$$

Convergence in probability implies that $q(t, \epsilon) \to 0$ as $t \to \infty$, for each fixed $\epsilon > 0$. One can then study the rate at which this convergence takes place. Note that some authors refer to bounds on $q(t, \epsilon)$ as "high confidence" bounds. However, for the purposes of this book, we use the following definition, which is inspired by [93].

**Definition 2.22.** Suppose $\{Y_t\}$ is a stochastic process, and $\{f_t\}$ is a sequence of positive numbers. We say that

1. $Y_t = O(f_t)$ if $\{Y_t/f_t\}$ is bounded almost surely.
2. $Y_t = \Omega(f_t)$ if $Y_t$ is positive almost surely, and $\{f_t/Y_t\}$ is bounded almost surely.
3. $Y_t = \Theta(f_t)$ if $Y_t$ is both $O(f_t)$ and $\Omega(f_t)$.
4. $Y_t = o(f_t)$ if $Y_t/f_t \to 0$ almost surely as $t \to \infty$.

The next theorem is a modification of Theorem 2.23 that provides bounds on the *rate* of convergence.

**Theorem 2.24.** *Suppose $\{z_t\}, \{f_t\}, \{g_t\}, \{\alpha_t\}$ are stochastic processes defined on some probability space $(\Omega, \mathcal{F}, P)$, taking values in $[0, \infty)$, adapted to some filtration $\{\mathcal{F}_t\}$. Suppose further that*

$$E_t(z_{t+1}) \leq (1 + f_t)z_t + g_t - \alpha_t z_t \; \forall t, \tag{2.3.40}$$

*where*

$$\sum_{t=0}^{\infty} f_t(\omega) < \infty, \sum_{t=0}^{\infty} g_t(\omega) < \infty, \sum_{t=0}^{\infty} \alpha_t(\omega) = \infty.$$

*Then $z_t = o(t^{-\lambda})$ for every $\lambda \in (0, 1]$ such that there exists a finite $T > 0$ such that*

$$\alpha_t(\omega) - \lambda t^{-1} \geq 0 \; \forall t \geq T, \tag{2.3.41}$$

*and in addition*

$$\sum_{T=0}^{\infty} (t+1)^{\lambda} g_t(\omega) < \infty, \sum_{T=0}^{\infty} [\alpha_t(\omega) - \lambda t^{-1}] = \infty, \tag{2.3.42}$$

*where $T$ is defined in* (2.3.41)

The proof makes use of some ideas from [93].

*Proof.* Over the interval $(0, \infty)$, the map $t \mapsto t^{\lambda}$ is concave for $\lambda \in (0, 1)$. It follows from the "graph below the tangent" property of a concave function that

$$(t+1)^{\lambda} \leq t^{\lambda} + \lambda t^{\lambda - 1}. \tag{2.3.43}$$

Now a ready consequence of (2.3.43) is

$$1 \leq \left(\frac{t+1}{t}\right)^{\lambda} \leq 1 + \lambda t^{-1}, \ \forall t \geq T.$$

Now we follow the suggestion of [93, Lemma 1] by recasting (2.3.41) in terms of $t^{\lambda} z_t$.[5] If we multiply both sides of (2.3.41) by $(t+1)^{\lambda}$, and divide by $t^{\lambda}$ where appropriate, we get

$$E_t((t+1)^{\lambda} z_{t+1}) \leq (1+f_t)\left(\frac{t+1}{t}\right)^{\lambda} t^{\lambda} z_t + (t+1)^{\lambda} g_t - \alpha_t \left(\frac{t+1}{t}\right)^{\lambda} t^{\lambda} z_t, \ \forall t \geq T.$$

Now we observe that

$$-\alpha_t \left(\frac{t+1}{t}\right)^{\lambda} \leq -\alpha_t, \ \forall t \geq T,$$

$$(1+f_t)\left(\frac{t+1}{t}\right)^{\lambda} \leq (1+f_t)(1+\lambda t^{-1}) = 1 + f_t(1+\lambda t^{-1}) + \lambda t^{-1}, \ \forall t \geq T.$$

If we now define the modified quantity $\bar{z}_t = t^{\lambda} z_t$, then the above bound can be rewritten as

$$E_t(\bar{z}_{t+1}) \leq [1 + f_t(1+\lambda t^{-1})]\bar{z}_t + (t+1)^{\lambda} g_t - (\alpha_t - \lambda t^{-1})\bar{z}_t, \ \forall t \geq T. \tag{2.3.44}$$

Since $1 + \lambda t^{-1}$ is bounded over $t \geq T$, it is obvious that

$$\sum_{t=T}^{\infty} f_t < \infty \implies \sum_{t=T}^{\infty} f_t(1+\lambda t^{-1}) < \infty.$$

Moreover, by assumption, there exists a finite $T$ such that

$$\alpha_t - \lambda t^{-1} \geq 0, \ \forall t \geq T.$$

Since it is always permissible to analyze the inequality (2.3.41) starting at time $T$, we can apply Theorem 2.22 to (2.3.41), with $\eta(r) = r$, and deduce that $\bar{z}_t \to 0$ as $t \to \infty$. This is equivalent to $z_t = o(t^{-\lambda})$. □

## Notes and References

The topic of the convergence of stochastic processes is vast, and clearly what is presented here is just a tiny sliver of the subject. Our choice of topics is dictated by their applicability to problems of nonconvex optimization and to Reinforcement Learning.

The main references cited for probability and stochastic processes are [10, 44, 173]. For general topics in measure theory and/or real analysis, the reader is directed to [16, 43, 127].

Theorem 2.22, the Robbins-Siegmund theorem, is an extension of the standard result that a nonnegative supermartingale converges almost surely. For this reason, it is known as the "almost supermartingale" theorem. This theorem represents a refinement of an earlier theorem from [52], but was discovered independently. Theorem 2.23 makes use of the concept of a function of class $\mathcal{B}$. This concept is introduced in [52], but it had no name. The concept is defined precisely, and given a name, in [168]. Theorem 2.23 as presented here is stated in this form in [71], as is Theorem 2.24. The definition of the "rate of convergence," and its application to the problem at hand, is motivated by [93].

[5] Since $t^{-1}$ is undefined when $t = 0$, the bounds below apply when $t \geq 1$.

# Chapter 3

# Stochastic Approximation: Algorithms and Convergence

In this chapter we formulate and analyze several versions of Stochastic Approximation (SA), which is the common thread that binds the two distinct topics studied in this book, namely nonconex optimization, and Reinforcement Learning. We first state the problem under study; then we analyze the solution to the problem using a variety of methods.

## 3.1 An Overview of Stochastic Approximation

In this section, we state the core problem in Stochastic Approximation (SA). Suppose $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$ is some function. For the moment, no assumptions are made about the nature of $f(\cdot)$. Assumptions about $f(\cdot)$ are added as and when they are needed. The objective is to find a solution to the equation $\mathbf{f}(\boldsymbol{\theta}^*) = \mathbf{0}$.[1] The phrase "Stochastic approximation," as well as the first results, were introduced in a seminal paper by Robbins and Monro [123]. SA was introduced as an iterative technique for finding a solution $\boldsymbol{\theta}^*$ when only noisy measurements $\mathbf{f}(\cdot)$ are available. It is not necessary for the function to be "known" (e.g., in closed form). All that is required is that, given an argument $\boldsymbol{\theta} \in \mathbb{R}^d$, an "oracle" gives us a noise-corrupted measurement in the form

$$\mathbf{y}_{t+1} = \mathbf{f}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}, \tag{3.1.1}$$

where $\{\boldsymbol{\xi}_t\}_{t \geq 1}$ is a noise sequence. Assumptions about the nature of the noise sequence are introduced at appropriate places. For the moment, we focus on how the noisy measurements could be used to construct a sequence of approximations $\{\boldsymbol{\theta}_t\}$ that we hope would converge to a $\boldsymbol{\theta}^* \in \mathbb{R}^d$ such that $\mathbf{f}(\boldsymbol{\theta}^*) = \mathbf{0}$.

The standard implementation of the SA algorithm is as follows: One begins with an initial guess $\boldsymbol{\theta}_0$ which could either be deterministic or random. This guess is updated according to the rule

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_t \mathbf{y}_{t+1} = \boldsymbol{\theta}_t + \alpha_t [\mathbf{f}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}], \tag{3.1.2}$$

where $\{\alpha_t\}$ is either a prespecified sequence of real numbers, or a prespecified sequence of random variables. It is customary to assume that $\alpha_t \in (0, \infty)$ for all $t$. If $\alpha_t$ is random, then it is assumed that $\alpha_t > 0$ almost surely. In the optimization literature, $\alpha_t$ is referred to as the "step size," and could vary as a function of $t$, the iteration counter (also referred to as "time"). In the Machine Learning literature, it is common to choose a *fixed* value of $\alpha_t \equiv \alpha$, which is then referred to as the "learning rate."

In their original paper, Robbins and Monro made very restrictive assumptions regarding the nature of the function $\mathbf{f}(\cdot)$ and the noise sequence $\{\boldsymbol{\xi}_{t+1}\}$. These assumptions have been substantially relaxed by later

[1] Obviously, $\mathbf{0}$ can be replaced by any arbitrary element of $\mathbb{R}^d$.

researchers; hence we need not recapitulate the original assumptions. However, one fact that has remained (more or less) unchanged over the decades is the set of sufficient conditions for the algorithm to converge, given in the original paper. These are rightly known as the "Robbins-Monro conditions," and are stated as

$$\sum_{t=0}^{\infty} \alpha_t^2 < \infty, \tag{3.1.3}$$

and

$$\sum_{t=0}^{\infty} \alpha_t = \infty. \tag{3.1.4}$$

Usually both conditions are written together; but there is a reason for displaying them separately in this book. Specifically, in [52], it is shown that (3.1.3) alone is sufficient to ensure that the iterations $\{\boldsymbol{\theta}_t\}$ are bounded almost surely.[2] The addition of (3.1.4) to (3.1.3) then leads to the stronger conclusion that $\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*$ almost surely.

At this point one might ask: Why shouldn't the update formula be

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t \mathbf{y}_{t+1},$$

because any zero of $\mathbf{f}(\cdot)$ is also a zero of $-\mathbf{f}(\cdot)$? As we shall see below, the choice of the plus sign or the minus sign depends on the behavior of the function $\mathbf{f}(\cdot)$. Specifically, some convergence proofs of SA are based on the solution $\boldsymbol{\theta}^*$ being a globally asymptotically stable (GAS) equilibrium of the associated ODE

$$\dot{\boldsymbol{\theta}} = \mathbf{f}(\boldsymbol{\theta}). \tag{3.1.5}$$

Clearly, replacing $\mathbf{f}(\cdot)$ by $-\mathbf{f}(\cdot)$ destroys the GAS property. If the GAS property holds, then the formulation in (3.1.2) is the approrpiate one, as we shall see below.

While the SA algorithm as described above is intended to find a zero of a function, SA can also be used to address some related problems. Two of them are mentioned here, namely: Finding a fixed point, and finding a stationary point.

Suppose $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$ is some function. It is desired to find a *fixed point* of the map $\mathbf{g}$, that is, a vector $\boldsymbol{\theta}^*$ such that $\mathbf{g}(\boldsymbol{\theta}^*) = \boldsymbol{\theta}^*$, when only noisy measurements of $\mathbf{g}(\cdot)$ are available. Thus, at time $t$, given a $\boldsymbol{\theta} \in \mathbb{R}^d$, an oracle returns the noise-corrupted measurement $\mathbf{g}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}$. As shown in Chapter 5, computing the value of a Markov reward problem, or the value of a policy in a Markov Decision Problem (MDP), both fall into this category. This problem can be formulated as that finding a zero of the function $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{g}(\boldsymbol{\theta}) - \boldsymbol{\theta}$. If we were to substitute this expression into (3.1.2), we get what might be called the "fixed point version" of SA, namely

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_t[\mathbf{g}(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t + \boldsymbol{\xi}_{t+1}] = (1 - \alpha_t)\boldsymbol{\theta}_t + \alpha_t[\mathbf{g}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}]. \tag{3.1.6}$$

In this situation, the step size $\alpha_t$ is restricted to lie in $(0, 1)$, as opposed to $(0, \infty)$ as in (3.1.2). An advantage of this formulation is that $\boldsymbol{\theta}_{t+1}$ is a convex combination of the current guess $\boldsymbol{\theta}_t$ and the noisy measurement $\mathbf{g}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}$.

Another application is that of finding a stationary point of a $\mathcal{C}^1$ function $J : \mathbb{R}^d \to \mathbb{R}$, that is, finding a $\boldsymbol{\theta}^* \in \mathbb{R}^d$ such that $\nabla J(\boldsymbol{\theta}^*) = \mathbf{0}$. This application is studied in detail in Chapter 4. Since the objective is to solve $\nabla J(\boldsymbol{\theta}^*) = \mathbf{0}$, we can replace $\mathbf{y}_{t+1}$ in (3.1.1) by a **stochastic gradient** $\mathbf{h}_{t+1}$, which is a noisy approximation to $\nabla J(\boldsymbol{\theta}_t)$. With this definition, the stochastic approximation step becomes

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t \mathbf{h}_{t+1}. \tag{3.1.7}$$

This is a generalization of the familiar gradient descent method of (1.1.10), with the true gradient replaced by the stochastic gradient. For this reason, (3.1.7) is often referred to as **Stochastic Gradient Descent**

[2] This means that almost all sample paths of the stochastic process are bounded, though the bound would depend on the sample path.

**(SGD)**. It might be mentioned that SGD is the workhorse of contemporary optimization, and neural network training.

At this point it might appear that we are using Stochastic Approximation to address three distinct problems, namely: Finding a zero of a function, finding a fixed point of a map, and finding a stationary point of a function. In reality, these three problems are closely related. We have already seen that finding a fixed point of a map $\mathbf{g}$ is equivalent to finding a zero of a function $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{g}(\boldsymbol{\theta}) - \boldsymbol{\theta}$. Now it is shown that, under suitable conditions, finding a stationary point of a $\mathcal{C}^2$ map $J(\cdot)$ is equivalent to finding the fixed point of an associated contraction map.

Specifically, suppose $J(\cdot)$ is $\mathcal{C}^2$, and suppose further that there exist constants $0 < a \leq b < \infty$ such that

$$aI_d \leq \nabla^2 J(\boldsymbol{\theta}) \leq bI_d, \ \forall \boldsymbol{\theta} \in \mathbb{R}^d, \tag{3.1.8}$$

where for symmetric matrices $A$ and $B$, the notation $A \leq B$ denotes that $B - A$ is positive semidefinite. Note that any function $J(\cdot)$ satisfying (3.1.8) is strictly convex, though the converse is not always true. Now define

$$r = \frac{b+a}{2}, \rho = \frac{b-a}{b+a}, \tag{3.1.9}$$

and note that $\rho < 1$. With these definitions, it is now shown that the map $\boldsymbol{\theta} \mapsto \boldsymbol{\theta} - (1/r)\nabla J(\boldsymbol{\theta})$ is a contraction, with constant $\rho$. To see this, observe that a ready consequence of (3.1.8) is

$$\left(1 - \frac{b}{r}\right) I_d \leq I_d - (1/r)\nabla^2 J(\boldsymbol{\theta}) \leq \left(1 - \frac{a}{r}\right) I_d.$$

However, it is easy to verify that

$$1 - \frac{b}{r} = -\rho, 1 - \frac{a}{r} = \rho.$$

In turn this implies that

$$\|I_d - (1/r)\nabla^2 J(\boldsymbol{\theta})\|_S \leq \rho, \tag{3.1.10}$$

where $\|M\|_S$ denotes the largest singular value of $M$, which is also the the $\ell_2$-induced matrix norm of the matrix $M$, that is

$$\|M\|_S = \max_{\|\mathbf{v}\|_2 \leq 1} \|M\mathbf{v}\|_2.$$

Next, observe that

$$\nabla J(\boldsymbol{\theta}) = \nabla J(\boldsymbol{\phi}) + \int_0^1 \nabla^2 J(\boldsymbol{\phi} + \lambda(\boldsymbol{\theta} - \boldsymbol{\phi}))(\boldsymbol{\theta} - \boldsymbol{\phi}) \, d\lambda.$$

Hence

$$(\boldsymbol{\theta} - (1/r)\nabla J(\boldsymbol{\theta})) - [\boldsymbol{\phi} - (1/r)\nabla J(\boldsymbol{\phi})] = \int_0^1 (I_d - (1/r)\nabla^2 J(\boldsymbol{\phi} + \lambda(\boldsymbol{\theta} - \boldsymbol{\phi})))(\boldsymbol{\theta} - \boldsymbol{\phi}) \, d\lambda.$$

Now we can invoke (3.1.10) to conclude that

$$\left\| \int_0^1 (I_d - (1/r)\nabla^2 J(\boldsymbol{\phi} + \lambda(\boldsymbol{\theta} - \boldsymbol{\phi})))(\boldsymbol{\theta} - \boldsymbol{\phi}) \, d\lambda \right\|_2 \leq \rho \|\boldsymbol{\theta} - \boldsymbol{\phi}\|_2.$$

In turn this implies that

$$\|(\boldsymbol{\theta} - (1/r)\nabla J(\boldsymbol{\theta})) - [\boldsymbol{\phi} - ((1/r)\nabla J(\boldsymbol{\phi}))]\|_2 \leq \rho \|\boldsymbol{\theta} - \boldsymbol{\phi}\|_2.$$

This is the desired conclusion.

Now observe that solving $\nabla J(\boldsymbol{\theta}) = \mathbf{0}$ is the same as solving $(1/r)\nabla J(\boldsymbol{\theta}) = \mathbf{0}$. Clearly

$$(1/r)\nabla J(\boldsymbol{\theta}) = \mathbf{0} \iff \boldsymbol{\theta} = \boldsymbol{\theta} - (1/r)\nabla J(\boldsymbol{\theta}).$$

From the above discussion, we know that the map on the right side is a contraction. Hence finding a stationary point of a $J(\cdot)$ that satisfies (3.1.8) is equivalent to solving the above fixed-point problem. To complete the discussion, let us see what the iteration looks like, for finding a fixed point of the map $\boldsymbol{\theta} \mapsto \boldsymbol{\theta} - (1/r)\nabla J(\boldsymbol{\theta})$. We start with an initial guess $\boldsymbol{\theta}_0$, and then update it via

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - (1/r)\nabla J(\boldsymbol{\theta}_t).$$

This is called "fixed step size" gradient descent. Note that, in order to compute $r$, it is necessary only to have a *lower bound* for $a$, and an *upper bound* for $b$, in (3.1.8).

After the SA algorithm was introduced in [123], some generalizations and/or simplifications followed very quickly; see [174, 77, 45, 39]. An excellent survey can be found in [87]. Book-length treatments of SA can be found in [85, 9, 86, 23].

In (3.1.2), it is noteworthy that *every component* of $\boldsymbol{\theta}_t$ is updated at time $t$. For this reason, in this book we refer to the approach in (3.1.2) as **Synchronous Stochastic Approximation (SSA)**, though the terminology is not very standard. SSA can be contrasted with a situation whereby, at each step $t$, *only one* component of $\boldsymbol{\theta}_t$ is updated. This is known as **Asynchronous Stochastic Approximation( ASA)**, a term that was introduced in [159, 158, 22], and is by now standard terminology. An intermediate approach is to update, at each step $t$, *some but not necessarily all* components of $\boldsymbol{\theta}_t$. In this book, this is referred to a **Block Asynchronous Stochastic Approximation (BASA)**. Again, this terminology is not standard. We derive sufficient conditions for the convergence of SSA in Section 3.2, and for BASA in Section 3.3. ASA need not be studiede separately as it is a special case of BASA.

## 3.2 Convergence of Synchronous Stochastic Approximation

In this section, we study the convergence of the "synchronous" Stochastic Approximation (SA) algorithm (3.1.1) under a variety of conditions. Variants of the standard SA algorithm are studied in Section 3.3 and 3.4.

Some of the theorems in this section only establish the convergence of the SA algorithm, whereas other theorems also establish the *rate* of convergence. All of these theorems make use of the theorems proved in Section 2.3.

This section is organized as follows: Some theorems also make use of Lyapunov stability theory, introduced in Section 7.2, as well as a new result in "converse" Lyapunov theory, which is presented later in this section.

Note that the assumptions on the function $\mathbf{f}(\cdot)$ and on the noise sequence $\{\boldsymbol{\xi}_{t+1}\}$ are far more general than those in the original Robbins-Monro paper. Also, the conclusions are stronger. For example, the original paper studies only the scalar case ($d = 1$). Moreover, the convergence of the iterations to the desired limit is only *in the quadratic mean*, and hence *in probability*. The current "best practice" is to strive to prove *almost sure* convergence, which is stronger. The reason for desiring almost sure convergence is obvious: The application of any stochastic algorithm results in *a single sample path* of a stochastic process. It is therefore worthwhile to know that *almost all sample paths* reach the correct answer.

### 3.2.1 Convergence Theorems for SA via Lyapunov Theory

Now let us return to the problem at hand, namely to establish the convergence of the SA algorithm, aims to find a zero of a $\mathcal{C}^1$ function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$. One begins with a (possibly random) initial guess $\boldsymbol{\theta}_0$, after which the update rule is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_t[\mathbf{f}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}], \tag{3.2.1}$$

where $\alpha_t$ is a nonnegative-valued and possibly random step size, and $\boldsymbol{\xi}_{t+1}$ is the measurement error. We begin with the assumptions on the function $\mathbf{f}(\cdot)$ in (3.2.1).

(F1) The equation $\mathbf{f}(\boldsymbol{\theta}^*) = \mathbf{0}$ has a unique solution, which is assumed to be $\boldsymbol{\theta}^* = \mathbf{0}$, by shifting coordinates if necessary.

(F2) There is a constant $S$ such that

$$\|\mathbf{f}(\boldsymbol{\theta})\|_2 \leq S\|\boldsymbol{\theta}\|_2, \ \forall \boldsymbol{\theta} \in \mathbb{R}^d.$$

Note that (F2) is weaker than assuming that $\mathbf{f}(\cdot)$ is globally Lipschitz-continuous with constant $S$, which would be

$$\|\mathbf{f}(\boldsymbol{\theta}) - \mathbf{f}(\boldsymbol{\phi})\|_2 \leq S\|\boldsymbol{\theta} - \boldsymbol{\phi}\|_2, \ \forall \boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^d.$$

In effect, (F2) is the above relation with $\boldsymbol{\phi}$ set equal to $\mathbf{0}$ (or $\boldsymbol{\theta}^*$ in the general case). In particular, the function $\mathbf{f}(\cdot)$ need not even be continuous in order to satisy (F2). For example, the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(\theta) = \begin{cases} 3\theta, & \theta \in [0, 1), \\ 1 + \exp(\theta - 1), & \theta \in [1, \infty), \\ -f(-\theta), & \theta < 0. \end{cases}$$

Then $f(\cdot)$ is discontinuous at $\theta = \pm 1$, but still satisfies (F2).

Next we state the assumptions on the measurement error $\boldsymbol{\xi}_{t+1}$. Let us define

$$\mathbf{z}_t = E_t(\boldsymbol{\xi}_{t+1}), \quad \boldsymbol{\zeta}_{t+1} = \boldsymbol{\xi}_{t+1} - \mathbf{z}_t. \tag{3.2.2}$$

Then it follows from Theorem 2.5 that

$$E_t(\boldsymbol{\zeta}_{t+1}) = \mathbf{0}, \quad CV_t(\boldsymbol{\xi}_{t+1}) = CV_t(\boldsymbol{\zeta}_{t+1}), \quad E_t(\|\boldsymbol{\xi}_{t+1}\|_2^2) = \|\mathbf{z}_t\|_2^2 + CV_t(\boldsymbol{\zeta}_{t+1}). \tag{3.2.3}$$

One can think of $\mathbf{z}_t$ as the the predictable part of the measurement error, and $\boldsymbol{\zeta}_{t+1}$ as the unpredictable part. So if $\mathbf{z}_t = \mathbf{0}$ for all $t$, then the noisy measurement $\mathbf{y}_{t+1} = \mathbf{f}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}$ can be said to be "unbiased," because $E_t(\mathbf{y}_{t+1}) = \mathbf{f}(\boldsymbol{\theta}_t)$. However, any convergence theory cannot be restricted to this situation. As we shall see later, there are so-called "zeroth-order" or "derivative-free" methods for implementing SA, in which case the bias $\mathbf{z}_t$ need not equal zero. Our theory needs to be versatile enough to cater to this situation as well.

With this notation, we state the assumptions on the measurement error $\boldsymbol{\xi}_{t+1}$.

(N1) There exists a sequence of constants $\{B_t\}$ such that

$$\|E_t(\boldsymbol{\xi}_{t+1})\|_2 = \|\mathbf{z}_t\|_2 \leq B_t(1 + \|\boldsymbol{\theta}_t\|_2), \ \forall t \geq 0, \tag{3.2.4}$$

(N2) There exists a sequence of constants , $\{M_t\}$ such that

$$CV_t(\boldsymbol{\zeta}_{t+1}) = E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2) \leq M_t^2(1 + \|\boldsymbol{\theta}_t\|_2^2), \ \forall t \geq 0. \tag{3.2.5}$$

We begin our study with a bound that is very useful in its own right. The bound is taken from [12, Eq. (2.4)].

**Theorem 3.1.** *Suppose $J : \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{C}^1$, and suppose further that $\nabla J(\cdot)$ is L-Lipschitz continuous, that is,*

$$\|\nabla J(\boldsymbol{\theta}) - \nabla J(\boldsymbol{\phi})\|_2 \leq L\|\boldsymbol{\theta} - \boldsymbol{\phi}\|_2, \ \forall \boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^d. \tag{3.2.6}$$

*Then*

$$J(\boldsymbol{\theta} + \boldsymbol{\phi}) \leq J(\boldsymbol{\theta}) + \langle \nabla J(\boldsymbol{\theta}), \boldsymbol{\phi} \rangle + \frac{L}{2}\|\boldsymbol{\phi}\|_2^2. \tag{3.2.7}$$

*Proof.* Define $h : \mathbb{R} \to \mathbb{R}$ via

$$h(\lambda) := h(\boldsymbol{\theta} + \lambda\boldsymbol{\phi}).$$

Then $h(\cdot)$ is in $\mathcal{C}^1$, and

$$h(0) = J(\boldsymbol{\theta}), \quad h(1) = J(\boldsymbol{\phi}), \quad h'(\lambda) = \langle \nabla J(\boldsymbol{\theta} + \lambda\boldsymbol{\phi}), \boldsymbol{\phi} \rangle, \ \forall \lambda \in \mathbb{R}.$$

Now observe that $h(0) = J(\boldsymbol{\theta})$ and $h(1) = J(\boldsymbol{\theta} + \boldsymbol{\phi})$. Therefore

$$\begin{aligned} h(1) &= h(0) + \int_0^1 h'(\lambda)\, d\lambda \\ &= h(0) + \int_0^1 \langle \nabla J(\boldsymbol{\theta} + \lambda\boldsymbol{\phi}), \boldsymbol{\phi} \rangle\, d\lambda \\ &= h(0) + \int_0^1 \langle \nabla J(\boldsymbol{\theta}, \boldsymbol{\phi} \rangle\, d\lambda + \int_0^1 \langle [\nabla J(\boldsymbol{\theta} + \lambda\boldsymbol{\phi}) - \nabla J(\boldsymbol{\theta})], \boldsymbol{\phi} \rangle\, d\lambda. \end{aligned} \tag{3.2.8}$$

By the Liptschitz continuity of $\nabla J(\cdot)$, it follows that

$$\|\nabla J(\boldsymbol{\theta} + \lambda\boldsymbol{\phi}) - \nabla J(\boldsymbol{\theta})\|_2 \leq L\|\boldsymbol{\phi}\|_2.$$

Further, by Scwharz' inequality,

$$\int_0^1 \langle [\nabla J(\boldsymbol{\theta} + \lambda\boldsymbol{\phi}) - \nabla J(\boldsymbol{\theta})], \boldsymbol{\phi} \rangle\, d\lambda \leq L\|\boldsymbol{\phi}\|_2^2 \int_0^1 \lambda\, d\lambda = \frac{L}{2}\|\boldsymbol{\phi}\|_2^2.$$

Substituting this into (3.2.8) gives

$$h(1) \leq h(0) + \langle \nabla J(\boldsymbol{\theta}), \boldsymbol{\phi} \rangle + \frac{L}{2}\|\boldsymbol{\phi}\|_2^2.$$

This is the same as (3.2.7). □

**Remarks:**

1. The inequality (3.2.7) is well-known in convex analysis. For example the right inequality in [109, Eq. 2.1.9] becomes (3.2.7) after changing $f$ to $J$, $x$ to $\boldsymbol{\theta}$ and $y$ to $\boldsymbol{\theta} + \boldsymbol{\phi}$. Thus Theorem 3.1 does away with the assumption that the function $J(\cdot)$ is convex, which is a *huge* improvement. However, it is important to note that [109, Eq. 2.1.9] has *two parts*. The *left* inequality implies that the function $J(\cdot)$ is convex, as shown there. In the present case, there is no analog of the left inequality.

2. The theorem follows readily from Taylor's theorem if $J(\cdot)$ is $\mathcal{C}^2$ and not just $\mathcal{C}^1$. This can be seen as follows: The assumption that $\nabla J(\cdot)$ is $L$-Lipschitz-continuous implies that $\nabla J(\cdot)$ is absolutely continuous, in the sense defined in [127]; see in particular the Remark at the bottom of page of 122. By the contents of [127, Section 6.4], it follows that $\nabla J(\cdot)$ is differentiable almost everywhere (i.e., everywhere except on a set of Lebesgue measure zero). Moreover, wherever $\nabla J(\cdot)$ is differentiable, it follows readily that $\|\nabla^2 J(\cdot)\|_S \leq L$. Here $\|A\|_S$ denotes the largest singular value of a matrix $A$. Note that $\|A\|_S$ also equals $\|A\|_{2\to 2}$, which is the matrix norm induced by the $\ell_2$-norm on vectors. Hence if were to strengthen the hypothesis of Theorem 3.1 to: $J \in \mathcal{C}^2$ (instead of $J \in \mathcal{C}^1$), and $\nabla J(\cdot)$ is $L$-Lipschitz continuous, then it would follow that $\|\nabla^2 J(\boldsymbol{\theta})\|_S \leq L$ for all $\boldsymbol{\theta}$. In such a case, Taylor's theorem would imply that, for each $\boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^d$, there exists a $\lambda \in (0, 1)$ such that

$$J(\boldsymbol{\theta} + \boldsymbol{\phi}) = J(\boldsymbol{\theta}) + \langle \nabla J(\boldsymbol{\theta}), \boldsymbol{\phi} \rangle + \frac{1}{2}\boldsymbol{\phi}^\top \nabla^2 J(\boldsymbol{\theta} + \lambda\boldsymbol{\phi})\boldsymbol{\phi}.$$

   This would in turn imply (3.2.7). a the contribution of [12] is to weaken the hypothesis from $J \in \mathcal{C}^2$ to $J \in \mathcal{C}^1$.

Now we present a sufficient condition for the convergence of the SA algorithm of (3.2.1), which involves the existence of a "Lyapunov function" $V : \mathbb{R}^d \to \mathbb{R}$ that satisfies some conditions. The concept of a Lyapunov function is introduced in Section 7.2 in the context of the stability of ODEs. In particular, let us associate an ODE

$$\dot{\boldsymbol{\theta}} = \mathbf{f}(\boldsymbol{\theta})$$

with the function $\mathbf{f}(\cdot)$ whose zero we are trying to find. Now suppose $V : \mathbb{R}^d \to \mathbb{R}_+$ is an $\mathcal{C}^1$ function with gradient $\nabla V$. Then the function $\dot{V} : \mathbb{R}^d \to \mathbb{R}$ associated with $V$ and the ODE above is defined by (cf. (7.2.9)):

$$\dot{V}(\boldsymbol{\theta}) := \langle \nabla V(\boldsymbol{\theta}), \mathbf{f}(\boldsymbol{\theta}) \rangle.$$

Now we state the standard assumptions on the Lyapunov function.

(L1) $\nabla V$ is $\mathcal{C}^1$ and $L$-Lipschitz continuous, and $\nabla V(\mathbf{0}) = 0$.

(L2) There exist positive constants $a, b$ such that

$$a\|\boldsymbol{\theta}\|_2^2 \leq V(\boldsymbol{\theta}) \leq b\|\boldsymbol{\theta}\|_2^2, \ \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{3.2.9}$$

To avoid a lot of repetition, we state a **standing assumption**:

(S) Assumptions (F1), (F2), (N1), (N2), (L1), (L2) hold.

Now we state our results on the convergence of the SA algorithm of (3.2.1). The first theorem gives sufficient conditions for the almost sure convergence of $\boldsymbol{\theta}_t$ to $\mathbf{0}$, but does not give any information on the *rate* of convergence. By strengthening the assumptions on $\dot{V}(\cdot)$, we derive bounds on the rate of convergence in the next theorem and its corollary.

**Theorem 3.2.** *Suppose that Assumptions (S) hold.*

1. *Suppose that $\dot{V}(\boldsymbol{\theta}) \leq 0$ for all $\boldsymbol{\theta}$, and that*

$$\sum_{t=0}^{\infty} \alpha_t^2 < \infty, \quad \sum_{t=0}^{\infty} \alpha_t B_t < \infty, \quad \sum_{t=0}^{\infty} \alpha_t^2 M_t^2 < \infty, \tag{3.2.10}$$

   *Then $\{V(\boldsymbol{\theta}_t)\}$ and $\{\|\boldsymbol{\theta}_t\|_2\}$ are bounded, and in addition, $V(\boldsymbol{\theta}_t)$ converges to some random variable as $t \to \infty$.*

2. *Suppose that, in addition to* (3.2.10)*, it is also the case that*

$$\sum_{t=0}^{\infty} \alpha_t = \infty, \tag{3.2.11}$$

   *and in addition, there exists a function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ belonging to Class $\mathcal{B}$ such that*

$$\dot{V}(\boldsymbol{\theta}) \leq -\psi(\|\boldsymbol{\theta}\|_2^2), \ \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{3.2.12}$$

   *Then $V(\boldsymbol{\theta}_t) \to 0$ and $\boldsymbol{\theta}_t \to \mathbf{0}$ as $t \to \infty$.*

*Proof.* Applying Theorem 3.1 to the function $V(\cdot)$, and making use of the updating formula (3.2.1) leads to

$$V(\boldsymbol{\theta}_{t+1}) \leq V(\boldsymbol{\theta}_t) + \alpha_t \langle \nabla V(\boldsymbol{\theta}_t), \mathbf{f}(\boldsymbol{\theta}_t) \rangle + \alpha_t \langle \nabla V(\boldsymbol{\theta}_t), \boldsymbol{\xi}_{t+1} \rangle + \alpha_t^2 \frac{L}{2} \|\mathbf{f}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}\|_2^2.$$

Applying $E_t(\cdot)$ to both sides, using (3.2.2) and (3.2.3), and applying the definition of $\dot{V}(\cdot)$, gives

$$\begin{aligned} E_t(V(\boldsymbol{\theta}_{t+1})) &\leq V(\boldsymbol{\theta}_t) + \alpha_t \dot{V}(\boldsymbol{\theta}_t) + \alpha_t \langle \nabla V(\boldsymbol{\theta}_t), \mathbf{z}_t \rangle \\ &+ \alpha_t^2 \frac{L}{2} [\|\mathbf{f}(\boldsymbol{\theta}_t)\|_2^2 + 2\langle \mathbf{f}(\boldsymbol{\theta}_t), \mathbf{z}_t \rangle + \|\mathbf{z}_t\|_2^2 + E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2)]. \end{aligned} \tag{3.2.13}$$

Now we observe that

$$\|\mathbf{f}(\boldsymbol{\theta}_t)\|_2 \leq S\|\boldsymbol{\theta}_t\|_2, \quad \|\nabla V(\boldsymbol{\theta}_t)\|_2 \leq L\|\boldsymbol{\theta}_t\|_2.$$

Substituting these bounds into (3.2.13), and invoking the assumptions (N1) and (N2), gives a bound in the form

$$\begin{aligned} E_t(V(\boldsymbol{\theta}_{t+1})) &\leq V(\boldsymbol{\theta}_t) + \alpha_t B_t L \|\boldsymbol{\theta}\|_2 (1 + \|\boldsymbol{\theta}\|_2) + \alpha_t^2 \frac{L}{2} [S^2 \|\boldsymbol{\theta}\|_2^2 + 2SB_t \|\boldsymbol{\theta}\|_2 (1 + \|\boldsymbol{\theta}\|_2)] \\ &+ \alpha_t^2 \frac{L}{2} [B_t^2 (1 + \|\boldsymbol{\theta}\|_2^2) + M_t^2 (1 + \|\boldsymbol{\theta}\|_2^2)] + \alpha_t \dot{V}(\boldsymbol{\theta}_t). \end{aligned}$$

We proceed to simplify the above inequality in stages. The first step is to incorporate the bound

$$\|\boldsymbol{\theta}\|_2 \leq \frac{1 + \|\boldsymbol{\theta}\|_2^2}{2}.$$

This gives

$$\begin{aligned} E_t(V(\boldsymbol{\theta}_{t+1})) &\leq V(\boldsymbol{\theta}_t) + \alpha_t B_t L (0.5 + 1.5 \|\boldsymbol{\theta}\|_2^2) + \alpha_t^2 \frac{L}{2} [S^2 \|\boldsymbol{\theta}\|_2^2 + SB_t (1 + 3\|\boldsymbol{\theta}\|_2^2)] \\ &+ \alpha_t^2 \frac{L}{2} [B_t^2 (1 + \|\boldsymbol{\theta}\|_2^2) + M_t^2 (1 + \|\boldsymbol{\theta}\|_2^2)] + \alpha_t \dot{V}(\boldsymbol{\theta}_t). \end{aligned}$$

The last step is to bound $\|\boldsymbol{\theta}\|_2^2$ by $(1/a)V(\boldsymbol{\theta}_t)$. Ths leads to the final form of the bound

$$E_t(V(\boldsymbol{\theta}_{t+1})) \leq (1 + f_t) V(\boldsymbol{\theta}_t) + g_t + \alpha_t \dot{V}(\boldsymbol{\theta}_t). \tag{3.2.14}$$

where

$$\begin{aligned} f_t &= \frac{1.5}{a} \alpha_t B_t L + \alpha_t^2 \frac{L}{2a} [S^2 + 3SB_t + B_t^2 + M_t^2], \\ g_t &= \frac{0.5}{a} \alpha_t B_t L + \alpha_t^2 \frac{L}{2a} [SB_t + B_t^2 + M_t^2]. \end{aligned}$$

This bound is in the form to which we can apply Theorem 2.23, if it can be established that the two sequences $\{f_t\}$ (not to be confused with $\mathbf{f}$) and $\{g_t\}$ are summable. Leaving aside various constant terms, both $f_t$ and $g_t$ involve these five terms:

$$\alpha_t^2,\ \alpha_t B_t,\ \alpha_t^2 B_t,\ \alpha_t^2 B_t^2,\ \alpha_t^2 M_t^2.$$

From (3.2.10), we know that the sequences $\{\alpha_t^2\}$, $\{\alpha_t B_t\}$, and $\{\alpha_t^2 M_t^2\}$ are all summable. Now, any summable sequence is also square-summable. Hence $\{\alpha_t^2 B_t^2\}$ is summable. Finally, since $\{\alpha_t^2\}$ is summable, it is evident that $\alpha_t$ is bounded. This, coupled with the summability of $\{\alpha_t B_t\}$, shows that $\{\alpha_t^2 B_t\}$ is also summable. Thus, if $\dot{V}(\boldsymbol{\theta}_t) \leq 0$ for all $\boldsymbol{\theta}_t$, then Item 1 of Theorem 2.23 applies, and Item 1 of the conclusions follows.

Now we come to Item 2 of the conclusions. For this purpose, define $\eta : \mathbb{R}_+ \to \mathbb{R}_+$ via

$$\eta(r) := \inf_{(r/b) \leq x \leq (r/a)} \psi(x).$$

Now (3.2.9) implies that

$$(1/b) V(\boldsymbol{\theta}_t) \leq \|\boldsymbol{\theta}\|_2 \leq (1/a) V(\boldsymbol{\theta}_t).$$

Therefore it is immediate that

$$-\alpha_t \psi(\|\boldsymbol{\theta}\|_2) \leq -\alpha_t \eta(V(\boldsymbol{\theta}_t)).$$

Moreover, since $\psi(\cdot)$ is a function of Class $\mathcal{B}$, so is $\eta(\cdot)$, as is easy to verify. Hence, in (3.2.14), we can replace the term $+\alpha_t \dot{V}(\boldsymbol{\theta}_t)$ by $-\alpha_t \eta(V(\boldsymbol{\theta}_t))$, and apply Item 2 of Theorem 2.23. This leads to Item 2 of the conclusions. □

**Theorem 3.3.** *Suppose that Assumptions (S), and* (3.2.10) *and* (3.2.11) *hold. Suppose further that there exists a constant $c > 0$ such that*

$$\dot{V}(\boldsymbol{\theta}) \leq -c \|\boldsymbol{\theta}\|_2^2,\ \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{3.2.15}$$

*Further, suppose there exist constants $\gamma > 0$ and $\delta \geq 0$ such that*

$$B_t = O(t^{-\gamma}), \quad M_t = O(t^{\delta}),$$

*where we take $\gamma = 1$ if $B_t = 0$ for all sufficiently large $t$, and $\delta = 0$ if $M_t$ is bounded. Choose the step-size sequence $\{\alpha_t\}$ as $O(t^{-(1-\phi)})$ and $\Omega(t^{-(1-C)})$ where $\phi$ is chosen to satisfy*

$$0 < \phi < \min\{0.5 - \delta, \gamma\}, \tag{3.2.16}$$

*and $C \in (0, \phi]$. Define*

$$\nu := \min\{1 - 2(\phi + \delta), \gamma - \phi\}. \tag{3.2.17}$$

*Then $\|\boldsymbol{\theta}_t\|_2^2 = o(t^{-\lambda})$ for every $\lambda \in (0, \nu)$. In particular, by choosing $\phi$ very small, it follows that $\|\boldsymbol{\theta}_t\|_2^2 = o(t^{-\lambda})$*

$$\lambda < \min\{1 - 2\delta, \gamma\}. \tag{3.2.18}$$

*Proof.* Since most of the hard work is already done in proving Theorem 3.2, the proof is only sketched. Since (3.2.15) is assumed to hold, in (3.2.14), the term $-\alpha_t \dot{V}(\boldsymbol{\theta}_t)$ can be replaced by $-\alpha_t c \|\boldsymbol{\theta}\|_2^2$, and then by $-\alpha_t (c/b) V(\boldsymbol{\theta}_t)$. Now one can apply Theorem 2.24 to obtain the result. □

**Corollary 3.1.** *Suppose that Assumptions (S), and* (3.2.10) *and* (3.2.11) *hold. Suppose further that there exists a constant $c > 0$ such that* (3.2.15) *holds. Finally, suppose $\mathbf{z}_t = \mathbf{0}$ almost surely, and there exists a finite constant $M$ such that*

$$CV_t \leq M^2 (1 + \|\boldsymbol{\theta}_t\|_2^2), \ \forall t. \tag{3.2.19}$$

*Then, by choosing $\phi = O(t^{-(1-\epsilon)})$ with $\epsilon > 0$ arbitrarily small, we can ensure that $V(\boldsymbol{\theta}_t), \|\boldsymbol{\theta}_t\|_2^2$ are $o(t^{-\lambda})$ for all $\lambda < 1$.*

The proof is omitted as it is easy.

### 3.2.2 Some Applications

In this subsection, we apply Theorems 3.2 and 3.3 to establish the convergencd of the SA algoritm in two specific problems, namely:

- When $\mathbf{f}(\cdot)$ is "passive," and
- When it is desired to compute the fixed point of a contraction $\mathbf{g}(\cdot)$.

In each case, it is assumed that only noisy measurements are available.

The first application, namely to find a solution of $\mathbf{f}(\boldsymbol{\theta}^*) = \mathbf{0}$ *even when* $\mathbf{f}(\cdot)$ is not necessarily continuous (except at $\mathbf{0}$), was one of the motivations in the seminal paper [52]. Note that, if $\mathbf{f}(\cdot)$ is discontinuous (except at $\mathbf{0}$), the ODE approach requires some modifications, because the ODE $\dot{\boldsymbol{\theta}} = \mathbf{f}(\boldsymbol{\theta})$ has solutions only in the Fillippov sense in general. The martingale approach pursued here does *not* become more complex when $\mathbf{f}(\cdot)$ is discontinuous.

The second application is to find a fixed point of a contractive map $\mathbf{g}(\cdot)$, when only noisy measurements of the function are available. In this situation, the "natural" iteration

$$\boldsymbol{\theta}_{t+1} = \mathbf{g}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1},$$

where $\boldsymbol{\xi}_{t+1}$ is the measurement error, does not work in general.

There is yet another application, which is to find a stationary point of a $\mathcal{C}^1$ map $J : \mathbb{R}^d \to \mathbb{R}$. In other words, it is desired to find a solution to $\nabla J(\boldsymbol{\theta}^*) = \mathbf{0}$. In principle this is the same as the first application, with $\mathbf{f}(\cdot) = \nabla J(\cdot)$. However, there are some special wrinkles. Hence this problem is studied separately in Chapter 4.

First, we show that the SA algorithm of (3.1.2) converges when the function $\mathbf{f}(\cdot)$ is "passive," which is made precise in Definition 3.1 below. This approach was pioneered in [52]. The reader is cautioned that in [52], the SA algorithm uses a minus sign in front of $\alpha_t$, that is, it uses the formulation (3.1.5), and the definition of passivity is adjusted commensuretly.

**Definition 3.1.** Suppose $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$, and that $\boldsymbol{\theta}^* \in \mathbb{R}^d$. Then $\mathbf{f}$ is said to be **passive at** $\boldsymbol{\theta}^*$ if (i) $\mathbf{f}(\boldsymbol{\theta}^*) = \mathbf{0}$, and (ii) for all $0 < \epsilon < M < \infty$, we have that

$$\inf_{\epsilon \le \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \le M} -\langle \boldsymbol{\theta} - \boldsymbol{\theta}^*, \mathbf{f}(\boldsymbol{\theta}) \rangle > 0, \ \forall \epsilon > 0. \tag{3.2.20}$$

**Remark:** If we define a function $\eta(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ by

$$\eta(r) := \inf_{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 = r} -\langle \boldsymbol{\theta} - \boldsymbol{\theta}^*, \mathbf{f}(\boldsymbol{\theta}) \rangle, \tag{3.2.21}$$

then $\mathbf{f}(\cdot)$ is passive at $\boldsymbol{\theta}^*$ if and only if $\eta(\cdot)$ is a function of Class $\mathcal{B}$. Note that if $\mathbf{f}(\cdot)$ is continuous, then (3.2.20) can be replaced by

$$\langle \boldsymbol{\theta} - \boldsymbol{\theta}^*, \mathbf{f}(\boldsymbol{\theta}) \rangle < 0 \ \forall \boldsymbol{\theta} \neq \boldsymbol{\theta}^*. \tag{3.2.22}$$

In circuit theory, a nonlinear characteristic that satisfies (3.2.20) with $\boldsymbol{\theta}^* = \mathbf{0}$ would be called "passive," so we borrow that terminology. Also, (3.2.20) *does not rule out* the possibility that $\mathbf{f}(\boldsymbol{\theta}) \to \mathbf{0}$ as $\|\boldsymbol{\theta}\| \to \infty$. For instance, the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$-f(\theta) = \begin{cases} \theta, & \text{if } \theta \in [0, 1], \\ 1 - \exp(\theta - 1), & \text{if } \theta \ge 1, \\ -f(-\theta), & \text{if } \theta < 0. \end{cases}$$

Then $f(\cdot)$ is passive at zero, even though $f(\theta) \to 0$ when $|\theta| \to \infty$.

Equation (3.2.20) implies that $\langle \boldsymbol{\theta} - \boldsymbol{\theta}^*, \mathbf{f}(\boldsymbol{\theta}) \rangle < 0$ for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}^* \in \mathbb{R}^d$. This in turn implies that $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*$ is the *only* solution to $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{0}$. To find $\boldsymbol{\theta}^*$, suppose we have available only noisy measurements of $\mathbf{f}(\cdot)$. Specifically, suppose we can measure

$$\mathbf{y}_{t+1} = \mathbf{f}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1},$$

where the measurement error satisfies assumptions (N1) and (N2). To determine $\boldsymbol{\theta}^*$, we use the SA iterations defined by (3.1.2). Suppose now that the step sizes $\alpha_t$ are positive, and satisfy the standard Robbins-Monro conditions hold, namely:

$$\sum_{t=1}^{\infty} \alpha_t^2 < \infty. \tag{3.2.23}$$

$$\sum_{t=1}^{\infty} \alpha_t = \infty, \tag{3.2.24}$$

It is shown now that $\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*$ almost surely as $t \to \infty$.

**Theorem 3.4.** *Suppose that Assumptions (F2), (N1) and (N2) hold, and in addition,* $\mathbf{f}(\cdot)$ *is passive at* $\boldsymbol{\theta}^*$. *Under these assumptions, we have the following conclusions:*

1. *If* (3.2.10) *holds, then the sequence* $\{\boldsymbol{\theta}_t\}$ *is bounded almost surely.*
2. *If* (3.2.11) *holds in addition to* (3.2.10)*, then* $\boldsymbol{\theta}_t \to \mathbf{0}$ *almost surely as* $t \to \infty$.

*Proof.* The proof follows readily by applying Theorem 3.2 with the Lyapunov function $V(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$. In this case,

$$\dot{V}(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \mathbf{f}(\boldsymbol{\theta} \rangle.$$

Thus

$$\dot{V}(\boldsymbol{\theta}) \le -\eta(\|\boldsymbol{\theta}\|_2),$$

where $\eta(\cdot)$ is defined in (3.2.21). □

Next, we study the use of SA to find a fixed point of a contractive map when only noisy measurements are available. Specifically, suppose $P \in \mathbb{R}^{d\times d}$ is a positive definite matrix, set $M = P^\top P$, and define the vector norm

$$\|\mathbf{v}\|_M := (\mathbf{v}^\top M\mathbf{v})^{1/2} = \|P\mathbf{v}\|_2, \tag{3.2.25}$$

Now suppose $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$ is a map such that (i) there is a unique $\boldsymbol{\theta}^*$ such that $\mathbf{g}(\boldsymbol{\theta}^*) = \boldsymbol{\theta}^*$, and (ii) there exists a constant $\rho < 1$ such that

$$\|\mathbf{g}(\boldsymbol{\theta}) - \boldsymbol{\theta}^*\|_M \le \rho\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_M,\ \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{3.2.26}$$

An easy way to ensure that (3.2.26) holds is to assume that $\mathbf{g}$ is a contraction with respect to $\|\cdot\|_M$, i.e., that

$$\|\mathbf{g}(\boldsymbol{\theta}) - \mathbf{g}(\boldsymbol{\phi})\|_M \le \rho\|\boldsymbol{\theta} - \boldsymbol{\phi}\|_M,\ \forall \boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^d.$$

Observe that (3.2.26) is just the above contraction condition with $\boldsymbol{\phi}$ set to the fixed point $\boldsymbol{\theta}^*$.

To find $\boldsymbol{\theta}^*$, we apply the fixed point version of the SA algorithm, namely

$$\boldsymbol{\theta}_{t+1} = (1 - \alpha_t)\boldsymbol{\theta}_t + \alpha_t[\mathbf{g}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}]. \tag{3.2.27}$$

This is the standard fixed point iteration with measurement errors.

**Theorem 3.5.** *Suppose $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$ has a unique fixed point $\boldsymbol{\theta}^* = \mathbf{0}$, and that (3.2.26) holds for some $\rho < 1$. Suppose further that Assumptions (N1) and (N2) hold. Under these conditions, Finally, suppose that (3.2.10) and (3.2.11) hold. Suppose there exist constants $\gamma > 0$ and $\delta \ge 0$ such that*

$$B_t = O(t^{-\gamma}), \quad M_t = O(t^\delta),$$

*where we take $\gamma = 1$ if $B_t = 0$ for all sufficiently large $t$, and $\delta = 0$ if $M_t$ is bounded. Choose the step-size sequence $\{\alpha_t\}$ as $O(t^{-(1-\phi)})$ and $\Omega(t^{-(1-C)})$ where $\phi$ is chosen to satisfy*

$$0 < \phi < \min\{0.5 - \delta, \gamma\}, \tag{3.2.28}$$

*and $C \in (0, \phi]$. Define*

$$\nu := \min\{1 - 2(\phi + \delta), \gamma - \phi\}. \tag{3.2.29}$$

*Then $\|\boldsymbol{\theta}_t\|_2^2 = o(t^{-\lambda})$ for every $\lambda \in (0, \nu)$. In particular, by choosing $\phi$ very small, it follows that $\|\boldsymbol{\theta}_t\|_2^2 = o(t^{-\lambda})$ whenever*

$$\lambda < \min\{1 - 2\delta, \gamma\}. \tag{3.2.30}$$

*Proof.* The proof is based on Theorem 3.3. Let us define the Lyapunov function

$$V(\boldsymbol{\theta}) := \frac{1}{2}\|\boldsymbol{\theta}\|_M^2 = \frac{1}{2}\boldsymbol{\theta}^\top M\boldsymbol{\theta}.$$

Observe that

$$\mathbf{f}(\boldsymbol{\theta}) = -\boldsymbol{\theta} + \mathbf{g}(\boldsymbol{\theta}).$$

Therefore

$$\dot{V}(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top M\mathbf{f}(\boldsymbol{\theta}) = -\|\boldsymbol{\theta}\|_M^2 + \boldsymbol{\theta}^\top M\mathbf{g}(\boldsymbol{\theta}).$$

However

$$\boldsymbol{\theta}^\top M\mathbf{g}(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top P^\top P\mathbf{g}(\boldsymbol{\theta}) \le \|P\boldsymbol{\theta}\|_2 \cdot \|P\mathbf{g}(\boldsymbol{\theta})\|_2 \le \rho\|\boldsymbol{\theta}\|_M^2.$$

Therefore

$$\dot{V} \le -(1 - \rho)\|\boldsymbol{\theta}\|_M^2 = -2(1 - \rho)V(\boldsymbol{\theta})$$

The rest of the details are as in Theorem 3.3, and the conclusions partain to $\|\boldsymbol{\theta}_t\|_M^2$. However, since $M$ is positive definite, the same bounds also apply to $\|\boldsymbol{\theta}_t\|_2^2$. □

### 3.2.3 Existence of Suitable Lyapunov Functions

Theorems 3.2 and 3.3 are quite powerful, *provided* there exists a suitable Lyapunov function that satisfies Assumptions (L1) and (L2). So-called "converse" Lyapunov theory gives conditions under which a suitable Lyapunov exists. In this subsection, we state and prove one such theorem for global exponential stability, which can be used in conjunction with Theorem 3.3.

There is a part of Lyapunov stability theory known as "converse" theory. The idea here is to show that, under suitable conditions of the function $\mathbf{f}$ in (3.1.5), if $\boldsymbol{\theta}^*$ is Globally Exponentially Stable (GES), then there exists a Lyapunov function $V$ satisfying the conditions of Theorem 3.3. Many standard books on Lyapunov stability theory do not include these theorems. The contents of this subsection are taken from [168].

As a prelude, we state and prove an inequality known as Gronwall's inequality, which turns an implicit inequality into an explicit inequality

**Lemma 3.1.** ***(Gronwall's Inequality)*** *Suppose $a : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous, and that $b, c \geq 0$ are constants. Under these conditions,*

$$a(t) \leq b + c \int_0^t a(\tau)\, d\tau,\ \forall t \geq 0 \tag{3.2.31}$$

*implies that*

$$a(t) \leq b \exp(ct),\ \forall t \geq 0. \tag{3.2.32}$$

*Proof.* Define

$$d(t) = b + c \int_0^t a(\tau)\, d\tau,$$

and observe that $d(t) \geq 0$ for all $t$. Now (3.2.31) states that $a(t) \leq d(t)$ for all $t$. Next

$$\dot{d}(t) = ca(t) \leq cd(t),\ \forall t \geq 0, \tag{3.2.33}$$

$$\begin{aligned} \frac{d}{dt}[d(t)\exp(-ct)] &= \dot{d}(t)\exp(-ct) - cd(t)\exp(-ct) \\ &= \exp(-ct)[\dot{d}(t) - cd(t)] \leq 0,\ \forall t \geq 0, \end{aligned}$$

from (3.2.33). Hence

$$d(t)\exp(-ct) \leq d(0) = b, \text{ or } d(t) \leq b\exp(ct)\ \forall t \geq 0.$$

Now the bound (3.2.32) follows from $a(t) \leq d(t)$ for all $t$. □

Now we state the new converse theorem. First we state the assumptions on the function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$.

(F1) The equation $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{0}$ has a unique solution $\boldsymbol{\theta}^*$.

(F2) The function $\mathbf{f}$ is twice continuously differentiable, and is globally Lipschitz-continuous with constant $L$. Thus

$$\|\mathbf{f}(\boldsymbol{\theta}) - \mathbf{f}(\boldsymbol{\phi})\|_2 \leq L\|\boldsymbol{\theta} - \boldsymbol{\phi}\|_2,\ \forall \boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^d. \tag{3.2.34}$$

Note that, as a consequence of this assumption, for each $\boldsymbol{\theta} \in \mathbb{R}^d$ there is a unique function $\mathbf{s}(\cdot, \boldsymbol{\theta})$ that satisfies the ODE

$$\frac{d\mathbf{s}(t, \boldsymbol{\theta})}{dt} = \mathbf{f}(\mathbf{s}(t, \boldsymbol{\theta})), \mathbf{s}(0, \boldsymbol{\theta}) = \boldsymbol{\theta}. \tag{3.2.35}$$

(F3) The equilibrium $\boldsymbol{\theta}^*$ of the ODE $\dot{\boldsymbol{\theta}} = \mathbf{f}(\boldsymbol{\theta})$ is globally exponentially stable. Thus there exist constants $\mu \geq 1, \gamma > 0$ such that

$$\|\mathbf{s}(t, \boldsymbol{\theta}) - \boldsymbol{\theta}^*\|_2 \leq \mu\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \exp(-\gamma t),\ \forall t \geq 0,\ \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{3.2.36}$$

(F4) There is a finite constant $K$ such that

$$\|\nabla^2 f_i(\boldsymbol{\theta})\|_S \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq K, \ \forall i \in [d], \ \forall \boldsymbol{\theta} \in \mathbb{R}^d, \tag{3.2.37}$$

where $[d]$ denotes the set $\{1, \ldots, d\}$, and $\|\cdot\|_S$ denotes the spectral norm of a matrix, i.e., its largest singular value. A consequence of (3.2.37) is that

$$\left| \frac{\partial^2 f_i(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right| \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq K, \ \forall i, j, k \in [d], \ \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{3.2.38}$$

**Theorem 3.6.** *Suppose Assumptions (F1)–(F4) hold. Under these hypotheses, there exists a $\mathcal{C}^2$ function $V : \mathbb{R}^d \to \mathbb{R}_+$ such that $V$ and its "derivative" $\dot{V} : \mathbb{R}^d \to \mathbb{R}$ together satisfy the following conditions: There exist positive constants $a, b, c$ and a finite constant $M$ such that*

$$c_1 \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \leq V(\boldsymbol{\theta}) \leq c_2 \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2, \dot{V}(\boldsymbol{\theta}) \leq -c_3 \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2, \ \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{3.2.39}$$

$$\|\nabla^2 V(\boldsymbol{\theta})\|_S \leq 2M, \ \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{3.2.40}$$

**Remark:** The existence of a Lyapunov function $V$ that satisfies (3.2.39) is quite standard. Indeed, the usual choice is

$$V(\boldsymbol{\theta}) := \int_0^\infty \|\mathbf{s}(t, \boldsymbol{\theta})\|_2^2 dt. \tag{3.2.41}$$

However, for this choice of $V$, no conclusions can be drawn about the behavior of the gradient $\nabla V$ nor the Hessian $\nabla^2 V$. In [34], the authors introduce a completely different Lyapunov function of the form

$$V(\boldsymbol{\theta}) := \int_0^T e^{2\kappa\tau} \|\mathbf{s}(\tau, \boldsymbol{\theta}) - \boldsymbol{\theta}^*\|_2^2 \, d\tau, \tag{3.2.42}$$

where $0 < \kappa < \gamma$ is arbitrary, and $T$ is any *finite* number such that

$$\frac{\ln \mu}{\gamma - \kappa} \leq T < \infty,$$

where $\mu, \gamma$ are defined in (3.2.36). For this choice of Lyapunov function, it is shown in [34] that there exists a finite constant $L'$ such that

$$\|\nabla V(\boldsymbol{\theta})\|_2 \leq L' \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2. \tag{3.2.43}$$

Therefore, the Lyapunov function $V$ of (3.2.41) "looks quadratic," while the Lyapunov function $V$ of (3.2.42) and its gradient both "look quadratic." Now Theorem 3.6 extends the theory further by showing that, if (F4) holds, then the Lyapunov function $V$, gradient $\nabla V$, and Hessian $\nabla^2 V$, all "look quadratic." As we shall see, (F4) is the key assumption that allows us to extend the converse Lyapunov theory of [34]. In turn this leads to a simple proof of Theorem 3.6.

*Proof.* Following in [34], define the Lyapunov function candidate $V$ as in (3.2.42). Then, as shown in [34], $V$ satisfies (3.2.39) and (3.2.43). The latter is not of any concern to us. So we focus on proving (3.2.40).

Note that the solution function $\mathbf{s}(\cdot, \boldsymbol{\theta})$ satisfies

$$\mathbf{s}(t, \boldsymbol{\theta}) = \boldsymbol{\theta} + \int_0^t \mathbf{f}(\mathbf{s}(\tau, \boldsymbol{\theta})) \, d\tau. \tag{3.2.44}$$

Therefore

$$\nabla_{\boldsymbol{\theta}} \mathbf{s}(t, \boldsymbol{\theta}) = I + \int_0^t \nabla_{\boldsymbol{\theta}} \mathbf{f}(\mathbf{s}(\tau, \boldsymbol{\theta})) \, d\tau. \tag{3.2.45}$$

Next, the chain rule gives

$$\nabla_{\boldsymbol{\theta}} \mathbf{f}(\mathbf{s}(\tau, \boldsymbol{\theta})) = \nabla_{\boldsymbol{\phi}} \mathbf{f}(\boldsymbol{\phi})|_{\boldsymbol{\phi} = \mathbf{s}(\tau, \boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \mathbf{s}(\tau, \boldsymbol{\theta}).$$

Now the global Lipschitz continuity of $\mathbf{f}$ implies that

$$\|\nabla_{\boldsymbol{\phi}}\mathbf{f}(\mathbf{s}(\tau,\boldsymbol{\phi}))\|_S \leq L,\ \forall\boldsymbol{\phi},\ \forall\tau.$$

Therefore (3.2.45) leads to (after dropping the subscript $\boldsymbol{\theta}$)

$$\|\nabla\mathbf{s}(t,\boldsymbol{\theta})\|_S \leq 1 + \int_0^t L\|\nabla\mathbf{s}(\tau,\boldsymbol{\theta})\|_S\, d\tau.$$

Now Gronwall's inequality of Lemma 3.1 leads to the bound

$$\|\nabla\mathbf{s}(t,\boldsymbol{\theta})\|_S \leq \exp(Lt),\ \forall t,\ \forall\boldsymbol{\theta}. \tag{3.2.46}$$

Next we proceed to find a bound on the second partial derivatives. It follows from (3.2.44) that

$$\frac{\partial s_i(t,\boldsymbol{\theta})}{\partial\theta_j} = \delta_{ij} + \int_0^t \frac{\partial f_i(\mathbf{s}(\tau,\boldsymbol{\theta}))}{\partial\theta_j}\, d\tau,$$

where $\delta_{ij}$ is the Kronecker delta. Next,

$$\frac{\partial^2 s_i(t,\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k} = \int_0^t \frac{\partial^2 f_i(\mathbf{s}(\tau,\boldsymbol{\theta}))}{\partial\theta_j\partial\theta_k}\, d\tau. \tag{3.2.47}$$

We will use (3.2.47) later. Next, expand $V(\boldsymbol{\theta})$ as

$$V(\boldsymbol{\theta}) = \int_0^T e^{2\kappa\tau}\sum_{i=1}^d [s_i(\tau,\boldsymbol{\theta}) - \theta_i^*]^2\, d\tau.$$

Thus

$$\frac{\partial V(\boldsymbol{\theta})}{\partial\theta_j} = \int_0^T 2e^{2\kappa\tau}\sum_{i=1}^d [s_i(\tau,\boldsymbol{\theta}) - \theta_i^*]\frac{\partial s_i(\tau,\boldsymbol{\theta})}{\partial\theta_j}\, d\tau,$$

$$\begin{aligned}\frac{\partial^2 V(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k} &= \int_0^T 2e^{2\kappa\tau}\sum_{i=1}^d \frac{\partial s_i(\tau,\boldsymbol{\theta})}{\partial\theta_k}\frac{\partial s_i(\tau,\boldsymbol{\theta})}{\partial\theta_j}\, d\tau \\ &+ \int_0^T 2e^{2\kappa\tau}\sum_{i=1}^d [s_i(\tau,\boldsymbol{\theta}) - \theta_i^*]\frac{\partial^2 s_i(\tau,\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\, d\tau \\ &= I_1 + I_2,\end{aligned}$$

where

$$I_1 = \int_0^T 2e^{2\kappa\tau}\sum_{i=1}^d \frac{\partial s_i(\tau,\boldsymbol{\theta})}{\partial\theta_k}\frac{\partial s_i(\tau,\boldsymbol{\theta})}{\partial\theta_j}\, d\tau, \tag{3.2.48}$$

$$I_2 = \int_0^T 2e^{2\kappa\tau}\sum_{i=1}^d [s_i(\tau,\boldsymbol{\theta}) - \theta_i^*]\frac{\partial^2 s_i(\tau,\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\, d\tau. \tag{3.2.49}$$

We will prove the boundedness of each integral separately. Note that, as a consequence of (3.2.49), we have

$$\left|\frac{\partial s_i(\tau,\boldsymbol{\theta})}{\partial\theta_k}\right|, \left|\frac{\partial s_i(\tau,\boldsymbol{\theta})}{\partial\theta_j}\right| \leq \|\nabla\mathbf{s}(\tau,\boldsymbol{\theta})\|_S \leq \exp L\tau,\ \forall\tau, i, j, k.$$

So the first integral is bounded by

$$|I_1| \leq \int_0^T 2de^{2\kappa\tau}e^{2L\tau}\,d\tau =: C_1 < \infty$$

for some constant $C_1$, whose precise value need not concern us. So we concentrate on showing that, under Assumption (F4), $I_2$ is also bounded globally.

Towards this end, we begin by observing that

$$\|\mathbf{s}(t,\boldsymbol{\theta})\|_2 \geq e^{-Lt}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2,\ \forall t \geq 0.$$

The proof is elementary and is omitted. In particular

$$\|\mathbf{s}(t,\boldsymbol{\theta})\|_2 \geq e^{-LT}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2,\ \forall t \in [0,T]. \tag{3.2.50}$$

Now we estimate the entity $\partial^2 f_i(\mathbf{s}(\tau,\boldsymbol{\theta}))/\partial\theta_j\partial\theta_k$ in (3.2.47). Note that

$$\frac{\partial f_i}{\partial\theta_j}(\mathbf{s}(\tau,\boldsymbol{\theta})) = \sum_{l=1}^d \left.\frac{\partial f_i(\boldsymbol{\phi})}{\partial\phi_l}\right|_{\boldsymbol{\phi}=\mathbf{s}(\tau,\boldsymbol{\theta})} \frac{\partial s_l(\tau,\boldsymbol{\theta})}{\partial\theta_j},$$

$$\begin{aligned}\frac{\partial^2 f_i(\mathbf{s}(\tau,\boldsymbol{\theta}))}{\partial\theta_j\partial\theta_k} &= \sum_{l=1}^d \left.\frac{\partial f_i(\boldsymbol{\phi})}{\partial\phi_l}\right|_{\boldsymbol{\phi}=\mathbf{s}(\tau,\boldsymbol{\theta})} \frac{\partial^2 s_l(\tau,\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k} \\ &+ \sum_{l=1}^d \frac{\partial}{\partial\theta_k}\left[\left.\frac{\partial f_i(\boldsymbol{\phi})}{\partial\phi_l}\right|_{\boldsymbol{\phi}=\mathbf{s}(\tau,\boldsymbol{\theta})}\right]\frac{\partial s_l(\tau,\boldsymbol{\theta})}{\partial\theta_j}.\end{aligned} \tag{3.2.51}$$

The second term can be expanded as

$$\sum_{l=1}^d\left[\sum_{r=1}^d \left.\frac{\partial^2 f_i(\boldsymbol{\phi})}{\partial\phi_l\partial\phi_r}\right|_{\boldsymbol{\phi}=\mathbf{s}(\tau,\boldsymbol{\theta})}\frac{\partial s_r(\tau,\boldsymbol{\theta})}{\partial\theta_k}\right]\frac{\partial s_l(\tau,\boldsymbol{\theta})}{\partial\theta_j}$$

Now Assumption (F4) and the bound (3.2.50) together imply that

$$\left|\left.\frac{\partial^2 f_i(\boldsymbol{\phi})}{\partial\phi_l\partial\phi_r}\right|_{\boldsymbol{\phi}=\mathbf{s}(\tau,\boldsymbol{\theta})}\right| \leq \|\nabla^2 f_i(\mathbf{s}(\tau,\boldsymbol{\theta}))\|_S \leq \frac{K}{\|\mathbf{s}(\tau,\boldsymbol{\theta})-\boldsymbol{\theta}^*\|_2} \leq \frac{Ke^{LT}}{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2},\ \forall\tau\in[0,T].$$

Also, as shown in (3.2.46),

$$\left|\frac{\partial s_r(\tau,\boldsymbol{\theta})}{\partial\theta_k}\right|, \left|\frac{\partial s_l(\tau,\boldsymbol{\theta})}{\partial\theta_j}\right| \leq \|\nabla\mathbf{s}(\tau,\boldsymbol{\theta})\|_S \leq e^{L\tau} \leq e^{LT},\ \forall\tau\in[0,T].$$

Next, the global Lipschitz continuity of $\mathbf{f}$ implies that

$$\left|\frac{\partial f_i(\boldsymbol{\phi})}{\partial\phi_l}\right| \leq L.$$

Substituting all of these bounds into (3.2.47) gives

$$\begin{aligned}\left|\frac{\partial^2 s_i(t,\boldsymbol{\theta})}{\partial\theta_j\partial\theta_l}\right| &\leq \int_0^t L\sum_{l=1}^d\left|\frac{\partial^2 s_l(\tau,\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\right|\,d\tau + \int_0^t\sum_{l=1}^d\sum_{r=1}^d\frac{Ke^{LT}e^{L\tau}e^{L\tau}}{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2}\,dt \\ &\leq C_2 + \int_0^t L\sum_{l=1}^d\left|\frac{\partial^2 s_l(\tau,\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\right|\,d\tau,\end{aligned} \tag{3.2.52}$$

where

$$C_2 = \frac{d^2 TKe^{3LT}}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2}$$

is inversely proportional to $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2$. Now define

$$h_{jk}(t, \boldsymbol{\theta}) := \sum_{i=1}^{d} \left| \frac{\partial^2 s_i(t, \boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right|.$$

Note that the right side of (3.2.52) does not depend on $i$. Therefore (3.2.52) implies that

$$\begin{aligned} h_{jk}(t, \boldsymbol{\theta}) &\leq \sum_{i=1}^{d} \left[ C_2 + \int_0^t L \left| \sum_{l=1}^{d} \frac{\partial^2 s_i(\tau, \boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right| \right] \\ &\leq C_2 d + \int_0^t L d h_{jk}(\tau, \boldsymbol{\theta})\, d\tau. \end{aligned}$$

So by Gronwall's inequality

$$h_{jk}(t, \boldsymbol{\theta}) \leq C_2 d e^{LdT},\ \forall t \in [0, T].$$

Since $h_{jk}$ is a sum, each individual component must also be smaller than $h_{jk}$ in magnitude. Thus

$$\left| \frac{\partial^2 s_i(t, \boldsymbol{\theta})}{\partial \theta_j \partial \theta_l} \right| \leq C_2 d e^{LdT} \leq \frac{C_3}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2}$$

for a suitable constant $C_3$. Therefore we have established that the Hessian of each $s_i$ decays as $\boldsymbol{\theta}$ gets farther away from $\boldsymbol{\theta}^*$. Now we return to $I_2$ as defined in (3.2.49), and observe that, as a consequence of Assumption (F3) of global exponential stability, we have

$$|s_i(t, \boldsymbol{\theta}) - \theta_i^*| \leq \|\mathbf{s}(t, \boldsymbol{\theta}) - \boldsymbol{\theta}^*\|_2 \leq \mu \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2,\ \forall t \geq 0.$$

Now in the definition of $I_2$, we get the bound

$$|s_i(t, \boldsymbol{\theta}) - \theta_i^*| \cdot \left| \frac{\partial^2 s_i(\tau, \boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right| \leq \mu \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \cdot \frac{C_3}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2} = \mu C_3.$$

Since the integrand in (3.2.49) is bounded and $T$ is finite, it follows that $I_2$ is also bounded. This finally leads to the desired conclusion that $\|\nabla^2 V\|_S$ is globally bounded. □

Note that in the above proof, the finiteness of the constant $T$ is crucial. The traditional Lyapunov function of the form (3.2.41) may not be suitable for the present purposes.

## 3.3 Block Asynchronous Stochastic Approximation

In this section, we study the problem of finding a fixed point of a map $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$ which is a contraction with respect to the $\ell_\infty$-norm. As shown in Chapter 5, this problem arises when it is desired to determine the value of a Markov Reward Process. Now, Theorem 3.5 establishes the convergence of the SA algorithm when the function $\mathbf{g}$ is a contraction wih respect to an *inner product* norm; however this theorem does not apply to the $\ell_\infty$-norm. Hence a distinct approach is needed.

The stochastic approximation algorithm studied in Section 3.2 can perhaps be termed as "fully synchronous" (or just "synchronous") because *every component* of the current guess $\boldsymbol{\theta}_t$ is updated at time $t+1$. At the other extreme lie "asynchronous" SA algorithms, wherein *exactly one* component of $\boldsymbol{\theta}_t$ is updated at time $t$. This is the approach used in temporal difference learning, and $Q$-learning, which are discussed

in subsequent chapters. In-between lies "Block Asynchronous Stochastic Approximation (BASA)," wherein, at each instant $t$, the indices belonging to a subset $S(t) \subseteq [d]$ of the components of $\boldsymbol{\theta}_t$ are updated. By choosing $S(t) = [d]$ we revert to fully synchronous SA, whereas if $S(t)$ is a singleton set at each $t$, then BASA becomes ASA (asynchronous SA). The elements of the set $S(t)$ can be chosen at random, and the cardinality of $S(t)$ (that is, the number of coordinates of $\boldsymbol{\theta}_t$ that are updated at time $t+1$) can vary from one time to another. Also, the step sizes need not be the same for each element of $S(t)$. To highlight this point, we switch notation and use $\{\beta_t\}$ for a fixed sequence of step sizes, while $\{\alpha_{t,i}\}$ denotes the sequence of step sizes for component $i$ at time $t$.

The contents of this section are based on [72].

### 3.3.1 Problem Formulation

If $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$ is a map and it is desired to find a fixed point of $\mathbf{g}(\cdot)$, then we can use the fixed-point version of SA, which is

$$\boldsymbol{\theta}_{t+1} = (\mathbf{1}_d - \boldsymbol{\alpha}_t) \circ \boldsymbol{\theta}_t + \boldsymbol{\alpha}_t \circ [\mathbf{g}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}]. \tag{3.3.1}$$

Here $\mathbf{1}_d$ denotes the column vector of $d$ ones, and $\circ$ denotes the Hadamard product.[3] In this case, it is customary to restrict $\boldsymbol{\alpha}_t$ to belong to $(0,1)^d$ instead of $(0,\infty)^d$. Then each component of $\boldsymbol{\theta}_{t+1}$ is a convex combination of the corresponding components of $\boldsymbol{\theta}_t$ and the noisy measurement of $\mathbf{g}(\boldsymbol{\theta}_t)$.

Next we discuss various options for the step size vector $\boldsymbol{\alpha}_t$, which is allowed to be random. In all cases, it is assumed that there is a *scalar deterministic* sequence $\{\beta_t\}$ taking values in $(0,1)$. We now discuss three commonly used variants of SA, namely: synchronous (also called fully synchronous), asynchronous, and block asynchronous. In **synchronous SA**, one chooses $\boldsymbol{\alpha}_t = \beta_t \mathbf{1}_d$. Thus, in (3.3.1), *the same* step size $\beta_t$ is applied to *every* component of $\boldsymbol{\theta}_t$. In block asynchronous SA (or BASA), there are $d$ different $\{0,1\}$-valued stochastic processes, denoted by $\kappa_t^i, i \in [d]$, called the "update" processes. Then the $i$-th component of $\boldsymbol{\theta}_t$ is updated only if $\kappa_t^i = 1$. To put it another way, define the "update set" as

$$S_t := \{i \in [d] : \kappa_t^i = 1\}.$$

Then $\alpha_t^i = 0$ if $i \notin S_t$. However, this raises the question as to what $\alpha_t^i$ is for $i \in S_t$. Two options are suggested in the literature, known as the "global" clock and the "local" clock respectively. This distinction was first suggested in [22]. If a global clock is used, then $\alpha_t^i = \beta_t$. To define the step size when a local clock is used, first define

$$\nu_t^i := \sum_{\tau=0}^{t} \kappa_t^i. \tag{3.3.2}$$

Thus $\nu_t^i$ counts the number of times that $\theta_t^i$ is updated, and is referred to as the "counter" process. Then the step size is defined as

$$\alpha_t^i := \beta_{\nu_t^i}. \tag{3.3.3}$$

The distinction between global and local clocks can be briefly summarized as follows: When a global clock is used, every component of $\boldsymbol{\theta}_t$ that gets updated has exactly the same step size, namely $\beta_t$, while the other components have a step size of zero. When a local clock is used, among the components of $\boldsymbol{\theta}_t$ that get updated at time $t$, different components may have different step sizes. An important variant of BASA is asynchronous SA (ASA). This phrase was apparently first used in [158], in the context of proving the convergence of the $Q$-learning algorithm from Reinforcement Learning (RL). In ASA, *exactly one* component of $\boldsymbol{\theta}_t$ is updated at each $t$. This can be represented as follows: Let $\{N_t\}$ be an integer-valued stochastic process taking values in $[d]$. Then, at time $t$, the update set $S_t$ is the singleton $\{N_t\}$. The counter process $\nu_t^i$ is now defined via

$$\nu_t^i = \sum_{\tau=0}^{t} I_{\{N_\tau = i\}},$$

[3] Recall that if $\mathbf{a}, \mathbf{b}$ are vectors of equal dimension, then their **Hadamard product** $\mathbf{c} = \mathbf{a} \circ \mathbf{b}$ is defined by $c_i := a_i b_i$ for all $i$.

where $I$ denotes the indicator process. The step size can either be $\beta_t$ if a global clock is used, or $\beta_{\nu_t^i}$ if a local clock is used. In [22], the author analyzes the convergence of ASA with both global as well as local clocks. In the $Q$-learning algorithm introduced in [172], the update is asynchronous (one component at a time) and a global clock is used. In [158], where the phrase ASA was first introduced, the convergence of ASA is proved under some assumptions which include $Q$-learning as a special case. Accordingly, the author uses a global clock in the formulation of ASA. In [46], the authors use a local clock to study the rate of convergence of $Q$-learning.

Next we discuss the assumptions made on the error vector $\boldsymbol{\xi}_{t+1}$. It is assumed that there exist sequences of constants $B_t$ and $M_t$ such that

$$\|E_t(\boldsymbol{\xi}_{t+1})\|_2 \leq B_t(1 + \|\boldsymbol{\theta}_t\|_2), \ \forall t. \tag{3.3.4}$$

$$CV_t(\boldsymbol{\xi}_{t+1}) \leq M_t(1 + \|\boldsymbol{\theta}_t\|_2^2), \ \forall t. \tag{3.3.5}$$

These are the same as the assumptions in Section 3.2, and Theorems 3.2 and 3.3 establish the convergence of *synchronous* (or full-coordinate update) SA under these assumptions.

### 3.3.2 Intermittent Updating: Convergence and Rates

The key distinguishing feature of BASA is that each component of $\boldsymbol{\theta}_t$ gets updated in an "intermittent" fashion. In other words, a component gets updated at some steps, but not at other steps. Before tackling the convergence of BASA in $\mathbb{R}^d$, in the present subsection we state and prove results analogous to Theorems 3.2 and 3.3 for the scalar case with intermittent updating.

The problem setup is as follows: The recurrence relationship is

$$w_{t+1} = (1 - \alpha_t\kappa_t)w_t + \alpha_t\kappa_t\xi_{t+1}, \tag{3.3.6}$$

where $\{w_t\}$ is an $\mathbb{R}$-valued stochastic process of interest, $\{\xi_t\}$ is the measurement error (or "noise"), $\{\alpha_t\}$ is a $(0,1)$-valued stochastic process called the "step size" process, and $\{\kappa_t\}$ is a $\{0,1\}$-valued stochastic process called the "update" process. Clearly, if $\kappa_t = 0$, then $w_{t+1} = w_t$, irrespective of the value of $\alpha_t$; therefore $w_{t+1}$ is updated only at those $t$ for which $\kappa_t = 1$. This is the rationale for the name. With the update process $\{\kappa_t\}$, as before we associate a "counter" process $\{\nu_t\}$, defined by

$$\nu_t = \sum_{s=0}^{t} \kappa_s. \tag{3.3.7}$$

Thus $\nu_t$ is the number of times up to and including time $t$ at which $w_t$ is updated. We also define

$$\nu^{-1}(\tau) := \min\{t \in \mathbb{N} : \nu_t = \tau\}, \ \forall \tau \geq 1. \tag{3.3.8}$$

Then $\nu^{-1}(\cdot)$ is well-defined, and

$$\nu(\nu^{-1}(\tau)) = \tau, \nu^{-1}(\nu_t) \leq t, \nu^{-1}(\tau) \leq \tau - 1. \tag{3.3.9}$$

The last inequality arises from the fact that there are $t+1$ terms in (3.3.7). Also, $\kappa_t = 1$ only when $t = \nu^{-1}(\tau)$ for some $\tau$, and is zero for other values of $t$. Hence, in (3.3.6), if $t = \nu^{-1}(\tau)$ for some $\tau$, then $w_t$ gets updated to $w_{t+1}$, and

$$w_{t+1} = w_{t+2} = \cdots = w_{\nu^{-1}(\tau+1)}, \tag{3.3.10}$$

at which time $w$ gets updated again. Thus $w_t$ is a "piecewise-constant" process, remaining constant between updates. This suggests that we can transform the independent variable from $t$ to $\tau$. Define

$$x_\tau := w_{\nu^{-1}(\tau)}, \zeta_{\tau+1} := \xi_{\nu^{-1}(\tau)+1}, \ \forall \tau \geq 1, \tag{3.3.11}$$

with the convention that $x_1 = w_0$. Note that the convention is consistent whether $\nu_0 = 1$ or not (as can be easily verified). Also we define

$$b_\tau := \alpha_t \kappa_t,$$

whenever $t = \nu^{-1}(\tau)$ for some $\tau$. With these definitions, (3.3.6) is equivalent to

$$x_{\tau+1} = (1 - b_\tau)x_\tau + b_\tau \zeta_{\tau+1}, \ \forall \tau \geq 1, \tag{3.3.12}$$

Note that, in (3.3.12), $b_\tau$ is a random variable for all $\tau \geq 1$, and that there is no $b_0$. To analyze the behavior of (3.3.12), we introduce some preliminary concepts. Let $\mathcal{F}_t$ be the $\sigma$-algebra generated by $w_0, \kappa_0^t, \xi_1^t$. With the change in time indices, define $\{\mathcal{G}_\tau\}$, where $\mathcal{G}_\tau = \mathcal{F}_{\nu^{-1}(\tau)}$, whenever $t = \nu^{-1}(\tau)$ for some $\tau$. Then it is easy to see that $\{\mathcal{G}_\tau\}$ is also a filtration, and that

$$E(x_\tau | \mathcal{G}_\tau) = E_t(w_t | \mathcal{F}_t)$$

whenever $t = \nu^{-1}(\tau)$ for some $\tau$. Hence we can mimic the earlier notation and denote $E(X|\mathcal{G}_\tau)$ by $E_\tau(X)$. Also, if it is assumed that original step size $\alpha_t$ belongs to $\mathcal{M}(\mathcal{F}_t)$, then $b_\tau \in \mathcal{M}(\mathcal{F}_t) = \mathcal{M}(\mathcal{F}_{\nu^{-1}(\tau)}) = \mathcal{M}(\mathcal{G}_\tau)$. The assumption implies that, while the step $\alpha_t$ may be random, it only makes use of the information available up to and including step $t$.

Now we present a general convergence result for (3.3.12). Observe that $\{w_t\}$ is a "piecewise-constant version" of $\{x_\tau\}$. Hence if some conclusions are established for the $x$-process, they are also established for the $w$-process, after adjusting for the time change from $t$ to $\tau$.

**Theorem 3.7.** *Consider the recursion* (3.3.12). *Suppose there exist constants* $B_t, M_t$ *such that*

$$|E_t(\xi_{t+1})| \leq B_t(1 + |w_t|) \ \forall t \geq 0, \tag{3.3.13}$$

$$CV_t(\xi_{t+1}) \leq M_t^2(1 + w_t^2), \ \forall t \geq 0. \tag{3.3.14}$$

*Define*

$$f_\tau = b_\tau^2(1 + 2\mu_{\nu^{-1}(\tau)}^2 + M_{\nu^{-1}(\tau)}^2) + 3b_\tau \mu_{\nu^{-1}(\tau)}, \tag{3.3.15}$$

$$g_\tau = b_\tau^2(2\mu_{\nu^{-1}(\tau)}^2 + M_{\nu^{-1}(\tau)}^2) + b_\tau \mu_{\nu^{-1}(\tau)}. \tag{3.3.16}$$

*Then we have the following conclusions:*

1. *If*

$$\sum_{\tau=1}^{\infty} f_\tau < \infty, \sum_{\tau=1}^{\infty} g_\tau < \infty, \tag{3.3.17}$$

   *then* $x_\tau$ *is bounded almost surely.*

2. *If, in addition to* (3.3.17), *we also have*

$$\sum_{\tau=1}^{\infty} b_\tau = \infty, \tag{3.3.18}$$

   *then* $x_\tau \to 0$ *as* $\tau \to \infty$.

3. *If both* (3.3.17) *and* (3.3.18) *hold, then* $x_\tau = o(\tau^{-\lambda})$ *for every* $\lambda < 1$ *such that*

$$\sum_{\tau=1}^{\infty} (\tau + 1)^\lambda g_\tau < \infty, \tag{3.3.19}$$

$$\sum_{\tau=1}^{\infty} [b_\tau - \lambda \tau^{-1}] = \infty, \tag{3.3.20}$$

   *and in addition, there exists a* $T < \infty$ *such that*

$$b_\tau - \lambda \tau^{-1} \geq 0 \ \forall \tau \geq T. \tag{3.3.21}$$

*Proof.* The proof consists of reformulating the bounds on the error $\boldsymbol{\xi}_{t+1}$ in such a way that Theorems 3.2 and 3.3 apply. By assumption, we have that

$$|E_t(\xi_{t+1})| \leq B_t(1 + |w_t|) \; \forall t.$$

In particular, when $t = \nu^{-1}(\tau)$, we have that $\zeta_{\tau+1} = \xi_{t+1}$, and

$$|E_\tau(\zeta_{\tau+1})| = |E_t(\xi_{t+1})| \leq B_t(1 + |w_t|) = \mu_{\nu^{-1}(\tau)}(1 + |x_\tau|).$$

It follows in an entirely analogous manner that

$$CV_\tau(\zeta_{\tau+1}) \leq M_{\nu^{-1}(\tau)}(1 + x_\tau^2).$$

With these observations, we see that Theorems 3.2 and 3.3 apply to (3.3.12), with the only changes being that (i) the stochastic process is scalar-valued and not vector-valued, (ii) the time index is denoted by $\tau$ and not $t$, and (iii) $B_t, M_t$ are replaced by $\mu_{\nu^{-1}(\tau)}, M_{\nu^{-1}(\tau)}$ respectively. Now the conclusions of the theorem follow from Theorems 3.2 and 3.3. □

Now we reprise the two commonly used approaches for choosing the step size, known as a "global clock" and a "local clock" respectively. This distinction was apparently first introduced in [22]. In each case, there is a *deterministic* sequence $\{\beta_t\}_{t\geq 0}$ of step sizes. If a global clock is used, then $\alpha_t = \beta_t$ at each update, so that $b_\tau = \beta_{\nu^{-1}(\tau)}$. If a local clock is used, then $\alpha_t = \beta_{\nu_t}$, so that then $b_\tau = \beta_{\tau-1}$ . The extra $-1$ in the subscript is to ensure consistency in notation. To illustrate, suppose $\kappa_t = 1$ for all $t$. Then $\nu_t = t + 1$, and $\nu^{-1}(\tau) = \tau - 1$.

Now we begin our analysis of (3.3.12) with the two types of clocks. Now that Theorem 3.7 is established, the challenge is to determine when (3.3.18) through (3.3.21) (as appropriate) hold for the two choices of step sizes, namely global vs. local clocks.

Towards this end, we introduce a few assumptions regarding the update process.

(U1) $\nu_t \to \infty$ as $t \to \infty$ almost surely.

(U2) There exists a random variable $r$ such that

$$\frac{\nu_t}{t} \to r \text{ as } t \to \infty, \text{ a.s..} \tag{3.3.22}$$

Observe that both assumptions are sample-pathwise. Thus (U2) implies (U1).

We begin by stating the convergence results when a local clock is used.

**Theorem 3.8.** *Suppose a local clock is used, so that $\alpha_t = \beta_{\nu_t}$, so that $b_\tau = \beta_{\tau-1}$. Suppose further that Assumption (U1) holds, and moreover*

*(a) $\{B_t\}$ is nonincreasing; that is, $\mu_{t+1} \leq B_t$, $\forall t$.*

*(b) $M_t$ is uniformly bounded, say by $M$.*

*With these assumptions,*

1. *If*

$$\sum_{t=0}^{\infty} \beta_t^2 < \infty, \sum_{t=0}^{\infty} \beta_t B_t < \infty, \tag{3.3.23}$$

*then $\{x_\tau\}$ is bounded almost surely, and $\{w_t\}$ is bounded almost surely.*

2. *If, in addition*

$$\sum_{t=0}^{\infty} \beta_t = \infty, \tag{3.3.24}$$

*then $x_\tau \to 0$ as $t \to \infty$ almost surely, and $w_t \to 0$ as $t \to \infty$ almost surely.*

3. *Suppose* $\beta_t = O(t^{-(1-\phi)})$, *for some* $\phi > 0$, *and* $\beta_t = \Omega(t^{-(1-C)})$ *for some* $C \in (0, \phi]$. *Suppose that* $B_t = O(t^{-\epsilon})$ *for some* $\epsilon > 0$. *Then* $x_\tau \to 0$ *as* $\tau \to \infty$, *and* $w_t \to 0$ *as* $t \to \infty$, *for all* $\phi < \min\{0.5, \epsilon\}$. *Further,* $x_\tau = o(\tau^{-\lambda})$, *and* $w_t = o((\nu_t)^{-\lambda})$ *for all* $\lambda < \epsilon - \phi$. *In particular, if* $B_t = 0$ *for all* $t$, *then* $x_\tau = o(\tau^{-\lambda})$, *and* $w_t = o((\nu_t)^{-\lambda})$ *for all* $\lambda < 1$.

4. *If Assumption (U2) holds instead of (U1), then in the previous item,* $w_t = o((\nu_t)^{-\lambda})$ *can be replaced by* $w_t = o(t^{-\lambda})$.

*Proof.* The proof consists of showing that, under the stated hypotheses, the appropriate conditions in (3.3.17) through (3.3.21) hold.

Recall that $b_\tau = \beta_{\tau-1}$. Also, by Assumption (U1), $\nu_t \to \infty$ as $t \to \infty$, almost surely. Hence $\nu^{-1}(\tau)$ is well-defined for all $\tau \geq 1$.

Henceforth all arguments are along a particular sample path, and we omit the phrase "almost surely," and also do not display the argument $\omega \in \Omega$.

We first prove Item 1 of the theorem. Recall the definitions of $f_\tau$ and $g_\tau$ from (3.3.15) and (3.3.16) respectively. Item 1 is established if t is shown that (3.3.17) holds. For this purpose, note that $\mu_s \leq B_t$ if $s > t$, and $M_t \leq M$ for all $t$. We analyze each of the three terms comprising $f_\tau$. First,

$$\sum_{\tau=1}^{\infty} b_\tau^2 = \sum_{\tau=1}^{\infty} \beta_{\tau-1}^2 = \sum_{t=0}^{\infty} \beta_t^2 < \infty.$$

Next, since $M_t \leq M$ for all $t$, we have that

$$\sum_{\tau=1}^{\infty} b_\tau^2 M_{\nu^{-1}(\tau)}^2 \leq M^2 \sum_{\tau=1}^{\infty} b_\tau^2 < \infty.$$

Finally,

$$\sum_{\tau=1}^{\infty} b_\tau \mu_{\nu^{-1}(\tau)} \leq \sum_{\tau=1}^{\infty} \beta_{\tau-1}\mu_{\tau-1} = \sum_{t=0}^{\infty} \beta_t B_t < \infty.$$

Here we use the fact that $\nu^{-1}(\tau) \geq \tau - 1$, so that $\mu_{\nu^{-1}(\tau)} \leq \mu_{\tau-1}$. Thus it follows from (3.3.15) that $\{f_\tau\} \in \ell_1$, which is the first half of (3.3.17). Next, since $\{b_\tau \mu_{\nu^{-1}(\tau)}\} \in \ell_1$, so is $\{b_\tau^2 \mu_{\nu^{-1}(\tau)}^2\}$. Hence it follows from (3.3.16) that $\{g_\tau\} \in \ell_1$, which is the second half of (3.3.17). This establishes that $\{x_\tau\}$ is bounded, which in turn implies that $\{w_t\}$ is bounded.

To prove Item 2, note that

$$\sum_{\tau=1}^{\infty} b_\tau = \sum_{\tau=0}^{\infty} \beta_\tau = \infty.$$

Hence (3.3.18) holds, and $x_\tau \to 0$ as $\tau \to \infty$, which in turn implies that $w_t \to 0$ as $t \to \infty$.

Finally we come to the rates of convergence. Recall that $B_t = O(t^{-\epsilon})$ while $M_t$ is bounded by $M$. Also, $\beta_t$ is chosen to be $O(t^{-(1-\phi)})$ and $\Omega(t^{-(1-C)})$. From the above, it is clear that

$$f_\tau = O(\tau^{-2+2\phi}) + O(\tau^{-1+\phi-\epsilon}).$$

Hence (3.3.17) holds if

$$-2 + 2\phi < -1 \text{ and } -1 + \phi - \epsilon < -1, \text{ or } \phi < \min\{0.5, \epsilon\}.$$

Next, from the definition of $g_\tau$ in (3.3.16), it follows that

$$(\nu^{-1}(\tau) + 1))^\lambda g_\tau \leq (\nu^{-1}(\tau + 1))^\lambda g_\tau = O(\tau^{-1+\phi-\epsilon+\lambda}).$$

Hence (3.3.19) holds if

$$-1 + \phi - \epsilon + \lambda < -1 \implies \lambda < \epsilon - \phi.$$

Combining everything shows that $x_\tau = o(\tau^{-\lambda})$ whenever

$$\phi < \min\{0.5, \epsilon\}, \lambda < \epsilon - \phi.$$

If $B_t = 0$ for all $t$, then $\epsilon$ can be chosen to be arbitrarily large. However, the limiting factor is that the argument in Theorem 3.3 holds only for $\lambda \leq 1$. Hence $x_\tau = o(\tau^{-\lambda})$ whenever

$$\phi < 0.5, \lambda < 1.$$

Now suppose Assumption (U2) holds, and fix some $\epsilon > 0$. Then along almost all sample paths, for sufficiently large $T$ we have that $\nu_t/t \geq r - \epsilon$ for all $t \geq T$. Thus, whenever $t \geq T$, we have that

$$\nu_t \geq rt \implies o((\nu_t)^{-\lambda}) \leq o((rt)^{-\lambda}) = o(t^{-\lambda}).$$

Thus $w_t$ has the same rate of convergence as $x_\tau$. □

Since the analysis can commence after a finite number of iterations, it is easy to see that Assumption (a) above can be replaced by the following: $\{B_t\}$ is eventually nonincreasing; that is, there exists a $T < \infty$ such that

$$\mu_{t+1} \leq B_t, \ \forall t \geq T.$$

Next we state a result when a global clocks is used. Theorem 3.9 below is not directly comparable to Theorem 3.8 above. Specifically, in Theorem 3.8, the bias coefficient $B_t$ is assumed to be non increasing, and the variance bound $M_t^2$ is assumed to bounded uniformly with respect to $t$. However, the step sizes are constrained only by the requirement that various summations are finite. In contrast, in Theorem 3.9, there are no assumptions regarding $B_t$ and $\mathcal{M}_t$, but the step size sequence $\{\beta_t\}$ is assumed to be nonincreasing.

**Theorem 3.9.** *Suppose a global clock is used, so that $\alpha_t = \beta_t$ whenever $t = \nu^{-1}(\tau)$ for some $\tau$ and as a result $b_\tau = \beta_{\nu^{-1}(\tau)}$. Suppose further that Assumption (U2) holds. Finally, suppose that $\beta_t$ is nonincreasing, so that $\beta_{t+1} \leq \beta_t$ for all $t$. Under these assumptions,*

1. *If* (3.3.23) *holds, and in addition*

$$\sum_{t=0}^{\infty} \beta_t^2 M_t^2 < \infty, \tag{3.3.25}$$

   *then $\{w_t\}$ is bounded almost surely.*

2. *If, in addition,* (3.3.24) *holds, then $w_t \to 0$ as $t \to \infty$ almost surely.*

3. *Suppose in addition that $\beta_t = O(t^{-(1-\phi)})$, for some $\phi > 0$, and $\beta_t = \Omega(t^{-(1-C)})$ for some $C \in (0, \phi]$. Suppose that $B_t = O(t^{-\epsilon})$ for some $\epsilon > 0$, and $M_t = O(t^\delta)$ for some $\delta \geq 0$. Then $w_t \to 0$ as $t \to \infty$ whenever*

$$\phi < \min\{0.5 - \delta, \epsilon\}.$$

   *Moreover, $w_t = o(t^{-\lambda})$ for all $\lambda < \epsilon - \phi$. In particular, if $B_t = 0$ for all $t$, then $w_t = o(t^{-\lambda})$ for all $\lambda < 1$.*

The proof of Theorem 3.9 makes use of the following auxiliary lemma.

**Lemma 3.2.** *Suppose the update process $\{\kappa_t\}$ satisfies Assumption (U2). Suppose $\{\beta_t\}$ is an $\mathbb{R}_+$-valued sequence of deterministic constants such that $\beta_{t+1} \leq \beta_t$ for all $t$, and in addition,* (3.3.24) *holds. Then*

$$\sum_{\tau=1}^{\infty} \beta_{\nu^{-1}(\tau)} = \sum_{t=0}^{\infty} \beta_t \kappa_t = \infty. \tag{3.3.26}$$

*Proof.* We begin by showing that there exists an integer $M$ such that, whenever $2^k > M$, we have

$$\frac{1}{2^k}\left(\sum_{t=2^k+1}^{2^{k+1}} \kappa_t\right) \geq \frac{r}{2}. \tag{3.3.27}$$

By assumption, the ratio $\nu_t/t \to r$ as $t \to \infty$, where $r$ could depend on the sample path (though the dependence on $\omega$ is not displayed). So we can define $\epsilon = r/2$, and choose an integer $M$ such that

$$\left|\frac{1}{T}\sum_{t=0}^{T-1}\kappa_t - r\right| = \left|\frac{1}{T}\sum_{t=0}^{T-1}(\kappa_t - r)\right| < \frac{\epsilon}{3},\ \forall T \geq M.$$

Thus, if $2^k > M$, we have that

$$\begin{aligned}\left|\frac{1}{2^k}\sum_{t=2^k+1}^{2^{k+1}}(\kappa_t - r)\right| &\leq \left|\frac{1}{2^k}\sum_{t=1}^{2^{k+1}}(\kappa_t - r)\right| + \left|\frac{1}{2^k}\sum_{t=1}^{2^k}(\kappa_t - r)\right| \\ &< \frac{2}{3}\epsilon + \frac{1}{3}\epsilon = \epsilon = \frac{r}{2}.\end{aligned}$$

Next, suppose that $\beta_{t+1} \leq \beta_t$ for all $t$. (If this holds only for all *sufficiently large* $t$, we just start all the summations from the time when the above holds.)

$$\begin{aligned}\sum_{t=0}^{\infty}\beta_t\kappa_t &\geq \sum_{k=1}^{\infty}\left(\sum_{t=2^k+1}^{2^{k+1}}\beta_t\kappa_t\right) \geq \sum_{k=1}^{\infty}\left(\sum_{t=2^k+1}^{2^{k+1}}\beta_{2^{k+1}}\kappa_t\right) \\ &= \sum_{k=1}^{\infty}\beta_{2^{k+1}}\left(\sum_{t=2^k+1}^{2^{k+1}}\kappa_t\right) \geq \sum_{k=1}^{\infty}\beta_{2^{k+1}}2^k\frac{r}{2} = \frac{r}{4}\sum_{k=1}^{\infty}\beta_{2^{k+1}}2^{k+1} \\ &= \frac{r}{4}\sum_{k=1}^{\infty}\sum_{t=2^{k+1}+1}^{2^{k+2}}\beta_{2^{k+1}} \geq \frac{r}{4}\sum_{k=1}^{\infty}\sum_{t=2^{k+1}+1}^{2^{k+2}}\beta_t = \frac{r}{4}\sum_{k=5}^{\infty}\beta_t = \infty.\end{aligned}$$

This is the desired conclusion. □

*Proof.* **Of Theorem 3.9:** Recall that a global clock is used, so that $b_\tau = \beta_{\nu^{-1}(\tau)}$. Hence

$$\begin{aligned}\sum_{\tau=1}^{\infty} f_\tau &= \sum_{\tau=1}^{\infty}[\beta^2_{\nu^{-1}(\tau)} + \beta^2_{\nu^{-1}(\tau)}M^2_{\nu^{-1}(\tau)} + \beta_{\nu^{-1}(\tau)}\mu_{\nu^{-1}(\tau)}] \\ &= \sum_{t=0}^{\infty}[\beta_t^2 + \beta_t M_t^2 + \beta_t B_t] < \infty\end{aligned}$$

Via entirely similar reasoning, it follows that $\{g_\tau\} \in \ell_1$. Hence (3.3.17) holds, and Item 1 follows.

To prove Item 2, it is necessary to establish (3.3.18), which in this case becomes

$$\sum_{\tau=1}^{\infty}\beta_{\nu^{-1}(\tau)} = \sum_{\tau=0}^{\infty} b_\tau = \infty.$$

This is (3.3.18). Hence Item 2 follows.

Finally we come to the rates of convergence. The only difference is that now $M_t = O(t^\delta)$ whereas it was bounded in Theorem 3.8. To avoid tedious repetition, we indicate only the changed steps. The only change is that now

$$f_\tau = O(\tau^{-2+2\phi}) + O(\tau^{-2+2\phi+2\delta}) + O(\tau^{-1+\phi-\epsilon}).$$

Hence (3.3.17) holds if

$$-2+2\phi<-1,-2+2\phi+2\delta<-1, \text{ and } -1+\phi-\epsilon<-1,$$

or

$$\phi<\min\{0.5-\delta,\epsilon\}.$$

Next, from the definition of $g_\tau$ in (3.3.16), it follows that

$$(\nu^{-1}(\tau)+1))^\lambda g_\tau \leq (\nu^{-1}(\tau+1))^\lambda g_\tau = O(\tau^{-1+\phi-\epsilon+\lambda}).$$

Hence (3.3.19) holds if

$$-1+\phi-\epsilon+\lambda<-1 \implies \lambda<\epsilon-\phi.$$

Hence $x_\tau = o(\tau^{-\lambda})$ and $w_t = o(t^{-\lambda})$ whenever

$$\phi<\min\{0.5-\delta,\epsilon\}, \lambda<\epsilon-\phi.$$

If $B_t = 0$ for all $t$, then we can choose $\epsilon$ to be arbitrarily large, and we are left with

$$\phi<0.5-\delta, \lambda<1.$$

□

### 3.3.3 Boundedness of Iterations

Next, we give a precise statement of the class of fixed point problems to be studied. In this subsection, it is shown that the iterations are bounded (almost surely), while in the next subsection, the convergence of the iterations is established, together with the rate of convergence. The boundedness of the iterations is established under far more general conditions than the convergence. More details are given at the appropriate place.

Let $\mathbb{N}$ denote the set of natural numbers including zero, and let $\mathbf{h} : \mathbb{N} \times (\mathbb{R}^d)^{\mathbb{N}} \to (\mathbb{R}^d)^{\mathbb{N}}$ denote a **measurement function**. Thus $\mathbf{h}$ maps $\mathbb{R}^d$-valued sequences into $\mathbb{R}^d$-valued sequences. The objective is to determine a fixed point of this map when only noisy measurements of $\mathbf{h}$ are available at each time $t$. Specifically, define

$$\boldsymbol{\eta}_t = \mathbf{h}(t, \boldsymbol{\theta}_0^t). \tag{3.3.28}$$

Suppose that, at time $t+1$, the learner has access to a vector $\boldsymbol{\eta}_t + \boldsymbol{\xi}_{t+1}$, where $\boldsymbol{\xi}_{t+1}$ denotes the measurement error. The objective is to determine a sequence $\boldsymbol{\pi}^* \in (\mathbb{R}^d)^{\mathbb{N}}$ (if it exists) such that

$$\mathbf{h}(\boldsymbol{\pi}^*) = \boldsymbol{\pi}^*,$$

using only the noise-corrupted measurements of $\boldsymbol{\eta}_t$.

To facilitate this, a few assumptions are made regarding the map $\mathbf{h}$. First, the map $\mathbf{h}$ is assumed to be **nonanticipative**[4] and to have **finite memory**. The nonanticipativeness of $\mathbf{h}$ means that

$$\boldsymbol{\theta}_0^\infty, \boldsymbol{\phi}_0^\infty \in (\mathbb{R}^d)^{\mathbb{N}}, \boldsymbol{\theta}_0^t = \boldsymbol{\phi}_0^t \implies \mathbf{h}(\tau, \boldsymbol{\theta}_0^\infty) = \mathbf{h}(\tau, \boldsymbol{\phi}_0^\infty), 0 \leq \tau \leq t. \tag{3.3.29}$$

In other words, $\mathbf{h}(t, \boldsymbol{\theta}_0^\infty)$ depends only on $\boldsymbol{\theta}_0^t$. The finite memory of $\mathbf{h}$ means that there exists a finite constant $\Delta$ which does not depend on $t$, such that $\mathbf{h}(t, \boldsymbol{\theta}_0^t)$ further depends only on $\boldsymbol{\theta}_{t-\Delta+1}^t$. With slightly sloppy notation, this can be written as

$$\mathbf{h}(t, \boldsymbol{\theta}_0^t) = \mathbf{h}(t, \boldsymbol{\theta}_{t-\Delta+1}^t), \ \forall t \geq \Delta, \ \forall \boldsymbol{\theta}_0^\infty \in (\mathbb{R}^d)^{\mathbb{N}}. \tag{3.3.30}$$

[4] In control and system theory, such a function is also referred to as "causal."

This formulation incorporates the possibility of "delayed information" of the form

$$\eta_{t,i} = g_i(\theta_1(t - \Delta_1(t)), \cdots, \theta_d(t - \Delta_d(t))), \tag{3.3.31}$$

where $\Delta_1(t), \cdots, \Delta_d(t)$ are delays that could depend on $t$. The only requirement is that each $\Delta_j(t) \leq \Delta$ for some finite $\Delta$. This formulation is analogous to [158, Eq. (2)] and [22, Eq. (1.4)], which is slightly more general in that they require only that $t - \Delta_i(t) \to \infty$ as $t \to \infty$, for each index $i \in [d]$. In particular, if $\mathbf{h}$ is "memoryless" in the sense that, for some function $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$, we have

$$\mathbf{h}(t, \boldsymbol{\theta}_0^t) = \mathbf{g}(\boldsymbol{\theta}_t), \tag{3.3.32}$$

then we can take $\Delta = 1$. Note that, if $\mathbf{h}$ is of the form (3.3.32), then the problem at hand becomes one of finding a fixed point in $\mathbb{R}^d$ of the map $\mathbf{g}$, gives noisy measurements of $\mathbf{g}$ at eath time step.

To proceed further, it is assumed that the measurement function satisfies the following assumption:

(F1) There exist an integer $\Delta \geq 1$ and a constant $\gamma \in (0, 1)$ such that

$$\|\mathbf{h}(t, \boldsymbol{\psi}_{t-\Delta+1}^t) - \mathbf{h}(t, \boldsymbol{\phi}_{t-\Delta+1}^t)\|_\infty \leq \gamma \|\boldsymbol{\psi}_{t-\Delta+1}^t - \boldsymbol{\phi}_{t-\Delta+1}^t\|_\infty, \ \forall t \geq \Delta, \ \forall \boldsymbol{\psi}_0^\infty, \boldsymbol{\phi}_0^\infty \in (\mathbb{R}^d)^{\mathbb{N}}. \tag{3.3.33}$$

This assumption means that the map $\boldsymbol{\theta}_{t-\Delta+1}^t \mapsto \mathbf{h}(t, \boldsymbol{\theta}_{t-\Delta+1}^t)$ is a contraction with respect to $\|\cdot\|_\infty$. In case $\Delta = 1$ and $\mathbf{h}$ is of the form (3.3.32), Assumption (F1) says that the map $\mathbf{g}$ is a contraction.

Now we discuss a few implications of Assumption (F1).

(F2) By repeatedly applying (3.3.33) over blocks of width $\Delta$, one can conclude that

$$\|\mathbf{h}(t, \boldsymbol{\psi}_{t-\Delta+1}^t) - \mathbf{h}(t, \boldsymbol{\phi}_{t-\Delta+1}^t)\|_\infty \leq \gamma^{\lfloor t/\Delta \rfloor} \|\boldsymbol{\psi}_0^{\Delta-1} - \boldsymbol{\phi}_0^{\Delta-1}\|_\infty, \ \forall \boldsymbol{\psi}_0^\infty, \boldsymbol{\phi}_0^\infty \in (\mathbb{R}^d)^{\mathbb{N}}. \tag{3.3.34}$$

Therefore, for every sequence $\boldsymbol{\phi}_0^\infty$, the iterations $\mathbf{h}(t, \boldsymbol{\phi}_0^t)$ converge to a unique fixed point $\boldsymbol{\pi}^*$. In particular, if we let $(\boldsymbol{\pi}^*)_0^\infty$ denote the sequence whose value is $\boldsymbol{\pi}^*$ for every $t$, then it follows that

$$\|\mathbf{h}(t, (\boldsymbol{\pi}^*)_0^t) - \boldsymbol{\pi}^*\|_\infty \leq C_0 \gamma^{\lfloor t/\Delta \rfloor}, \ \forall t, \tag{3.3.35}$$

for some constant $C_0$.

(F3) The following also follows from Assumption (F1): There exist constants $\rho < 1$ and $c_1' > 0$ such that

$$\|\mathbf{h}(t, \boldsymbol{\phi}_0^t)\|_\infty \leq \rho \max\{c_1', \|\boldsymbol{\phi}_0^t\|_\infty\}, \ \forall \boldsymbol{\phi} \in (\mathbb{R}^d)^{\mathbb{N}}, t \geq 0. \tag{3.3.36}$$

In order to determine $\boldsymbol{\pi}^*$ in (F2), we use BASA. Specifically, we choose $\boldsymbol{\theta}_0$ as we wish (either deterministically or at random). At time $t$, we update $\boldsymbol{\theta}_t$ to $\boldsymbol{\theta}_{t+1}$ according to

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{\alpha}_t \circ [\boldsymbol{\eta}_t + \boldsymbol{\xi}_{t+1}], \tag{3.3.37}$$

where $\boldsymbol{\alpha}_t$ is the *vector of step sizes* belonging to $[0, 1)^d$, $\boldsymbol{\xi}_{t+1}$ is the measurement noise vector belonging to $\mathbb{R}^d$, and $\circ$ denotes the Hadamard product. We are interested in studying two questions:

(Q1) Under what conditions is the sequence of iterations $\{\boldsymbol{\theta}_t\}$ bounded almost surely?

(Q2) Under what conditions does the sequence of iterations $\{\boldsymbol{\theta}_t\}$ converge to $\boldsymbol{\pi}^*$ as $t \to \infty$?

Question (Q1) is addressed in this subsection, whereas Question (Q2) is addressed in the next.

In order to study the above two questions, we make some assumptions about various entities in (3.3.37). Let $\mathcal{F}_t$ denote the $\sigma$-algebra generated by the random variables $\boldsymbol{\theta}_0$, $\boldsymbol{\xi}_1^t$, and $\alpha_{0,i}^{t,i}$ for $i \in [d]$. Then it is clear that $\{\mathcal{F}_t\}$ is a filtration. As before, we denote $E(X|\mathcal{F}_t)$ by $E_t(X)$.

The first set of assumptions is on the noise.

(N1) There exists a finite constant $c_1'$ and a sequence of constants $\{B_t\}$ such that

$$\|E_t(\boldsymbol{\xi}_{t+1})\|_2 \leq c_1' B_t(1 + \|\boldsymbol{\theta}_0^t\|_\infty), \ \forall t \geq 0. \tag{3.3.38}$$

(N2) There exists a finite constant $c_2'$ and a sequence of constants $\{M_t\}$ such that

$$CV_t(\boldsymbol{\xi}_{t+1}) \leq c_2' M_t^2(1 + \|\boldsymbol{\theta}_0^t\|_\infty^2), \ \forall t \geq 0, \tag{3.3.39}$$

where, as before,

$$CV_t(\boldsymbol{\xi}_{t+1}) = E_t(\|\boldsymbol{\xi}_{t+1} - E_t(\boldsymbol{\xi}_{t+1})\|_2^2)$$

Before proceeding further, let us compare the conditions (3.3.38) and (3.3.39) with their counterparts (3.2.4) and (3.2.5) in Theorem 3.2. It can be seen that the above two requirements are more liberal (i.e., less restrictive) than in Theorem 3.2, because the quantity $\|\boldsymbol{\theta}_t\|_2$ is replaced by $\|\boldsymbol{\theta}_0^t\|_\infty$. Hence, in (3.3.38) and (3.3.39), the bounds are more loose. However, Theorems 3.10 and 3.11 in the next subsection apply only to *contractive* mappings. Hence Theorems 3.10 and 3.11 complement Theorem 3.2, and do not subsume it.

The next set of assumptions is on the step size sequence.

(S1) The random step size sequences $\{\alpha_{t,i}\}$ and the sequences $\{B_t\}$, $\{M_t^2\}$ and satisfy (almost surely)

$$\sum_{t=0}^{\infty} \alpha_{t,i}^2 < \infty, \sum_{t=0}^{\infty} M_t^2 \alpha_{t,i}^2 < \infty, \sum_{t=0}^{\infty} B_t \alpha_{t,i} < \infty, \ \forall i \in [d]. \tag{3.3.40}$$

(S2) The random step size sequence $\{\alpha_{t,i}\}$ satisfies (almost surely)

$$\sum_{t=0}^{\infty} \alpha_{t,i} = \infty, \text{ a.s., } \forall i \in [d]. \tag{3.3.41}$$

With these assumptions in place, we state the main result of this subsection, namely, the almost sure boundedness of the iterations. In the next subsection, we state and prove the convergence of the iterations, under more restrictive assumptions.

**Theorem 3.10.** *Suppose that Assumptions (N1) and (N2) about the noise sequence, (S1) and (S2) about the step size sequence, and (F1) about the function* $\mathbf{h}$ *hold, and that* $\boldsymbol{\theta}_{t+1}$ *is defined via* (3.3.37)*. Then* $\sup_t \|\boldsymbol{\theta}_t\|_\infty < \infty$ *almost surely.*

The proof of the theorem is fairly long and involves several preliminary results and observations.

To aid in proving the results, we introduce a sequence of "renormalizing constants." This is similar to the technique used in [158]. For $t \geq 0$, define

$$\Lambda_t := \max\{\|\boldsymbol{\theta}_0^t\|_\infty, c_1'\}, \tag{3.3.42}$$

where $c_1'$ is defined in (3.3.28). With this definition, it follows from (3.3.36) that $\boldsymbol{\eta}_t = \mathbf{h}(t, \boldsymbol{\theta}_0^t)$ satisfies

$$\|\boldsymbol{\eta}_t\|_\infty \leq \rho \Lambda_t, \ \forall t. \tag{3.3.43}$$

Define $\boldsymbol{\zeta}_{t+1} = Ł_t^{-1} \boldsymbol{\xi}_{t+1}$ for all $t \geq 0$. Now observe that $Ł_t^{-1} \leq c_1^{-1}$, and $Ł_t^{-1} \leq (\|\boldsymbol{\theta}_0^t\|_\infty)^{-1}$. Hence

$$\|E_t(\zeta_{t+1,i})\|_\infty \leq c_1' B_t(c_1^{-1} + 1) =: c_2 B_t, \tag{3.3.44}$$

where $c_2 = c_1'(c_1^{-1} + 1)$. In particular, the above implies that

$$|E_t(\zeta_{t+1,i})| \leq c_2 B_t, \ \forall t \geq 0. \tag{3.3.45}$$

Similarly

$$CV_t(\zeta_{t+1,i}) \leq c_3 M_t^2, \ \forall t \geq 0, \tag{3.3.46}$$

for some constant $c_3$.

If we compare (3.3.44) with (3.3.38), and (3.3.45) with (3.3.39), we see that the bounds for the "modified" error $\boldsymbol{\zeta}_{t+1}$ are simpler than those for $\boldsymbol{\xi}_{t+1}$. Specifically, the right side of both (3.3.44) and (3.3.45) are bounded with respect to $\boldsymbol{\theta}_0^t$ for each $t$, though they may be unbounded as functions of $t$. In contrast, the right sides of (3.3.38) an (3.3.39) are permitted to be functions of $\|\boldsymbol{\theta}_0^t\|_\infty$.

Though the next result is quite obvious, we state it separately, because it is used repeatedly in the sequel.

**Lemma 3.3.** *For $i \in [d]$ and $0 \leq s \leq k < \infty$, define the doubly-indexed stochastic process*

$$D_i(s, k+1) = \sum_{t=s}^{k} \Big[ \prod_{r=t+1}^{k} (1 - \alpha_{r,i}) \Big] \alpha_{t,i} \zeta_{t+1,i}, \tag{3.3.47}$$

*where an empty product is taken as 1. Then $\{D_i(s,k)\}$ satisfies the recursion*

$$D_i(s, k+1) = (1 - \alpha_{k,i}) D_i(s,k) + \alpha_{k,i} \zeta_{k+1,i}, D_i(s,s) = 0. \tag{3.3.48}$$

*In the other direction,* (3.3.47) *gives a closed-form solution for the recursion* (3.3.48).

Recall that $\mathbb{N}$ denotes the set of non-negative integers $\{0, 1, 2, \ldots, \}$. The next lemma is basically the same as [158, Lemma 2].

**Lemma 3.4.** *There exists $\Omega_1 \subset \Omega$ with $P(\Omega_1) = 1$ and $r_1^* : \Omega_1 \times (0,1) \to \mathbb{N}$ such that*

$$|D_i(s, k+1)(\omega)| \leq \epsilon, \ \forall k \geq s \geq r_1^*(\omega, \epsilon). \tag{3.3.49}$$

*Proof.* Let $\epsilon > 0$ be given. It follows from Lemma 3.3 that $D_i$ satisfies the recursion

$$D_i(0, t+1) = (1 - \alpha_{t,i}) D_i(0,t) + \alpha_{t,i} \zeta_{t+1,i}$$

with $D_i(0,0) = 0$. Let us fix an index $i \in [d]$, and invoke (3.3.45) and (3.3.46). Then it follows from (3.3.46) that

$$CV_t(\zeta_{t+1,i}) \leq c_3 M_t^2,$$

and (3.3.45) also holds. Now, if Assumptions (S1) and (S2) also hold, then all the hypotheses needed to apply Theorem 3.7 are in place. Therefore $D_i(0, k+1)$ converges to zero almost surely. This holds for each $i \in [d]$ Therefore, if we define

$$\Omega_1 = \{\omega \in \Omega_1 : D_i(0, k+1)(\omega) \to 0 \text{ as } t \to \infty \ \forall i \in [d]\},$$

then $P(\Omega_1) = 1$. We can see that for $\omega \in \Omega_1$ we can choose $r_1^*(\omega, \epsilon)$ such that $\forall k \geq r_1^*(\omega, \epsilon), i \in [d]$ we have

$$|D_i(0, k+1)(\omega)| \leq \tfrac{1}{2}\epsilon.$$

To proceed further, we suppress the argument $\omega$ in the interests of clarity. Observe from (3.3.47) that, whenever $s \leq k$ we have

$$\begin{aligned}
D_i(s, k+1) &= \sum_{t=s}^{k} \Big[ \prod_{r=t+1}^{k} (1 - \alpha_{r,i}) \Big] \alpha_{t,i} \zeta_{t+1,i} && (3.3.50)\\
&= \sum_{t=0}^{k} \Big[ \prod_{r=t+1}^{k} (1 - \alpha_{r,i}) \Big] \alpha_{t,i} \zeta_{t+1,i} - \sum_{t=0}^{s-1} \Big[ \prod_{r=t+1}^{k} (1 - \alpha_{r,i}) \Big] \alpha_{t,i} \zeta_{t+1,i} && (3.3.51)\\
&= D_i(0, k+1) - \Big[ \prod_{r=s}^{k} (1 - \alpha_{r,i}) \Big] \sum_{t=0}^{s-1} \Big[ \prod_{r=t+1}^{s-1} (1 - \alpha_{r,i}) \Big] \alpha_{t,i} \zeta_{t+1,i} && (3.3.52)\\
&= D_i(0, k+1) - \Big[ \prod_{r=s}^{k} (1 - \alpha_{r,i}) \Big] D_i(0,s). && (3.3.53)
\end{aligned}$$

Since $1-\alpha_{r,i} \in (0,1)$ for all $r,i$, it follows that the product also belongs to $(0,1)$. Therefore

$$|D_i(s,k+1)| \leq |D_i(0,k+1)| + |D_i(0,s)| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

This is the desired conclusion. □

**Lemma 3.5.** *There exists $\Omega_2 \subset \Omega$ with $P(\Omega_2) = 1$ and $r_2^* : \Omega_1 \times \mathbb{N} \times (0,1) \to \mathbb{N}$ such that*

$$\prod_{s=j}^{k}(1-\alpha_{s,i}(\omega)) \leq \epsilon, \ \forall\, k \geq r_2^*(\omega,j,\epsilon), i \in [d], \omega \in \Omega_2. \tag{3.3.54}$$

*Proof.* In view of the assumption (S2), if we define

$$\Omega_2 = \left\{ \omega \in \Omega : \sum_{s=j}^{\infty} \alpha_{t,i}(\omega) = \infty \ \forall i \in [d] \right\},$$

then $P(\Omega_2) = 1$. For all $\omega \in \Omega_2$, we have

$$\sum_{s=j}^{\infty} \alpha_{t,i}(\omega) = \infty.$$

Using the elementary inequality $(1-x) \leq \exp\{-x\}$ for all $x \in [0,\infty)$, it follows that

$$\prod_{s=j}^{k}(1-\alpha_{t,i}(\omega)) \leq \exp\left\{ -\sum_{s=j}^{k} \alpha_{t,i}(\omega) \right\}.$$

Hence for $\omega \in \Omega_2$, $\prod_{s=j}^{k}(1-\alpha_{t,i}(\omega))$ converges to zero as $k \to \infty$. Thus we can choose $r_2^*(\omega,j,\epsilon)$ with the required property. □

In the rest of this section, we will fix $\omega \in \Omega_1 \cap \Omega_2$, the functions $r_1^*$, $r_2^*$ obtained in Lemma 3.4 and Lemma 3.5 respectively and prove that if (F1) holds, then $\|\boldsymbol{\theta}_t(\omega)\|_\infty$ is bounded, which proves Theorem 3.7.

Let us rewrite the updating rule (3.3.37) as

$$\theta_{t+1,i} = (1-\alpha_{t,i})\theta_{t,i} + \alpha_{t,i}(\eta_{t,i} + \Lambda_t \zeta_{t+1,i}), i \in [d],\, t \geq 0, \tag{3.3.55}$$

By recursively invoking (3.3.55) for $k \in [0,t]$, we get

$$\theta_{t+1,i} = A_{t+1,i} + B_{t+1,i} + C_{t+1,i} \tag{3.3.56}$$

where

$$A_{t+1,i} = \Big[\prod_{k=0}^{t}(1-\alpha_{k,i})\Big]\theta_{0,i}, \tag{3.3.57}$$

$$B_{t+1,i} = \sum_{k=0}^{t}\Big[\prod_{r=k+1}^{t}(1-\alpha_{r,i})\Big]\alpha_{k,i}\eta_{k,i}, \tag{3.3.58}$$

$$C_{t+1,i} = \sum_{k=0}^{t}\Big[\prod_{r=k+1}^{t}(1-\alpha_{r,i})\Big]\alpha_{k,i}\Lambda_k\zeta_{k+1,i}. \tag{3.3.59}$$

**Lemma 3.6.** *For $i \in [d]$ ,*

$$|C_{t+1,i}| \leq \Lambda_t \sup_{0 \leq r \leq t} |D_i(r,t+1)|. \tag{3.3.60}$$

*Proof.* We begin by establishing an alternate expression for $C_{k,i}$, namely

$$C_{t+1,i} = Ł_0 D_i(0,t+1) + \sum_{k=1}^{t} (Ł_k - Ł_{k-1}) D_i(k,t+1), \tag{3.3.61}$$

where $D_i(\cdot,\cdot)$ is defined in (3.3.47). For this purpose, observe from Lemma 3.3 that $C_{t+1,i}$ satisfies

$$C_{t+1,i} = Ł_t \alpha_{t,i} \zeta_{t+1,i} + (1-\alpha_{t,i}) C_{t,i} = Ł_t D_i(t,t+1) + (1-\alpha_{t,i}) C_{t,i}, \tag{3.3.62}$$

because $\alpha_{t,i}\zeta_{t+1,i} = D_i(t,t+1)$ due to (3.3.48) with $s=t$. The proof of (3.3.61) is by induction. It is evident from (3.3.59) that

$$C_{1,i} = Ł_0 \alpha_{0,1} \zeta_{1,i} = Ł_0 D_i(0,1).$$

Thus (3.3.61) holds when $t=0$. Now suppose by way of induction that

$$C_{t,i} = Ł_0 D_i(0,t) + \sum_{k=1}^{t-1} (Ł_k - Ł_{k-1}) D_i(k,t). \tag{3.3.63}$$

Using this assumption, and the recursion (3.3.62), we establish (3.3.61).

Substituting from (3.3.63) into (3.3.62) gives

$$C_{t+1,i} = Ł_t D_i(t,t+1) + Ł_0 (1-\alpha_{t,i}) D_i(0,t) + (1-\alpha_{t,i}) \sum_{k=1}^{t-1} (Ł_k - Ł_{k-1}) D_i(k,t). \tag{3.3.64}$$

Now (3.3.47) implies that

$$(1-\alpha_{t,i}) D_i(k,t) = D_i(k,t+1) - \alpha_{t,i}\zeta_{t+1,i} = D_i(k,t+1) - D_i(t,t+1).$$

Therefore the summation in (3.3.64) becomes

$$\begin{aligned} \sum_{k=1}^{t-1} (Ł_k - Ł_{k-1})(1-\alpha_{t,i}) D_i(k,t) &= \sum_{k=1}^{t-1} (Ł_k - Ł_{k-1}) D_i(k,t) \\ &- D_i(t,t+1) \sum_{k=1}^{t-1} (Ł_k - Ł_{k-1}) = S_1 + S_2 \text{ say.} \end{aligned}$$

Then $S_2$ is just a telescoping sum and equals

$$S_2 = -Ł_{t-1} D_i(t,t+1) + Ł_0 D_i(t,t+1).$$

The second term in (3.3.64) equals

$$Ł_0 (1-\alpha_{t,i}) D_i(0,t) = Ł_0 [D_i(0,t+1) - \alpha_{t,i}\zeta_{t+1,i}] = Ł_0 D_i(0,t+1) - Ł_0 D_i(t,t+1).$$

Putting everything together and observing that the term $Ł_0 D_i(t,t+1)$ cancels out gives

$$C_{t+1,i} = Ł_0 D_i(0,t+1) + (Ł_t - Ł_{t-1}) D_i(t,t+1) + \sum_{k=1}^{t-1} (Ł_k - Ł_{k-1}) D_i(k,t).$$

This is the same as (3.3.64) with $t+1$ replacing $t$. This completes the induction step and thus (3.3.61) holds. Using the fact that $Ł_t \geq Ł_{t-1}$, the desired bound (3.3.60) follows readily. □

*Proof.* (Of Theorem 3.7) As per the statement of the theorem, we assume that (F1) holds. We need to prove that

$$\sup_{t\geq 0} Ł_t < \infty.$$

Define

$$\delta = \min\{\frac{1-\rho}{2\rho}, \frac{1}{2}\},$$

and observe that, as a consequence, we have that $\rho(1+2\delta) \leq 1$. Choose $r_1^* = r_1^*(\delta)$ as in Lemma 3.4 such that

$$|D_i(s, k+1)| \leq \delta \; \forall k \geq s \geq r_1^*, \; \forall i \in [d].$$

It is now shown that

$$Ł_t \leq (1+2\delta) Ł_{r_1^*} \; \forall t, \; \forall i \in [d]. \tag{3.3.65}$$

By the monotonicity of $\{Ł_t\}$, it is already known that $Ł_t \leq Ł_{r_1^*}$ for $t \leq r_1^*$. Hence, once (3.3.65) is established, it will follow that

$$\sup_{0\leq t<\infty} \Lambda_t \leq (1+2\delta)\Lambda_{r_1^*}.$$

The proof of (3.3.65) is by induction on $t$. Accordingly, suppose (3.3.65) holds for $t \leq k$. Using (3.3.60), we have

$$|C_{k+1,i}| \leq \delta Ł_k \leq Ł_{r_1^*}\delta(1+2\delta). \tag{3.3.66}$$

It is easy to see from its definition that

$$|A_{k+1,i}| \leq Ł_{r_1^*}\Big[\prod_{s=0}^{k}(1-\alpha_{s,i})\Big]$$

Using the induction hypothesis that $Ł_t \leq (1+2\delta)Ł_{r_1^*}$ for $t \leq k$, we have

$$\begin{aligned}
|B_{k+1,i}| &\leq \sum_{s=0}^{k}\Big[\prod_{r=s+1}^{k}(1-\alpha_{r,i})\Big]\alpha_{s,i}|\eta_{s,i}| \\
&\leq \sum_{s=0}^{k}\Big[\prod_{r=s+1}^{k}(1-\alpha_{r,i})\Big]\alpha_{s,i}\rho Ł_s \\
&\leq \rho(1+2\delta)Ł_{r_1^*}\sum_{s=0}^{k}\Big[\prod_{r=s+1}^{k}(1-\alpha_{r,i})\Big]\alpha_{s,i} \\
&\leq Ł_{r_1^*}\sum_{s=0}^{k}\Big[\prod_{r=s+1}^{k}(1-\alpha_{r,i})\Big]\alpha_{s,i},
\end{aligned}$$

because $\rho(1+2\delta) \leq 1$. Also, the following identity is easy to prove by induction.

$$\Big[\prod_{s=0}^{k}(1-\alpha_{s,i})\Big] + \sum_{s=0}^{k}\Big[\prod_{r=s+1}^{k}(1-\alpha_{r,i})\Big]\alpha_{s,i} = 1 \; \forall k < \infty \tag{3.3.67}$$

Combining these bounds gives

$$|A_{k+1,i}| + |B_{k+1,i}| \leq Ł_{r_1^*}.$$

Combining this with (3.3.56) and (3.3.66) leads to

$$\theta_{k+1,i} \leq Ł_{r_1^*}(1+\delta(1+2\delta)) \leq Ł_{r_1^*}(1+2\delta).$$

Therefore $\|\boldsymbol{\theta}_{k+1}\|_\infty \leq Ł_{r_1^*}(1+2\delta)$, and

$$Ł_{k+1} = \max\{\|\boldsymbol{\theta}_{k+1}\|_\infty, Ł_k\} \leq Ł_{r_1^*}(1+2\delta).$$

This proves the induction hypothesis and completes the proof of Theorem 3.7. □

### 3.3.4 Convergence of Iterations with Rates

In this subsection, we further study the iteration sequence (3.3.37), under a variety of Block (or Batch) updating schemes, corresponding to various choices of the step sizes. Whereas the almost sure boundedness of the iterations is established in the previous subsection, in this subsection we prove that the iterations converge to the desired fixed point $\boldsymbol{\pi}^*$. Then we also find bounds on the rate of convergence.

We study three specific methods for choosing the step size vector $\boldsymbol{\alpha}_t$ in (3.3.37). Within the first two methods, we further divide into local clocks and global clocks. However, in the third method, we permit only the use of a global clock, for reasons to be specified.

#### Convergence Theorem

The overall plan is to follow up Theorem 3.10, which establishes the almost sure boundedness of the iterations, with a stronger result showing that the iterations converge almost surely to $\boldsymbol{\pi}^*$, the fixed point of the map $\mathbf{h}$. This convergence is established under the same assumptions as in Theorem 3.10. In particular, the step size sequence is assumed to satisfy (S1) and (S2). Having done this, we then study conditions under which (S1) and (S2) hold for each of the three methods for choosing the step sizes.

**Theorem 3.11.** *Suppose that Assumptions (N1) and (N2) about the noise sequence, (S1) and (S2) about the step size sequence, and (F1) about the function* $\mathbf{h}$ *hold, and that* $\boldsymbol{\theta}_{t+1}$ *is defined via* (3.3.37). *Then* $\boldsymbol{\theta}_t \to \boldsymbol{\pi}^*$ *as* $t \to \infty$ *almost surely, where* $\boldsymbol{\pi}^*$ *is defined in (F2).*

*Proof.* From (3.3.56), we have an expression for $\boldsymbol{\theta}_{t+1,i}$, where $A_{t+1,i}$, $B_{t+1,i}$ and $C_{t+1,i}$ are given by (3.3.57), (3.3.58) and (3.3.59) respectively. Also, by changing notation from $k$ to $t$ and $s$ to $k$ in (3.3.67), and multiplying both sides by $\pi_i^*$, we can write

$$\pi_i^* = \Big[\prod_{k=0}^{t}(1-\alpha_{k,i})\Big]\pi_i^* + \left\{\sum_{k=0}^{t}\Big[\prod_{r=k+1}^{t}(1-\alpha_{r,i})\Big]\alpha_{k,i}\right\}\pi_i^*, \ \forall t.$$

Substituting from these formulas gives

$$\theta_{t+1,i} - \pi_i^* = \bar{A}_{t+1,i} + \bar{B}_{t+1,i} + C_{t+1,i}, \tag{3.3.68}$$

where

$$\bar{A}_{t+1,i} = \prod_{k=0}^{t}(1-\alpha_{k,i})(\theta_{0,i} - \pi_i^*), \tag{3.3.69}$$

$$\bar{B}_{t+1,i} = \Big[\prod_{r=k+1}^{t}(1-\alpha_{r,i})\Big]\alpha_{k,i}(\eta_{k,i} - \pi_i^*), \tag{3.3.70}$$

and $C_{t+1,i}$ is as in (3.3.59). It is shown in turn that each of these quantities approaches zero as $t \to \infty$.

First, from Assumption (S2), it follows that[5]

$$\prod_{k=0}^{t}(1-\alpha_{k,i}) \to 0 \text{ as } t \to \infty.$$

Since $\theta_{0,i} - \pi_i^*$ is a constant along each sample path, $\bar{A}_{t+1,i}$ approaches zero.

Second, by combining (3.3.34) and (3.3.35) in Property (F2), it follows that

$$|\eta_{t,i} - \pi_i^*| \le \gamma^{\lfloor t/\Delta \rfloor}\|\boldsymbol{\theta}_0^{\Delta} - (\boldsymbol{\pi}^*)^{\Delta}\|_{\infty} \le C_1\gamma^{\lfloor t/\Delta \rfloor}$$

[5]We omit the phrase "almost surely" in these arguments.

for some constant $C_1$ (which depends on the sample path). Thus

$$\sum_{r=0}^{\infty} |\eta_{t,i} - \pi_i^*| < \infty$$

along almost all sample paths. Now it follows from (3.3.70) that

$$\begin{aligned} |\bar{B}_{t+1,i}| &\leq \Big[ \prod_{r=k+1}^{t} (1-\alpha_{r,i}) \Big] \alpha_{k,i} |\eta_{k,i} - \pi_i^*| \\ &\leq \Big[ \prod_{r=k+1}^{t} (1-\alpha_{r,i}) \Big] \alpha_{k,i} C_1 \gamma^{\lfloor t/\Delta \rfloor} =: L_{t+1,i}. \end{aligned} \tag{3.3.71}$$

Let $L_{t+1,i}$ denote the right side of this inequality. Then it follows from Lemma 3.3 that $L_{t+1,i}$ satisfies the recursion

$$L_{t+1,i} = (1-\alpha_{t,i}) L_{t,i} + \alpha_{t,i} C_1 \gamma^{\lfloor t/\Delta \rfloor}. \tag{3.3.72}$$

The convergence of $L_{t+1,i}$ to zero can be proved using Theorem 3.7. Since the quantity $C_1 \gamma^{\lfloor t/\Delta \rfloor}$ is deterministic, its mean is itself and its variance is zero. So in (3.3.13) and (3.3.14), we can define

$$B_t^L := C_1 \gamma^{\lfloor t/\Delta \rfloor}, M_t^L := 0 \; \forall t.$$

We can substitute these definitions into (3.3.15) and (3.3.16), and define

$$f_\tau^L = b_\tau^2 (1 + 2\mu_{\nu^{-1}(\tau)}^2) + 3 b_\tau \mu_{\nu^{-1}(\tau)}, \tag{3.3.73}$$

$$g_\tau^L = b_\tau^2 (2\mu_{\nu^{-1}(\tau)}^2) + b_\tau \mu_{\nu^{-1}(\tau)}. \tag{3.3.74}$$

Since $\alpha_t \in [0,1]$ and the sequence $\{B_t^L\}$ is summable (because $\gamma < 1$), and $M_t^L \equiv 0$, (3.3.17) is satisfied. Also, by Assumption (S2), (3.3.18) is satisfied. Hence $L_{t+1,i} \to 0$ as $t \to \infty$, which in turn implies that $\bar{B}_{t+1,i} \to 0$ as $t \to \infty$.

Finally, we come to $C_{t+1,i}$. It is evident from (3.3.59) and Lemma 3.3 that $C_{t+1,i}$ satisfies the recursion

$$C_{t+1,i} = (1-\alpha_{t,i}) C_{t,i} + \alpha_{t,i} Ł_t \zeta_{t,i}. \tag{3.3.75}$$

Now observe that $Ł_t$ is bounded, and the rescaled error signal $\zeta_{t+1,i}$ satisfies (3.3.45) and (3.3.46). Hence, if $Ł^*$ is a bound for $Ł_t$, then it follows from (3.3.45) and (3.3.46) that

$$|E_t(Ł_t \zeta_{t+1,i})| \leq c_2 Ł^* B_t, \; \forall t \geq 0, CV_t(Ł_t \zeta_{t+1,i}) \leq c_3 Ł^* M_t^2, \; \forall t \geq 0, \tag{3.3.76}$$

Hence, when Assumptions (S1) and (S2) hold, it follows from Theorem 3.7 that $C_{t+1,i} \to 0$ as $t \to \infty$. □

Next, we describe three different ways of choosing the update processes $\{\kappa_{t,i}\}$.

**Bernoulli Updating:** For each $i \in [d]$, choose a rate $b_i \in (0,1]$, and let $\{\kappa_{t,i}\}$ be a Bernoulli process such that

$$\Pr\{\kappa_{t,i} = 1\} = b_i, \; \forall t.$$

Moreover, the processes $\{\kappa_{t,i}\}$ and $\{\kappa_{t,j}\}$ are independent whenever $i \neq j$. Let $\nu_{t,i}$, the counter process for coordinate $i$, be defined as usual. Then it is easy to see that $\nu_{t,i}/t \to b_i$ as $t \to \infty$, for each $i \in [d]$. Thus Assumption (U2) is satisfied for each $i \in [d]$.

**Markovian Updating:** Suppose $\{Y_t\}$ is a sample path of an irreducible Markov process on the state space $[d]$. Define the update process $\{\kappa_{t,i}\}$ by

$$\kappa_{t,i} = I_{\{Y_t = i\}} = \begin{cases} 1, & \text{if } Y_t = i, \\ 0, & \text{if } Y_t \neq i. \end{cases}$$

Let $\boldsymbol{\mu}$ denote the stationary distribution of the Markov process. Then the ratio $\nu_{t,i}/t \to \mu_i$ as $t \to \infty$, for each $i \in [d]$. Hence once again Assumption (U2) holds.

**Batch Markovian Updating:** This is an extension of the above. Instead of a single Markovian sample path, there are $N$ different sample paths, denoted by $\{Y_t^n\}$ where $n \in [N]$. Each sample path $\{Y_t^n\}$ comes an irreducible Markov process over the state space $[d]$, and the dynamics of different Markov processes could be different (though there does not seem to be any advantage to doing this). The update process is now given by

$$\kappa_{t,i} = \sum_{n \in [N]} I_{\{Y_t^n = i\}}.$$

Define the counter process $\nu_{t,i}$ as before, and let $\boldsymbol{\mu}^n$ denote the stationary distribution of the $n$-th Markov process. Then

$$\frac{\nu_{t,i}}{t} \to \sum_{n \in [N]} \mu_i^n.$$

Hence once again Assumption (U2) holds.

Now we establish convergence rates under each of the above updating methods (and indeed, any method such that Assumption (U2) is satisfied). The proof of Theorem 3.11 gives us a hint on how this can be done. Specifically, each of the entities $\bar{A}_{t+1,i}, L_{t+1,i}, C_{t+1,i}$ satisfies a stochastic recursion, whose rate of convergence can be established using Theorems 3.8 and 3.9. These theorems apply to *scalar-valued* stochastic processes with intermittent updating. In principle, when updating $\boldsymbol{\theta}_t$, we could use a mixture of global and local clocks for different components. However, in our view, this would be quite unnatural. Instead, it is assumed that for *every* component, either a global clock or a local clock is used. Recall also the bounds (3.3.38) and (3.3.39) on the error $\boldsymbol{\xi}_{t+1}$.

**Theorem 3.12.** *Suppose a local clock is used, so that $\alpha_{t,i} = \beta_{\nu_{t,i}}$ for each $i$ that is updated at time $t$. Suppose that $\{B_t\}$ is nonincreasing; that is, $\mu_{t+1} \leq B_t$, $\forall t$, and $M_t$ is uniformly bounded, say by $M$. Suppose in addition that $\beta_t = O(t^{-(1-\phi)})$, for some $\phi > 0$, and $\beta_t = \Omega(t^{-(1-C)})$ for some $C \in (0, \phi]$. Suppose that $B_t = O(t^{-\epsilon})$ for some $\epsilon > 0$. Then $\boldsymbol{\theta}_\tau \to 0$ as $\tau \to \infty$ for all $\phi < \min\{0.5, \epsilon\}$. Further, $\boldsymbol{\theta}_\tau = o(\tau^{-\lambda})$ for all $\lambda < \epsilon - \phi$. In particular, if $B_t = 0$ for all $t$, then $\boldsymbol{\theta}_\tau = o(\tau^{-\lambda})$ for all $\lambda < 1$.*

The proof of the rate of convergence uses Item (3) of Theorem 3.7. In the proof, let us ignore the index $i$ wherever possible, because the subsequent analysis applies to each index $i$. Recall that $\bar{A}_{t+1,i}$ is defined in (3.3.69). Since $\ln(1-x) \leq -x$ for all $x \in (0,1)$, it follows that

$$\ln \prod_{k=0}^{t} (1 - \alpha_{k,i}) \leq -\sum_{k=0}^{t} \alpha_{k,i},$$

where $\alpha_{k,i} = 0$ unless there is an update at time $k$. Now, since a local clock is used, we have that $\alpha_{k,i} = \beta_{\nu_{k,i}}$ whenever there is an update at time $k$. Therefore

$$\sum_{k=0}^{t} \alpha_{k,i} = \sum_{s=0}^{\nu_{t,i}} \beta_s$$

Now, if Assumption (U2) holds (which it does for each of the three types of updating considered), it follows that $\nu_{t,i} \approx t/r$ for large $t$. Thus, if $\beta_\tau = \Omega(\tau^{-(1-C)})$, then we can reason as follows:

$$\sum_{s=0}^{\nu_t} \beta_s \approx \sum_{s=0}^{t/r} s^{-(1-C)} \approx (t/r)^C.$$

Therefore, for large enough $t$, we have that

$$\prod_{k=0}^{t} (1 - \alpha_k) \leq \exp(-(t/r)^C).$$

It follows from (3.3.69) that $\bar{A}_{t+1,i} \to 0$ *geometrically fast.*

Next we come to $\bar{B}_{t+1,i}$, which is bounded by $L_{t+1,i}$, as defined in (3.3.72). Recall the definitions (3.3.73) and (3.3.74) for the sequences $\{f_\tau^L\}$ and $\{g_\tau^L\}$. Then (3.3.17) and (3.3.18) will hold whenever $C > 0$. Since Assumption (U2) holds, we have that

$$\mu^L_{\nu^{-1}(\tau)} = C_1 \gamma^{\lfloor \nu^{-1}(\tau)/\Delta \rfloor} \leq C_2 \gamma^{r'\tau}$$

for suitable constants $C_2$ and $r'$. The point to note is that the sequence $\{C_2\gamma^{r'\tau}\}$ is a *geometrically convergent* sequence because $\gamma < 1$. Therefore (3.3.19) holds for *every* $\lambda > 0$. Also, (3.3.20) holds for all $C > 0$. Hence it follows from Item (3) of Theorem 3.7 that $L_{t+1,i} = o(t^{-\lambda})$ for every $\lambda > 0$.

This leaves only $C_{t+1,i}$. We already know that $C_{t+1,i}$ satisfies the recursion (3.3.75). Moreover, the modified error sequence $\{\mathrm{L}_t\zeta_{t,i}\}$ satisfies (3.3.76). The estimates for the rate of convergence now follow from Item (3) of Theorem 3.7, and need not be discussed again.

**Theorem 3.13.** *Suppose a global clock is used, so that $\alpha_{t,i} = \beta_{t,i}$ whenever the $i$-th component of $\boldsymbol{\theta}_t$ is updated. Suppose that $\beta_t$ is nonincreasing, so that $\beta_{t+1} \leq \beta_t$ for all $t$. Suppose in addition that $\beta_t = O(t^{-(1-\phi)})$, for some $\phi > 0$, and $\beta_t = \Omega(t^{-(1-C)})$ for some $C \in (0, \phi]$. Suppose that $B_t = O(t^{-\epsilon})$ for some $\epsilon > 0$, and $M_t = O(t^\delta)$ for some $\delta \geq 0$. Then $\boldsymbol{\theta}_t \to 0$ as $t \to \infty$ whenever*

$$\phi < \min\{0.5 - \delta, \epsilon\}.$$

*Moreover, $\boldsymbol{\theta}_t = o(t^{-\lambda})$ for all $\lambda < \epsilon - \phi$. In particular, if $B_t = 0$ for all $t$, then $\boldsymbol{\theta}_t = o(t^{-\lambda})$ for all $\lambda < 1$.*

The proof is omitted as it is very similar to that of Theorem 3.12.

## 3.4 Variants of Standard Stochastic Approximation

### 3.4.1 Averaged Stochastic Approximation

Papers by Ruppert [128], Polyak [116], Polyak and Juditsky [117] and Nemirovski et al. [106].

An important variant of standard SA is the so-called "averaged" SA, pioneered in [128, 116] and developed further in [117, 106]. The idea is simply to *average* the iterations of a standard SA algorithm. Specifically, let $\{\boldsymbol{\theta}_t\}$ denote the sequence produced by an SA algorithm (the specific nature of which is not important for the moment), and define

$$\bar{\boldsymbol{\theta}}_t = \frac{1}{t} \sum_{\tau=1}^{t} \boldsymbol{\theta}_\tau.$$

Note that $\bar{\boldsymbol{\theta}}_t$ can be computed iteratively from $\boldsymbol{\theta}_t$ via

$$\bar{\boldsymbol{\theta}}_{t+1} = \frac{t}{t+1}\bar{\boldsymbol{\theta}}_t + \frac{1}{t+1}\boldsymbol{\theta}_{t+1}.$$

Hence "averaging" can also be viewed as a two-step iterative algorithm: The first step is to generate $\boldsymbol{\theta}_{t+1}$ from $\boldsymbol{\theta}_t$, and the second step is to generate $\bar{\boldsymbol{\theta}}_{t+1}$ from $\bar{\boldsymbol{\theta}}_t$ and $\boldsymbol{\theta}_{t+1}$ as above. In [128, 116], the asymptotic covariance of the matrix $t^{-1}(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^\top \in \mathbb{R}^{d\times d}$ is computed. For linear stochastic approximation problems, it is shown that this quantity converges to the "lowest possible" covariance matrix, which depends on the (unknown) parameters of the problem. The key point is that the iterative algorithm does not assume knowledge of these parameters, but achieves "asymptotically optimal" scaled covariance as $t \to \infty$. In [117], it is shown that the quantity $t^{-1}(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^\top$ is asymptotically multivariate normal, and of course, the covariance matrix is again "optimal." In [106], the analysis is extended to convex objective functions, thus relaxing the assumption of strong convexity assumed in [117] and its predecessors.

### 3.4.2 Two Time Scale Stochastic Approximation

Papers by Borkar [21], and by Lakshminarayanan and Bhatnagar [88]

### 3.4.3 Finite-Time Stochastic Approximation

Some relevant papers are [170, 141, 13, 32, 120, 56, 57]. Forthcoming survey paper by Chen and Maguluri.

### 3.4.4 Markovian Stochastic Approximation

Some relevant references are [90, 33, 120, 19, 140].

## Notes and References

The Stochastic Approximation method is introduced in a seminal paper by Robbins and Monro [123] with that same title, for finding a solution to a scalar equation $f(x) = 0$, where $f : \mathbb{R} \to \mathbb{R}$, when only noisy measurements of $f(\cdot)$ are available.

While [123] is seminal, the results are quite restrictive by today's standards:

- The function $f(\cdot)$ is globally bounded; see [123, Eq. (5)].
- The measurement error $\xi_t$ is assumed to have zero conditional mean; see [123, Eq. (3)].
- The measurement error $\xi_t$ is assumed to be bounded; see [123, Eq. (4)].
- The convergence is only in probability, and not almost sure.

Despite all of these restrictions, the paper can be credited with having started a new area of research. Interestingly, the paper does not have a single reference, suggesting that there was very little by way of precedent for the method. The phrase "Stochastic Approximation" comes from this paper, as to the conditions (3.1.3) and (3.1.4).

Shortly after the publication of [123], Kiefer and Wolfowitz [77] extended the results of Robbins-Monro for finding a stationary point of a smooth function $J : \mathbb{R} \to \mathbb{R}$. For this purpose, they replaced the true gradiend $J'(\cdot)$ by a first-order approximation of the form (4.2.10) (in the next chapter). They realized two technical challenges posed by their formulation, namely: The measurement error $\xi_t$ is "biased" in that its conditional expectation is *not* zero, and its conditional variance grows without bound as $t$ increased. In [17], Blum extended the approach of Kiefer-Wolfowitz to maps $J : \mathbb{R}^d \to \mathbb{R}^d$. His formulation also had the same technical difficulties as Kiefer-Wolfowitz. In [45], Dvoretzky presents a formulation of stochastic approximation that contains both the Robbins-Monro and Kiefer-Wolfowitz formulations as special cases.

In all these cases, the authors suggest various workarounds, but not a general theory. Moreover, the convergence is only in probability, and not almost sure (with the exception of [17]). In the opinion of the present author, the first approach that was capable of being generalized further is given gy Gladyshev [52]. His approach consisted of carrying out an affine transformation of the stochastic process $\{\boldsymbol{\theta}_t\}$ in such a way that the transformed process is a nonnegative supermartingale, and hence converged almost surely to a limit. By inverting the affine transformation, it followed that $\boldsymbol{\theta}_t$ also converged almost surely to a limit. Some further analysis established that the limit was indeed the desired solution. In the original paper, the function $\mathbf{f}(\cdot)$ is assumed to be "passive," that is, there exists a function $c(\cdot)$ belonging to Class $\mathcal{B}$ such that

$$\langle f(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \geq c(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2), \ \forall \boldsymbol{\theta} \in \mathbb{R}^d,$$

where $\boldsymbol{\theta}^*$ is the unique solution of $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{0}$. Subsequent analysis in [168] showed that the key attribute of $\mathbf{f}(\cdot)$ is not passivity, but the global asymptotic stability of the associated ODE $\dot{\boldsymbol{\theta}} = \mathbf{f}(\boldsymbol{\theta})$. If $\mathbf{f}(\cdot)$ is passive, then $V(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$ is a suitable Lyapunov function. In addition to this, Gladyshev was the first to establish

a "division of labor" whereby the square summability of the step sizes is sufficient to ensure the almost sure boundedness of the iterations, while the additional assumption that the sum of the step sizes diverges ensures convergence to the desired limit.

A slightly later, but independent, development is the Robbins-Siegmund theorem [124]. It partially subsumes the results of Gladyshev; moreover, the theory applies even to the situation where the step sizes $\alpha_t$ are random, whereas the theory of Gladyshev does not. In the opinion of the present author, Gladyshev's work is not sufficiently well-known in the Western research community, and that most contemporary researchers cite the Robbins-Siegmund theorem. Interestingly, in [9, pp. 343–344], the authors give a simple proof of the Robbins-Siegmund theorem, and use it to prove the almost sure convergence of the SA algorithm under the passivity condition used in [52]. However, since the proof in [9] uses the notion of a stopping time process, which is not used in this book, we give the original (and longer) proof for the Robbins-Siegmund theorem.

The convergence theorems for standard SA are taken from [70, 71]. Some recent results on Stochastic approximation can be found in [69, 99].

The next stage in the evolution of Stochastic Approximation theory is the formulation of the so-called ODE approach. This approach began in the early 1970s in the erstwhile Soviet Union and in the Western world. In the USSR, among the first papers were [101, 102], in which the author derived sufficient conditions under which the solutions of a *stochastic* difference equation can be approximated by the solution of an associated *deterministic* difference equation. In the general case, the approximation is uniformly good over a finite interval. If it is assumed that solution of the original stochastic difference equation is bounded over time, then the approximation is uniformly good over an infinite time interval. In these references, the step size is fixed, and successive measurement errors are assumed to be independent. In [38], Derevitskii and Fradkov bound the error between the trajectories of the stochastic process $\{\boldsymbol{\theta}_t\}$ and the solutions of the associated ODE $\dot{\boldsymbol{\theta}} = \mathbf{f}(\boldsymbol{\theta})$, (as opposed to another difference equation as in Meerkov's work). They assume that the noise sequence $\{\boldsymbol{\xi}_{t+1}\}$ is i.i.d., but permit *time-varying* step sizes. Their theory works whether or not the step sizes approach zero. In [84], Kushner analyzes the Kiefer-Wolfowitz version of Stochastic Approximation; that is, he tackles the fact that the error is biased and has variance that grows without bound as $t$ increases. He permits the errors to be correlated (unlike [38]), and derives an expression for the limiting behavior in terms of an *integral* equation.[6] However, Kushner establishes only convergence in probability. This work is followed by Ljung in [94, 95]. In these references, as in [84], the noise $\boldsymbol{\xi}_{t+1}$ is allowed both to be biased and also to have unbounded variance. But unlike in Kushner's paper, Ljung establishes almost sure convergence. Ljung explicitly mentions the limit ODE in [94]. In [95], he shows that the square summability of the step size sequence can be relaxed, if the noise sequence has finite moments of order greater than two.

In Section 3.3, we have analyzed Asynchronous SA as well as Block Asynchronous SA (BASA). Perhaps the first paper to analyze the behavior of stochastic algorithms when the vector $\boldsymbol{\theta}_t$ is updated in an asynchronous fashion is [159]. The first papers to study Stochastic Approximation when only one component of $\boldsymbol{\theta}_t$ is updated at each $t$ (that is, ASA) are [158] and [66]. In [158], the emphasis is on proving the convergence of the $Q$-learning algorithm, which is introduced here in Chapter 6, while in [66], the emphasis is on proving the convergence of a version of the $TD(\lambda)$ algorithm for computing the value of a Markov Reward Process. This algorithm is also introduced in Chapter 6. In the present Section 3.3, we have abstracted the essence of the proof in [158], and have also permitted *Block* updating. The material in Section 3.3 on Block Asynchronous Stochastic Approximation (BASA) is taken from [73, 72].

Notes and References for the material in Section 3.4 will be added once the section is written.

The approach taken here for proving the convergence of the SA algorithm is based on the Robbins-Siegmund theorem (Theorem 2.22), and might be referred to as the "martingale approach." Note that there is another very popular technique for analyzing the convergence of the SA algorithm, known popularly as the "ODE method." Since the ODE approach is also widely used, we give a very brief summary of some of the key aspects of this approach.

[6] Obviously, the integral equation can be equivalently expressed as an ODE, and vice versa. But every other paper mentions a limit ODE, and not a limit integral equation.

In the ODE method, the key step is to show that, as time progresses, the sample path of the iterations $\{\boldsymbol{\theta}_t\}$ begins to resemble the *deterministic* solution trajectory of an associated ODE

$$\dot{\boldsymbol{\theta}} = \mathbf{f}(\boldsymbol{\theta}).$$

Some relevant references for the ODE approach are [38, 94, 95, 85, 103, 9, 86, 7, 22, 24]. If the SA iterations $\{\boldsymbol{\theta}_t\}$ are bounded almost surely (a property called "stability"), and a few technical assumptions hold, then the iterations $\boldsymbol{\theta}_t$ converge to the *set of solutions* of the equations $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{0}$. In particular, if $\boldsymbol{\theta}^*$ is the unique globally attractive equilibrium of this associated ODE, then it can be shown that $\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*$ almost surely as $t \to \infty$, again under suitable technical assumptions. The books [7, 22, 24] describe the ODE approach in full generality, and the interested reader may consult these authoritative resources.

The ODE approach is more general than the "martingale" approach put forward here, in that the ODE approach is applicable even when the equation $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{0}$ has multiple solutions. However, much of the theory is based on the assumption that the SA iterations $\{\boldsymbol{\theta}_t\}$ are bounded almost surely. Often this latter assumption can b e validated using different methods.

A major breakthrough in the ODE approach is contained in the paper [25], in which the *almost sure boundedness of the iterations is a conclusion and not a part of the hypotheses.* Specifically, the authors define another vector field $\mathbf{f}_\infty$ as follows:

$$\mathbf{f}_\infty(\boldsymbol{\theta}) := \lim_{r\to\infty} \frac{\mathbf{f}(r\boldsymbol{\theta})}{r},\ \forall \boldsymbol{\theta} \in \mathbb{R}^d.$$

It is assumed that $\mathbf{0}$ is a globally asymptotically stable equilibrium of the associated ODE

$$\dot{\boldsymbol{\theta}} = \mathbf{f}_\infty(\boldsymbol{\theta}).$$

If this assumption holds, then the authors prove that $\boldsymbol{\theta}_t$ converges to the (unique) solution $\boldsymbol{\theta}^*$ almost surely as $t \to \infty$.

While it is undoubtedly a major improvement to make the almost sure boundedness of the iterations as a conclusion and not a hypothesis, the above assumption contains a subtle limitation of the approach. Specifically, if the function $\mathbf{f}(\cdot)$ *grows sublinearly* in the sense that

$$\lim_{r\to\infty} \frac{\mathbf{f}(r\boldsymbol{\theta})}{r} \equiv \mathbf{0},\ \forall \boldsymbol{\theta} \in \mathbb{R}^d,$$

then it is clear that $\mathbf{f}_\infty(\boldsymbol{\theta}) \equiv \mathbf{0}$ for all $\boldsymbol{\theta}$, and the associated ODE cannot have $\mathbf{0}$ as a globally asymptotically stable equilibrium. In particular, if the function $\mathbf{f}(\cdot)$ is globally bounded, then the results of [25] do not apply. In contrast, the martingale approach can cope with sublinearly growing functions $\mathbf{f}(\cdot)$ without any difficulties.

To summarize, the martingale approach and the ODE approach have complementary strengths and weaknesses. Much of the time, both approaches are applicable to the problem at hand. But there are some situations where one approach is applicable but not the other.

# Chapter 4

# Applications to Optimization

In this chapter, we apply the ideas introduced in the preceding chapters to:

- Identify some important classes of nonconvex functions to which these ideas are applicable.
- State several commonly-used algorithms for both convex and nonconvex optimization.
- State and prove theorems on the convergence of these algorithms, as well as their rates of convergence to a solution.

## 4.1 Some Invex Functions

The reader is reminded that in this book, we study only *unconstrained* optimization; see Section 1.1.1. Also, throughout this chapter, we make two "standing" assumptions, which are standard in the literature. Note that $J(\cdot)$ denotes the objective function.

(J1) $J(\cdot)$ is $\mathcal{C}^1$, and $\nabla J(\cdot)$ is globally Lipschitz-continuous with constant $L$.

(J2) $J(\cdot)$ is bounded below. Thus

$$J^* := \inf_{\boldsymbol{\theta} \in \mathbb{R}^d} J(\boldsymbol{\theta}) > -\infty.$$

However, it is *not* assumed that the infimum $J^*$ is attained. For instance, the function $J(\theta) = \exp(-\theta)$ satisfies the standing assumptions. Hereafter, to simplify notation, we replace $J(\cdot)$ by $J(\cdot) - J^*$, which enables us to assume that $J^* = 0$, without any loss of generality. When the infimum is indeed attained, we define the set

$$S_J := \{\boldsymbol{\theta} \in \mathbb{R}^d : J(\boldsymbol{\theta}) = 0\}, \tag{4.1.1}$$

and observe that it is a closed, nonempty set. Moreover, the quantity

$$\rho_J(\boldsymbol{\theta}) := \inf_{\boldsymbol{\phi} \in S_J} \|\boldsymbol{\phi} - \boldsymbol{\theta}\|_2 \tag{4.1.2}$$

is well-defined, and is referred as the "distance" to $S_J$.

The algorithms studied in this chapter are stochastic versions of the gradient descent algorithm, and various versions of momentum-based algorithms. The deterministic versions of these algorithms are very briefly discussed in Section 1.1.3. Indeed, Stochastic Gradient Descent (SGD) is the most widely used method for training very large neural networks. The material in Section 4.2.1 is motivated by this application.

The topic of this chapter is nonconvex optimization. Specifically, we are interested in finding (if possible) *global* mimizers of an objective function $J(\cdot)$. As stated in Lemma 1.1, if $J(\cdot)$ is convex, then *every* stationary point (i.e., a $\boldsymbol{\theta}^*$ such that $\nabla J(\boldsymbol{\theta}^*) = \mathbf{0}$) is a global minimizer. The converse is always true, for *any* $\mathcal{C}^1$ objective

function, convex or nonconvex. That is, if $\boldsymbol{\theta}^*$ is a global (or even local) minimizer of $J(\cdot)$, then $\nabla J(\boldsymbol{\theta}^*) = \mathbf{0}$. Thus, for smooth convex objective functions $J(\cdot)$, we have that $\nabla J(\boldsymbol{\theta}^*) = \mathbf{0}$ is both a necessary as well as a sufficient condition for $\boldsymbol{\theta}^*$ to be a global minimizer.

Over the decades, several attempts have been made to find classes of functions that satisfy the property that every stationary point is also a global minimumizer. Obviously the intent is to go beyond just convex functions. Among many such classes of functions, we will focus on one specific class, known as "invex" functions.

**Definition 4.1.** A $\mathcal{C}^1$ function $J : \mathbb{R}^d \to \mathbb{R}$ is said to be **invex** if there exists a map $\boldsymbol{\eta} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ such that

$$J(\boldsymbol{\phi}) \geq J(\boldsymbol{\theta}) + \langle \boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{\phi}), \nabla J(\boldsymbol{\theta}) \rangle, \ \forall \boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^d. \tag{4.1.3}$$

One can also say that $J(\cdot)$ is invex *with respect to* a particular function $\boldsymbol{\eta}(\cdot)$ if (4.1.3) holds, because the bound might hold for one choice of $\boldsymbol{\eta}(\cdot)$ but not another. It is also possible to modify the definition slightly and say that $J(\cdot)$ is **invex on** $S \subseteq \mathbb{R}^d$ if (i) $\boldsymbol{\eta} : S \times S \to \mathbb{R}^d$, and (ii) (4.1.4) holds only for all $(\boldsymbol{\theta}, \boldsymbol{\phi}) \in S \times S$.

Observe that if $J_1$ and $J_2$ are invex *with the same function* $\boldsymbol{\eta}(\cdot)$, then so is $c_1 J_1 + c_2 J_2$ for any nonnegative constants $c_1, c_2$. Thus, for a fixed function $\boldsymbol{\eta}(\cdot)$, the set of functions that are invex with respect to $\boldsymbol{\eta}(\cdot)$ is a convex cone. Since we won't use the concept of a convex cone in this book, we do not pursue this matter further.

The above definition is introduced in [59]. However, the phrase "invex" is not used therein, but is introduced in [35]. Note that if $J(\cdot)$ is convex, then we can take

$$\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{\phi}) := \boldsymbol{\phi} - \boldsymbol{\theta}.$$

However, we shall see below several examples of invex functions that are not convex.

It is obvious from (4.1.3) that if $J(\cdot)$ is invex, and if $\nabla J(\boldsymbol{\theta}^*) = \mathbf{0}$, then $J(\boldsymbol{\theta}) \geq J(\boldsymbol{\theta}^*)$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$. Thus $\boldsymbol{\theta}^*$ is a global minimizer of $J(\cdot)$. Note that the nature of the function $\boldsymbol{\eta}(\cdot)$ plays no role in this observation. More generally, if $J(\cdot)$ is invex on $S$, and if $\boldsymbol{\theta}^* \in S$ satisfies $\nabla J(\boldsymbol{\theta}^*) = \mathbf{0}$, then $\boldsymbol{\theta}^*$ is a minimizer of $J(\cdot)$ *over the set* $S$. A remarkable result from [36] states that the converse is also true: If every stationary point of $J(\cdot)$ is also a global minimizer, then *there exists* a function $\boldsymbol{\eta}(\cdot)$ such that $J(\cdot)$ is an invex function with respect to $\boldsymbol{\eta}(\cdot)$. See [36, Eq. (9)] and the text thereafter. These results derive $\boldsymbol{\eta}(\cdot)$ in terms of a Lagrangian dual problem, and thus do not readily lead to "explicit" formulas for $\boldsymbol{\eta}(\cdot)$.

This result suggests that we should be studying the minimization of invex functions. However, in the present context, the invexity property alone is not sufficient. This is because we wish to establish not merely that every stationary point is also a global minimizer, but something more, namely that the SGD algorithm converges for all functions of a particular class. For this purpose, we introduce two other classes of functions, denoted by (PL) and (KL'). Both classes are subsets of the class of invex functions. Thus, by the results of [36], for each such function there exists a corresponding function $\boldsymbol{\eta}(\cdot)$ such that (4.1.3) holds. However, it is not straightforward to actually compute $\boldsymbol{\eta}(\cdot)$. For the theory below, this is not a limitation, because the function $\boldsymbol{\eta}(\cdot)$ does not play any role. Thus, at least for the purposes of this book, the results of [36] are strictly of academic interest. For functions of Class (KL'), we establish the convergence of SGD, but without any rates. For functions of Class (PL), we not only establish the convergence of SGD, but also derive estimates on the *rate* of convergence.

For a comprehensive discussion of invexity, Class (PL), and Class (KL) (a forerunner of Class (KL')), the reader is directed to [74]. For applications to the convergence of SGD, the reader may consult [70, 71].

**Definition 4.2.** Suppose $J : \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{C}^1$ and satisfies the standing assumptions (J1) and (J2). Assume without loss of generality that $J^*$, the infimum of $J(\cdot)$, equals zero.

(PL) The function $J(\cdot)$ is said to belong to the class (PL) if there exists a constant $K$ such that

$$\|\nabla J(\boldsymbol{\theta})\|_2^2 \geq K J(\boldsymbol{\theta}), \ \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{4.1.4}$$

(KL') The function $J(\cdot)$ is said to belong to the class (PL) if there exists a function $\psi(\cdot)$ of Class $\mathcal{B}$ such that

$$\|\nabla J(\boldsymbol{\theta})\|_2 \geq \psi(J(\boldsymbol{\theta})),\ \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{4.1.5}$$

Now we discuss the origin and the significance of these concepts.

PL stands for the Polyak-Łojasiewicz condition. In [114], Polyak introduced (4.1.4), and showed that it is sufficient to ensure that iterations converge at a "linear" (or geometric) rate to a global minimum, whether or not $J(\cdot)$ is convex. Note that (4.1.4) can also be rewritten as

$$\|\nabla J(\boldsymbol{\theta})\|_2 \geq K^{1/2}[J(\boldsymbol{\theta})]^{1/2},\ \forall \boldsymbol{\theta} \in \mathbb{R}^d.$$

To place the (PL) property in context, let us recall the definition of strong convexity. The function $J(\cdot)$ is said to be $R$**-strongly convex** if there exists a constant $R > 0$ such that

$$J(\boldsymbol{\phi}) \geq J(\boldsymbol{\theta}) + \langle \nabla J(\boldsymbol{\theta}), \boldsymbol{\phi} - \boldsymbol{\theta} \rangle + \frac{R}{2}\|\boldsymbol{\phi} - \boldsymbol{\theta}\|_2^2.$$

See for example [109, Section 2.1.3]. In this case, $J(\cdot)$ has a unique global minimizer, call it $\boldsymbol{\theta}^*$. Again, let us assume that $J^* = J(\boldsymbol{\theta}^*) = 0$. Then we can apply [108, Eq. (2.1.24)] with $f = J$, $x = \boldsymbol{\theta}^*$, $y = \boldsymbol{\theta}$, and $\mu = R$, which gives

$$J(\boldsymbol{\theta}) \leq \frac{1}{2R}\|\nabla J(\boldsymbol{\theta})\|_2^2.$$

Thus an $R$-strongly convex function satisfies (PL) with $K = 2R$. Therefore one can think of the (PL) property as a generalization of this particular property of strongly convex functions.

As shown in Lemma 4.1 below, whenever $J(\cdot)$ is $\mathcal{C}^1$, $\nabla J(\cdot)$ is $L$-Lipschitz continuous, and $J^*$ is a lower bound for $J(\cdot)$, it is the case that

$$\|\nabla J(\boldsymbol{\theta})\|_2^2 \leq 2L(J(\boldsymbol{\theta}) - J^*).$$

In particular, by redefining $J(\cdot)$ if necessary, we can take $J^* = 0$, in which case we have $\|\nabla J(\boldsymbol{\theta})\|_2^2 \leq 2LJ(\boldsymbol{\theta})$. The PL condition is the *inverse* of the above observation, in the sense that $\|\nabla J(\boldsymbol{\theta})\|_2^2$ is bounded *below* by a constant multiple of $J(\boldsymbol{\theta})$.

On the other hand, the class (PL) is strictly larger than the class of strongly convex functions; it also contains some *nonconvex* functions.

**Example 4.1.** Define

$$J(\theta) = \theta^2 + 3\sin^2\theta.$$

A plot of $\theta$ versus $J(\theta)$ is shown in Figure 4.1. The figure shows both $J(\theta)$ as well as the ratio $(\nabla J(\theta))^2/J(\theta)$ as functions of $\theta$.[1] Since $J(\cdot)$ is an even function, the plot is shown only for $\theta \geq 0$. It can be verified numerically that $J(\cdot)$ is not convex, but satisfies the (PL) property with $K = 0.3511$.

In [96], Łojasiewicz introduced a more general condition

$$\|J(\boldsymbol{\theta})\|_2 \geq C[J(\boldsymbol{\theta})]^r,\ \forall \boldsymbol{\theta} \in \mathbb{R}^d, \tag{4.1.6}$$

for some constant $C$ and some exponent $r \in [1/2, 1)$. He also showed that (4.1.6) holds for real algebraic varieties in a neighborhood of critical points.

In [83], Kurdyka proposed a more general inequality than (4.1.6), namely: There exist a constant $c > 0$ and a function $v : [0, c) \to \mathbb{R}$ which is $\mathcal{C}^1$ on $(0, c)$, such that $v'(x) > 0$ for all $x \in (0, c)$, and

$$\|\nabla(v \circ J)(\boldsymbol{\theta})\|_2 \geq 1,\ \forall \boldsymbol{\theta} \in J^{-1}(0, c), \tag{4.1.7}$$

where (only on this occasion) $\circ$ denotes the composition of two functions. By applying the chain rule, one can rewrite (4.1.7) as

$$\|\nabla J(\boldsymbol{\theta})\|_2 \geq [v'(J(\boldsymbol{\theta})]^{-1}. \tag{4.1.8}$$

[1] Since $d = 1$, we can use $J'(\theta)$ instead of $\nabla J(\theta)$. But we use $\nabla J(\theta)$ to be consistent.

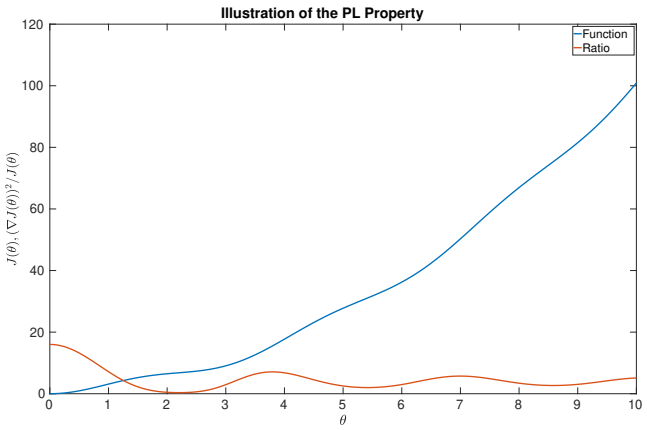



Figure 4.1: An Example of a function in the Class (PL): $J(\theta) = \theta^2 + 3\sin^2\theta$

In particular, if $v(x) = x^{1-r}$ for some $r \in (0,1)$, then (4.1.8) becomes (4.1.6) with $C = 1/(1-r)$. For this reason, (4.1.8) is sometimes referred to as the Kurdyka-Łojasiewicz (KL) inequality. See for example [18]. In our case, we don't require the right side to be a differentiable function; rather we require only that it be a function of Class $\mathcal{B}$ of $J(\boldsymbol{\theta})$. Hence we choose to call this condition as (KL'), to suggest that it is similar to, but weaker than, the KL condition.

**Example 4.2.** Consider an even function $J : \mathbb{R} \to \mathbb{R}$ defined by

$$J * \theta) = \begin{cases} \theta^2 + 4\sin^2\theta, & 0 \le \theta \le 5, \\ J(5) + 0.5J'(5)(1 - \exp(-2(\theta - 5))), & \theta > 5, \\ J(-\theta), & \theta < 0. \end{cases}$$

A plot of $J(\theta)$ and of $(\nabla J(\theta))^2/J(\theta)$ are shown in Figure 4.2. Again, since $J(\cdot)$ is an even function, the plot is shown only for $\theta \ge 0$. From this it can be seen (and it is also readily verified) that, though the ratio $(\nabla J(\theta))^2/J(\theta) \to 0$ as $\theta \to \infty$, the ratio is never actually zero. Thus $(\nabla J(\theta))^2/J(\theta)$ is a function of Class $\mathcal{B}$. As a result, this function satisfies the property (KL').

It is clear from the definition of both (PL) and (KL') classes that, if $\nabla J(\boldsymbol{\theta}) = \mathbf{0}$, then $J(\boldsymbol{\theta}) = 0$, which is the global minimum. Hence, by the result of [36], every Class (PL) function and every (KL') function is invex. This proof is rather indirect, and it would be desirable to have a more direct proof of this fact.

## 4.2 Review of Some Standard Algorithms

In this section, we briefly survey a few standard algorithms for convex optimization. The convergence of these algorithms is not discussed, as (for the most part) the convergence can be inferred from the results for nonconvex optimization in Section 4.3 The reader is directed to an excellent survey paper [26] that discusses many issues not covered here, with an emphasis on applications to machine learning.

### 4.2.1 Stochastic Gradient Descent

Recall the Gradient Descent algorithm, also known as steepest descent, described in (1.1.10). In the **Stochastic Gradient Descent (SGD)** algorithm, the true gradient $\nabla J(\boldsymbol{\theta}_t)$ is replaced by a *random* vector $\mathbf{h}_{t+1}$,

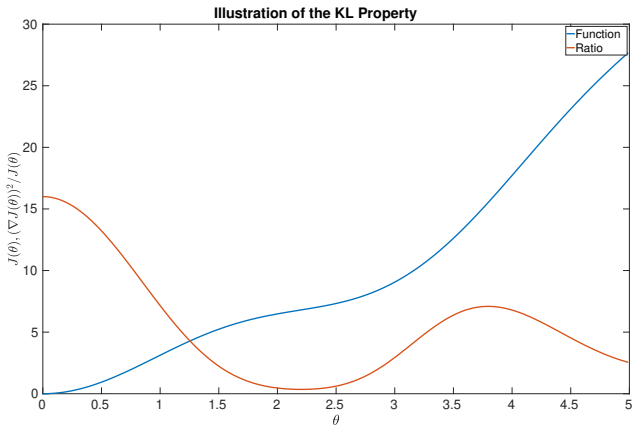



Figure 4.2: An example of a function of Class (KL')

which is supposed to approximate $\nabla J(\boldsymbol{\theta}_t)$. Thus (1.1.10) gets replaced by

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t \mathbf{h}_{t+1}. \tag{4.2.1}$$

Usually the step size sequence is deterministic and predetermined. However, some variations are possible, which we discuss next. It is noteworthy that the phrase "stochastic gradient" is used with two different meanings in the literature. Both of them are discussed here.

Much of the literature addresses the following specific type of optimization problem: Suppose $\mathcal{X}$ is some set, and $\pi$ is some probability measure on $\mathcal{X}$. Suppose further that $f : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}$ is a $\mathcal{C}^1$ function, and define the objective function

$$J(\boldsymbol{\theta}) := E_{x\sim\pi}[f(x,\boldsymbol{\theta})] = \int_{\mathcal{X}} f(x,\boldsymbol{\theta})\, \pi(dx). \tag{4.2.2}$$

One can ensure that the above integral is well-defined by imposing some reasonable assumptions on the function $f$ and/or the probability measure $\pi$. In order to minimize $J(\cdot)$, it becomes necessary to compute the gradient $\nabla J(\boldsymbol{\theta})$. This raises the question as to when

$$\nabla J(\boldsymbol{\theta}) = E_{x\sim\pi}[\nabla_{\boldsymbol{\theta}} f(x,\boldsymbol{\theta})]? \tag{4.2.3}$$

In other words, when it is permissible to interchange differentiation and integration in (4.2.3)? If $\mathcal{X}$ is a finite set, then this is automatic, because the expectation with respect to $x$ is just a finite summation. If $\mathcal{X}$ is an infinite set, this is not automatic. However, in the practically important case where $f(x,\boldsymbol{\theta})$ is a *convex* function of $\boldsymbol{\theta}$ for almost all $x$, (4.2.3) holds with a few additional technical assumptions. The reader is directed to [126, Eq. (11)] and [134, Eq. (7.1270], which give the required equality. These results are not stated here, as that would take us too far afield.

A typical application where $J(\cdot)$ has the form (4.2.2) would be neural network training. Suppose $\mathbf{x} \in \mathbb{R}^n$ is the input to the network, $y \in \mathbb{R}$ the desired output with input $\mathbf{x}$ (the label), and $\boldsymbol{\theta}$ is the set of "weights" or adjustable parameters in the network. A neural network "architecture" defines family of maps $H(\cdot,\boldsymbol{\theta}) : \mathbb{R}^n \to \mathbb{R}$ for each $\boldsymbol{\theta} \in \mathbb{R}^d$. Finally, there is a "loss function" $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$; quite often $L(y,z) = |y-z|^2$. The training data consists of labelled pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$. To choose the weight vector optimally, one minimizes

$$J(\boldsymbol{\theta}) := \frac{1}{m}\sum_{i=1}^m L(y_i, H(\mathbf{x}_i,\boldsymbol{\theta})).$$

To put this problem within the framework of (4.2.2), we can define $\mathcal{X}$ to be the finite set $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_m, y_m)\}$, and choose $\pi$ to be the uniform distribution on $\mathcal{X}$.

Next we discuss three approaches to approximating $\nabla J(\boldsymbol{\theta})$ when $J(\cdot)$ is as in (4.2.2). As a part of this, we introduce three phrases that are widely used in the world of optimization and ML. Further details can be found in [26, Section 3.3]. As a part of this, we introduce *one of the two usages* of the phrase Stochastic Gradient: the other usage is introduced in Section 4.3.

**Stochastic Gradient:** At step $t$, choose a random element $x_{t+1} \in \mathcal{X}$ with distribution $\pi$. To permit adaptive sampling, it is *not assumed* that $x_{t+1}$ is independent of the preceding samples $(x_1, \cdots, x_t)$. Then the search direction $\mathbf{h}_{t+1}$ is set equal to

$$\mathbf{h}_{t+1} = \nabla_{\boldsymbol{\theta}} f(x_{t+1}, \boldsymbol{\theta}_t). \tag{4.2.4}$$

Since $x_{t+1}$ follows the distribution $\pi$, the expected value of the above quantity is

$$E_{x_{t+1} \sim \pi}[\mathbf{h}_{t+1}] = E_{x_{t+1} \sim \pi}[\nabla_{\boldsymbol{\theta}} f(x_{t+1}, \boldsymbol{\theta}_t)].$$

If the sufficient conditions from [126, 134] hold, then the above expected value is indeed the true gradient $\nabla J(\boldsymbol{\theta}_t)$. This is the justification for this approach.

**Batch Update:** In this case,

$$\mathbf{h}_{t+1} = \nabla J(\boldsymbol{\theta}_t)$$

as computed in (4.2.3). If $\mathcal{X}$ is finite, say $|\mathcal{X}| = n$, then the above computation involves adding $n$ different individual gradients $\nabla J(x_i, \boldsymbol{\theta}_t)$ over $x_i \in \mathcal{X}$. If $n$ is large, the computation can be quite expensive. However, there is no approximation involved.

**Minibatch Update:** This approach is intermediate between the above two approaches. At step $t$, an integer $N_t$ (possibly random) is chosen, and $N_t$ samples $x_j, j \in [N_t]$ are chosen from $\mathcal{X}$. The analysis is simplest if these samples are drawn independently with distribution $\pi$, after replacement. Then

$$\mathbf{h}_{t+1} = \frac{1}{N_t} \sum_{j=1}^{N_t} \nabla_{\boldsymbol{\theta}} f(x_j, \boldsymbol{\theta}_t). \tag{4.2.5}$$

If there are repeated samples, then the corresponding terms are summed more than once in the above equation. As with the stochastic gradient approach, we have that

$$E_{x_{t+1} \sim \pi}[\mathbf{h}_{t+1}] = \nabla J(\boldsymbol{\theta}_t)$$

under suitable conditions.

Until now, we have focused on objective functions of the form (4.2.2), and ways to approximate its gradient by random sampling. We have also not catered to the possibility of *errors in the computation* of the gradients, which can be modelled as additive noise. Next we discuss approximation methods that apply to general $\mathcal{C}^1$ objective functions, with possibly noisy computations of the gradients. There are two parts to this: (i) Constructing approximations to the true gradient, and (ii) selecting which components of the current guess $\boldsymbol{\theta}_t$ are to be updated at step $t$. We discuss these two topics in the opposite order. That is, we begin by discussing some popular methods of choosing coordinates to be updated, assuming that the true gradient, corrupted by additive noise, is available. It will be obvious that the same selection strategies can also be applied to any stochastic gradient as well.

The first of these methods is referred to as "Coordinate Gradient Descent" as in [175] and elsewhere, but also sometimes as "stochastic gradient," thus possibly leading to confusion with (4.2.4).

**Coordinate Gradient Descent:** Suppose that, at step $t$, the current guess is $\boldsymbol{\theta}_t$, and suppose that the learner has access to a (possibly noise-corrupted) measurement $\nabla J(\boldsymbol{\theta}) + \boldsymbol{\xi}_{t+1}$. An index $i \in [d]$ is chosen at random with a uniform probability, and the search direction is defined as

$$\mathbf{h}_{t+1} = d\mathbf{e}_i \circ [\nabla J(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}], \tag{4.2.6}$$

where $\mathbf{e}_i$ denotes the $i$-elementary unit vector, and $\circ$ denotes the Hadamard or componentwise product.[2] Even if $\boldsymbol{\xi}_{t+1} \equiv \mathbf{0}$, $\mathbf{h}_{t+1}$ is still random due to the choice of $i$. The factor of $d$ is to ensure that the conditional expectation with respect to $\boldsymbol{\theta}_t$ of $\mathbf{h}_{t+1}$ equals the true gradient $\nabla J(\boldsymbol{\theta}_t)$ plus the expectation of $\boldsymbol{\xi}_{t+1}$. If this $\mathbf{h}_{t+1}$ is substituted into (4.2.1), it is obvious that only the $i$-th component of $\boldsymbol{\theta}_t$ is updated at time $t$, and all other components remain the same.

An excellent survey of coordinate gradient descent for convex objective functions is found in [175], and some results for nonconvex objective functions are found in [165]. It is worth pointing out that, in these references and many others, the error term $\boldsymbol{\xi}_{t+1}$ is assumed to be zero. Thus the only source of randomness is the coordinate to be updated. Much of the detailed analysis carried out in these papers would not be applicable in the presence of measurement errors.

One can also apply this philosophy of updating only one (possibly randomly chosen) coordinate at a time to stochastic approximation as in (4.2.1). Note that the ability to cope with noisy measurements is a key strength of SA. This leads to the update formula

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_t \mathbf{e}_i \circ [\mathbf{f}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}]. \tag{4.2.7}$$

In such a case, it is common to refer to this approach as **Asynchronous SA** or ASA. This terminology was apparently introduced in [158]. The approach is studied further in [22]. In particular, a distinction between using a "global clock" and a "local clock" for componentwise updating is introduced in that reference.

**Block Coordinate Gradient Descent:** A variant of the above is to carry out "block" updating. At each time, a possibly random subset $S_t \subseteq [d]$ is selected. Define

$$\mathbf{e}_{S_t} := \sum_{i \in S_t} \mathbf{e}_i.$$

Then the vector $\mathbf{h}_{t+1}$ is defined as

$$\mathbf{h}_{t+1} := \frac{d}{|S_t|} \mathbf{e}_{S_t} \circ [\nabla J(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}]. \tag{4.2.8}$$

This implies that, at time $t$, only the components of $\boldsymbol{\theta}_t$, $i \in S_t$ are updated, and the rest are unchanged. As above, block updating can also be incorporated in the SA algorithm of (4.2.1), as follows:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{\alpha}_t \circ \mathbf{e}_{S_t} \circ [\mathbf{f}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}], \tag{4.2.9}$$

where $\boldsymbol{\alpha}_t$ is now a *vector* of step sizes. Thus, while only those components $i \in S_t$ are updated, different updated components could have different step sizes. This topic is discussed in Section 3.3. The reader is referred to [73, 72] for a detailed treatment.

**Gradients Using Only Function Evaluations:** Next we discuss some approaches to generating approximate gradients that make use of only function evaluations. The first such approach is in [77], which is for the case $d = 1$, and requires two function evaluations per iteration. Subsequently Blum [17] presented an approach for the case $d > 1$, which requires $d + 1$ evaluations per iteration. When $d$ is large, this approach is clearly impractical. A significant improvement came in[138], in which a method called "simultaneous perturbation stochastic approximation" (SPSA) was introduced, which requires only *two* function evaluations, irrespective of the dimension $d$. However, the proof of convergence of SPSA given in [138] requires many assumptions. These are simplified in [31]. An "optimal" version of SPSA is introduced in [129], and is described below.

For each index $t + 1$, suppose $\Delta_{t+1,i}, i \in [d]$ are $d$ different and pairwise independent **Rademacher variables**.[3] Moreover, suppose that $\Delta_{t+1,i}, i \in [d]$ are all independent (not just conditionally independent)

[2] If $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, then $\mathbf{c} = \mathbf{a} \circ \mathbf{b}$ belongs to $\mathbb{R}^d$ and is defined via $c_i = a_i b_i$ for all $i$.

[3] Recall that Rademacher random variables assume values in $\{-1, 1\}$ and are independent of each other.

of the $\sigma$-algebra $\mathcal{F}_t$ for each $t$. Let $\mathbf{\Delta}_{t+1} \in \{-1,1\}^d$ denote the vector of Rademacher variables at time $t+1$. Then the search direction $\mathbf{h}_{t+1}$ in (4.2.1) is defined componentwise, via

$$h_{t+1,i} = \frac{[J(\boldsymbol{\theta}_t + c_t\mathbf{\Delta}_{t+1}) + \xi^+_{t+1,i}] - [J(\boldsymbol{\theta}_t - c_t\mathbf{\Delta}_{t+1}) - \xi^-_{t+1,i}]}{2c_t\Delta_{t+1,i}}, \tag{4.2.10}$$

where $\xi^+_{t+1,1}, \cdots, \xi^+_{t+1,d}$, $\xi^-_{t+1,1}, \cdots, \xi^-_{t+1,d}$ represent the measurement errors. A similar idea is used in [110], except that the bipolar vector $\mathbf{\Delta}_{t+1}$ is replaced by a random Gaussian vector $\boldsymbol{\eta}_{t+1}$ in $\mathbb{R}^d$. As can be seen from the literature, one of the key steps in analyzing SPSA is to find tail probability estimates of the quantity $\|\boldsymbol{\eta}_{t+1}\|_2/|\eta_{t+1,i}|$. If $\boldsymbol{\eta}_{t+1}$ is Gaussian, then this ratio can be arbitrarily large, albeit with small probability. However, with Rademacher perturbations, the ratio $\|\mathbf{\Delta}_{t+1}\|_2/|\Delta_{t+1,i}|$ always equals $\sqrt{d}$. This observation considerably simplifies the analysis. An excellent survey of this topic can be found in [91], which discusses other approaches not mentioned here.

The original SPSA envisages only two measurements per iteration, and the resulting estimate of $\nabla J(\boldsymbol{\theta}_t)$ has bias $O(c_t)$ and conditional variance $O(1/c_t^2)$. However, it is possible to take more measurements and reduce the bias of the estimate, while retaining the same bound on the conditional variance. Specifically, if $k+1$ measurements are taken, then the bias is $O(c_t^k)$ (which converges to zero more quickly), while the conditional variance remains as $O(1/c_t^2)$. See [112] and the references therein.

The framework discussed until now addresses *additive* measurement errors. Now we present a more general framework is proposed that is capable of handling not only additive measurement errors, but also *multiplicative* errors, and others. This treatment is taken from [53]. In that paper, three (closely related) algorithms are proposed in this paper, out of which only the second one is detailed here, in the interests of brevity.

The set-up is as follows: Suppose $f : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ is a $\mathcal{C}^1$ function, and $\pi$ is a (possibly unknown) probability measure on $\mathbb{R}^p$. The objective function is as in (4.2.2), namely

$$J(\boldsymbol{\theta}) = \int_{\mathbb{R}^n} f(\mathbf{w}, \boldsymbol{\theta})\ \pi(d\mathbf{w}) = E_{\mathbf{w}\sim\pi}[f(\mathbf{w}, \boldsymbol{\theta})].$$

There is also a probability distribution $P$ on $\mathbb{R}^d$, chosen by the learner, whose role is to generate an i.i.d. sequence of perturbations $\{\mathbf{\Delta}_t\}_{t\geq 1}$. In addition, there two i.i.d. sequences $\{\mathbf{w}^+_t\}_{t\geq 0}$, and $\{\mathbf{w}^-_t\}_{t\geq 0}$, with distribution $\pi$. To update the current guess $\boldsymbol{\theta}_t$, one undertakes the following steps. As with the other derivative-free methods, there are two sequences: $\{\alpha_t\}$ of step sizes, and $\{c_t\}$ of increments. At time $t$, the perturbation vector $\mathbf{\Delta}_{t+1}$ is known, so one can define

$$\mathbf{x}^+_{t+1} = \boldsymbol{\theta}_t + c_t\mathbf{\Delta}_{t+1}, \quad \mathbf{x}^-_{t+1} = \boldsymbol{\theta}_t - c_t\mathbf{\Delta}_{t+1}.$$

The measurements available to the learner at time $t$ consist of the pair

$$y^+_{t+1} = f(\mathbf{w}^+_t, \mathbf{x}^+_{t+1}) + \xi^+_{t+1}, \ \ y^-_{t+1} = f(\mathbf{w}^-_t, \mathbf{x}^-_{t+1}) + \xi^-_{t+1},$$

where $\xi^+_{t+1}, \xi^-_{t+1}$ are measurement errors. The last step is to define the stochastic gradient $\mathbf{h}_{t+1}$. This is stated in terms of a sequence of "kernel functions" $K_t : \mathbb{R}^d \to \mathbb{R}^d$ that satisfy, for each $t$

$$\int_{\mathbb{R}^d} K_t(\mathbf{z})\ P(d\mathbf{z}) = \mathbf{0}, \quad \int_{\mathbb{R}^d} K_t(\mathbf{z})\mathbf{z}^\top\ P(d\mathbf{z}) = I_d, \quad \int_{\mathbb{R}^d} \|K_t(\mathbf{z})\|_2^2\ P(d\mathbf{z}) < \infty.$$

With this notation, the stochastic gradient $\mathbf{h}_{t+1}$ is defined as

$$\mathbf{h}_{t+1} = \frac{y^+_{t+1} - y^-_{t+1}}{2c_t} K_t(\mathbf{\Delta}_{t+1}),$$

with the update rule as in (4.2.1), namely

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t\mathbf{h}_{t+1},$$

Note that the choice

$$K_t(\mathbf{z}) = (1/z_1, \cdots, 1/z_d)$$

gives the standard Kiefer-Wolfowitz-Blum approach, presented here as (4.2.10). However, it is clear that the present scheme offers considerably more flexibility.

In order to analyze the behavior of the algorithm, it is assumed in [53] that

1. $J(\cdot)$ is a strongly convex function of $\boldsymbol{\theta}$, and

2. There is a constant $L$ such that $\nabla_{\boldsymbol{\theta}} f(\mathbf{w}, \boldsymbol{\theta})$ is $L$-Lipschitz continuous for each $\mathbf{w} \in \mathbb{R}^n$.

In particular, Item 1 means that $J(\cdot)$ has a unique global minimizer $\boldsymbol{\theta}^*$. Under these assumptions, [53, Theorem 1] gives sufficient conditions for $\boldsymbol{\theta}_t$ to converge to $\boldsymbol{\theta}^*$ in the mean-squared sense, and almost surely. The reader is directed to [53] for more details.

We conclude this subsection by discussing some universal *lower* bounds on the achievable performance of gradient-based optimization methods. These results are taken from [4], but stated in the present notation. The authors study an objective function $J : \mathbb{R}^d \to \mathbb{R}$ with a globally Lipschitz-continuous gradient [4, Eq. (3)]. Further, it is assumed that

$$E_t(\mathbf{h}_{t+1}) = \nabla J(\boldsymbol{\theta}_t),$$

and that there is a finite constant $M$ such that

$$CV_t(\mathbf{h}_{t+1}) \leq M^2.$$

See [4, Eq. (2)]. Thus the stochastic gradient is assumed to provide an *unbiased* estimate of the true gradient. Moreover, the conditional variance of the stochastic gradient is assumed to be bounded, both as a function of $t$ and as a function of $\boldsymbol{\theta}_t$. These assumptions are the same as (3.2.4) and (3.2.5) with $\mu_t = 0$ for all $t$, and $M_t^2 \leq M^2$ for all $t$. Hence they are more restrictive than the assumptions made in this book, namely (3.2.4) and (3.2.5). Even under these restrictive assumptions, it is shown that, in the case where $J(\cdot)$ is convex, achieving $\|\nabla J(\boldsymbol{\theta}_t)\|_2 \leq \epsilon$ requires $\Omega(\epsilon^{-2})$ iterations in the worst case; see [4, Section 1.1]. For an *arbitrary* nonconvex function, the bound goes up to $\Omega(\epsilon^{-4})$.[4] Therefore, if we wish to find a $T$ such that

$$\|\nabla J(\boldsymbol{\theta}_t)\|_2 \leq \epsilon, \ \forall t \geq T,$$

then $T = \Omega(\epsilon^{-2})$ for convex functions, and $T = \Omega(\epsilon^{-4})$ for nonconvex functions. We can turn this around to get a bound on the best achievable rate of convergence. If $T = \Omega(\epsilon^{-k})$, then $\epsilon = \Omega(T^{-1/k})$ in the worst case. Hence $\|\nabla J(\boldsymbol{\theta}_t)\|_2 = \Omega(t^{-1/2})$ if $J(\cdot)$ is convex, and $\|\nabla J(\boldsymbol{\theta}_t)\|_2 = \Omega(t^{-1/4})$ if $J(\cdot)$ is a general nonconvex function. The assumptions in [4] are the same as (3.2.4) and (3.2.5) with $\mu_t = 0$ for all $t$, and $M_t^2 \leq M^2$ for all $t$. One of the contributions of the paper [70] is to show that when the function $J(\cdot)$ belongs to class (PL), then $\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = o(t^{-\lambda})$ and $J(\boldsymbol{\theta}_t) = o(t^{-\lambda})$ for all $\lambda < 1$. These bounds are practically the same as the lower bounds in [4]. The details are presented in Section 4.3. It is important to remind the reader that the "universal" lower bound $\|\nabla J(\boldsymbol{\theta}_t)\|_2 = \Omega(t^{-1/4})$ applies for *arbitrary* nonconvex functions. But if $J(\cdot)$ is restricted to satisfy Property (PL), then, as mentioned above, the achievable performance improves to $\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = o(t^{-\lambda})$ and $J(\boldsymbol{\theta}_t) = o(t^{-\lambda})$ for all $\lambda < 1$.

### 4.2.2 Momentum-Based Methods

The phrase "momentum-based" is somewhat vague, but refers to methods wherein the search direction at step $t$ depends not only on the current guess $\boldsymbol{\theta}_t$, but also on the previous guess $\boldsymbol{\theta}_{t-1}$.

It should be mentioned that, in the early 1960s, a class of optimization algorithms were introduced, known as "conjugate gradient" methods. There were purely deterministic in nature, and were distinguished by the fact that the "search direction" (basically $\mathbf{h}_{t+1}$ in (4.2.1), but deterministic) is a linear combination

[4] There are some additional technical assumptions which are not repeated here.

of the current gradient $\nabla J(\boldsymbol{\theta}_t)$ and the previous gradient $\nabla J(\boldsymbol{\theta}_{t-1})$. Momentum-based based are different in that "past" iterations enter through $\boldsymbol{\theta}_{t-1}$ and not $\nabla J(\boldsymbol{\theta}_{t-1})$. A good summary of classical conjugate gradient methods can be found in [51, Section 5.3] and in [115, Section 3.2]. Moreover, in Polyak's book, the relationships between two types of methods are explored.

The Heavy Ball (HB) method, introduced in [113], is one of earliest "momentum-based" methods for optimization. The algorithm introduced in [113] is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha \nabla J(\boldsymbol{\theta}_t) + \mu(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}). \tag{4.2.11}$$

It is shown by Polyak that, if $J(\boldsymbol{\theta})$ is quadratic of the form $(1/2)\boldsymbol{\theta}^\top A\boldsymbol{\theta} + \langle \mathbf{v}, \boldsymbol{\theta}\rangle + c$ for some positive definite matrix $A$, vector $\mathbf{v}$ and constant $c$, then the HB method requires $1/\sqrt{R}$ fewer iterations compared to the gradient descent method, provided $\mu$ is chosen as $(\sqrt{R}-1)/(\sqrt{R}+1)$, where $R$ denotes the condition number of $A$.

A subsequent and widely-used momentum-based method is Nesterov's Accelerated Gradient (NAG) method [107]. In [143], NAG is reformulated in a manner that brings out the similarities as well as the differences with HB. Specifically, the NAG algorithm can be written as

$$\mathbf{v}_{t+1}^N = \mu_t \mathbf{v}_t^N - \alpha_t \nabla J[\boldsymbol{\theta}_t + \mu_t \mathbf{v}_t^N], \tag{4.2.12}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \mathbf{v}_{t+1}^N. \tag{4.2.13}$$

These two equations can be combined into the single equation

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t \nabla J[\boldsymbol{\theta}_t + \mu_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1})] + \mu_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}). \tag{4.2.14}$$

This can be compared with the HB formulation (4.2.11), namely

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t \nabla J(\boldsymbol{\theta}_t) + \mu_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}). \tag{4.2.15}$$

In other words, in NAG the gradient is computed *after* the momentum correction term $\mu_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1})$ is added to $\boldsymbol{\theta}_t$. It is shown in [109, Section 2.2] that when $J(\cdot)$ is a smooth convex function with a Lipschitz-continuous gradient, NAG converges to the minimum at the rate of $O(t^{-2})$. Moreover, no gradient-based algorithm can achieve a faster rate. A more precise statement and references are needed. More details can be found in [26, Section 7]. The paper [143] also shows that NAG can be deployed successfully for training deep neural networks.

Another relevant reference is [8], in which an alternate momentum-based method is proposed, namely

$$\mathbf{v}_{t+1}^B = \mu_t \mathbf{v}_t^B - \alpha_t \nabla J(\Theta_t),, \tag{4.2.16}$$

$$\begin{aligned} \Theta_{t+1} &= \Theta_t + (1+\mu_t)\mathbf{v}_{t+1}^B - \mu_{t-1}\mathbf{v}_t^B \\ &= \Theta_t + \mu_t\mu_{t-1}\mathbf{v}_t^B + (1+\mu_t)\alpha_t \nabla J(\Theta_t). \end{aligned} \tag{4.2.17}$$

If started with the initial guess $\boldsymbol{\theta}_0 = \mathbf{0}$, the trajectory of this algorithm matches that in [143] (which is just a reformulation of NAG) both at the start and in the final phase of local convergence to the solution. But the formulation in [8] is closer to Polyak's HB compared to NAG, because the gradient $\nabla J(\cdot)$ is computed at the current guess $\Theta_t$, and not a shifted version of it.

It has been mentioned in previous chapters that the behavior of the SA algorithm can be analyzed by studying the stability properties of an associated ODE. The same is true of momentum-based methods as well. In the case of momentum-based methods, the associated ODEs are *second order* in $\boldsymbol{\theta}$. Also, the analysis based on ODEs does not always apply when the measurements are noisy. With those caveats, we briefly summarize a few relevant papers. The behavior of NAG is analyzed in [142], when the step size $\alpha_t$ is held constant, while the momentum coefficient $\mu_t$ varies with time. It is shown that the "optimal" schedule for $\mu_t$ is $\mu_t = (t+2)/(t+5)$. Another paper along the same lines is [5]. Similarly, the papers [2, 6] analyze the Heavy Ball algorithm from an ODE standpoint. Note that there is no measurement error in these papers.

## 4.3 Stochastic Gradient Descent

In the previous section, we discussed (but did not really analyze) several standard gradient-based methods for finding a stationary point of a given function. In all of the formulations, there was no provision for mdasuremsent errors. In the the remainder of this chapter, we analyze the more general situation where measurement errors are permitted, and establish both the convergence as well as the *rates* of convergence of various algorithms, under suitable hypothesses.

In this section we carry out our analysis of the SGD algorithm of (4.2.1), which is repeated here for the reader's convenience:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t \mathbf{h}_{t+1}, \tag{4.3.1}$$

The main tools we use to carry out this analysis are Theorems 2.23 and 2.24.

In order to analyze the convergence of (4.3.1), we recall the standing assumptions on $J(\cdot)$, namely:

(J1) $J(\cdot)$ is $\mathcal{C}^1$, and $\nabla J(\cdot)$ is globally Lipschitz-continuous with constant $L$.

(J2) $J(\cdot)$ is bounded below. Thus

$$J^* := \inf_{\boldsymbol{\theta} \in \mathbb{R}^d} J(\boldsymbol{\theta}) > -\infty.$$

Note that it is *not* assumed that the infimum is actually attained, nor that the minimizer is unique if the minimum is attained.

Before proceeding further, we present a very useful consequence of of Assumptions (J1) and (J2).

**Lemma 4.1.** *Suppose (J1) and (J2) hold. Then*

$$\|\nabla J(\boldsymbol{\theta})\|_2^2 \leq 2L[J(\boldsymbol{\theta}) - J^*]. \tag{4.3.2}$$

*Proof.* By applying [12, Eq. (2.4)], stated here as Theorem 3.1, to $J(\boldsymbol{\theta})$, it follows that, for every $\boldsymbol{\phi}, \boldsymbol{\theta} \in \mathbb{R}^d$, we have

$$J^* \leq J(\boldsymbol{\phi}) \leq J(\boldsymbol{\theta}) + \langle \nabla J(\boldsymbol{\theta}), \boldsymbol{\phi} - \boldsymbol{\theta} \rangle + \frac{L}{2} \|\boldsymbol{\phi} - \boldsymbol{\theta}\|_2^2.$$

Now choose $\boldsymbol{\phi} = \boldsymbol{\theta} - (1/L)\nabla J(\boldsymbol{\theta})$. This leads to

$$J^* \leq J(\boldsymbol{\theta}) - \frac{1}{L}\|\nabla J(\boldsymbol{\theta})\|_2^2 + \frac{1}{2L}\|\nabla J(\boldsymbol{\theta})\|_2^2 = J(\boldsymbol{\theta}) - \frac{1}{2L}\|\nabla J(\boldsymbol{\theta})\|_2^2.$$

This is the same as (4.3.2). □

As pointed out in the first Remark after Theorem 3.1, the bound (3.2.7) is well-known for convex functions; however, Theorem 3.1 extends the bound to *nonconvex* functions. Similarly, in the present setting, (4.3.2) is also well-known for convex functions; but the contribution of Lemma 4.1 is to show that convexity is not needed.

Also, we introduce one more property, named (NSC), that the function $J(\cdot)$ is expected to satisfy. This property consists of the following assumptions, taken together.

1. The function $J(\cdot)$ attains its infimum. Therefore the set $S_J$ defined in (4.1.1) is nonempty.

2. The function $J(\cdot)$ has compact level sets. For every constant $c \in (0, \infty)$, the level set

$$L_J(c) := \{\boldsymbol{\theta} \in \mathbb{R}^d : J(\boldsymbol{\theta}) \leq c\}$$

is compact.

3. There exists a number $r > 0$ and a continuous function $\eta : [0, r] \to \mathbb{R}_+$ such that $\eta(0) = 0$, and

$$\rho(\boldsymbol{\theta}) \leq \eta(J(\boldsymbol{\theta}) - J^*), \ \forall \boldsymbol{\theta} \in L_J(r), \tag{4.3.3}$$

where $\rho(\boldsymbol{\theta})$ is defined as

$$\rho(\boldsymbol{\theta}) := \inf_{\boldsymbol{\phi} \in S_J} \|\boldsymbol{\theta} - \boldsymbol{\phi}\|_2.$$

and equals the distance from $\boldsymbol{\theta}$ to the set $S_J$.

The acronym (NSC) stands for "near strong convexity," or "nearly strongly convex," depending on the syntax. Recall from Definition 1.3 that $J(\cdot)$ is said to be $R$-**strongly convex** if

$$J(\boldsymbol{\theta}) \geq J(\boldsymbol{\phi}) + \langle \nabla J(\boldsymbol{\phi}), \boldsymbol{\theta} - \boldsymbol{\phi} \rangle + \frac{R}{2} \|\boldsymbol{\theta} - \boldsymbol{\phi}\|_2^2, \ \forall \boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^d.$$

Note that the equation above is slightly different from (1.1.9), but is equivalent to it. Now, if $J(\cdot)$ is $R$-strongly convex, it has a unique global minimizer, which can be denoted as $\boldsymbol{\theta}^*$. Next, if we substitute $\boldsymbol{\phi} = \boldsymbol{\theta}^*$ in the above equation, we get

$$J(\boldsymbol{\theta}) \geq J^* - \frac{R}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2.$$

If we now observe that $\rho(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2$, the above inequality can be rewritten as

$$\rho(\boldsymbol{\theta}) \leq \sqrt{\frac{2(J(\boldsymbol{\theta}) - J^*)}{R}}.$$

Thus every strongly convex function satisfies (NSC), but the converse is not true in general.

It is obvious that, if (NSC) is satisfied, then $J(\boldsymbol{\theta}_t) \to 0$ as $t \to \infty$ implies that $\rho(\boldsymbol{\theta}_t) \to 0$ as $t \to \infty$. Thus, whenever $J(\cdot)$ satisfies (NSC), and we are able to establish that $J(\boldsymbol{\theta}_t) \to J^*$ as $t \to \infty$, it follows automatically that $\rho(\boldsymbol{\theta}_t) \to 0$ as $t \to \infty$. In other words, the convergence of $J(\boldsymbol{\theta}_t)$ to its minimum value, coupled with (NSC), implies that $\boldsymbol{\theta}_t$ converges to the set $S_J$.

With these preliminaries out of the way, we can begin to analyze the Stochastic Gradient Descent algorithm described in (4.2.1). Recall that $\mathbf{h}_{t+1}$ in (4.2.1) is the stochastic gradient. To characterize it, define

$$\mathbf{z}_t = E_t(\mathbf{h}_{t+1}), \quad \mathbf{x}_t = \mathbf{z}_t - \nabla J(\boldsymbol{\theta}_t), \quad \boldsymbol{\zeta}_{t+1} = \mathbf{h}_{t+1} - \mathbf{z}_t. \tag{4.3.4}$$

One can think of $\mathbf{z}_t$ as the 'predictable" part of the stochastic gradient $\mathbf{h}_{t+1}$, that is, the best approximation at time $t$ of $\mathbf{h}_{t+1}$. In view of this interpretation, it ready follows that $\mathbf{x}_t$ can be thought of as the **bias** of the stochastic gradient. The rationale is that, ideally, we would want the search direction to be the true gradient $\nabla J(\boldsymbol{\theta}_t)$; therefore the difference $\mathbf{z}_t$ and $\nabla J(\boldsymbol{\theta}_t)$ is the bias.

The last equation in (4.3.4) implies that $E_t(\boldsymbol{\zeta}_{t+1}) = \mathbf{0}$. Therefore

$$E_t(\|\mathbf{h}_{t+1}\|_2^2) = \|\mathbf{z}_t\|_2^2 + E_t\|\boldsymbol{\zeta}_{t+1}\|_2^2. \tag{4.3.5}$$

Now we state our assumptions on the quantities $\mathbf{x}_t$ and $\boldsymbol{\zeta}_{t+1}$. The assumptions on these quantities are similar to the assumptions (3.2.4) and (3.2.5) on the additive noise in Stochastic Approximation. Specifically, it is assumed that there exist sequences of constants $\{B_t\}$ and $\{M_t\}$ such that

$$\|\mathbf{x}_t\|_2 \leq B_t[1 + \|\nabla J(\boldsymbol{\theta}_t)\|_2], \ \forall \boldsymbol{\theta}_t \in \mathbb{R}^d, \ \forall t, \tag{4.3.6}$$

$$E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2) \leq M_t^2[1 + J(\boldsymbol{\theta}_t)], \ \forall \boldsymbol{\theta}_t \in \mathbb{R}^d, \ \forall t. \tag{4.3.7}$$

Now we briefly discuss the significance of these assumptions.

1. Note that (4.3.6) permits the stochastic gradient to be a *biased* estimate of $\nabla J(\boldsymbol{\theta}_t)$. This by itself is not unusual. In several papers, assumptions of the form (4.3.6) occur, but *without the* $\|\boldsymbol{\theta}_t\|_2$ *term.* We now give an example of a situation where the presence of this term arises naturally. Consider the "Coordinate Gradient Descent" algorithm described in (4.2.8). In the traditional approach, every coordinate is sampled uniformly at random, which explains the presence of the factor $d$ in the equation. Now consider an "off-policy" type of coordinate sampling, in which, at time $t$, the coordinates are sampled with a probability distribution $\boldsymbol{\phi}_t$, which *need not equal* the uniform distribution. However, $\boldsymbol{\phi}_t \to \mathbf{u}_d$ as $t \to \infty$, where $\mathbf{u}_d$ is the uniform distribution on a set of $d$ elements. To analyze this case, let $I_t$ denote the coordinate chosen to be updated at time $t$. Then

$$I_t = i \text{ w.p. } \phi_{t,i}.$$

Hence the stochastic gradient can be computed as

$$\mathbf{h}_{t+1} = d[\nabla J(\boldsymbol{\theta}_t)] \circ \mathbf{e}_{I_t} \text{ w.p. } \phi_{t,i},$$

To estimate the quantity $\|\mathbf{x}_t\|_2$ where $\mathbf{x}_t = E_t(\mathbf{h}_{t+1}) - \nabla J(\boldsymbol{\theta}_t)$, we use the notation $g_i$ for $[\nabla J(\boldsymbol{\theta}_t)]_i$, for brevity. Then

$$[\mathbf{h}_{t+1} - \nabla J(\boldsymbol{\theta}_t)]_i = \begin{cases} (d-1)g_i, & \text{w.p. } \phi_{t,i}, \\ -g_i, & \text{w.p. } \phi_{t,j}, j \neq i. \end{cases}$$

Therefore, with $\mathbf{x}_t = E_t(\mathbf{h}_{t+1} - \nabla J(\boldsymbol{\theta}_t))$ as earlier, we have that

$$\begin{aligned} x_{t,i} &= (d-1)g_i\phi_{t,i} - \sum_{j \neq i} g_i\phi_{t,j} = dg_i\phi_{t,i} - g_i \sum_{j=1}^{d} \phi_{t,j} \\ &= (d\phi_{t,i} - 1)g_i = d(\phi_{t,i} - u_i)g_i, \end{aligned}$$

where $u_i = 1/d$ is the $i$-th component of the uniform distribution (for each $i$). Summing over $i$ leads to

$$\begin{aligned} \|\mathbf{x}_t\|_1 &= d \sum_{i=1}^{d} |(\phi_{t,i} - u_i)| \cdot |g_i| \\ &\leq d\|\boldsymbol{\phi}_t - \mathbf{u}_d\|_1 \|\nabla J(\boldsymbol{\theta}_t)\|_\infty, \end{aligned}$$

where $\|\boldsymbol{\phi}_t - \mathbf{u}_d\|_1$ denotes the $\ell_1$ distance between $\boldsymbol{\phi}_t$ and $\mathbf{u}_d$. Next, after observing that $\|\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_2 \leq \|\mathbf{v}\|_1$, we arrive finally at

$$\|\mathbf{x}_t\|_2 \leq d\|\boldsymbol{\phi}_t - \mathbf{u}\|_1 \|\nabla J(\boldsymbol{\theta}_t)\|_2,$$

which is a special case of (4.3.6). Note that, when the "off-policy" sampling probability distribution is not the uniform distribution, the presence of the term $\|\nabla J(\boldsymbol{\theta}_t)\|_2$ in (4.3.6) is unavoidable.

2. Next we discuss (4.3.7). One can compare (4.3.7) with the so-called Expected Smoothness condition proposed as Assumption 2 in [75], namely

$$E_t(\|\mathbf{h}_{t+1}\|_2^2) \leq 2AJ(\boldsymbol{\theta}_t) + B\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 + C, \tag{4.3.8}$$

for suitable constants $A, B, C$. This is proposed as "the weakest assumption" for analyzing the convergence of SGD for nonconvex functions. If $J(\cdot)$ satisfies Assumptions (J1) and (J2), then we can apply Lemma 4.1. As a result, the term $B\|\nabla J(\boldsymbol{\theta}_t)\|_2^2$ can be bounded by $2BLJ(\boldsymbol{\theta}_t)$, resulting in

$$E_t(\|\mathbf{h}_{t+1}\|_2^2) \leq 2(A + BL)J(\boldsymbol{\theta}_t) + C \leq M(1 + J(\boldsymbol{\theta}_t)), \tag{4.3.9}$$

where

$$M = \max\{2(A + BL), C\}.$$

Thus, for functions $J(\cdot)$ satisfying Assumptions (J1) and (J2), the present assumption (4.3.7) is weaker than (4.3.8). Also, the various constants in (4.3.8) are bounded with respect to $t$, whereas in (4.3.7), the bound $M_t$ is allowed to be unbounded with respect to $t$. As shown long ago in [77], permitting the variance to be unbounded with time is an essential feature in analyzing SGD based on function evaluations alone.

With this background, we state the first convergence result, which *does not have* any conclusions about the rate of convergence. As always, these bounds and conclusions hold almost surely. Not surprisingly, the statement of the theorem bears a strong resemblance to Theorem 3.2, as does the proof. However, in the interests of making each chapter self-contained, the proof is given in its entirety.

**Theorem 4.1.** *Suppose the objective function $J(\cdot)$ satisfies the standing assumptions (J1) and (J2), and that the stochastic gradient $\mathbf{h}_{t+1}$ satisfies* (4.3.6) *and* (4.3.7)*. With these assumptions, we have the following conclusions;*

1. *Suppose*

$$\sum_{t=0}^{\infty} \alpha_t^2 < \infty, \quad \sum_{t=0}^{\infty} \alpha_t B_t < \infty, \quad \sum_{t=0}^{\infty} \alpha_t^2 M_t^2 < \infty. \tag{4.3.10}$$

   *Then $\{\nabla J(\boldsymbol{\theta}_t)\}$ and $\{J(\boldsymbol{\theta}_t)\}$ are bounded, and in addition, $J(\boldsymbol{\theta}_t)$ converges to some random variable as $t \to \infty$.*

2. *If in addition*

$$\sum_{t=0}^{\infty} \alpha_t = \infty, \tag{4.3.11}$$

   *then*

$$\liminf_{t\to\infty} \|\nabla J(\boldsymbol{\theta}_t)\|_2 = 0. \tag{4.3.12}$$

3. *If in addition $J(\cdot)$ satisfies (KL'), then $J(\boldsymbol{\theta}_t) \to 0$ and $\nabla J(\boldsymbol{\theta}_t) \to \mathbf{0}$ as $t \to \infty$.*

4. *Suppose that in addition to (KL'), $J(\cdot)$ also satisfies (NSC), and that* (4.3.10) *and* (4.3.11) *both hold. Then $\rho(\boldsymbol{\theta}_t) \to 0$ as $t \to \infty$.*

*Proof.* The proof is based on Theorem 2.23. It follows from applying Theorem 3.1 to (4.3.1) that

$$J(\boldsymbol{\theta}_{t+1}) \le J(\boldsymbol{\theta}_t) - \alpha_t \langle \nabla J(\boldsymbol{\theta}_t), \mathbf{h}_{t+1} \rangle + \frac{\alpha_t^2 L}{2} \|\mathbf{h}_{t+1}\|_2^2. \tag{4.3.13}$$

Applying the operator $E_t$ to both sides, using the definitions in (4.3.4), and applying (4.3.5), gives

$$E_t(J(\boldsymbol{\theta}_{t+1})) \le J(\boldsymbol{\theta}_t) - \alpha_t \langle \nabla J(\boldsymbol{\theta}_t), \mathbf{z}_t \rangle + \frac{\alpha_t^2 L}{2} [\|\mathbf{z}_t\|_2^2 + E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2). \tag{4.3.14}$$

We will bound each term separately, repeatedly using (4.3.6), (4.3.7), Schwarz' inequality, and the obvious inequality

$$2a \le 1 + a^2, \ \forall a \in \mathbb{R}.$$

First,

$$\begin{aligned} \langle \nabla J(\boldsymbol{\theta}_t), \mathbf{z}_t \rangle &= \|\nabla J(\boldsymbol{\theta}_t)\|_2^2 + \langle \nabla J(\boldsymbol{\theta}_t), \mathbf{x}_t \rangle \\ &\ge \|\nabla J(\boldsymbol{\theta}_t)\|_2^2 - \|\nabla J(\boldsymbol{\theta}_t)\|_2 \cdot \|\mathbf{x}_t\|_2. \end{aligned}$$

Now

$$\begin{aligned}\|\nabla J(\boldsymbol{\theta}_t)\|_2 \cdot \|\mathbf{x}_t\|_2 &\leq B_t\|\nabla J(\boldsymbol{\theta}_t)\|_2[1+\|\nabla J(\boldsymbol{\theta}_t)\|_2] \\ &= B_t\|\nabla J(\boldsymbol{\theta}_t)\|_2 + B_t\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 \\ &\leq 0.5B_t + 1.5B_t\|\nabla J(\boldsymbol{\theta}_t)\|_2^2] \qquad (4.3.15) \\ &\leq B_t + 2B_t\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 \leq B_t + 4B_tLJ(\boldsymbol{\theta}_t). \qquad (4.3.16)\end{aligned}$$

In the last equation we have replaced 0.5 by 1 just to avoid dealing with fractions, and have also used (4.3.2). Hence

$$\begin{aligned}-\alpha_t\langle\nabla J(\boldsymbol{\theta}_t), \mathbf{z}_t\rangle &\leq -\alpha_t\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 + \alpha_t\|\nabla J(\boldsymbol{\theta}_t)\|_2 \cdot \|\mathbf{x}_t\|_2 \\ &\leq -\alpha_t\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 + \alpha_tB_t + 4\alpha_tB_tLJ(\boldsymbol{\theta}_t).\end{aligned}$$

Next,

$$\begin{aligned}\|\mathbf{z}_t\|_2^2 &\leq \|\nabla J(\boldsymbol{\theta}_t)\|_2^2 + 2\|\nabla J(\boldsymbol{\theta}_t)\|_2 \cdot \|\mathbf{x}_t\|_2 + \|\mathbf{x}_t\|_2^2 \\ &\leq \|\nabla J(\boldsymbol{\theta}_t)\|_2^2 + B_t + 3B_t\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 + \|\mathbf{x}_t\|_2^2 \\ &\leq B_t + 2L(1+3B_t)J(\boldsymbol{\theta}_t) + \|\mathbf{x}_t\|_2^2.\end{aligned}$$

Note that here we use the tighter estimate from (4.3.15). Next,

$$\begin{aligned}\|\mathbf{x}_t\|_2^2 &\leq B_t^2[1+\|\nabla J(\boldsymbol{\theta}_t)\|_2]^2 = B_t^2[1+2\|\nabla J(\boldsymbol{\theta}_t)\|_2 + \|\nabla J(\boldsymbol{\theta}_t)\|_2^2] \\ &\leq 2B_t^2[1+\|\nabla J(\boldsymbol{\theta}_t)\|_2^2] \leq 2B_t^2[1+2LJ(\boldsymbol{\theta}_t)].\end{aligned}$$

Substituting into the above gives the bound

$$\|\mathbf{z}_t\|_2^2 \leq B_t + 2B_t^2 + 2L(1+3B_t+2B_t^2)J(\boldsymbol{\theta}_t).$$

Finally, by assumption (4.3.7),

$$E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2) \leq M_t^2[1+2LJ(\boldsymbol{\theta}_t)].$$

Substituting these bounds into (4.3.14) gives a bound to which Theorem 2.23 can be applied, namely:

$$E_t(J(\boldsymbol{\theta}_{t+1})) \leq (1+f_t)J(\boldsymbol{\theta}_t) + g_t - \alpha_t\|\nabla J(\boldsymbol{\theta}_t)\|_2^2, \qquad (4.3.17)$$

where

$$f_t = 2L[2\alpha_tB_t + \frac{L}{2}\alpha_t^2(1+3B_t+2B_t^2) + \alpha_t^2M_t^2], \qquad (4.3.18)$$

$$g_t = \alpha_tB_t + \frac{L}{2}\alpha_t^2(B_t+2B_t^2+M_t^2). \qquad (4.3.19)$$

Now it is straight-forward to verify that the conditions in (4.3.10) suffice to establish that both sequences $\{f_t\}$ and $\{g_t\}$ are summable. There are five different terms occuring in (4.3.18) and (4.3.19), namely

$$\alpha_t^2,\ \alpha_tB_t,\ \alpha_t^2B_t,\ \alpha_t^2B_t^2,\ \alpha_t^2M_t^2.$$

Now (4.3.10) states that $\{\alpha_t^2\}$, $\{\alpha_tB_t\}$ and $\{\alpha_t^2M_t^2\}$ are summable. The first condition implies that $\alpha_t$ is bounded, which implies that $\{\alpha_t^2B_t\}$ is also summable. Finally, since every summable sequence is also square-summable ($\ell_1$ is a subset of $\ell_2$), $\{\alpha_t^2B_t^2\}$ is also summable. Since all the conditions needed to apply Item 1 of Theorem 2.23 hold, it follows that $\{J(\boldsymbol{\theta}_t)\}$ is bounded and converges to some random variable. Now (4.3.2) implies that $\nabla J(\boldsymbol{\theta}_t)$ is also bounded. This establishes the Item 1 of the theorem.

To prove Item 2, note that if property (KL') holds, then Item 2 of Theorem 2.23 applies, and $J(\boldsymbol{\theta}_t) \to 0$ as $t \to \infty$.

Finally, Item 3 is a ready consequence of $J(\boldsymbol{\theta}_t) \to 0$ and property (NSC). □

Next we strengthen Assumption (KL') to (PL), and prove an estimate for the *rate* of convergence.

**Theorem 4.2.** *Let various symbols be as in Theorem 4.1. Suppose $J(\cdot)$ satisfies the standing assumptions (J1) and (J2), and also property (PL), and that* (4.3.10) *and* (4.3.11) *hold. Further, suppose there exist constants $\gamma > 0$ and $\delta \geq 0$ such that*

$$B_t = O(t^{-\gamma}), \quad M_t = O(t^{\delta}), \ \forall t \geq 1,$$

*where we take $\gamma = 1$ if $B_t = 0$ for all sufficiently large $t$, and $\delta = 0$ if $M_t$ is bounded. Choose the step-size sequence $\{\alpha_t\}$ as $O(t^{-(1-\phi)})$ and $\Omega(t^{-(1-C)})$ where $\phi$ and $C$ are chosen to satisfy*

$$0 < \phi < \min\{0.5 - \delta, \gamma\}, \quad C \in (0, \phi].$$

*Define*

$$\nu := \min\{1 - 2(\phi + \delta), \gamma - \phi\}. \tag{4.3.20}$$

*Then $\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = o(t^{-\lambda})$ and $J(\boldsymbol{\theta}_t) = o(t^{-\lambda})$ for every $\lambda \in (0, \nu)$. In particular, by choosing $\phi$ very small, it follows that $\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = o(t^{-\lambda})$ and $J(\boldsymbol{\theta}_t) = o(t^{-\lambda})$ whenever*

$$\lambda < \min\{1 - 2\delta, \gamma\}. \tag{4.3.21}$$

*Proof.* Recall the bound (4.3.17) and the definitions of $f_t, g_t$ from (4.3.18) and (4.3.19) respectively. Replacing the property (KL') by property (PL) allows us to replace the term $-\alpha_t \|\nabla J(\boldsymbol{\theta}_t)\|_2^2$ in (4.3.17) by $-\alpha_t K J(\boldsymbol{\theta}_t)$. This makes Theorem 2.24 applicable to the resulting bound. Under the stated hypotheses, it readily follows that

$$\alpha_t^2 = O(t^{-2+2\phi}), \alpha_t^2 M_t^2 = O(t^{-2+2(\phi+\delta)}), \alpha_t B_t = O(t^{-1+\phi-\gamma}).$$

Now define $\nu$ as in (4.3.20). Then each of the above three terms is $O(t^{-(1+\nu)})$, while both $\{\alpha_t^2 B_t^2\}$ and $\{\alpha_t^2 B_t\}$ decay even faster. Hence, with $\nu$ defined as in (4.3.20),

$$f_t, g_t = O(t^{-(1+\nu)}),$$

and both sequences are summable.

Now we are in a position to apply Theorem 2.24. We can conclude that $J(\boldsymbol{\theta}_t) = o(t^{-\lambda})$ whenever $2\alpha_t - \lambda t^{-1} \geq 0$ for sufficiently large $t$, and

$$\{(t+1)^{\lambda} g_t\} \in \ell_1,$$

$$\sum_{t=1}^{\infty} [2\alpha_t - \lambda t^{-1}] = \infty. \tag{4.3.22}$$

Now observe that $2\alpha_t = \Omega(t^{-(1-C)})$, and $C > 0$. Choose a contant $D$ such that $2\alpha_t \geq D t^{-(1-C)}$ for sufficiently large $t$. Then, whatever be the value of $\lambda$, it is clear that

$$D t^{-(1-C)} - \lambda t^{-1} \geq 0$$

for sufficiently large $t$. Also, since $C > 0$, it is evident that $\alpha_t$ decays more slowly than $\lambda t^{-1}$. Hence (4.3.22) is satisfied. Thus the last step of the proof is to determine conditions under which $\{(t+1)^{\lambda} g_t\} \in \ell_1$. Since $g_t = O(t^{-(1+\nu)})$, it follows that $(t+1)^{\lambda} g_t = O(t^{-(1+\nu-\lambda)})$, which is summable if $\lambda < \nu$. Hence it follows that $J(\boldsymbol{\theta}_t) = o(t^{-\lambda})$ whenever $\lambda < \nu$.

To prove the last statement, observe that, while there is an upper bound on $\phi$, namely $\min\{0.5 - \delta, \gamma\}$, there is no lower bound. So we can choose $\phi = \epsilon$, a very small number. This leads to

$$\lambda < \nu = \min\{1 - 2\delta - 2\epsilon, \gamma - \epsilon\}.$$

But since $\epsilon$ can be made arbitrarily small, this translates to (4.3.21). □

**Corollary 4.1.** *Suppose all hypotheses of Theorem 4.2 hold. In particular, if $B_t = 0$ for all large enough $t$ in (4.3.6), and $M_t$ in (4.3.7) is bounded with respect to $t$, then $\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = o(t^{-\lambda})$ and $J(\boldsymbol{\theta}_t) = o(t^{-\lambda})$ for all $\lambda < 1$.*

The proof is immediate from Theorem 4.2. With $B_t = 0$, one can take $\gamma = 1$, and with $M_t$ being bounded, one can take $\delta = 0$. Substituting these into (4.3.21) leads to the desired conclusion.

**Remark:** It is worthwhile to compare the content of Corollary 4.1 with the bounds from [4], as summarized in Section 4.2.1. In that paper, it is assumed that $\mathbf{z}_t = \nabla J(\boldsymbol{\theta}_t)$, and that there is a finite constant $M$ such that $CV_t(\mathbf{h}_{t+1}) \leq M^2$; see [4, Eq. (2)]. In the present notation, this is the same as saying that $B_t = 0$ for all $t$, and that $M_t = M$ for all $t$. With these assumptions on the stochastic gradient, it is shown that for an arbitrary convex function, the best achievable rate for a convex objective function is that $\|\nabla J(\boldsymbol{\theta}_t)\|_2 = O(t^{-1/2})$. Now suppose a function $J(\cdot)$ satisfies both Standing Assumptions (J1), (J2) and the (PL) property. Thus there exists a constant $K$ such that

$$KJ(\boldsymbol{\theta}_t) \leq \|\nabla J(\boldsymbol{\theta}_t)\|_2^2 \leq 2LJ(\boldsymbol{\theta}_t).$$

Then, as per Corollary 4.1, it follows that $J(\boldsymbol{\theta}_t) = o(t^{-\lambda})$ and $\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = o(t^{-\lambda})$ for every $\lambda < 1$. There is virtually no difference between $O(t^{-1})$ and $o(t^{-\lambda})$ for all $\lambda < 1$. Thus our results extend the bounds from [4] from convex functions to a somewhat larger class, namely those that satisfy Assumption (S3) as well as the Polyak-Łojasiewicz condition.

Next, we study stochastic gradient methods based on function evaluations alone. The Simultaneous Perturbation SA (SPSA), described in (4.2.10), is typical of this approach. In this equation, two function evaluations are used at each step; however, there exist approaches that use only one function evaluation at each step. For the stochastic gradient of (4.2.10), the quantities $B_t$ and $M_t$ satisfy

$$B_t = O(c_t), \quad M_t^2 = (1/c_t^2). \tag{4.3.23}$$

A more general approach, somewhat reminiscent of the Runge-Kutta method, is proposed in [112], wherein $k+1$ function evaluations are used at each step, leading to

$$B_t = O(c_t^k), \quad M_t^2 = (1/c_t^2), \tag{4.3.24}$$

which reduces to the above when $k = 1$. This observation raises the question as to whether there is an "optimal" choice of the "increment" $c_t$, so as to achieve the fastest convergence. Specifically, suppose we choose $c_t = \Theta(t^s)$ for some exponent $s$. What is the choice of $s$ that maximizes the bound $\nu$ in (3.1.2)?

**Corollary 4.2.** *Suppose all hypotheses of Theorem 4.2 hold. Suppose $B_t$, $M_t$ satisfy (4.3.23) for arbitrary increment $c_t$, and that $c_t = \Theta(t^{-1})$. Then the optimal choice for the exponent $s$ is $1/3$. Then, with $\alpha_t = O(t^{-(1-\phi)})$, by choosing $\phi = \epsilon > 0$ arbitrarily small, and $s = (1-\epsilon)/3$, we get*

$$J(\boldsymbol{\theta}_t), \|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = o(t^{-\lambda}), \ \forall \lambda < 1/3. \tag{4.3.25}$$

*More generally, suppose $B_t$, $M_t$ satisfy (4.3.24) for arbitrary increment $c_t$. Then, with $\alpha_t = O(t^{-(1-\phi)})$, by choosing $\phi = \epsilon > 0$ arbitrarily small, and $s = (1-\epsilon)/(k+2)$, we get*

$$J(\boldsymbol{\theta}_t), \|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = o(t^{-\lambda}), \ \forall \lambda < k/(k+2). \tag{4.3.26}$$

*Proof.* With $c_t = O(t^{-s})$, it is already known from [77] that

$$B_t = O(c_t) = O(t^{-s}), \quad M_t^2 = O(1/c_t^2) = O(t^{2s}).$$

Hence we can apply Theorem 4.2 with $\gamma = s, \delta = 2s$. Then the rate of convergence becomes $o(t^{-\lambda})$ whenever $\lambda \in (0, \nu)$, and

$$\nu = \min\{1 - 2(\phi + s), s - \phi\}.$$

To motivate the proof, we depict these two inequalities and the "optimal" choice of $s$ for the case $k = 1$. Figure 4.3 depicts the two inequalities

$$1 - 2(\phi + s) \geq 0, s - \phi \geq 0,$$

or

$$\phi + s \leq 0.5, \phi \leq s.$$

The blue line depicts when both parts of the minimum defining $\nu$ are equal, namely $3s + \phi = 1$. Along this line, $\mu$ is maximum when $s = 1/3$ and $\phi = 0$, where $\mu = 1/3$. In reality the inequalities should be strict. Hence, for arbitrarily small $\epsilon > 0$, we can choose

$$\phi = \epsilon, \quad s = \frac{1-\epsilon}{3}, \quad \mu = \frac{1}{3} - \frac{4\epsilon}{3}.$$

But since $\epsilon$ is arbitrary, this works out to $\mu < 1/3$. Hence (4.3.25) follows. In the case of general $k$, we have

$$1 - 2(\phi + s) = ks - \phi, \text{ or } (k+2)s + \phi = 1.$$

So by choosing $\phi = \epsilon$, we get

$$s = \frac{1-\epsilon}{k+2}, \quad \mu = \frac{k(1-\epsilon)}{k+2} - \epsilon = \frac{k}{k+2} - \epsilon\frac{2k+2}{k+2}.$$

Again, since $\epsilon$ is arbitrary, (4.3.26) follows. □

It is worth noting that, when $k+1$ function evaluations are carried out, not only is the convergence rate faster, but the step sizes also become larger ($O(t^{k/(k+2)})$).

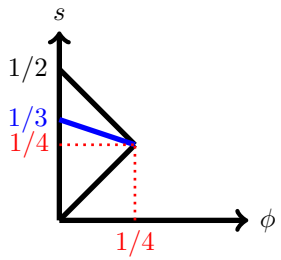



Figure 4.3: Feasible combinations of $(\phi, s)$

**Remarks:** Now we discuss the significance of Corollary 4.2 and its relationship to previously known results.

1. The analysis in [4] on the achievable rates of convergence applies only when the stochastic gradient is unbiased ($B_t = 0$ for all $t$), and its conditional variance is bounded. When only function evaluations are used to construct a stochastic gradient, these assumptions do not hold. Corollary 4.2 partially fills this gap.

2. In [110], the authors study what would be called Simultaneous Perturbation SA with two measurements (but with a Gaussian perturbation vector instead of Rademacher perturbations). It is shown that the iterations converge at the rate $J(\boldsymbol{\theta}_t) = O(t^{-1/2})$. However, there is no error in the measurements, and the objective function is restricted to be convex. In contrast, in the present situation, a rate of $o(t^{-\lambda})$ is achieved for $\lambda < 1/3$ even in the presence of measurement errors, and for a class of nonconvex objective funtions. Moreover, by choosing $k = 2$ in the approach of [112], that is, by carrying out *three* function evaluations at each step, the rate goes up to $\lambda < 1/2$, the same as in [110]. By letting $k \to \infty$, one can make $\lambda$ arbitrarily close to one. In the view of the author, this last observation is only of theoretical interest.

## 4.4 A Unified Theory for Momentum-Based Methods

In this section, we set up a general class of momentum-based algorithms that includes both the Stochastic Heavy Ball (SHB) and the Stochastic Nesterov Accelerated Gradient (SNAG) algorithms as special cases, with suitable choices of the parameters. Then we state and prove sufficient conditions for the convergence of the general algorithm. Obviously, these sufficient conditions would then guarantee the SHB algorithm. However, the theory does not apply to the standard version of SNAG, in which the momentum parameter approaches one. Rather, it applies to a variant of SNAG.

Recall that the problem is to minimize a $\mathcal{C}^1$ objective function $J : \mathbb{R}^d \to \mathbb{R}$. As before, it is assumed that $J(\cdot)$ satisfies assumptions (J1) and (J2) stated earlier.

### 4.4.1 A Unified Momentum-Based Algorithm

The iterative algorithm is not based on updating $\boldsymbol{\theta}_t$ directly. Rather, it is defined in terms of two auxiliary vectors, denoted here by $\mathbf{w}_t$ and $\mathbf{v}_t$. The relationship between $\mathbf{w}_t$ and $\boldsymbol{\theta}_t$ is given by

$$\mathbf{w}_t = \boldsymbol{\theta}_t + \epsilon_t \mathbf{v}_t. \tag{4.4.1}$$

The general algorithm consists of updating formulas for $\mathbf{w}_t$ and $\mathbf{v}_t$, as follows:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + a_t \mathbf{v}_t - b_t \alpha_t \mathbf{h}_{t+1}, \tag{4.4.2}$$

$$\mathbf{v}_{t+1} = \mu_t \mathbf{v}_t - \alpha_t \mathbf{h}_{t+1}, \tag{4.4.3}$$

where, as always, $\alpha_t$ is the step size, while $\mu_t$ is known as the **momentum parameter**. In addition, $\{a_t\}$, $\{b_t\}$, $\{\epsilon_t\}$ are sequences of real constants that can be adjusted to make (4.4.2)–(4.4.3) mimic various standard algorithms. Usually they are viewed differently from the sequences $\{\mu_t\}$ and $\{\alpha_t\}$. Further, $\mathbf{h}_{t+1}$ is a random vector that is an approximation to $\nabla J(\mathbf{w}_t)$ (note, not necessarily to $\nabla J(\boldsymbol{\theta}_t)$), known as the **stochastic gradient**. All of our analysis pertains to the behavior of $\mathbf{w}_t$ and $\mathbf{v}_t$. However, the conclusions can be translated back to the behavior of the original argument variable $\boldsymbol{\theta}_t$, using (4.4.1).

Now it is shown that both SHB and SNAG are special cases of (4.4.2) and (4.4.3) for suitable choices of the various constants. Since Stochastic Gradient Descent (SGD) is a special case of SHB, it too is a special case of the above algorithm. However, SGD is not a momentum-based algorithm.

In the present context, the objective is to solve the equation $\nabla J(\boldsymbol{\theta}) = \mathbf{0}$ using noisy measurements of the gradient. Recall from (4.2.11) that the general formulation of SHB studied here is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \mu_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) - \alpha_t \mathbf{h}_{t+1}, \tag{4.4.4}$$

where $\alpha_t$ is the step size, $\mu_t$ is the momentum parameter, and $\mathbf{h}_{t+1}$ is a random approximation to $\nabla J(\boldsymbol{\theta}_t)$. The Heavy Ball method was first introduced in [113], where both $\alpha_t$ and $\mu_t$ are fixed constants.

To put (4.4.4) in the form (4.4.2)–(4.4.3), define

$$\mathbf{v}_t := \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}, \mathbf{w}_t = \boldsymbol{\theta}_t. \tag{4.4.5}$$

With these definitions, it is easy to show that the update equations for $\mathbf{w}_t$ and $\mathbf{v}_t$ are

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \mu_t \mathbf{v}_t - \alpha_t \mathbf{h}_{t+1}, \tag{4.4.6}$$

$$\mathbf{v}_{t+1} = \mu_t \mathbf{v}_t - \alpha_t \mathbf{h}_{t+1}. \tag{4.4.7}$$

These equations are of the form (4.4.2)–(4.4.3) if we define

$$\epsilon_t = 0, a_t = \mu_t, b_t = 1.$$

Moreover, since $\mathbf{w}_t = \boldsymbol{\theta}_t$, the stochastic gradient $\mathbf{h}_{t+1}$ is a random approximation to $\nabla J(\boldsymbol{\theta}_t)$.

The Nesterov Accelerated Gradient (NSG) algorithm was introduced in [107]. In the current notation, with the possibility of the gradient being stochastic, and the momentum coefficient being allowed to vary with $t$, it can be stated as follows (following [143, Eqs. (3)–(4)]):

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \mu_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) - \alpha_t \mathbf{h}_{t+1}, \tag{4.4.8}$$

where $\mathbf{h}_{t+1}$ is a random approximation of $\nabla J(\boldsymbol{\theta}_t + \mu_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}))$, and not $\nabla J(\boldsymbol{\theta}_t)$. We analyze (4.4.8) using the reformulation in [8, Eqs. (6)–(7)], stated here as (4.2.16) and (4.2.17). To accommodate the shift in the argument of $\nabla J(\cdot)$, we proceed as follows: Define

$$\mathbf{v}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}, \mathbf{w}_t = \boldsymbol{\theta}_t + \mu_t \mathbf{v}_t. \tag{4.4.9}$$

Then the updating formulas are given by [8, Eqs. (6)–(7)] as

$$\mathbf{v}_{t+1} = \mu_t \mathbf{v}_t - \alpha_t \mathbf{h}_{t+1}, \tag{4.4.10}$$

which is the same as (4.4.7), and

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \mu_{t+1}\mu_t \mathbf{v}_t - (1 + \mu_{t+1})\alpha_t \mathbf{h}_{t+1}. \tag{4.4.11}$$

These equations are of the form (4.4.2) and (4.4.3) with

$$\epsilon_t = 1 \; \forall t, a_t = \mu_{t+1}\mu_t, b_t = 1 + \mu_{t+1}. \tag{4.4.12}$$

Once again, as can be seen from (4.2.17), the random search direction $\mathbf{h}_{t+1}$ is an approximation to $\nabla J(\mathbf{w}_t)$.

Finally, since SGD is a special case of SHB with $\mu_t \equiv 0$ for all $t$, it too is a special case of the general algorithm (4.4.2)–(4.4.3).

In this context we mention [92] and its predecessors [135, 177] which present a "Stochastic Unified Momentum (SUM)" algorithm. In the paper [92], the objective function is of the form

$$J(\boldsymbol{\theta}) = E_{\mathbf{w} \sim P} F(\boldsymbol{\theta}, \mathbf{w}).$$

The SUM algorithm consists of two coupled equations (in their notation):

$$m_t = \mu m_{t-1} - \eta_t g_t, \quad x_{t+1} = x_t - \lambda \eta_t g_t + (1 - \lambda) m_t.$$

Other than the fact that the momentum coefficient $\mu$ is constant, the only difference between the above, and (4.4.3)–(4.4.1), is that the above has a "convex combination" of two terms, which is absent in our formulation. But this is a minor detail. Hence it is not claimed that our unified algorithm itself is more general. Rather, the generality is in the conclusions. We can prove a stronger form of convergence, under conditions that are analogous to the standard Robbins-Monro conditions.

### 4.4.2 Literature Review

Next, we present a very brief review of the relevant results from the literature on SHB and SNAG, to provide a point of departure to compare the results in this sectio against those. A more detailed review of momentum-based algorithms is given in [121, Section 1.1].

After the publication of the two seminal papers [113] and [107], a great deal of analysis has been carried out on these algorithms. The approach adopted in this section is to analyze momentum-based algorithms using the contents of Section 2.3, what might be called the "almost supermartingale" approach. However, there is considerable literature on the asymptotic behavior of the ODEs on $\mathbb{R}^d$ associated with these algorithms. Whereas the ODE associated with SGD (described in (4.3.1)) is of first-order, the ODEs associated with the SHB and SNAG methods are of second-order, due to the presence of the "delay" terms. The ODE associated with NAG is analyzed in detail in [142], when the step size $\alpha$ is held constant, while the momentum coefficient

$\mu_t \to 1$ as $t \to \infty$. This is consistent with the standard formulation of SNAG, whereas in SHB, the momentum parameter is constant while the step size varies with time. In [142], it is shown that the "optimal" schedule is $\mu_t = (t+2)/(t+5)$. In [5], the rate of convergence of this ODE is analyzed further by imposing additional structure on $J(\cdot)$, such as the Kurdyka-Łojasiewicz property. It is shown that, in certain situations, it is possible for classical steepest descent method to outperform NAG. The second-order ODE associated with HB is analyzed in [2, 6], when $J(\cdot)$ satisfies the Polyak-Łojasiewicz property. In all of the above formulations, it is assumed that the "stochastic gradient" $\mathbf{h}_{t+1}$ equals the true gradient $\nabla J(\boldsymbol{\theta}_t)$; thus these models do not allow for measurement errors. Hence the analysis applies only to HB or NAG, not SHB or SNAG.

Now we come to more recent papers on SHB, which *do* permit measurement errors. In much of the literature, it is assumed that $J(\cdot)$ is *convex*; here we replace convexity by the weaker properties (PL) and (KL'). Moreover, in many papers, attention is focused in the *convergence in expectation*, or *convrgence in probability* of various algorithms. In the review paper [26], the emphasis is almost exclusively on convergence in expectation. SHB and SNAG are discussed in [26, Section 7]. In other papers, the conclusions are even weaker: It is shown only that

$$\lim_{t\to\infty} \min_{1\le\tau\le t} E[\|\nabla J(\boldsymbol{\theta}_\tau)\|_2^2 = 0. \tag{4.4.13}$$

The above conclusion is weaker than

$$\liminf_{t\to\infty} E[\|\nabla J(\boldsymbol{\theta}_t)\|_2^2] = 0. \tag{4.4.14}$$

This is because, if $E[\|\nabla J(\boldsymbol{\theta}_t)\|_2^2] = 0$ for some $t$, then (4.4.13) holds, but not necessarily (4.4.14). Basically (4.4.13) is forward-looking, while (4.4.14) is backward-looking. Similarly, the conclusion that

$$\min_{1\le\tau\le t} \|\nabla J(\boldsymbol{\theta}_t)\|_2^2 \to 0$$

in probability is a weaker conclusion that

$$\liminf_{t\to\infty} \|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = 0,$$

where again the convergence is in probability.

Other research on the convergence of HB (without establishing almost sure convergence) is summarized very well on page 3 of [130] and Section 1.1 of [93].

In [50], the authors analyze the HB algorithm where $\mathbf{h}_{t+1} = \nabla J(\boldsymbol{\theta}_t)$; thus there is no provision for measurement noise, so that the algorithm being analyzed is HB and not SHB. The function $J(\cdot)$ is assumed to be convex, and to have a globally Lipschitz-continuous gradient. The authors *do not* show that $J(\boldsymbol{\theta}_t)$ converges to the global minimum of $J(\cdot)$. Rather, they show that *the average of the first t iterations* converges to the minimum value of the function $J(\cdot)$. This is somewhat in the same spirit as the papers [117, 68], in which the authors show that the average of the first $t$ iterations of $\boldsymbol{\theta}_t$ converges to the minimizer of $J(\cdot)$. In [49], the authors study the SHB for some classes of nonconvex functions. It is assumed that the stochastic gradient is unbiased, i.e., that $E_t(\mathbf{h}_{t+1}) = \nabla J(\boldsymbol{\theta}_t)$. The iterations are shown to converge to a minimum, but at the cost of "uniformly elliptic bounds" on the measurement error $\boldsymbol{\zeta}_{t+1}$, which are very restrictive.

Now we discuss in detail a couple of papers that are most closely related to the present subsection. In this context, it is very useful to know that the algorithm converges to the desired limit *almost surely*. This is because any stochastic algorithm generates *one sample path* of a stochastic process, and it is therefore essential to know that almost all sample paths converge to the desired answer. However, there are only a handful of papers that establish the almost-sure convergence of SHB and/or SNAG. These are discussed in detail in this subsection.

In [130], the objective function is an expected value, of the form ([130, Eq. (1)])

$$J(\boldsymbol{\theta}) = E_{\mathbf{w}\sim P} F(\boldsymbol{\theta}, \mathbf{w}).$$

The function $F(\cdot, \mathbf{w})$ is convex for each $\mathbf{w}$, and its gradient is Lipschitz-continuous with constant $L_{\mathbf{w}} \leq L$ for all $\mathbf{w}$. Thus the same holds for $J(\cdot)$ as well. The stochastic gradient is chosen as ([130, Eq. (SHB)])

$$\mathbf{h}_{t+1} = \nabla_{\boldsymbol{\theta}_t} F(\mathbf{w}_{t+1}, \boldsymbol{\theta}_t),$$

where $\mathbf{w}_{t+1}$ is chosen i.i.d. with distribution $P$. Effectively this means that the stochastic gradient is unbiased. Also, it is assumed that, for some constant $\sigma^2$, the conditional variance $CV_t(\mathbf{h}_{t+1})$ of the stochastic gradient is bounded by ([130, Eq. (5)])

$$CV_t(\mathbf{h}_{t+1}) \leq 4L(J(\boldsymbol{\theta}_t) - J^*) + \sigma^2,$$

where $J^*$ is the infimum of $J(\cdot)$. In [130] the authors study the SHB with time-varying parameter $\mu_t$, namely

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t \mathbf{h}_{t+1} + \mu_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}), \tag{4.4.15}$$

It is suggested how to convert (4.4.15) above into *two* equations, which do not contain any "delayed" terms. Specifically, the authors iteratively define

$$\lambda_{t+1} = \frac{\lambda_t}{\mu_t} - 1, \eta_t = (1 + \lambda_{t+1})\alpha_t \tag{4.4.16}$$

In the above, the quantity $\lambda_0$ is not specified and is chosen by the user. They then define[5]

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{h}_{t+1}, \tag{4.4.17}$$

$$\boldsymbol{\theta}_{t+1} = \frac{\lambda_{t+1}}{1 + \lambda_{t+1}} \boldsymbol{\theta}_t + \frac{1}{1 + \lambda_{t+1}} \mathbf{w}_{t+1}. \tag{4.4.18}$$

Then $\boldsymbol{\theta}_{t+1}$ satisfies (4.4.15).

The convergence of (4.4.17)–(4.4.18) is established under [130, Condition 1], namely the sequence $\{\eta_t\}$ is decreasing, and moreover

$$\sum_{t=0}^{\infty} \eta_t = \infty, \quad \sum_{t=0}^{\infty} \eta_t^2 \sigma^2 < \infty, \quad \sum_{t=1}^{\infty} \frac{\eta_t}{\sum_{\tau=0}^{t-1} \eta_\tau} = \infty. \tag{4.4.19}$$

Thus in [130] the original step size sequence $\{\alpha_t\}$ and momentum sequence $\{\mu_t\}$ are replaced by the "synthetic" step size sequence $\{\eta_t\}$, and the convergence conditions are stated in terms of $\eta_t$. It is shown that, in general, $J(\boldsymbol{\theta}_t) \to J^*$ where $J^*$ is the minimum value of $J(\cdot)$, at a rate of $O(t^{-1/2})$. In the "over-parametrized" case, the rate improves to $O(t^{-1})$. Moreover, the iterations $\boldsymbol{\theta}_t$ converge to a minimizer of $J(\cdot)$.

Now we give our interpretation of the results in [130]. There are two restrictive features of these results. First, the conditions (4.4.19) are more stringent than the standard Robbins-Monro conditions, namely

$$\sum_{t=0}^{\infty} \eta_t^2 < \infty, \quad \sum_{t=0}^{\infty} \eta_t = \infty, \tag{4.4.20}$$

Compared to (4.4.20), there are two extra assumptions in (4.4.19), namely: (i) the synthetic step size $\eta_t$ is decreasing, and (ii) the summation of $\eta_t / \sum_{\tau=0}^{t-1} \eta_\tau$ is divergent. Since $S_t$ is an increasing sequence, the divergence of this summation is a more restrictive assumption than the second Robbins-Monro condition in (4.4.20)

The second challenge in this approach is that, given the *original* step size and momentum sequences, there is no easy way to verify whether (4.4.19) is satisfied. This is why, in [130, Theorem 8], the authors

[5]To facilitate a comparison with the original paper, we use the same symbol $\mathbf{w}_t$. However, their quantity $\mathbf{w}_t$ is closer to our $\mathbf{u}_t$ defined in Section 3.3.2.

*begin with* the sequence $\{\eta_t\}$, which appears to us to be somewhat unnatural. If $\mu_t = \mu$, a fixed constant, for all $t$, then a possible solution to (4.4.16) is

$$\lambda_t = \lambda_0 = \frac{\mu}{1-\mu},\ \forall t, \eta_t = \frac{1}{1-\mu}\alpha_t,\ \forall t.$$

Since $\eta_t$ is a constant multiple of $\alpha_t$, if $\{\alpha_t\}$ satisfies (4.4.20), then so does $\{\eta_t\}$. However, if $\mu_t$ varies as a function of $t$, this approach will not work. Specifically, it is shown in Section 4.5.3 that, if the momentum coefficient $\mu_t$ is monotonically decreasing, then $\lambda_t \to \infty$ as $t \to \infty$. Consequently, a Robbins-Monro like assumpion such as $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$ *need not imply* that $\sum_{t=0}^{\infty} \eta_t^2 < \infty$. In the other direction, if $\mu_t$ is monotonically increasing but bounded away from 1, them there exsts a finite $T$ such that $1 + \lambda_{t+1} < 0$ for all $t \geq T$, thus causing the "step size" $\eta_t$ to become negative, which is absurd.

In contrast, the approach proposed here can handle the case where not just the momentum parameter $\mu_t$ is time-varying, but *all* parameters vary with $t$. Moreover, the conditions for convergence reduce to the familiar Robbins-Monro conditions if the stochastic gradient is unbiased and has finite variance (even if the parameters vary with $t$). In the more general case where the stochastic gradient is biased, and/or the conditional variance of the stochastic gradient grows without bound as a function of $t$, the conditions for convergence are those in Theorem 4.1. As we have seen earlier, this formulation allows us to handle the so-called zeroth-order methods, wherein the stochastic gradient is computed using only noisy measurements of the objective function.

Next we come to [93]. The analysis in [130] is applicable only to *convex* objective functions. In [93], the authors prove results that are applicable to arbitrary nonconvex functions that have a Lipschitz-continuous gradient. However, for nonconvex funtions, they can prove only that

$$\lim_{t\to\infty} \min_{0\leq\tau\leq t} \|\nabla J(\boldsymbol{\theta}_\tau)\|_2^2 = 0. \tag{4.4.21}$$

Clearly, this is a weaker conclusion than $\nabla J(\boldsymbol{\theta}_t) \to \mathbf{0}$ as $t \to \infty$. To prove that conclusion, they assume that $J(\cdot)$ is *strongly convex*. They also relax the bound on the conditional variance of the stochastic gradient to the so-called Expected Smoothness assumption of [75], namely

$$E_t(\|\mathbf{h}_{t+1}\|_2^2) \leq 2AJ(\boldsymbol{\theta}_t) + B\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 + C, \tag{4.4.22}$$

for suitable constants $A, B, C$. This is proposed in [75] as "the weakest assumption" for analyzing the convergence of SGD or SHB for nonconvex functions. However, unlike in [130], these authors assume that the momentum term is a constant, that is, $\mu_t = \mu\ \forall t$.

After the brief literature review, we now compare our results to those of [93]. Throughout, we replace the variance bound (4.4.22) by weaker bound (4.4.28). We also permit the momentum parameter $\mu_t$ to vary with $t$, which is not possible in the method of proof used in [93]. When no convexity of any type is assumed, and the only assumption is that $\nabla J(\cdot)$ is Lipschitz-continuous, we are able to show that

$$\liminf_{t\to\infty} \|\nabla J(\boldsymbol{\theta}_t)\|_2 = 0. \tag{4.4.23}$$

Given any sequence of nonnegative numbers $\{x_t\}$, it is easy to show that

$$\liminf_{t\to\infty} x_t = 0 \implies \lim_{t\to\infty} \min_{0\leq\tau\leq t} x_\tau = 0,$$

but the converse need not be true. (Suppose $x_T = 0$ for some $T$ but $x_t \geq \epsilon > 0$ for all $t > T$.) Hence our conclusion (4.4.23) is stronger than (4.4.21). Next, we permit a mild form of nonconvexity (namely the KL or PL properties). In this more general setting, we nevertheless derive the almost sure convergence of the iterations, when the Robbins-Monro or Kiefer-Wolfowitz-Blum conditions are satisfied.

Now let us return to [92]. In that paper, it is assumed that the stochastic gradient is unbiased and has uniformly bounded variance, whereas we permit a more general type of stochastic gradient, which satisfies

(4.4.27)–(4.4.28). Our conclusions are also stronger. Under the Robbins-Monro or Kiefer-Wolfowitz-Blum conditions, when $J(\cdot)$ satisfies the (KL) property, we deduce that $\boldsymbol{\theta}_t$ converges almost surely to the set of minimizers. When $J(\cdot)$ satisfies the stronger (PL) property, we can bound the rate of convergence. Finally, if the only assumption is that $\nabla J(\cdot)$ is Lipschitz-continuous, we are able to show that

$$\liminf_{t\to\infty} \|\nabla J(\boldsymbol{\theta}_t)\|_2 = 0.$$

In contrast, in [92], the authors show only that

$$\lim_{t\to\infty} \min_{1\leq\tau\leq t} E[\|\nabla J(\boldsymbol{\theta}_\tau)\|_2^2 = 0.$$

This is a weaker conclusion, as shown above.

### 4.4.3 Statements of Main Theorems

In this subsection we state the main theorems concerning the unified momentum approach. The proofs are given in the next subsection.

#### Assumptions on the Stochastic Gradient

Let $\mathcal{F}_t$ denote the $\sigma$-algebra generated by $\boldsymbol{\theta}_0, \mathbf{h}_1^t$, where $\mathbf{h}_1^t$ denotes $(\mathbf{h}_1, \cdots \mathbf{h}_t)$; note that there is no $\mathbf{h}_0$. As before, for an $\mathbb{R}^d$-valued random variable $X$, let $E_t(X)$ denote the **conditional expectation** $E(X|\mathcal{F}_t)$, and let $CV_t(X)$ denote its **conditional variance** defined by

$$CV_t(X) = E_t(\|X - E_t(X)\|_2^2) = E_t(\|X\|_2^2) - \|E_t(X)\|_2^2. \tag{4.4.24}$$

With these notational conventions in place we state the assumptions on $\mathbf{h}_{t+1}$. We begin by defining

$$\mathbf{z}_t = E_t(\mathbf{h}_{t+1}), \quad \mathbf{x}_t = \mathbf{z}_t - \nabla J(\mathbf{w}_t), \quad \boldsymbol{\zeta}_{t+1} = \mathbf{h}_{t+1} - \mathbf{z}_t. \tag{4.4.25}$$

Thus $\mathbf{x}_t$ denotes the "bias" of the stochastic gradient. If $\mathbf{h}_{t+1}$ is an unbiased estimate of $\nabla J(\mathbf{w}_t)$, then $\mathbf{x}_t = \mathbf{0}$. Most papers in the literature assume that $\mathbf{x}_t = \mathbf{0}$, but our objective here is specifically to *permit* biased estimates. This is necessary to analyze the situation where the stochastic gradient is obtained using function valuations alone. The last equation in (4.4.25) implies that $E_t(\boldsymbol{\zeta}_{t+1}) = \mathbf{0}$. Therefore

$$E_t(\|\mathbf{h}_{t+1}\|_2^2) = \|\mathbf{z}_t\|_2^2 + E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2). \tag{4.4.26}$$

With these definitions, the assumption on the stochastic gradient is that there exist sequences of constants $\{B_t\}$ and $\{M_t\}$ such that

$$\|\mathbf{x}_t\|_2 \leq B_t[1 + \|\nabla J(\mathbf{w}_t)\|_2], \ \forall \boldsymbol{\theta}_t \in \mathbb{R}^d, \ \forall t, \tag{4.4.27}$$

$$E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2) \leq M_t^2[1 + J(\mathbf{w}_t)], \ \forall \boldsymbol{\theta}_t \in \mathbb{R}^d, \ \forall t. \tag{4.4.28}$$

Equation (4.4.27) states that the stochastic gradient $\mathbf{h}_{t+1}$ can be biased, but the extent of the bias has to be bounded by a constant plus the norm of the gradient. As we will see in subsequent sections, while $B_t$ is permitted to be nonzero, eventually it has to approach zero; in other words, the stochastic gradient has to be "asymptotically unbiased." In contrast, (4.4.28) states that the conditional variance of the stochastic gradient can grow as a function of the iteration counter $t$. This feature is essential to permit the analysis of so-called zeroth-order methods, where only a small number (often just two) of function evaluations are used to construct $\mathbf{h}_{t+1}$.

**Assumptions on the Constants**

Aside from the step length $\alpha_t$, there are four constants in the algorithm (4.4.2)–(4.4.3). The assumptions on these constants are as follows: There exist constants $\bar{a}, \underline{b}, \bar{b}, \bar{\mu}, \bar{\epsilon}$ such that, for all $t$, we have

$$0 \leq a_t \leq \bar{a}, 0 < \underline{b} \leq b_t \leq \bar{b}, 0 \leq \mu_t \leq \bar{\mu} < 1, |\epsilon_t| \leq \bar{\epsilon} < \infty. \tag{4.4.29}$$

Now we discuss a few implications of the above bounds. First, $a_t$ is always nonnegative and bounded above. Second, $b_t$ is bounded both below and above by positive constants. Third, the momentum coefficient $\mu_t$ can equal zero, but is bounded away from 1. Finally, $\epsilon_t$ can be either positive or negative, but is bounded in magnitude. Observe that when SHB is formulated as a special case of (4.4.2)–(4.4.3), the assumptions in (4.4.29) hold. As for SNAG, in the *traditional* formulation, the momentum parameter $\mu_t \uparrow 1$ as $t \to \infty$. Hence the assumptions in (4.4.29) do not hold. What is analyzed here is a nonstandard version of SNAG in which (4.4.29) hold. The version of SNAG analyzed in [93] is even more restrictive in that $\mu_t$ is a fixed constant less than one.

Two ready consequences of these assumptions are that, if we define

$$k_t := \frac{a_t}{(1-\mu_t)}, \bar{k} := \frac{\bar{a}}{1-\bar{\mu}}, \tag{4.4.30}$$

then

$$k_t \in [0, \bar{k}], b_t + k_{t+1} \in [\underline{b}, \bar{b} + \bar{a}/(1-\bar{\mu})]. \tag{4.4.31}$$

A key assumption is this: Define $\delta_t := k_{t+1} - k_t$. Then

$$\delta_t \to 0 \text{ as } t \to \infty. \tag{4.4.32}$$

Note that there are no restrictions on the *sign* of $\delta_t$. This assumption is readily satisfied if both $\{a_t\}$ and $\{\mu_t\}$ converge to some limits. The assumption allows us to transform the variables in (4.4.2)–(4.4.3) in such a way that the resulting transformed equations are "asymptotically decoupled." More details can be found below.

In our analysis, it is quite permissible to allow *all five* constants $a_t$, $b_t$, $\epsilon_t$, $\mu_t$, $\alpha_t$ to be random variables. In this case, the bounds in (4.4.29) and (4.4.31) hold almost surely. If we define $\mathcal{F}_t$ to be the $\sigma$-algebra generated by $\boldsymbol{\theta}_0$ and $\mathbf{h}_1^t$, then all of these constants need to belong to $\mathcal{M}(\mathcal{F}_t)$, the set of random variables that are measurable with respect to $\mathcal{F}_t$. In particular, in (4.4.12), we see that $\epsilon_t = \mu_{t+1}\mu_t$. Thus, in order to incorporate the approach of [8] in the present framework, we must assume that $\mu_{t+1} \in \mathcal{M}(\mathcal{F}_t)$, i.e., that $\{\mu_t\}$ is a **predictable** process.

With these assumptions out of the way, we now state the two main theorems regarding the convergence of the general algorithm (4.4.2) and (4.4.3)), and several corollaries thereof. In brief, when the objective function $J(\cdot)$ satisfies the (KL') property, and the analogs of the Kiefer-Wolfowitz-Blum conditions are satisfied (see (4.4.33) and (4.4.34) below), then the algorithm converges almost surely. If the hypothesis on $J(\cdot)$ is strengthened to (PL) from (KL'), then we can also derive *bounds* on the rate of convergence.

Theorem 4.3 below shows that the unified momentum algorithm converges under the same conditions as in Theorem 4.1.

**Theorem 4.3.** *Suppose that the various constants satisfy the assumptions in* (4.4.29)*, while the objective function $J(\cdot)$ satisfies Standing Assumptions (J1) and (J2). Further, suppose the stochastic gradient $\mathbf{h}_{t+1}$ satisfies the assumptions* (4.4.27)–(4.4.28)*. With these assumptions, we can state the following:*

1. *Suppose*

$$\sum_{t=0}^{\infty} \alpha_t^2 < \infty, \quad \sum_{t=0}^{\infty} \alpha_t B_t < \infty, \quad \sum_{t=0}^{\infty} \alpha_t^2 M_t^2 < \infty. \tag{4.4.33}$$

*Then $\{\nabla J(\boldsymbol{\theta}_t)\}$ and $\{J(\boldsymbol{\theta}_t)\}$ are bounded, and in addition, $J(\boldsymbol{\theta}_t)$ converges almost surely to some random variable as $t \to \infty$.*

2. *If in addition*

$$\sum_{t=0}^{\infty} \alpha_t = \infty, \tag{4.4.34}$$

*then*

$$\liminf_{t\to\infty} \|\nabla J(\boldsymbol{\theta}_t)\|_2 = 0. \tag{4.4.35}$$

3. *If, in addition to* (4.4.33) *and* (4.4.34)*, the function $J(\cdot)$ satisfies (KL'), then $J(\boldsymbol{\theta}_t) \to 0$ and $\nabla J(\boldsymbol{\theta}_t) \to \mathbf{0}$ as $t \to \infty$, where both convergences are in the almost sure sense.*

4. *Suppose that in addition to (KL'), $J(\cdot)$ also satisfies (NSC), and that* (4.4.33) *and* (4.4.34) *both hold. Then $\rho(\boldsymbol{\theta}_t) \to 0$ almost surely as $t \to \infty$.*

Now we state some useful corollaries of the above theorem.

**Corollary 4.3.** *Suppose that the various constants satisfy the assumptions in* (4.4.29)*, while the objective function $J(\cdot)$ satisfies Standing Assumptions (S1) and (S2). Further, suppose the stochastic gradient $\mathbf{h}_{t+1}$ satisfies the assumptions* (4.4.27)–(4.4.28)*, with $B_t = 0$ for all $t$, and $M_t^2 \leq M^2$ for all $t$ for some fixed constant $M$. With these assumptions, we can state the following:*

1. *Suppose*

$$\sum_{t=0}^{\infty} \alpha_t^2 < \infty. \tag{4.4.36}$$

*Then $\{\nabla J(\boldsymbol{\theta}_t)\}$ and $\{J(\boldsymbol{\theta}_t)\}$ are bounded, and in addition, $J(\boldsymbol{\theta}_t)$ converges almost surely to some random variable as $t \to \infty$.*

2. *If in addition* (4.4.34) *holds, then*

$$\liminf_{t\to\infty} \|\nabla J(\boldsymbol{\theta}_t)\|_2 = 0.$$

3. *If in addition $J(\cdot)$ satisfies (KL'), then $J(\boldsymbol{\theta}_t) \to 0$ and $\nabla J(\boldsymbol{\theta}_t) \to \mathbf{0}$ as $t \to \infty$, where both convergences are in the almost sure sense.*

4. *Suppose that in addition to (KL'), $J(\cdot)$ also satisfies (NSC), and that* (4.4.36) *and* (4.4.34) *both hold. Then $\rho(\boldsymbol{\theta}_t) \to 0$ almost surely as $t \to \infty$.*

Note that (4.4.36) and (4.4.33) are the familiar Robbins-Monro conditions introduced in [123]. Thus, when the stochastic gradient is unbiased and has bounded variance, the conditions for the convergence of the general algorithm (4.4.2)–(4.4.3) are the familiar ones for SGD, as shown in [71].

**Corollary 4.4.** *Under the assumptions of Theorem 4.3, suppose further that there exists a sequences of constants $c_t$ (known as the "increment") such that $B_t = O(c_t)$, and $M_t^2 = O(1/c_t^2)$. With these assumptions, we can state the following:*

1. *Suppose*

$$\sum_{t=0}^{\infty} \alpha_t^2 < \infty, \quad \sum_{t=0}^{\infty} \alpha_t c_t < \infty, \quad \sum_{t=0}^{\infty} (\alpha_t^2)/(c_t^2) < \infty. \tag{4.4.37}$$

*Then $\{\nabla J(\boldsymbol{\theta}_t)\}$ and $\{J(\boldsymbol{\theta}_t)\}$ are bounded, and in addition, $J(\boldsymbol{\theta}_t)$ converges almost surely to some random variable as $t \to \infty$.*

2. *If in addition $J(\cdot)$ satisfies (KL'), and* (4.4.34) *holds, then $J(\boldsymbol{\theta}_t) \to 0$ and $\nabla J(\boldsymbol{\theta}_t) \to \mathbf{0}$ as $t \to \infty$, where both convergences are in the almost sure sense.*

3. *Suppose that in addition to (KL'), $J(\cdot)$ also satisfies (NSC), and that* (4.4.36) *and* (4.4.34) *both hold. Then $\rho(\boldsymbol{\theta}_t) \to 0$ almost surely as $t \to \infty$.*

Thus the point of these two corollaries is to show that the conditions for the convergence of the unified algorithm in Theorem 4.3 are exactly the same as those for the convergence of the SGD algorithm in Theorem 4.1, even in the presence of time-varying momentum terms. In contrast, as shown in Section 4.5.3, the previously known sufficient conditions for convergence given in [130] are more restrictive.

Corollary 4.4 is relevant when the stochastic gradient is obtained using only function evaluations, and no gradient computations. It can be thought as the counterpart of Corollary 4.2 to the unified momentum algorithm.

The objective of the next theorem is to show that if the hypothesis (KL') is strengthened to (PL), then it is possible to obtain bounds on the *rate* of convergenc e.

**Theorem 4.4.** *Let various symbols be as in Theorem 4.3. Suppose $J(\cdot)$ satisfies the standing assumptions (S1) and (S2) and also property (PL), and that* (4.4.37) *and* (4.4.34) *hold. Further, suppose there exist constants $\gamma > 0$ and $\delta \geq 0$ such that*

$$B_t = O(t^{-\gamma}), \quad M_t = O(t^{\delta}), \ \forall t \geq 1,$$

*where we take $\gamma = 1$ if $B_t = 0$ for all sufficiently large $t$, and $\delta = 0$ if $M_t$ is bounded. Choose the step-size sequence $\{\alpha_t\}$ as $O(t^{-(1-\phi)})$ and $\Omega(t^{-(1-C)})$ where $\phi$ and $C$ are chosen to satisfy*

$$0 < \phi < \min\{0.5 - \delta, \gamma\}, \quad C \in (0, \phi]. \tag{4.4.38}$$

*Define*

$$\nu := \min\{1 - 2(\phi + \delta), \gamma - \phi\}. \tag{4.4.39}$$

*Then $\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = o(t^{-\lambda})$ and $J(\boldsymbol{\theta}_t) = o(t^{-\lambda})$ for every $\lambda \in (0, \nu)$. In particular, by choosing $\phi$ very small, it follows that $\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = o(t^{-\lambda})$ and $J(\boldsymbol{\theta}_t) = o(t^{-\lambda})$ whenever*

$$\lambda < \min\{1 - 2\delta, \gamma\}. \tag{4.4.40}$$

### 4.4.4 Proofs of the Main Results

In this subsection, we present the proofs of the theorems in the previous subsection.

#### Transformation of Variables

The convergence analysis of (4.4.2)–(4.4.3) is based on carrying out a linear transformation of the variables such that the resulting equations are "nearly" decoupled, and are *exactly* decoupled if all terms $a_t, b_t, \epsilon_t, \mu_t$ are constant. In contrast, in [130], the authors propose a linear transformation that achieves exact decoupling even when $\mu_t$ varies with $t$. As shown in Section 4.4.5, this approach is *untenable* when $\mu_t$ is monotonic, either decreasing or increasing. In contrast, our approach does not suffer from such limitations. Moreover, as shown in the results stated in Section 4.4.3, our conditions for the convergence of the algorithm in (4.4.2)–(4.4.3) are natural generalizations of the familiar Robbins-Monro [123] or the Kiefer-Wolfowitz-Blum [77, 17] conditions, unlike in [130].

Let us rewrite (4.4.2)–(4.4.3) as

$$\begin{bmatrix} \mathbf{w}_{t+1} \\ \mathbf{v}_{t+1} \end{bmatrix} = \begin{bmatrix} I & a_t I \\ 0 & \mu_t I \end{bmatrix} \begin{bmatrix} \mathbf{w}_t \\ \mathbf{v}_t \end{bmatrix} - \begin{bmatrix} b_t I \\ I \end{bmatrix} \alpha_t \mathbf{h}_{t+1}, \tag{4.4.41}$$

where each $I$ denotes $I_{d \times d}$. Define

$$A_t = \begin{bmatrix} I & a_t I \\ 0 & \mu_t I \end{bmatrix}, \Lambda_t = \begin{bmatrix} I & 0 \\ 0 & \mu_t I \end{bmatrix}. \tag{4.4.42}$$

Then $A_t$ is the coefficient matrix in (4.4.41) and $\Lambda_t$ is the matrix of the eigenvalues of $A_t$. In order to diagonalize $A_t$ into $\Lambda_t$, we compute the matrix of eigenvectors of $A_t$, as follows:

$$Z_t = \begin{bmatrix} I & -\frac{a_t}{1-\mu_t} I \\ 0 & \mu_t I \end{bmatrix} = \begin{bmatrix} I & -k_t I \\ 0 & \mu_t I \end{bmatrix}, Z_t^{-1} = \begin{bmatrix} I & k_t I \\ 0 & \mu_t I \end{bmatrix}, \tag{4.4.43}$$

where

$$k_t := \frac{a_t}{1-\mu_t}. \tag{4.4.44}$$

Then $Z_t^{-1} A_t Z_t = \Lambda_t$. Next, define

$$\begin{bmatrix} \mathbf{u}_t \\ \mathbf{v}_t \end{bmatrix} := Z_t^{-1} \begin{bmatrix} \mathbf{w}_t \\ \mathbf{v}_t \end{bmatrix} = \begin{bmatrix} \mathbf{w}_t + k_t \mathbf{v}_t \\ \mathbf{v}_t \end{bmatrix}. \tag{4.4.45}$$

Here we take advantage of the fact that the bottom block of $Z_t^{-1}$ is $[0 \;\; I]$. Hence, in effect, $\mathbf{w}_t$ is replaced by $\mathbf{u}_t$, but $\mathbf{v}_t$ is left unaltered. Hence the update equation for $\mathbf{v}_t$ also remains as (4.4.3).

Next we compute the update equation for $\mathbf{u}_t$.

$$\mathbf{u}_{t+1} = \mathbf{w}_{t+1} + k_{t+1}\mathbf{v}_{t+1} = \mathbf{w}_{t+1} + k_t \mathbf{v}_{t+1} + \delta_t \mathbf{v}_{t+1}, \tag{4.4.46}$$

where

$$\delta_t = k_{t+1} - k_t = \frac{a_{t+1}}{1-\mu_{t+1}} - \frac{a_t}{1-\mu_t}. \tag{4.4.47}$$

Now observe that

$$\mathbf{w}_{t+1} + k_t \mathbf{v}_{t+1} = \mathbf{w}_t + a_t \mathbf{v}_t - b_t \alpha_t \mathbf{h}_{t+1} + k_t \mu_t \mathbf{v}_t - k_t \alpha_t \mathbf{h}_{t+1}.$$

However

$$k_t \mu_t + a_t = a_t \left( \frac{\mu_t}{1-\mu_t} + 1 \right) = \frac{a_t}{1-\mu_t} = k_t.$$

Hence we can write

$$\mathbf{w}_{t+1} + k_t \mathbf{v}_{t+1} = \mathbf{w}_t + k_t \mathbf{v}_t - \alpha_t (b_t + k_t) \mathbf{h}_{t+1} = \mathbf{u}_t - \alpha_t (b_t + k_t) \mathbf{h}_{t+1}. \tag{4.4.48}$$

The last term in (4.4.46) becomes

$$\delta_t \mathbf{v}_{t+1} = \delta_t \mu_t \mathbf{v}_t - \delta_t \alpha_t \mathbf{h}_{t+1}. \tag{4.4.49}$$

Substituting from (4.4.48) and (4.4.49) into (4.4.46) gives the final form of the update equation for $\mathbf{u}_t$.

$$\begin{aligned} \mathbf{u}_{t+1} &= \mathbf{u}_t + \delta_t \mu_t \mathbf{v}_t - (b_t + k_t + \delta_t) \alpha_t \mathbf{h}_{t+1} \\ &= \mathbf{u}_t + \delta_t \mu_t \mathbf{v}_t - (b_t + k_{t+1}) \alpha_t \mathbf{h}_{t+1}, \end{aligned} \tag{4.4.50}$$

while the updating equation for $\mathbf{v}_t$ remains as before, namely

$$\mathbf{v}_{t+1} = \mu_t \mathbf{v}_t - \alpha_t \mathbf{h}_{t+1}. \tag{4.4.51}$$

These are the two equations whose behavior is analyzed in the remainder of the subsection. Based on the analysis, we make inferences about the behavior $\mathbf{w}_t$, and eventually, $\boldsymbol{\theta}_t$. Note that these two equations are *not* decoupled in general, due to the presence of the term $\delta_t \mu_t \mathbf{v}_t$ in (4.4.50). However, in the special case where both $a_t$ and $\mu_t$ are constant, then $\delta_t = 0$ for all $t$, and the equations are indeed decoupled. This is the approach used in [121] to study the SHB algorithm when $\mu_t$ is constant. More generally, if both $a_t$ and $\mu_t$ converge to some some constants as $t \to \infty$, then $\delta_t \to 0$ as $t \to \infty$, and the equations become "asymptotically decoupled." We can draw some useful conclusions when $\delta_t \to 0$ as $t \to \infty$.

**Proof of Theorem 4.3**

The proof of Theorem 4.3 is based on applying the Robbins-Siegmund theorem stated here as Theorem 2.22 to the "Lyapunov function"

$$V_t := J(\mathbf{u}_t) + \|\mathbf{v}_t\|_2^2. \tag{4.4.52}$$

The reason for calling it a "Lyapunov function" is that its conditional expectation obeys the conditions of the Robbins-Siegmund theorem; this in turn allows us to deduce the convergence of $V_t$ to zero almost surely. We will find an upper bound for $E_t(V_t)$ in the form

$$E_t(V_t) \leq V_t + f_t V_t + g_t - \frac{1-\bar{\mu}^2}{2} - \frac{\alpha_t \bar{b}}{2}\|\nabla J(\mathbf{u}_t)\|_2^2 - F_t, \tag{4.4.53}$$

where the sequences $\{f_t\}, \{g_t\}$ are nonnegative and summable, and $F_t$ is a quadratic form in $\nabla J(\mathbf{u}_t)$ and $\|\mathbf{v}_t\|_2$ which is positive definite *for sufficiently large* $t$, say for all $t \geq T$. (In case these entities are random, these conditions hold almost surely), Since we can always start our analysis at time $T$, we can neglect the term $-F_t$ for all $t \geq T$, and apply Theorem 2.22.

Going forward, we will avoid a lot of cumbersome notation if we agree to refer to a nonnegative sequence $\{z_t\}$ as a Well-Behaved Function (`WBF`) if there exist *nonnegative summable* sequences $\{f_t\}, \{g_t\}$ such that

$$z_t \leq g_t + f_t V_t, \ \forall t \geq 0. \tag{4.4.54}$$

In case the various entities are random, the assumptions (inequality and summability) hold almost surely. Clearly the sum of two `WBF` is again a `WBF`, and a `WBF` multiplied by a bounded sequence is again a `WBF`. Therefore any `WBF` can be absorbed into the terms $g_t + f_t V_t$, and it is not necessary to keep careful track of them.

Bounding $E_t(V_{t+1})$ involves several intricate computations. For this purpose, it is now shown that the first two conditions in (4.4.33) imply that

$$\sum_{t=0}^{\infty} \alpha_t^2 B_t < \infty, \quad \sum_{t=0}^{\infty} \alpha_t^2 B_t^2 < \infty. \tag{4.4.55}$$

The proof of this claim is as follows: The first bound in (4.4.33) implies in particular that $\alpha_t \to 0$ as $t \to \infty$, and hence $\alpha_t$ is bounded, say by $\bar{\alpha}$. Therefore

$$\sum_{t=0}^{\infty} \alpha_t^2 B_t \leq \bar{\alpha} \sum_{t=0}^{\infty} \alpha_t B_t < \infty.$$

This is the first bound in (4.4.55). As for the second bound, recall that every (absolutely) summable sequence is also square summable. Therefore we can append the two bounds in (4.4.55) to the three bounds in (4.4.33).

The first step in proceeding further is to reformulate the bounds (4.4.27) and (4.4.28), which are stated in terms of $\nabla J(\mathbf{w}_t)$, in terms of $\nabla J(\mathbf{u}_t)$. Accordingly, we modify (4.4.25) by defining

$$\bar{\mathbf{x}}_t = \mathbf{z}_t - \nabla J(\mathbf{u}_t) = E_t(\mathbf{h}_{t+1}) - \nabla J(\mathbf{u}_t). \tag{4.4.56}$$

The objectives are to find bounds for $\|\bar{\mathbf{x}}_t\|_2^2$ and $E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2)$ in terms of $V_t$. Throughout we use the bound (4.4.19), namely $\|\nabla J(\mathbf{u}_t)\|_2^2 \leq 2LJ(\mathbf{u}_t)$. We also make repeated use of the obvious inequalities

$$x \leq (1+x^2)/2, xy \leq (x^2+y^2)/2, \ \forall x, y \in \mathbb{R}. \tag{4.4.57}$$

We begin with a bound for $\|\bar{\mathbf{x}}_t\|_2$. Observe that

$$\bar{\mathbf{x}}_t = \mathbf{z}_t - \nabla J(\mathbf{u}_t) = \mathbf{x}_t + \nabla J(\mathbf{w}_t) - \nabla J(\mathbf{u}_t)$$

Hence it follows from (4.4.56) and (4.4.27) that

$$\begin{aligned}\|\bar{\mathbf{x}}_t\|_2 &\le \|\mathbf{x}_t\|_2 + L\|\mathbf{w}_t - \mathbf{u}_t\|_2 \\ &\le B_t(1 + \|\nabla J(\mathbf{w}_t)\|_2) + Lk_t\|\mathbf{v}_t\|_2 \\ &\le B_t(1 + \|\nabla J(\mathbf{u}_t)\|_2 + Lk_t\|\mathbf{v}_t\|_2) + Lk_t\|\mathbf{v}_t\|_2 \\ &\le B_t(1 + \|\nabla J(\mathbf{u}_t)\|_2 + L\bar{k}\|\mathbf{v}_t\|_2) + L\bar{k}\|\mathbf{v}_t\|_2.\end{aligned} \tag{4.4.58}$$

Next, we can find a bound for $\|\bar{\mathbf{x}}_t\|_2^2$ starting from (4.4.58), and arrive at

$$\begin{aligned}\|\bar{\mathbf{x}}_t\|_2^2 \le\ & B_t^2(1 + \|\nabla J(\mathbf{u}_t)\|_2 + L\bar{k}\|\mathbf{v}_t\|_2)^2 \\ & + 2B_tL\bar{k}(1 + \|\nabla J(\mathbf{u}_t)\|_2 + L\bar{k}\|\mathbf{v}_t\|_2) \cdot \|\mathbf{v}_t\|_2 + (L\bar{k})^2\|\mathbf{v}_t\|_2^2.\end{aligned} \tag{4.4.59}$$

Note that terms of the form $\|\nabla J(\mathbf{u}_t)\|_2$, $\|\mathbf{v}_t\|_2$, $\|\nabla J(\mathbf{u}_t)\|_2^2$, $\|\nabla J(\mathbf{u}_t)\|_2 \cdot \|\mathbf{v}_t\|_2$ and $\|\mathbf{v}_t\|_2^2$ can be bounded by terms of the form $C_1 + C_2V_t$ for suitable constants $C_1$ and $C_2$. Clearly $\|\mathbf{v}_t\|_2^2 \le V_t$. The rest can be bounded repeatedly using (4.4.57). Specifically

$$\begin{aligned}\|\nabla J(\mathbf{u}_t)\|_2 \le \frac{1}{2}(1 + \|\nabla J(\mathbf{u}_t)\|_2^2) &\le \frac{1}{2} + LJ(\mathbf{u}_t) \le \frac{1}{2} + LV_t, \\ \|\mathbf{v}_t\|_2 \le \frac{1}{2}(1 + \|\mathbf{v}_t\|_2^2) &\le \frac{1}{2}(1 + V_t), \\ \|\nabla J(\mathbf{u}_t)\|_2^2 &\le 2LJ(\mathbf{u}_t) \le 2LV_t, \\ \|\nabla J(\mathbf{u}_t)\|_2 \cdot \|\mathbf{v}_t\|_2 \le \frac{1}{2}(\|\nabla J(\mathbf{u}_t)\|_2^2 + \|\mathbf{v}_t\|_2^2) & \\ \le LJ(\mathbf{u}_t) + \frac{1}{2}\|\mathbf{v}_t\|_2^2) &\le \max\{L, 1/2\}V_t.\end{aligned}$$

Applying all these bounds to (4.4.58) shows that

$$\|\bar{\mathbf{x}}_t\|_2^2 \le B_t^2(D_{11} + D_{12}V_t) + B_t(D_{21} + D_{22}V_t) + (L\bar{k})^2\|\mathbf{v}_t\|_2^2, \tag{4.4.60}$$

for suitable constants $D_{11}$ through $D_{22}$.

For future use, we also bound $\|\mathbf{z}_t\|_2^2$. Since $\mathbf{z}_t = \bar{\mathbf{x}}_t + \nabla J(\mathbf{u}_t)$, we can write

$$\begin{aligned}\|\mathbf{z}_t\|_2^2 &\le \|\bar{\mathbf{x}}_t\|_2^2 + 2\|\bar{\mathbf{x}}_t\|_2 \cdot \|\nabla J(\mathbf{u}_t)\|_2 + \|\nabla J(\mathbf{u}_t)\|_2^2 \\ &\le \|\bar{\mathbf{x}}_t\|_2^2 + [\|\bar{\mathbf{x}}_t\|_2^2 + \|\nabla J(\mathbf{u}_t)\|_2^2] + \|\nabla J(\mathbf{u}_t)\|_2^2 \\ &= 2\|\bar{\mathbf{x}}_t\|_2^2 + 4LJ(\mathbf{u}_t).\end{aligned} \tag{4.4.61}$$

Since $J(\mathbf{u}_t) \le V_t$, we can substitute from (4.4.60) into (4.4.61) to obtain the bound

$$\begin{aligned}\|\mathbf{z}_t\|_2^2 &\le B_t^2(D_{11} + D_{12}V_t) + B_t(D_{21} + D_{22}V_t) + 4LJ(\mathbf{u}_t) + (L\bar{k})^2\|\mathbf{v}_t\|_2^2 \\ &\le B_t^2(D_{11} + D_{12}V_t) + B_t(D_{21} + D_{22}V_t) + \max\{4L, (L\bar{k})^2\}V_t.\end{aligned} \tag{4.4.62}$$

With these bounds in place, we now proceed to prove (4.4.52). Clearly

$$E_t(V_{t+1}) = E_t(J(\mathbf{u}_{t+1})) + E_t(\|\mathbf{v}_{t+1}\|_2^2). \tag{4.4.63}$$

So we bound each of these two terms individually. First, it follows from (4.4.51) that

$$\|\mathbf{v}_{t+1}\|_2^2 = \|\mu_t\mathbf{v}_t - \alpha_t\mathbf{h}_{t+1}\|_2^2 = \mu_t^2\|\mathbf{v}_t\|_2^2 - 2\alpha_t\mu_t\langle\mathbf{v}_t, \mathbf{h}_{t+1}\rangle + \alpha_t^2\|\mathbf{h}_{t+1}\|_2^2.$$

Therefore, from (4.4.26), we get

$$E_t(\|\mathbf{v}_{t+1}\|_2^2) = \mu_t^2\|\mathbf{v}_t\|_2^2 - 2\alpha_t\mu_t\langle\mathbf{v}_t, \mathbf{z}_t\rangle + \alpha_t^2[\|\mathbf{z}_t\|_2^2 + E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2)]. \tag{4.4.64}$$

We can estimate the last two terms separately.

First,

$$-2\alpha_t\mu_t\langle \mathbf{v}_t, \mathbf{z}_t\rangle \leq 2\alpha_t\mu_t\|\mathbf{v}_t\|_2 \cdot \|\mathbf{z}_t\|_2 \leq 2\alpha_t\bar{\mu}\|\mathbf{v}_t\|_2 \cdot \|\mathbf{z}_t\|_2.$$

Note that the bound is unaffected by the presence or the absence of the minus sign in front of the inner product. Further,

$$2\alpha_t\bar{\mu}\|\mathbf{v}_t\|_2 \cdot \|\mathbf{z}_t\|_2 \leq 2\alpha_t\bar{\mu}\|\mathbf{v}_t\|_2 \cdot [\|\bar{\mathbf{x}}_t\|_2 + \|\nabla J(\mathbf{u}_t)\|_2].$$

Substituting for $\|\bar{\mathbf{x}}_t\|_2$ from (4.4.58), and recalling that $\{\alpha_t B_t\}$ is a summable sequence leads to the observation that

$$2\alpha_t\bar{\mu}\|\mathbf{v}_t\|_2 \cdot \|\bar{\mathbf{x}}_t\|_2 = \texttt{WBF} + 2\alpha_t\bar{\mu}L\bar{k}\|\mathbf{v}_t\|_2^2 + 2\alpha_t\bar{\mu}\|\mathbf{v}_t\|_2 \cdot \|\nabla J(\mathbf{u}_t)\|_2, \tag{4.4.65}$$

where "WBF" denotes a well-behaved function, defined in (4.4.54). Therefore it is not necessary to write it out in detail. As a result

$$2\alpha_t\bar{\mu}\|\mathbf{v}_t\|_2 \cdot \|\mathbf{z}_t\|_2 = \texttt{WBF} + 2\alpha_t\bar{\mu}L\bar{k}\|\mathbf{v}_t\|_2^2 + 4\alpha_t\bar{\mu}\|\mathbf{v}_t\|_2 \cdot \|\nabla J(\mathbf{u}_t)\|_2, \tag{4.4.66}$$

Next we bound the last term.

$$\alpha_t^2[\|\mathbf{z}_t\|_2^2 + E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2)] = \alpha_t^2\|\mathbf{z}_t\|_2^2 + \alpha_t^2 E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2)]. \tag{4.4.67}$$

We already have a bound for $\|\mathbf{z}_t\|_2^2$, namely (4.4.62). As discussed earlier, the hypothesis (4.4.33) implies (4.4.55). Therefore the term $\alpha_t^2\|\mathbf{z}_t\|_2^2$ is a WBF. So let us focus on $E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2)]$. There is a bound on this quantity in (4.4.28), but it is stated in terms $J(\mathbf{w}_t)$. The bound is now restated in terms of $J(\mathbf{u}_t)$, using Theorem 3.1. We know from (4.4.45) that $\mathbf{u}_t = \mathbf{w}_t + k_t\mathbf{v}_t$. So applying Theorem 3.1 gives

$$J(\mathbf{w}_t) = J(\mathbf{u}_t - k_t\mathbf{v}_t) \leq J(\mathbf{u}_t) - k_t\langle \nabla J(\mathbf{u}_t), \mathbf{v}_t\rangle + \frac{Lk_t^2}{2}\|\mathbf{v}_t\|_2^2. \tag{4.4.68}$$

Now Schwarz' inequality and (4.4.57) lead to

$$\begin{aligned} -k_t\langle \nabla J(\mathbf{u}_t), \mathbf{v}_t\rangle &\leq \frac{k_t}{2}[\|\nabla J(\mathbf{u}_t)\|_2^2 + \|\mathbf{v}_t\|_2^2] \\ &\leq \frac{k_t}{2}[2LJ(\mathbf{u}_t) + \|\mathbf{v}_t\|_2^2] \leq \frac{\bar{k}}{2}[2LJ(\mathbf{u}_t) + \|\mathbf{v}_t\|_2^2], \end{aligned} \tag{4.4.69}$$

This can be substituted into (4.4.68) to give

$$J(\mathbf{w}_t) \leq J(\mathbf{u}_t) + L\bar{k}J(\mathbf{u}_t) + \left[\frac{\bar{k}}{2} + \frac{L\bar{k}^2}{2}\right]\|\mathbf{v}_t\|_2^2 \leq D_3V_t, \tag{4.4.70}$$

where

$$D_3 = \max\left\{L\bar{k}, \frac{\bar{k}}{2} + \frac{L\bar{k}^2}{2}\right\}.$$

Therefore the bound in (4.4.28) can be reformulated as

$$\alpha_t^2 E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2)] \leq \alpha_t^2 D_3 V_t, \tag{4.4.71}$$

which is a WBF in view of the assumptions (4.4.33). Substituting all these bounds into (4.4.64) gives

$$E_t(\|\mathbf{v}_{t+1}\|_2^2) \leq \texttt{WBF} + \mu_t^2\|\mathbf{v}_t\|_2^2 + 2\alpha_t\bar{\mu}L\bar{k}\|\mathbf{v}_t\|_2^2 + 2\bar{\mu}\alpha_t\|\mathbf{v}_t\|_2 \cdot \|\nabla J(\mathbf{u}_t)\|_2. \tag{4.4.72}$$

Next we turn our attention to $E_t(J(\mathbf{u}_{t+1}))$. Recall from (4.4.50) that

$$\mathbf{u}_{t+1} = \mathbf{u}_t + \delta_t\mu_t\mathbf{v}_t - (b_t + k_{t+1})\alpha_t\mathbf{h}_{t+1}.$$

Therefore, by applying Lemma 3.4, we get

$$\begin{aligned} J(\mathbf{u}_{t+1}) &= J(\mathbf{u}_t + \delta_t\mu_t\mathbf{v}_t - (b_t + k_{t+1})\alpha_t\mathbf{h}_{t+1}) \\ &\leq J(\mathbf{u}_t) + \delta_t\mu_t\langle\nabla J(\mathbf{u}_t), \mathbf{v}_t\rangle - \alpha_t(b_t + k_{t+1})\langle\nabla J(\mathbf{u}_t), \mathbf{h}_{t+1}\rangle \\ &+ \frac{L}{2}\|\delta_t\mu_t\mathbf{v}_t - \alpha_t(b_t + k_{t+1})\mathbf{h}_{t+1}\|_2^2. \end{aligned} \tag{4.4.73}$$

From (4.4.22) we have that

$$\|\delta_t\mu_t\mathbf{v}_t - \alpha_t(b_t + k_{t+1})\mathbf{h}_{t+1}\|_2^2 = \|\delta_t\mu_t\mathbf{v}_t - \alpha_t(b_t + k_{t+1})\mathbf{z}_t - \alpha_t(b_t + k_{t+1})\boldsymbol{\zeta}_{t+1}\|_2^2.$$

However, since $E_t(\boldsymbol{\zeta}_{t+1}) = \mathbf{0}$, it follows that

$$\begin{aligned} E_t(\|\delta_t\mu_t\mathbf{v}_t - \alpha_t(b_t + k_{t+1})\mathbf{h}_{t+1}\|_2^2) &= \|\delta_t\mu_t\mathbf{v}_t - \alpha_t(b_t + k_{t+1})\mathbf{z}_t\|_2^2 \\ &+ \alpha_t^2(b_t + k_{t+1})^2 E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2). \end{aligned}$$

Applying $E_t(\cdot)$ to both sides of (4.4.73), and substituting the above relationship, gives

$$\begin{aligned} E_t(J(\mathbf{u}_{t+1})) &\leq J(\mathbf{u}_t) + \delta_t\mu_t\langle\nabla J(\mathbf{u}_t), \mathbf{v}_t\rangle - \alpha_t(b_t + k_{t+1})\langle\nabla J(\mathbf{u}_t), \mathbf{z}_t\rangle \\ &+ \frac{L}{2}\|\delta_t\mu_t\mathbf{v}_t - \alpha_t(b_t + k_{t+1})\mathbf{z}_t\|_2^2 + \frac{L}{2}\alpha_t^2(b_t + k_{t+1})^2 E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2). \end{aligned} \tag{4.4.74}$$

Now we analyze each of the terms in (4.4.74) individually. Before doing so, we replace several functions of $t$ by their bounds. Specifically

- $\delta_t$ could be positive or negative, but is assumed to converge to 0. Therefore $|\delta_t|$ is bounded, say by $\bar{\delta}$.
- $\mu_t \in [0, \bar{\mu}]$ where $\bar{\mu} < 1$.
- $b_t \in [\underline{b}, \bar{b}]$ where $0 < \underline{b} \leq \bar{b}$, and $k_t \in [0, \bar{k}]$. Therefore $b_t + k_{t+1} \in [\underline{b}, \bar{b} + \bar{k}]$.

With these observations, we have the following bounds:

$$\delta_t\mu_t\langle\nabla J(\mathbf{u}_t), \mathbf{v}_t\rangle \leq \bar{\delta}\bar{\mu}\|\nabla J(\mathbf{u}_t)\|_2 \cdot \|\mathbf{v}_t\|_2. \tag{4.4.75}$$

Next

$$\begin{aligned} -\alpha_t(b_t + k_{t+1})\langle\nabla J(\mathbf{u}_t), \mathbf{z}_t\rangle &= -\alpha_t(b_t + k_{t+1})\|\nabla J(\mathbf{u}_t)\|_2^2 \\ &- \alpha_t(b_t + k_{t+1})\langle\nabla J(\mathbf{u}_t), \bar{\mathbf{x}}_t\rangle \\ &\leq -\alpha_t\underline{b}\|\nabla J(\mathbf{u}_t)\|_2^2 \\ &+ \alpha_t(\bar{b} + \bar{k})\|\nabla J(\mathbf{u}_t)\|_2 \cdot \|\bar{\mathbf{x}}_t\|_2. \end{aligned} \tag{4.4.76}$$

To bound the last term on the right side of (4.4.76), we use the bound on $\|\bar{\mathbf{x}}_t\|_2$ from (4.4.58), and the summability of $\{\alpha_t B_t\}$. This gives

$$\alpha_t(\bar{b} + \bar{k})\|\nabla J(\mathbf{u}_t)\|_2 \cdot \|\bar{\mathbf{x}}_t\|_2 \leq \alpha_t L\bar{k}(\bar{b} + \bar{k})\|\nabla J(\mathbf{u}_t)\|_2 \cdot \|\mathbf{v}_t\|_2 + \texttt{WBF}. \tag{4.4.77}$$

Next we tackle the first quadratic term on the right side of (4.4.74).

$$\begin{aligned} \frac{L}{2}\|\delta_t\mu_t\mathbf{v}_t - \alpha_t(b_t + k_{t+1})\mathbf{z}_t\|_2^2 &= \frac{L}{2}\|\delta_t\mu_t\mathbf{v}_t\|_2^2 \\ &+ \frac{\alpha_t^2 L}{2}(b_t + k_{t+1})^2\|\mathbf{z}_t\|_2^2 \\ &- \alpha_t L(b_t + k_{t+1})\delta_t\mu_t\langle\mathbf{v}_t, \mathbf{z}_t\rangle. \end{aligned} \tag{4.4.78}$$

Each of the three terms can be analyzed individually.

$$\frac{L}{2}\|\delta_t\mu_t\mathbf{v}_t\|_2^2 \leq \frac{L\bar{\mu}^2\delta_t^2}{2}\|\mathbf{v}_t\|_2^2. \tag{4.4.79}$$

Next, from (4.4.62), it follows that

$$\frac{\alpha_t^2 L}{2}(b_t + k_{t+1})^2\|\mathbf{z}_t\|_2^2 = \texttt{WBF}. \tag{4.4.80}$$

Finally, we already have a bound for the cross-product term $\alpha_t\mu_t\langle\mathbf{v}_t, \mathbf{z}_t\rangle$ from (4.4.66). After multiplying this bound by $L\bar{\delta}(\bar{b}+\bar{k})$, we get

$$\begin{aligned}\alpha_t L(b_t + k_{t+1})\delta_t\mu_t\langle\mathbf{v}_t, \mathbf{z}_t\rangle \leq \texttt{WBF} &+ \alpha_t L^2\bar{k}\bar{\delta}(\bar{b}+\bar{k})\|\mathbf{v}_t\|_2^2 \\ &+ 2\alpha_t L\bar{\delta}(\bar{b}+\bar{k})\|\mathbf{v}_t\|_2 \cdot \|\nabla J(\mathbf{u}_t)\|_2.\end{aligned} \tag{4.4.81}$$

Now we can add up all these bounds. This gives

$$\begin{aligned}E(V_{t+1}) &= E_t(J(\mathbf{u}_{t+1})) + E_t(\|\mathbf{v}_{t+1}\|_2^2) \\ &\leq J(\mathbf{u}_t) + \|\mathbf{v}_t\|_2^2 - (1-\bar{\mu}^2)\|\mathbf{v}_t\|_2^2 - \alpha_t\underline{b}\|\nabla J(\mathbf{u}_t)\|_2^2 \\ &+ C_1\alpha_t\|\mathbf{v}_t\|_2^2 + C_2\alpha_t\|\mathbf{v}_t\|_2 \cdot \|\nabla J(\mathbf{u}_t)\|_2 + \texttt{WBF},\end{aligned} \tag{4.4.82}$$

where $C_1, C_2$ are some positive constants whose precise value is not important. Next, we can "borrow" half of each of the two negative terms in the above, and rewrite the bound as

$$E(V_{t+1}) \leq V_t - \frac{1-\bar{\mu}^2}{2}\|\mathbf{v}_t\|_2^2 - \alpha_t\frac{\underline{b}}{2}\|\nabla J(\mathbf{u}_t)\|_2^2 - F_t + \texttt{WBF}, \tag{4.4.83}$$

where $F_t$ is the quadratic form

$$F_t = \begin{bmatrix} \|\mathbf{v}_t\|_2 & \|\nabla J(\mathbf{u}_t)\|_2 \end{bmatrix} \begin{bmatrix} \frac{1-\bar{\mu}^2}{2} - \alpha_t C_1 & -\alpha_t(C_2/2) \\ -\alpha_t(C_2/2) & \alpha_t(\underline{b}/2) \end{bmatrix} \begin{bmatrix} \|\mathbf{v}_t\|_2 \\ \|\nabla J(\mathbf{u}_t)\|_2 \end{bmatrix}.$$

Let us define

$$K_t = \begin{bmatrix} \frac{1-\bar{\mu}^2}{2} - \alpha_t C_1 & -\alpha_t(C_2/2) \\ -\alpha_t(C_2/2) & \alpha_t(\underline{b}/2) \end{bmatrix}. \tag{4.4.84}$$

It is now shown that $K_t$ is a positive definite matrix, and $F_t$ is a positive definite form, for $t$ sufficiently large; specifically, there exists a $T < \infty$ such that $F_t \geq 0$ for all $t \geq$T. Suppose we succeed in proving this. Since we can always start our analysis of (4.4.71) starting at time $T$, we can write

$$E(V_{t+1}) \leq V_{t+1} - \left(\frac{1-\bar{\mu}^2}{2}\right)\|\mathbf{v}_t\|_2^2 - \alpha_t\|\nabla J(\mathbf{w}_t)\|_2^2 + \texttt{WBF}, \ \forall t \geq T. \tag{4.4.85}$$

In other words, the term $-F_t$ is gone. Now (4.4.85) is in a form to which the Robbins-Siegmund theorem (Lemma 3.2) can be applied. So let us now establish the positive definiteness of the quadratic form for sufficiently large $t$. Note that a symmetric $2 \times 2$ matrix is positive definite if its trace and its determinant are both positive. In this case

$$\mathrm{tr}(K_t) = \frac{1-\bar{\mu}^2}{2} - (C_1 - (\underline{b}/2))\alpha_t, \quad \det(K_t) = \frac{1-\bar{\mu}^2}{2}\frac{\underline{b}}{2}\alpha_t - C_3\alpha_t^2,$$

where $C_3$ is another constant. Since, by hypothesis, $\sum_{t=0}^{\infty}\alpha_t^2 < \infty$, it follows that $\alpha_t \to 0$ as $t \to \infty$. Hence the trace of $K_t$ is positive for sufficiently large $t$. Similarly, in the expression for the determinant of $K_t$, the positive term is linear in $\alpha_t$, whereas the negative term is quadratic in $\alpha_t$. Hence the determinant of $K_t$ is also positive for sufficiently large $t$. Hence we conclude that $K_t$ is a positive definite matrix for sufficiently large $t$.

With this observation, we can apply Theorem 2.22 to (4.4.85).

We begin wih Item 1. Note that all statements hold "almost surely," so this qualifier is not repeated each time. Suppose (4.4.85) holds. Then the following conclusions follow from Theorem 2.23:

- $J(\mathbf{u}_t) + \|\mathbf{v}_t\|_2^2$ is bounded. Moreover, there is a random variable $W$ such that $J(\mathbf{u}_t) + \|\mathbf{v}_t\|_2^2 \to W$ (almost surely) as $t \to \infty$.
- Further, almost surely

$$\sum_{t=0}^{\infty} \left( \frac{1 - \bar{\mu}^2}{2} \right) \|\mathbf{v}_t\|_2^2 + \alpha_t \|\nabla J(\mathbf{u}_t)\|_2^2 < \infty. \tag{4.4.86}$$

Since the summands in (4.4.86) are both nonnegative, and $(1 - \bar{\mu}^2)/2$ is just a constant, it follows that

$$\sum_{t=0}^{\infty} \|\mathbf{v}_t\|_2^2 < \infty, \tag{4.4.87}$$

$$\sum_{t=0}^{\infty} \alpha_t \|\nabla J(\mathbf{u}_t)\|_2^2 < \infty. \tag{4.4.88}$$

Now (4.4.87) implies that $\|\mathbf{v}_t\|_2^2 \to 0$ as $t \to \infty$, i.e., that $\mathbf{v}_t \to \mathbf{0}$ as $t \to \infty$. In turn, if $J(\mathbf{u}_t) + \|\mathbf{v}_t\|_2^2 \to X$, then $J(\mathbf{w}_t) \to X$ as $t \to \infty$.

Now recall from (4.4.1) that $\boldsymbol{\theta}_t = \mathbf{w}_t - \epsilon_t \mathbf{v}_t$. Since $J(\cdot)$ is continuous and $\mathbf{v}_t \to \mathbf{0}$, it follows that $J(\boldsymbol{\theta}_t) \to W$ as $t \to \infty$. The boundedness of $\{J(\boldsymbol{\theta}_t)\}$ follows from it being a convergent sequence. Finally, the boundedness of $\{\nabla J(\boldsymbol{\theta}_t)\}$ follows from Lemma 4.1. Thus we have established Item 1.

Next we address Item 2. Suppose (4.4.34) holds. Then it readily follows from (4.4.88) that

$$\liminf_{t \to \infty} \|\nabla J(\mathbf{u}_t)\|_2^2 = 0.$$

To translate this conclusion into the behavior of $\nabla J(\boldsymbol{\theta}_t)$, we proceed as follows: It follows from the definitions of $\mathbf{w}_t$ and $\mathbf{u}_t$ that

$$\boldsymbol{\theta}_t = \mathbf{u}_t - (k_t + \epsilon_t)\mathbf{v}_t.$$

Since $\epsilon_t$ and $k_t$ are bounded, $\mathbf{v}_t \to \mathbf{0}$ as $t \to \infty$, and $\nabla J(\cdot)$ is Lipschitz-continuous, it can be concluded that

$$\liminf_{t \to \infty} \|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = 0.$$

This is Item 2.

Next we address Item 3 of the theorem. The hypotheses are that, in addition to (4.4.33), (4.4.34) also holds, and $J(\cdot)$ satisfies Property (KL'). Then by definition there exists a function $\psi : \mathbb{R} \to \mathbb{R}$ in Class $\mathcal{B}$ such that $\|\nabla J(\boldsymbol{\theta}_t)\|_2 \geq \psi(J(\boldsymbol{\theta}_t))$. Recall that all the stochastic processes are defined on some underlying probability space $(\Omega, \Sigma, P)$. Define

$$\Omega_0 := \{\omega \in \Omega : J(\boldsymbol{\theta}(\omega)) \to W(\omega) \quad \& \quad \|\mathbf{v}_t(\omega)\|_2^2 \to 0\},$$

$$\Omega_1 := \{\omega \in \Omega : \sum_{t=0}^{\infty} \alpha_t(\omega) = \infty\}.$$

Note that if the step sizes are deterministic, then $\Omega_1 = \Omega$. Define $\Omega_2 = \Omega_0 \cap \Omega_1$, and note that $P(\Omega_2) = 1$, by Item 1.

The objective is to show that $W(\omega) = 0$ for all $\omega \in \Omega_2$. Once this is done, it would follow from Lemma 3.2 that

$$\|\nabla J(\boldsymbol{\theta}_t(\omega))\|_2 \leq [2LJ(\boldsymbol{\theta}_t(\omega))]^{1/2} \to 0 \text{ as } t \to \infty, \ \forall \omega \in \Omega_2.$$

Accordingly, suppose that, for some $\omega \in \Omega_0$, we have that $W(\omega) > 0$, say $W(\omega) = 2p$, where $p > 0$. Define

$$G(\omega) := \sup_t J(\boldsymbol{\theta}_t(\omega)).$$

Then $G(\omega) < \infty$ because $\{J(\boldsymbol{\theta}_t(\omega))\}$ is a convergent sequence. Define

$$q := \frac{1}{2} \inf_{p \leq r \leq G(\omega)} \psi(r).$$

Then $q > 0$ because $\psi(\cdot)$ is a function of Class $\mathcal{B}$. Now choose a $T_0 < \infty$ such that $J(\boldsymbol{\theta}(\omega)) \geq p$ for all $t \geq T_0$. By the (KL') property, it follows that

$$\|\nabla J(\boldsymbol{\theta}(\omega))\|_2 \geq 2q, \ \forall t \geq T_0.$$

Next, choose $T_1 < \infty$ such that $\|\mathbf{v}_t(\omega)\|_2 \leq q/L$ for all $t \geq T_1$, and define $T_2 = \min\{T_0, T_1\}$. Then it follows from the Lipschitz continuity of $\nabla J(\cdot)$ that

$$\|\nabla J(\mathbf{w}_t(\omega))\|_2 \geq \|\nabla J(\boldsymbol{\theta}_t(\omega))\|_2 - L\|\mathbf{v}_t(\omega)\|_2 \geq q, \ \forall t \geq T_2. \tag{4.4.89}$$

On the other hand, because $\omega \in \Omega_2$, we have that

$$\sum_{t=T_2} \alpha_t(\omega) = \infty. \tag{4.4.90}$$

Thus (4.4.89) and (4.4.90) together imply that

$$\sum_{t=T_2}^{\infty} \alpha_t \|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = \infty.$$

Since this contradicts (4.4.88), we conclude that no such $\omega \in \Omega_2$ can exist. In other words $W(\omega) = 0$ for all $\omega \in \Omega_2$. This establishes Item 3.

Item 4 is a ready consequence of Item 3 and Property (NSC). If $\{J(\boldsymbol{\theta}_t)\}$ is bounded, then the fact that $J(\cdot)$ has compact level sets means that $\{\boldsymbol{\theta}_t\}$ is bounded. Then the fact that $J(\boldsymbol{\theta}_t) \to 0$ as $t \to \infty$ implies that $\rho(\boldsymbol{\theta}_t) \to 0$ as $t \to \infty$; in other words, the distance from the iterate $\boldsymbol{\theta}_t$ to the set $S_J$ of global minima approaches zero. Note that it is *not* assumed that $S_J$ consists of a singleton.

This completes the proof of Theorem 4.3.

**Proof of Theorem 4.4**

The proof, based on Theorem 4.3, is basically the same as that of Theorem 4.4, except that we invoke Theorem 2.24 instead of Theorem 2.23.. The only difference is that the bound (4.4.84) holds only after some time $T$. Clearly this does not affect the *asymptotic* rate of convergence. Nevertheless, in the interests of completeness, the proof is *sketched* here.

The hypotheses on the various constants imply that

$$\alpha_t^2 = O(t^{-2+2\phi}), \alpha_t^2 M_t^2 = O(t^{-2+2(\phi+\delta)}), \alpha_t B_t = O(t^{-1+\phi-\gamma}),$$

while $\alpha_t^2 B_t$ and $\alpha_t^2 B_t^2$ decay faster than $\alpha_t B_t$. Hence both $\{f_t\}$ and $\{g_t\}$ are summable if

$$-2 + 2\phi < -1, -2 + 2(\phi + \delta) < -1, -1 + \phi - \gamma < -1.$$

The three inequalities are satisfied if $\phi$ satisfies (4.4.38). Next, let us define $\nu$ as in (4.4.39), and apply Theorem 2.24. This leads to the conclusion that $J(\mathbf{u}_t) + \|\mathbf{v}_t\|_2^2 = o(t^{-\lambda})$ for every $\lambda \in (0, \nu)$. In turn this means that, individually, both $J(\mathbf{u}_t)$ and $\|\mathbf{v}_t\|_2^2$ are $o(t^{-\lambda})$ for every $\lambda \in (0, \nu)$. Since $\boldsymbol{\theta}_t = \mathbf{u}_t - (k_t + \epsilon_t)\mathbf{v}_t$, and both $\epsilon_t$ and $k_t$ are bounded, this leads to $J(\boldsymbol{\theta}_t) = o(t^{-\lambda})$ for every $\lambda \in (0, \nu)$. Finally, the (PL) property leads to $\|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = o(t^{-\lambda})$ for every $\lambda \in (0, \nu)$. If we choose the step size sequence to decay very slowly, then the bound in (4.4.40) follows readily. This completes the proof of Theorem 4.4.

### 4.4.5 Nonviability of an Earlier Iterative Scheme

In this section, we analyze the behavior of the solutions of (4.4.16) introduced in [130], reproduced here for the convenience of the reader.

$$\lambda_{t+1} = \frac{\lambda_t}{\mu_t} - 1, \eta_t = (1 + \lambda_{t+1})\alpha_t$$

Recall from (4.4.20) that the convergence conditions in [130] involve the "synthetic" step size sequence $\{\eta_t\}$. The objective of this section is to show that this approach is not feasible. Specifically, if $\{\mu_t\}$ is a decreasing sequence, then $\lambda_t \to \infty$ as $t \to \infty$. Thus, even if the original step size sequence $\{\alpha_t\}$ is square-summable, the synthetic sequence of step sizes $\{\eta_t\}$ need not be. Thus the assumptions in (4.4.20) are strictly stronger than the standard Robbins-Monro conditions, which are the sufficient conditions used in Theorem 4.3. In the other direction, if $\{\mu_t\}$ is an increasing sequence bounded away from 1, eventually $1 + \lambda_{t+1} < 0$, thus leading to a negative step size $\eta_t$, which is absurd. Thus the point is that, while the approach in [130] is quite elegant, it is not practical.

We begin by presenting a "closed-form" formula for $\lambda_{t+1}$ as a function of the $\mu_t$ sequence. Write the first equation in (4.4.16) as

$$\begin{aligned}\lambda_{t+1} &= \frac{\lambda_t}{\mu_t} - 1 = \frac{\lambda_t - \lambda_{t-1}}{\mu_t} + \frac{\lambda_{t-1}}{\mu_t} - 1 \\ &= \frac{1}{\mu_t}(\lambda_t - \lambda_{t-1}) + \left(\frac{1}{\mu_t} - \frac{1}{\mu_{t-1}}\right)\lambda_{t-1} + \frac{\lambda_{t-1}}{\mu_{t-1}} - 1 \\ &= \lambda_t + \frac{1}{\mu_t}(\lambda_t - \lambda_{t-1}) + \left(\frac{1}{\mu_t} - \frac{1}{\mu_{t-1}}\right)\lambda_{t-1}.\end{aligned} \tag{4.4.91}$$

Therefore

$$\lambda_{t+1} - \lambda_t = \frac{1}{\mu_t}(\lambda_t - \lambda_{t-1}) + \left(\frac{1}{\mu_t} - \frac{1}{\mu_{t-1}}\right)\lambda_{t-1}. \tag{4.4.92}$$

It is easy to show by induction that a "closed-form" solution to (4.4.91) is

$$\lambda_{t+1} = \lambda_t + \left[\prod_{\tau=1}^{t} \frac{1}{\mu_\tau}\right](\lambda_1 - \lambda_0) + \sum_{\tau=1}^{t}\left[\prod_{s=\tau}^{t-1}\frac{1}{\mu_s}\right]\left(\frac{1}{\mu_\tau} - \frac{1}{\mu_{\tau-1}}\right)\lambda_{\tau-1}, \tag{4.4.93}$$

where empty products are taken as 1 and empty sums are taken as 0. Note that $\lambda_0$ is unspecified. So if we take $\lambda_0 = \mu_0/(1-\mu_0)$, then

$$\lambda_1 = \frac{\lambda_0}{\mu_0} - 1 = \frac{1}{1-\mu_0} - 1 = \frac{\mu_0}{1-\mu_0} = \lambda_0.$$

With this choice, the first term in (4.4.93) drops out; but this is not much of a simplification. Note also that if $\mu_t = \mu$ for all $t$, then $\lambda_t = \lambda_0$ for all $t$.

Now let us analyze the behavior of $\lambda_t$ in two specific situations: (i) $\{\mu_t\}$ is a strictly decreasing, i.e., $\mu_t < \mu_{t-1}$ for all $t$, and (ii) $\{\mu_t\}$ is strictly increasing, but bounded above by some $\bar{\mu} < 1$. In principle, the closed form solution (4.4.93) can be used to analyze arbitrary sequences $\{\mu_t\}$. However, the two situations studied here are perhaps the most natural.

**Lemma 4.2.** *Suppose $\lambda_0$ is chosen as $\mu_0/(1-\mu_0)$, so that $\lambda_1 = \lambda_0$. Suppose further that $\mu_t < \mu_{t-1}$ for all $t \geq 1$. Then $\lambda_t \to \infty$ as $t \to \infty$.*

*Proof.* The first step is to show that $\lambda_{t+1} > \lambda_t$ for all $t \geq 1$. The proof is by induction. First, for $t = 1$, we have that

$$\lambda_2 = \frac{\lambda_1}{\mu_1} - 1 > \frac{\lambda_1}{\mu_0} - 1 = \frac{1}{1-\mu_0} - 1 = \lambda_1.$$

Next suppose $\lambda_t > \lambda_{t-1}$. Then

$$\lambda_{t+1} = \frac{\lambda_t}{\mu_t} - 1 > \frac{\lambda_{t-1}}{\mu_{t-1}} - 1 = \lambda_t.$$

This completes the proof by induction.

Next, we invoke the recursion (4.4.91).

$$\lambda_{t+1} - \lambda_t = \frac{1}{\mu_t}(\lambda_t - \lambda_{t-1}) + \left(\frac{1}{\mu_t} - \frac{1}{\mu_{t-1}}\right)\lambda_{t-1}.$$

The fact that $\mu_t < \mu_{t-1}$ implies that

$$\left(\frac{1}{\mu_t} - \frac{1}{\mu_{t-1}}\right)\lambda_{t-1} > 0, \ \forall t \geq 2.$$

Hence

$$\lambda_{t+1} - \lambda_t > \frac{1}{\mu_t}(\lambda_t - \lambda_{t-1}) > \frac{1}{\mu_2}(\lambda_t - \lambda_{t-1}), \ \forall t \geq 2.$$

As a consequence we get

$$\lambda_{t+1} - \lambda_t > \left[\prod_{s=2}^{t} \frac{1}{\mu_s}\right](\lambda_2 - \lambda_1) > \left(\frac{1}{\mu_2}\right)^{t-1}(\lambda_2 - \lambda_1), \ \forall t \geq 2.$$

We can add the above bound for all $t$. Because it is a telescoping sum, we get

$$\lambda_{t+1} = \lambda_2 + \sum_{k=2}^{t}(\lambda_{k+1} - \lambda_k) \geq (\lambda_2 - \lambda_1)\sum_{k=2}^{t}\left(\frac{1}{\mu_2}\right)^{k-1} \to \infty \text{ as } t \to \infty.$$

□

**Lemma 4.3.** *Suppose $\mu_{t-1} < \mu_t < 1$ for all $t$, and that*

$$\prod_{\tau=2}^{t}\left(\frac{1}{\mu_\tau}\right) \to \infty \text{ as } t \to \infty. \tag{4.4.94}$$

*Then there exists a finite $t_0$ such that*

$$1 + \lambda_t < 0, \ \forall t \geq t_0. \tag{4.4.95}$$

*In particular, if $\mu_t \leq \bar{\mu} < 1$ for all $t$, then we can take*

$$t_0 = 3 + \log_{(1/\bar{\mu})}\frac{\lambda_1}{\lambda_1 - \lambda_2}. \tag{4.4.96}$$

*Proof.* Observe that

$$\lambda_2 = \frac{\lambda_1}{\mu_1} - 1 < \frac{\lambda_1}{\mu_0} - 1 = \lambda_1.$$

Now suppose that $\lambda_t < \lambda_{t-1}$. Then

$$\lambda_{t+1} = \frac{\lambda_t}{\mu_t} - 1 < \frac{\lambda_{t-1}}{\mu_{t-1}} - 1 = \lambda_t.$$

After observing that

$$\frac{1}{\mu_t} - \frac{1}{\mu_{t-1}} < 0,$$

we can rewrite(4.4.91) as

$$\lambda_t - \lambda_{t+1} = \frac{1}{\mu_t}(\lambda_{t-1} - \lambda_t) + \left(\frac{1}{\mu_{t-1}} - \frac{1}{\mu_t}\right)\lambda_{t-1}. \tag{4.4.97}$$

Now suppose $\lambda_\tau > 0$ for $\tau = 1, \cdots, t$. Then (4.4.97) implies that

$$\lambda_\tau - \lambda_{\tau+1} > \frac{1}{\mu_\tau}(\lambda_{\tau-1} - \lambda_\tau) > \left(\prod_{k=2}^{\tau}\frac{1}{\mu_k}\right)(\lambda_1 - \lambda_2). \tag{4.4.98}$$

$$\lambda_1 - \lambda_{t+1} = \sum_{\tau=1}^{t}(\lambda_\tau - \lambda_{\tau-1}) > \left[\sum_{\tau=1}^{t}\left(\prod_{k=2}^{t}\frac{1}{\mu_k}\right)\right](\lambda_1 - \lambda_2). \tag{4.4.99}$$

Consequently

$$\begin{aligned}\lambda_{t+1} &< \lambda_1 - \left[\sum_{\tau=1}^{t}\left(\prod_{k=2}^{\tau}\frac{1}{\mu_k}\right)\right](\lambda_1 - \lambda_2) \\ &< \lambda_1 - \left(\prod_{k=2}^{t}\frac{1}{\mu_k}\right)(\lambda_1 - \lambda_2).\end{aligned} \tag{4.4.100}$$

Now choose $T$ such that

$$\left(\prod_{k=2}^{T}\frac{1}{\mu_k}\right) > \frac{\lambda_1}{\lambda_1 - \lambda_2}. \tag{4.4.101}$$

This is possible in view of (4.4.94). Then there are two possibilities: (i) $\lambda_\tau > 0$ for $\tau = 2, \cdots, T$. Then $\lambda_{T+1} < 0$ by virtue of (4.4.101). In this case we have that

$$\lambda_{T+2} = \frac{\lambda_{T+1}}{\mu_{T+1}} - 1 < -1.$$

Therefore $\lambda_{T+2} + 1 < 0$. The argument can be repeated, to show that $\lambda_t + 1 < 0$ for all $t \geq T + 2$. Hence we can take $t_0 = T + 2$ in (4.4.101). (ii) There exists a $\tau$ between 2 and $T$ such that $\lambda_\tau \leq 0$. By the above reasoning, it follows that

$$\lambda_{\tau+1} = \frac{\lambda_\tau}{\mu_\tau} - 1 \leq -1 < 0.$$

Therefore

$$\lambda_{\tau+2} = \frac{\lambda_{\tau+1}}{\mu_{\tau+1}} - 1 < -1, \text{ or } \lambda_{\tau+2} + 1 < 0.$$

As above, this leads to the conclusion that $\lambda_t + 1 < 0$ for all $t \geq \tau + 2$. Since $\tau \leq T$, we can conclude as before that $\lambda_t + 1 < 0$ for all $t \geq T + 2$. Hence we can again take $t_0 = T + 2$ in (4.4.101).

To prove the last claim, suppose that $\mu_t \leq \bar{\mu} < 1$ for all $t$. Then we can replace (4.4.101) by

$$\left(\frac{1}{\bar{\mu}}\right)^{T-1} \geq \frac{\lambda_1}{\lambda_1 - \lambda_2}.$$

Solving for $T$ and choosing $t_0 = T + 2$ gives (4.4.96). □

## 4.5 Stochastic Algorithms with Block Updating

Until now, we have studied what might be called "full coordinate updating." Thus, in (4.3.1), it is assumed that, at each step $t$, *every* component of $\boldsymbol{\theta}_t$ is updated. Similarly, in (4.4.2)–(4.4.3), it is assumed that *every* component of both $\mathbf{w}_t$ and $\mathbf{v}_t$ are updated at each $t$. In Chapter 3, we studied various types of "block stochastic approximation" in which, at each step $t$, *some but not necessarily all* components of $\boldsymbol{\theta}_t$ are updated. This is referred to as **block updating**, though the terminology is not standard. We study three different types of block updating, which cater to the most commonly used methods. It is shown that, for the SGD algorithm, if the assumptions in Theorems 4.1 or 4.2 are satisfied, then the conclusions of these theorems continue to hold under each of these types of block updating. For the unified momentum algorithm, the situation is not so satisfactory. We are able to prove that, if block updating is used in the stochastic gradient term $\mathbf{h}_{t+1}$ in (4.4.2)–(4.4.3), but *not* in the delay or momentum term, then the conclusions of Theorems 4.3 and 4.4 continue to hold. At the moment, there are no results on what happens when block updating is applied also to the momentum terms.

### 4.5.1 Various Block Updating Schemes

Let $\mathbf{h}_{t+1}$ denote the stochastic gradient in (4.3.1). The updating method described in (4.3.1) is then the "full coordinate" update option. We refer to it as "Option 1." Now we describe three different options for block updating, which we call single coordinate, multiple coordinate, and Bernoulli updates. These are called Options 2, 3 and 4, and are denoted by $\mathbf{h}_{t+1}^{(k)}$ for $k = 2, 3, 4$. These updating schemes include most if not all of the widely used block updating methods.

**Option 1: Full Coordinate Update:** Let

$$\mathbf{h}_{t+1}^{(1)} = \mathbf{h}_{t+1}. \tag{4.5.1}$$

**Option 2: Single Coordinate Update:** This option is also referred to as "coordinate gradient descent" in [175] and studied further in [165]. At time $t$, choose an index $\kappa_t \in [d]$ at random with a uniform probability, and independently of previous choices. Let $\mathbf{e}_{\kappa_t}$ denote the elementary unit vector with a 1 as the $\kappa_t$-th component and zeros elsewhere. Then define

$$\mathbf{h}_{t+1}^{(2)} = d\mathbf{e}_{\kappa_t} \circ \mathbf{h}_{t+1}, \tag{4.5.2}$$

where $\circ$ denotes the Hadamard, or component-wise, product of two vectors of equal dimension. Thus

$$[\mathbf{h}_{t+1}^{(2)}]_j = \begin{cases} h_{t+1,i} & \text{if } j = i, \\ 0 & \text{if } j \neq i. \end{cases}$$

The factor $d$ arises because the likelihood that $\kappa_t$ equaling any one index $i \in [d]$ is $1/d$. In this option, if $\kappa_t = i \in [d]$ at step $t$, then only the $i$-th coordinate of $\boldsymbol{\theta}_t$ gets updated at time $t$. In other words,

$$\boldsymbol{\theta}_{t+1,j} = \boldsymbol{\theta}_{j,t}, \ \forall j \neq i.$$

**Option 3: Multiple Coordinate Update:** This option is just coordinate update along multiple coordinates chosen independently at random. At time $t$, choose $N$ different indices $\kappa_t^n$ from $[d]$ *with replacement*, with each choice being independent of the rest, and also of past choices. Moreover, each $\kappa_t^n$ is chosen from $[d]$ with uniform probability. Then define

$$\mathbf{h}_{t+1}^{(3)} := \frac{d}{N} \sum_{n=1}^{N} \mathbf{e}_{\kappa_t^n} \circ \mathbf{h}_{t+1}. \tag{4.5.3}$$

Because sampling is with replacement, the average number of times an index $i \in [d]$ gets selected for updating is is $N/d$; to normalize this, the multiplicative factor in (4.5.3) is the reciprocal of the average. In this option,

$\mathbf{h}_{t+1}^{(3)}$ can have *up to* $N$ nonzero components. Because the sampling is *with replacement*, there might be some duplicated samples. In such a case, the corresponding component of $\mathbf{h}_{t+1}$ simply gets counted multiple times in (4.5.3).

**Option 4: Bernoulli Update:** At time $t$, let $\{B_{t,i}, i \in [d]\}$ be independent Bernoulli processes with success rate $\rho_t$. Thus

$$\Pr\{B_{t,i} = 1\} = \rho_t, \ \forall i \in [d]. \tag{4.5.4}$$

It is permissible for the success probability $\rho_t$ to vary with time. However, at any one time, all components must have the same success probability. Then define

$$\mathbf{v}_t := \sum_{i=1}^{d} \mathbf{e}_i I_{\{B_{t,i}=1\}} \in \{0,1\}^d. \tag{4.5.5}$$

Thus $\mathbf{v}_t$ is a random vector, and $v_{t,i}$ equals 1 if $B_{t,i} = 1$, and equals 0 otherwise. Now define

$$\mathbf{h}_{t+1}^{(4)} = \frac{1}{\rho_t} \mathbf{v}_t \circ \mathbf{h}_{t+1}. \tag{4.5.6}$$

Note that, as with the other options, the factor $1/\rho_t$ is the reciprocal of the likelihood of a particular $i \in [d]$ being selected for updating. However, there is no *a priori* upper bound on the number of nonzero components of $\mathbf{h}_{t+1}^{(4)}$; the stochastic gradient $\mathbf{h}_{t+1}^{(4)}$ can have up to $d$ nonzero components. It is also possible that $B_{t,i} = 0$ for each $i$, in which case $\mathbf{v}_t = \mathbf{0}$ and $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$. But the *expected* number of nonzero components is $\rho_t d$.

### 4.5.2 Convergence of SGD with Block Updating

When the choice of the block update direction involves some random choices (such as $\kappa_t^n$ or $B_{t+1,i}$), the definition of the filtration $\{\mathcal{F}_t\}$ needs to be adjusted. In the case of Option 2 (coordinate updating), $\mathcal{F}_t$ is the $\sigma$-algebra generated by $\kappa_0^t$ in addition to $\boldsymbol{\theta}_0^t$ and $\mathbf{h}_1^t$. In the case of Option 3, $\kappa_0^t$ is replaced by the collection $\kappa_{0,i}^t$ for $i \in [N]$. Finally, in Option 4, $\kappa_0^t$ is replaced by $\mathbf{v}_0^t$.

The objectives of Lemma 4.4 below are: (i) to show that the conditional expectation $E_t(\mathbf{h}_{t+1}^{(k)}$ is the same for all four values of $K$, and (ii) to give explicit expressions for the conditional variance $CV_t(\mathbf{h}_{t+1}^{(k)}$ for each value of $k$. To reduce the notational burden, we denote $\mathbf{h}_{t+1}^{(1)}$ by just $\mathbf{h}_{t+1}$.

**Lemma 4.4.** *As in* (4.3.4)*, define*

$$\mathbf{z}_t = E_t(\mathbf{h}_{t+1}), \boldsymbol{\zeta}_{t+1} = \mathbf{h}_{t+1} - \mathbf{z}_t.$$

*Then*

$$E_t(\mathbf{h}_{t+1}^{(k)}) = E_t(\mathbf{h}_{t+1}^{(1)}) = \mathbf{z}_t, k = 2,3,4. \tag{4.5.7}$$

*Moreover,*

$$\begin{aligned} CV_t(\mathbf{h}_{t+1}^{(2)}) &= (d-1)\|\mathbf{z}_t\|_2^2 + dE_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2), \\ CV_t(\mathbf{h}_{t+1}^{(3)}) &= (d-1)\|\mathbf{z}_t\|_2^2 + dE_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2), \\ CV_t(\mathbf{h}_{t+1}^{(4)}) &= \frac{1-\rho_t}{\rho_t}\|\mathbf{z}_t\|_2^2 + \frac{1}{\rho_t}E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2). \end{aligned} \tag{4.5.8}$$

*Proof.* It is obvious that (4.5.7) is satisfied. Therefore, to compute the conditional variance of $\mathbf{h}_{t+1}^{(k)}$, it is necessary to compute the residual $\|\mathbf{h}_{t+1}^{(k)} - \mathbf{z}_t\|_2^2$, and then take its conditional expectation.

**Option 2:** Suppose that $\kappa_t = i$. Then

$$h_{t+1,j}^{(2)} = \begin{cases} d(z_{t,i} + \zeta_{t+1,i}), & \text{if } j = i, \\ 0, & \text{if } j \neq i, \end{cases}$$

$$h^{(2)}_{t+1,j} - z_{t,j} = \begin{cases} (d-1)z_{t,i} + d\zeta_{t+1,i}, & \text{if } j = i, \\ -z_{t,j}, & \text{if } j \neq i, \end{cases}$$

Therefore, conditioned on the event $\kappa_t = i$, we have that

$$\sum_{j=1}^{d}(h^{(2)}_{t+1,j} - z_{t,j})^2 = (d-1)^2 z^2_{t,i} + d^2\zeta^2_{t+1,i} + 2d(d-1)z_{t,i}\zeta_{t+1,i} + \sum_{j\neq i} z^2_{t,j},$$

Now we take the conditional expectation of the above quantity. For this purpose, we note that (i) each of the events $\kappa_t = i$ occurs with probability $1/d$, and (ii) $E_t(z_{t,i}\zeta_{t+1,i}) = 0$. Hence

$$\begin{aligned} E_t(\|\mathbf{h}^{(2)}_{t+1} - \mathbf{z}_t\|_2^2) &= \frac{1}{d}\sum_{i=1}^{d}\left((d-1)^2 z^2_{t,i} + \sum_{j\neq i} z^2_{t,j}\right) + \frac{1}{d}\sum_{i=1}^{d} E_t(d^2\zeta^2_{t+1,i}) \\ &= \frac{(d-1)^2 + (d-1)}{d}\|\mathbf{z}_t\|_2^2 + dE_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2) \\ &= (d-1)\|\mathbf{z}_t\|_2^2 + dE_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2). \end{aligned}$$

This gives the first equation in (4.5.8).

**Option 3:** Observe that $\mathbf{h}_{t+1}$ is the average of $N$ different quantities wherein the error terms $\boldsymbol{\zeta}^n_{t+1}, n \in [N]$ are independent. Therefore their variances just add up, giving the middle equation in (4.5.8).

**Option 4:** For notational simplicity, we just use $\rho$ in the place of $\rho_t$. In this case, each component $h_{t+1,i}$ equals $(1/\rho)(z_{t,i} + \zeta_{t+1,i})$ with probability $\rho$, and 0 with probability $1-\rho$. Thus $h_{t+1,i} - z_{t,i}$ equals $((1/\rho) - 1)z_{t,i} + (1/\rho)\zeta_{t+1,i}$ with probability $\rho$, and $-z_{t,i}$ with probability $1-\rho$. As can be easily verified, the conditional variance is $((1-\rho)/\rho)z^2_{t,i} + (1/\rho)E_t(\zeta^2_{t+1,i}))$ for each component. As the Bernoulli processes for each component are mutually independent, the variances simply add up. It follows that

$$CV_t(\mathbf{h}^{(4)}_{t+1}) = \frac{1-\rho}{\rho}\|\mathbf{z}_t\|_2^2 + \frac{1}{\rho}E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2),$$

which is the bottom equation in (4.5.8). □

With Lemma 4.4 in place, we can now state the following meta-theorem on the convergence of block-uptating applied to the SGD algorithm. To state the theorem, we define $\mathbf{x}_t$ as in (4.3.4), and study the SGD formulation

$$\boldsymbol{\theta}^{(k)}_{t+1} = \boldsymbol{\theta}^{(k)}_t - \alpha_t \mathbf{h}^{(k)}_{t+1}. \tag{4.5.9}$$

This formulation is just (4.3.1), with the "full coordinate" stochastic gradient $\mathbf{h}_{t+1}$ replaced by $\mathbf{h}^{(k)}_{t+1}$ for $k = 2, 3, 4$. As stated earlier, we denote $\mathbf{h}^{(1)}_{t+1}$ as $\mathbf{h}_{t+1}$.

**Theorem 4.5.** *Suppose the stochastic gradient $\mathbf{h}_{t+1}$ satisfies the bounds* (4.3.6) *and* (4.3.7)*. Further, suppose that when Option* 4 *is used, then*

$$\inf_t \rho_t =: \bar{\rho} > 0. \tag{4.5.10}$$

*Then the conclusions of Theorems 4.1 and 4.2 continue to hold for $\{\boldsymbol{\theta}^{(k)}_t\}$ for $k = 2, 3, 4$.*

*Proof.* For the update rule (4.5.9), one can just replace $\mathbf{h}_{t+1}$ by $\mathbf{h}^{(k)}_{t+1}$ in (4.3.13). Therefore, for $k = 2$ or $k = 3$, (4.3.14) gets replaced by

$$\begin{aligned} E_t(J(\boldsymbol{\theta}^{(k)}_{t+1})) &= J(\boldsymbol{\theta}_t) - \alpha_t\langle\nabla J(\boldsymbol{\theta}^{(k)}_t), \mathbf{z}_t\rangle + \frac{\alpha_t^2 L}{2}E_t(\|\mathbf{h}^{(k)}_{t+1}\|_2^2) \\ &= J(\boldsymbol{\theta}_t) - \alpha_t\langle\nabla J(\boldsymbol{\theta}^{(k)}_t), \mathbf{z}_t\rangle + \frac{\alpha_t^2 L}{2}[(d-1)\|\mathbf{z}_t\|_2^2 + dE_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2)] \\ &\leq J(\boldsymbol{\theta}^{(k)}_t) - \alpha_t\langle\nabla J(\boldsymbol{\theta}^{(k)}_t), \mathbf{z}_t\rangle + \frac{\alpha_t^2 dL}{2}[\|\mathbf{z}_t\|_2^2 + E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2)] \\ &= J(\boldsymbol{\theta}^{(k)}_t) - \alpha_t\langle\nabla J(\boldsymbol{\theta}^{(k)}_t), \mathbf{z}_t\rangle + \frac{\alpha_t^2 dL}{2}CV_t(\|\mathbf{h}_{t+1}\|_2^2). \end{aligned} \tag{4.5.11}$$

In deriving (4.5.11), we use two facts: First, all the stochastic gradients $\mathbf{h}_{t+1}^{(k)}$ have the same conditional expectation, and second, it follows from (4.5.8) that

$$CV_t(\|\mathbf{h}_{t+1}^{(k)}\|_2^2) = (d-1)\|\mathbf{z}_t\|_2^2 + dE_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2) \leq d[\|\mathbf{z}_t\|_2^2 + dE_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2)] = dCV_t(\|\mathbf{h}_{t+1}\|_2^2).$$

So in effect we have replaced $\alpha_t^2$ by $d\alpha_t^2$. The desired conclusions now follow readily. Next, when Option 4 is utilized, we have that

$$CV_t(\mathbf{h}_{t+1}^{(4)}) = \frac{1-\rho_t}{\rho_t}\|\mathbf{z}_t\|_2^2 + \frac{1}{\rho_t}E_t(\|\boldsymbol{\zeta}_{t+1}\|_2^2) \leq \frac{1-\bar{\rho}}{\bar{\rho}}CV_t(\|\mathbf{h}_{t+1}\|_2^2).$$

So once again the desired conclusion follows. □

### 4.5.3 Convergence of the Unified Momentum Algorithms

Next we study the unified momentum-based algorithms of (4.4.2)–(4.4.3), but with block updating. Specifically, suppose

$$\mathbf{w}_{t+1} = \mathbf{w}_t + a_t\mathbf{v}_t - b_t\alpha_t\mathbf{h}_{t+1}^{(k)}, \tag{4.5.12}$$

$$\mathbf{v}_{t+1} = \mu_t\mathbf{v}_t - \alpha_t\mathbf{h}_{t+1}^{(k)}, \tag{4.5.13}$$

where $\mathbf{h}_{t+1}^{(k)}$ denotes the block-updated stochastic gradient. Thus, at each step $t$, some but not all components of $\mathbf{h}_{t+1}^{(k)}$ will be zero. However, there is no block-updating in the other terms.

**Theorem 4.6.** *Suppose that the various constants satisfy the assumptions in* (4.4.29)*, while the objective function $J(\cdot)$ satisfies Standing Assumptions (J1) and (J2). Further, suppose the stochastic gradient $\mathbf{h}_{t+1}$ satisfies the assumptions* (4.4.27)–(4.4.28)*. Further, suppose that when Option* 4 *is used, then*

$$\inf_t \rho_t =: \bar{\rho} > 0. \tag{4.5.14}$$

*Then the conclusions of Theorems 4.3 and 4.4 continue to hold for $\{\boldsymbol{\theta}_t^{(k)}\}$ for $k = 2, 3, 4$.*

The proofs are omitted as they are obvious.

## Notes and References

As shown in Section 3.1, the problem of finding a stationary point of a $\mathcal{C}^1$ function $J(\cdot)$ is equivalent to finding a solution of $\nabla J(\boldsymbol{\theta}^*) = \mathbf{0}$. Hence all the discussion in Chapter 3 is also applicable here.

As mentioned in the Notes and References section of Chapter 1, methods such as steepest descent, conjugate gradient, and quasi-Newton etc. *using the exact gradient vector* were widely studied in the 1960s. However, the behavior of these algorithms when the true gradient was replaced by an approximate, or even stochastic, gradient commenced only in the 1970s. One of the early papers to study this approach is [118], in which the authors introduce a "pseudo-gradient" (which is the same as the present stochastic gradient) which, *on average*, has a negative inner product with the true gradient. From that beginning, optimization using a stochastic gradient has witnessed an explosion of papers. The objectives of these papers was mostly to relax the assumptions on the class of functions (from strongly convex or convex to something more general), and on the measurement errors (by permitting biased noise and/or noise whose conditional variance grows without bound at the iteration counter $t$ increases). The results in Section 4.3 are the most general available at present, and are taken from [70, 71].

The material in Section 4.5 is largely taken from [121], which also contains several numerical examples. There are several other papers that mention "block" updating, such as [176, 30, 97, 122]. However, the choice of the "blocks" to be udpated is far less general than it is in [121]. The discussion of momentum-based methods with time-varying parameters is taken from [169].

# Chapter 5

# Markov Decision Processes

A brief introduction to Reinforcement Learning was given in Section 1.2. A widely used mathematical formalism for studying problems in RL is Markov Decision Processes (MDPs) where the dynamics of the Markov process are not known, and must somehow be "inferred" on the fly. Before tackling that problem, we must first understand MDPs when the dynamics *are* known. That is the aim of the present chapter. In the interests of simplicity, the discussion is limited to the situation where the state and action spaces underlying the MDP are finite sets. MDPs where the underlying state space and/or action space is countable, or an arbitrary measurable space, are also of interest in some applications. For example, the situation where $\mathcal{X}$ and/or $\mathcal{U}$ are subsets of some Euclidean space $\mathbb{R}^d$ for some $d$ are also sometimes of interest. Two recent papers [60, 61] present some new techniques for addressing such problems. The latter paper also contains an extensive and relevant bibliography. However, we do not study the more general situations in these notes.

The topic of MDPs is quite well-studied, and there are several excellent books on the subject. The reader is directed to [119] for a comprehensive treatment of the subject, which also studies the case of infinite state and action spaces. The theory of MDPs is also studied in [145] and [148]. The book [27] contains several practical examples of MDPs.

## 5.1 Markov Reward Processes

Recall the introduction to Markov processes in Section 2.2. Further facts about Markov processes can be found in [131, 167].

Suppose $\mathcal{X}$ is a finite set of cardinality $n$, written as $\{x_1, \ldots, x_n\}$. If $\{X_t\}_{t \geq 0}$ is a stationary Markov process assuming values in $\mathcal{X}$, then the corresponding state transition matrix $A$ is defined by

$$a_{ij} = \Pr\{X_{t+1} = x_j | X_t = x_i\}. \tag{5.1.1}$$

Thus the $i$-th row of $A$ is the conditional probability vector of $X_{t+1}$ when $X_t = x_i$. Clearly the row sums of the matrix $A$ are all equal to one. Therefore the induced norm $\|A\|_{\infty \to \infty}$ also equals one.

Up to now there is nothing new beyond the contents of Section 2.2. Now suppose that there is a "reward" function $R : \mathcal{X} \to \mathbb{R}$ associated with each state. There is no consensus within the community about whether the reward corresponding to the state $X_t$ is paid at time $t$ as in [148], or time $t+1$, as in [119, 145]. It is assumed here that the reward is paid at time $t$, and is denoted by $R_t$; the modifications required to handle the other approach are easy and left to the reader. The reward $R_t$ can either be a deterministic function of $X_t$, or a random function. If $R_t$ is a deterministic function of $X_t$, then we have that $R_t = R(X_t)$ where $R$ is the reward function mapping $\mathcal{X}$ into (a finite subset of) $\mathbb{R}$. On the other hand, if $R_t$ is a random function of $X_t$, then one would have to provide the probability distribution of $R_t$ given $X_t$. Since $X_t$ has only $n$ different values, we would have to provide $n$ different probability distributions.

Two kinds of Markov reward processes are widely studied, namely: Discounted reward processes, and average reward processes. Each of these is studied in a separate subsection.

### 5.1.1 Discounted Reward Processes

a

To study discounted Markov Reward Processes, we choose a "discount factor" $\gamma \in (0,1)$. Suppose $x_i \in \mathcal{X}$ is the "starting state of interest." Then the **expected discounted future reward** $V(x_i)$ starting at time 0 in state $x_i$ is defined as

$$V(x_i) = E\left[\sum_{t=0}^{\infty} \gamma^t R_t | X_0 = x_i\right]. \tag{5.1.2}$$

We often just use "discounted reward" instead of the longer phrase. Note that, because the set $\mathcal{X}$ is finite, the reward function $R_t$ is bounded if it is a deterministic function of $X_t$. If $R_t$ is a random variable dependent on $X_t$, then it customary to assume that it is bounded. With these assumptions, because $\gamma < 1$, the above summation converges and is well-defined. The quantity $V(x_i)$ is referred to as the **value function** associated with $x_i$, and the vector

$$\mathbf{v} = [\ V(x_1) \quad \cdots \quad V(x_n)\ ]^\top, \tag{5.1.3}$$

is referred to as the **value vector**. Note that, throughout these notes, we view the value as both a *function* $V : \mathcal{X} \to \mathbb{R}$ as well as a *vector* $\mathbf{v} \in \mathbb{R}^n$. The relationship between the two is given by (5.1.3). We shall use whichever interpretation is more convenient in a given context.

This raises the question as to how the value function and/or value vector is to be determined. Define the vector $\mathbf{r} \in \mathbb{R}^n$ via

$$\mathbf{r} := [\ r_1 \quad \cdots \quad r_n\ ]^\top, \tag{5.1.4}$$

where, if $R_t$ is a random function of $X_t$, then

$$r_i := E[R_t | X_t = x_i]. \tag{5.1.5}$$

Of course, if $R_t$ is a deterministic function $R(X_t)$, then $r_i$ is just $R(x_i)$.

**Theorem 5.1.** *The vector* $\mathbf{v}$ *satisfies the recursive relationship*

$$\mathbf{v} = \mathbf{r} + \gamma A \mathbf{v}, \tag{5.1.6}$$

*or, in expanded form,*

$$V(x_i) = r_i + \gamma \sum_{j=1}^{n} a_{ij} V(x_j). \tag{5.1.7}$$

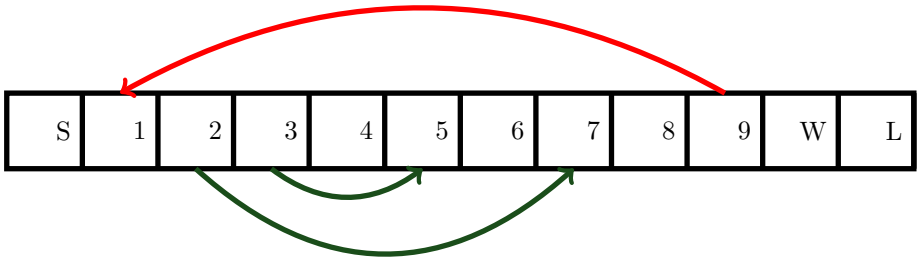
*Proof.* Let $x_i \in \mathcal{X}$ be arbitrary. Then by definition we have

$$V(x_i) = E\left[\sum_{t=0}^{\infty} \gamma^t R_t | X_0 = x_i\right] = r_i + E\left[\sum_{t=1}^{\infty} \gamma^t R_t | X_0 = x_i\right]. \tag{5.1.8}$$

However, if $X_0 = x_i$, then $X_1 = x_j$ with probability $a_{ij}$. Therefore we can write

$$\begin{aligned} E\left[\sum_{t=1}^{\infty} \gamma^t R_t | X_0 = x_i\right] &= \sum_{j=1}^{n} a_{ij} E\left[\sum_{t=1}^{\infty} \gamma^t R_t | X_1 = x_j\right] \\ &= \gamma \sum_{j=1}^{n} a_{ij} E\left[\sum_{t=0}^{\infty} \gamma^t R_t | X_0 = x_j\right] \\ &= \gamma \sum_{j=1}^{n} a_{ij} V(x_j). \end{aligned} \tag{5.1.9}$$

In the second step we use fact that the Markov process is stationary. Substituting from (5.1.9) into (5.1.8) gives the recursive relationship (5.1.7). □




**Example 5.1.**

We analyze the toy snakes and ladders game of Example 2.6. As shown therein, the state transition matrix of this game is given by

| | $S$ | 1 | 4 | 5 | 6 | 7 | 8 | $W$ | $L$ |
|---|---|---|---|---|---|---|---|---|---|
| $S$ | 0 | 0.25 | 0.25 | 0.25 | 0 | 0.25 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0.25 | 0.50 | 0 | 0.25 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0.25 | 0.25 | 0.25 | 0.25 | 0 | 0 |
| 5 | 0 | 0.25 | 0 | 0 | 0.25 | 0.25 | 0.25 | 0 | 0 |
| 6 | 0 | 0.25 | 0 | 0 | 0 | 0.25 | 0.25 | 0.25 | 0 |
| 7 | 0 | 0.25 | 0 | 0 | 0 | 0 | 0.25 | 0.25 | 0.25 |
| 8 | 0 | 0.25 | 0 | 0 | 0 | 0 | 0.25 | 0.25 | 0.25 |
| $W$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| $L$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

We define a random reward function for this problem, as follows: We set $R_t = f(X_{t+1})$, where $f$ is defined as follows: $f(W) = 5$, $f(L) = -2$, $f(x) = 0$ for all other states. However, there is an expected reward *depending on the state at the next time instant*. For example, if $X_0 = 6$, then the expected value of $R_0$ is 5/4, whereas if $X_0 = 7$ or $X_0 = 8$, then the expected value of $R_0$ is 3/4.

Now let us see how the implicit equation (5.1.6) can be solved to determine the value vector $\mathbf{v}$. Since the induced matrix norm $\|A\|_{\infty\to\infty} = 1$ and $\gamma < 1$, it follows that the matrix $I - \gamma A$ is nonsingular. Therefore, for every fixed assignment of rewards to states, there is a unique $\mathbf{v}$ that satisfies (5.1.6). In principle it is possible to deduce from (5.1.6) that

$$\mathbf{v} = (I - \gamma A)^{-1}\mathbf{r}. \tag{5.1.10}$$

The difficulty wth this formula however is that in most actual applications of Markov Decision Problems, the integer $n$ denoting the size of the state space $\mathcal{X}$ is quite large. Moreover, inverting a matrix has cubic complexity in the size of the matrix. Therefore it may not be practicable to invert the matrix $I - \gamma A$. So we are forced to look for alternate approaches. A feasible approach is provided by the Contraction Mapping Theorem (CMT), namely Theorem 7.1. With the contraction mapping theorem in hand, we can apply it to the problem of computing the value of a discounted Markov reward process.

**Theorem 5.2.** *The map $\mathbf{y} \mapsto T\mathbf{y} := \mathbf{r} + \gamma A\mathbf{y}$ is monotone and is a contraction with respect to the $\ell_\infty$-norm, with contraction constant $\gamma$.*

*Proof.* The first statement is that if $\mathbf{y}_1 \leq \mathbf{y}_2$ componentwise (and note that the vectors $\mathbf{y}_1, \mathbf{y}_2$ need not consist of only positive components), then $T\mathbf{y}_1 \leq T\mathbf{y}_2$. This is obvious from the fact that the matrix $A$ has only nonnegative components, so that $A\mathbf{y} \geq \mathbf{0}$ whenever $\mathbf{y} \geq \mathbf{0}$, where the inequalities are componentwise. Now suppose that $\mathbf{y}_1 \leq \mathbf{y}_2$. Then

$$\mathbf{y}_2 - \mathbf{y}_1 \geq \mathbf{0} \implies A(\mathbf{y}_2 - \mathbf{y}_1) \geq \mathbf{0}.$$

Therefore $A\mathbf{y}_1 \leq A\mathbf{y}_2$, which in turn implies that $T$ is monotone. For the second statement, note that, because the matrix $A$ is row-stochastic, the induced norm of $A$ with respect to $\|\cdot\|_\infty$ is equal to one. Therefore

$$\|T\mathbf{y}_1 - T\mathbf{y}_2\|_\infty = \|\gamma A(\mathbf{y}_1 - \mathbf{y}_2)\|_\infty \leq \gamma\|\mathbf{y}_1 - \mathbf{y}_2\|_\infty.$$

This completes the proof. □

Therefore one can solve (5.1.6) by repeated application of the contraction map $T$. In other words, we can choose some vector $\mathbf{y}^0$ arbitrarily, and then define

$$\mathbf{y}^{i+1} = \mathbf{r} + \gamma A\mathbf{y}^i.$$

Then the contraction mapping theorem tells us that $\mathbf{y}^i$ converges to the value vector $\mathbf{v}$. Moreover, from (7.1.3) one can estimate how far the current iteration is from the solution $\mathbf{v}$. Note that the contraction constant $\rho$ in the statement of the theorem can be taken as the discount factor $\gamma$. Define the constant

$$c := \|\mathbf{r} + \gamma A\mathbf{y}^0 - \mathbf{y}^0\|_\infty,$$

which measures how far away the initial guess $\mathbf{y}^0$ is from satisfying (5.1.6). Then we have the estimate

$$\|\mathbf{y}^i - \mathbf{v}\|_\infty \leq \frac{\gamma^i}{1-\gamma}c. \tag{5.1.11}$$

In this approach to finding the value function, each iteration has quadratic complexity in $n$, the size of the state space. Moreover, (5.1.11) can be used to decide how many iterations should be run to get an acceptable estimate for $\mathbf{v}$. This approach to determining $\mathbf{v}$ (albeit approximately) is known as "value iteration." Now, if we wish to find an approximation $\mathbf{v}_k$ to $\mathbf{v}$ that is accurate to within some prespecified accuracy $\epsilon$, then we need to ensure that

$$\frac{\gamma^k}{1-\gamma}c \leq \epsilon, \text{ or } k \geq \frac{\log(c/(\epsilon(1-\gamma)))}{\log(1/\gamma)} = \frac{\log((\epsilon(1-\gamma))/c)}{\log\gamma} =: N,$$

after routine calculations. Thus if we use $N$ iterations, then the complexity of value iteration is $O(Nn^2)$ as opposed to $O(n^3)$ for using (5.1.10). Hence the value iteration approach is preferable if $N \ll n$. To illustrate, let us choose typical values of $\gamma = 0.99$, $\epsilon = 10^{-4}$. If the initial mismatch $c = 5$, then $N = 1,535$. So if, for example, $n = 10^6$, then value iteration would be preferable. Note that the faster future rewards are discounted (i.e., the smaller $\gamma$ is), the faster the iterations will converge.

### 5.1.2 Average Reward Markov Processes

a

Now we discuss average reward Markov processes. As before, there is a Markov process $\{X_t\}_{t\geq 0}$ on a finite space $\mathcal{X}$ of cardinality $n$, with the state transition matrix $A \in [0,1]^{n\times n}$, and a reward function $R : \mathcal{X} \to \mathbb{R}$. If the reward is random, it is assumed that the reward is bounded almost surely (to avoid technicalities), and the symbol $r_i$ is used to denote the *expected value* of the reward to be paid at time $t$, when $X_t = x_i$.

The objective is to compute the **average reward**

$$c^* := \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} E[R(X_t)|X_0 \sim \boldsymbol{\phi}], \tag{5.1.12}$$

where $\boldsymbol{\phi} \in \mathbb{S}(\mathcal{X})$ is a probability distribution on $\mathcal{X}$. Compared with the definition (5.1.2) of the discounted reward, two points of contrast would strike us at once.

1. In (5.1.2), the existence of the sum is not in question, because $\gamma < 1$. However, in the present instance, there is no *a priori* reason to assume that the limit in (5.1.12) exists.

2. The value function $V$ in (5.1.2) is associated with an initial state $x_i$. It is implicit in the definition that $V(x_i)$ need not equal $V(x_j)$ if $x_i \neq x_j$. In (5.1.12), the initial state is replaced by an initial distribution $\boldsymbol{\phi}$, which is more general. However, we write $c^*$, instead of $c^*(\boldsymbol{\phi})$, suggesting that the limit, if it exists, is independent of $\boldsymbol{\phi}$.

Theorem 5.3 presents a simple sufficient condition to address both of the above observations.

**Theorem 5.3.** *Suppose $A$ is irreducible, and let $\boldsymbol{\mu}$ denote its unique stationary distribution. Then*

$$c^* = \boldsymbol{\mu}\mathbf{r} = E[R, \boldsymbol{\mu}], \ \forall \boldsymbol{\phi} \in \mathbb{S}(\mathcal{X}), \tag{5.1.13}$$

*where $\mathbf{r}$ is the reward vector defined in* (5.1.4)*.*

*Proof.* If $X_0 \sim \boldsymbol{\phi}$, then $X_t \sim \boldsymbol{\phi}A^t$. Therefore

$$E[R(X_t)|X_0 \sim \boldsymbol{\phi}] = \boldsymbol{\phi}A^t\mathbf{r}.$$

Also, as stated in Theorem 2.11, we have

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} A^t = \mathbf{1}_n\boldsymbol{\mu}.$$

Therefore

$$c^* = \boldsymbol{\phi}\left[\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} A^t\right]\mathbf{r} = \boldsymbol{\phi}\mathbf{1}_n\boldsymbol{\mu}\mathbf{r} = \boldsymbol{\mu}\mathbf{r} = E[R, \boldsymbol{\mu}], \tag{5.1.14}$$

because $\boldsymbol{\phi}\mathbf{1}_n = 1$. This is the desired result. □

Next we introduce an important concept known variously as the **bias** or the **transient reward**. For a discussion (albeit with "reward" replaced by "cost"), see [119, Section 8.2.3] or [3, Section 4.1].

**Definition 5.1.** Suppose $A$ is primitive,[1] and define $c^*$ as in (5.1.14) For each index $i$, the **transient reward** $J_i^* \in \mathbb{R}$ is defined as

$$J_i^* = \sum_{t=0}^{\infty}\{E[R(X_t|X_0 = x_i] - c^*\}. \tag{5.1.15}$$

*A priori* it is not clear why the sum in (5.1.15) is well-defined, because there is no averaging over time. It is now shown that the transient reward is indeed well-defined, and several explicit expressions are given for it.

**Theorem 5.4.** *Suppose $A$ is primitive, and let $\boldsymbol{\mu}$ denote its stationary distribution. Define $M := \mathbf{1}_n\boldsymbol{\mu} \in [0,1]^{n\times n}$, and $\mathbf{J}^* \in \mathbb{R}^n$ as $[J_i^*]$. Then the following statements are true:*

1. *The vector $\mathbf{J}^*$ is well-defined.*

2. *An explicit expression for $\mathbf{J}^*$ is given by*

$$\mathbf{J}^* = (I - A + M)^{-1}(I - M)\mathbf{r} = (I - A + M)^{-1}(\mathbf{r} - c^*\mathbf{1}_n). \tag{5.1.16}$$

[1] This is equivalent to assuming that $A$ is irreducible and aperiodic; see Theorem 2.10.

*3. The vector* $\mathbf{J}^*$ *satisfies the "Poisson equation"*

$$\mathbf{J} = \mathbf{r} - c^*\mathbf{1}_n + A\mathbf{J}. \tag{5.1.17}$$

*Moreover,* $\mathbf{J}^*$ *is the unique solution of* (5.1.17) *that satisfies*

$$\boldsymbol{\mu}\mathbf{J} = 0. \tag{5.1.18}$$

*Proof.* Note that $\boldsymbol{\mu}, \mathbf{1}_n$ are row and column eigenvectors of $A$ corresponding to the eivenvalue $\lambda = 1$, and that all other eigenvalues of $A$ have magnitude less than one. So if we define

$$A_2 = A - \mathbf{1}_n\boldsymbol{\mu} = A - M,$$

then the spectrum of $A_2$ is the same as that of $A$, except that the eigenvalue at 1 is replaced by 0. In particular, $\rho(A_2) < 1$, and as a consequence

$$\sum_{t=0}^{\infty} A_2^t = (I - A_2)^{-1} = (I - A + M)^{-1}. \tag{5.1.19}$$

Next, suppose $\mathbf{v} \in \mathbb{R}^n$ satisfies $\boldsymbol{\mu}\mathbf{v} = 0$. Then it is easy to verify that $A\mathbf{v} = A_2\mathbf{v}$, and moreover, $\boldsymbol{\mu}A_2\mathbf{v} = 0$. Repeated application of this relationship shows that $A^t\mathbf{v} = A_2^t\mathbf{v}$, for all $t \geq 1$. Therefore, for every such $\mathbf{v}$, we have that

$$\sum_{t=0}^{\infty} A^t\mathbf{v} = \sum_{t=0}^{\infty} A_2^t\mathbf{v} = (I - A + M)^{-1}\mathbf{v}. \tag{5.1.20}$$

Now in particular, choose

$$\mathbf{v} = \mathbf{r} - c^*\mathbf{1}_n = (I - M)\mathbf{r}.$$

Then it follows from (5.1.14) that $\boldsymbol{\mu}\mathbf{v} = 0$. Hence (5.1.20) implies that

$$\sum_{t=0}^{\infty} A^t(\mathbf{r} - c^*\mathbf{1}_n) = (I - A + M)^{-1}(I - M)\mathbf{r}.$$

To prove Statements 1 and 2, let $\mathbf{e}_i$ denote the $i$-th elementary basis vector. Then $X_0 = x_i$ is equivalent to $X_0 \sim \mathbf{e}_i^\top$. Then $X_t \sim \mathbf{e}_t^\top A^t$, and

$$J_i^* = \sum_{t=0}^{\infty} [\mathbf{e}_i^\top A^t\mathbf{r} - c^*],$$

$$\begin{aligned} \mathbf{J}^* &= \sum_{t=0}^{\infty}(A^t\mathbf{r} - c^*\mathbf{1}_n) = \sum_{t=0}^{\infty} A^t(\mathbf{r} - c^*\mathbf{1}_n) \\ &= (I - A + M)^{-1}(I - M)\mathbf{r}. \end{aligned} \tag{5.1.21}$$

Here we use the fact that $c^*\mathbf{1}_n = c^*A^t\mathbf{1}_n$ for all $t$. This establishes Statements 1 and 2.

Now we come to Statement 3. From (5.1.15), we get

$$\begin{aligned} J_i^* &= \sum_{t=0}^{\infty}\{E[R(X_t|X_0 = x_i] - c^*\} \\ &= r_i - c^* + \sum_{t=1}^{\infty}\{E[R(X_t|X_0 = x_i] - c^*\} \\ &= r_i - c^* + \sum_{j=1}^{n} a_{ij} \sum_{t=1}^{\infty}\{E[R(X_t|X_1 = x_j] - c^*\} \\ &= r_i - c^* + \sum_{j=1}^{n} a_{ij} J_j^*, \end{aligned}$$

which is just (5.1.17) written out in component form. Hence $\mathbf{J}^*$ is a particular solution of (5.1.17).

Finally, observe that if $J$ is another solution of (5.1.17), then $(\mathbf{J}^* - \mathbf{J}) = A(\mathbf{J}^* - \mathbf{J})$, which implies that $\mathbf{J} = \mathbf{J}^* + \alpha \mathbf{1}_n$ for some constant $\alpha$. Thus $\{\mathbf{J}^* + \alpha \mathbf{1}_n : \alpha \in \mathbb{R}\}$ is the set of all solutions to (5.1.17). Now, since $\boldsymbol{\mu}(\mathbf{r} - c^* \mathbf{1}_n) = 0$, it follows that

$$\boldsymbol{\mu}\mathbf{J}^* = \boldsymbol{\mu} \sum_{t=0}^{\infty} A^t(\mathbf{r} - c^* \mathbf{1}_n) = \sum_{t=0}^{\infty} \boldsymbol{\mu}(\mathbf{r} - c^* \mathbf{1}_n) = 0.$$

Moreover, if $\boldsymbol{\mu}(\mathbf{J}^* + \alpha \mathbf{1}_n) = 0$, then $\alpha = 0$. Hence $\mathbf{J}^*$ is the unique solution of (5.1.17) that satisfies $\boldsymbol{\mu}\mathbf{J} = 0$. □

It is possible to give an alternate proof of Statement 3, and we do so now. Suppose $\mathbf{J}^*$ is given by (5.1.21). Observe that

$$\boldsymbol{\mu}(I - A + M) = \boldsymbol{\mu}M = \boldsymbol{\mu}, \text{ or } \boldsymbol{\mu}(I - A + M)^{-1} = \boldsymbol{\mu}.$$

Also, $\boldsymbol{\mu}(I - M) = \mathbf{0}$. Therefore

$$\boldsymbol{\mu}\mathbf{J}^* = \boldsymbol{\mu}(I - A + M)^{-1}(I - M)\mathbf{r} = \boldsymbol{\mu}(I - M)\mathbf{r} = 0.$$

Next, (5.1.21) implies that

$$(I - A + M)\mathbf{J}^* = (I - M)\mathbf{r}.$$

However, $M\mathbf{J}^* = \mathbf{1}_n \boldsymbol{\mu}\mathbf{J}^* = \mathbf{0}$, and $(I - M)\mathbf{r} = \mathbf{r} - c^* \mathbf{1}_n$. Therefore

$$\mathbf{J}^* - A\mathbf{J}^* = \mathbf{r} - c^* \mathbf{1}_n.$$

This is just (5.1.17). The above derivation avoids infinite sums.

Let us now summarize the situation of discounted reward processes vis-a-vis average reward processes.

- The discounted reward is well-defined for *all* Markov reward processes, irrespective of the nature of the matrix $A$.
- If $A$ is *irreducible*, then the average reward is also well-defined. However, there is no guarantee that the transient reward is well-defined.
- If $A$ is not just irreducible but also *primitive*, then the transient reward is also well-defined.

## 5.2 Markov Decision Processes

### 5.2.1 Markov Decision Processes: Problem Set-Up

In a Markov process, the state $X_t$ evolves on its own, according to a predetermined state transition matrix. In contrast, in a MDP, there is also another variable called the "action" which affects the dynamics. Specifically, in addition to the state space $\mathcal{X}$, there is also a finite set of actions $\mathcal{U}$. Associated with each action $u_k \in \mathcal{U}$ is a corresponding state transition matrix $A^{u_k} = [a_{ij}^{u_k}]$. So at time $t$, if the state $X_t$ equals $x_i$, and an action $u_k \in \mathcal{U}$ is applied, then

$$\Pr\{X_{t+1} = x_j | X_t = x_i, U_t = u_k\} = a_{ij}^{u_k}, \ \forall x_j \in \mathcal{X}. \tag{5.2.1}$$

Obviously, for each fixed $u_k \in \mathcal{U}$, the corresponding state transition matrix $A^{u_k}$ is row-stochastic. In addition, there is also a "reward" function $R : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$. Note that in a Markov reward process, the reward depends only on the current state, whereas in a Markov decision process, the reward depends on both the current state as well as the action taken. As in Markov reward processes studied in Section 5.1, it is possible to permit $R$ to be a random function of $X_t$ and $U_t$ as opposed to a deterministic function. Moreover, to be consistent with the earlier convention, it is assumed that the reward $R(X_t, U_t)$ is paid at time $t$. Note that other authors assume that the reward is paid at time $t + 1$.

In the above definition, the assumption is that the set of permissible actions $\mathcal{U}$ *does not depend* on the current state $x_i$. One can imagine situations where this assumption may not be realistic. An example is provided by autonomous navigation in the midst of obstacles. Suppose there is a grid, some squares of which are occupied by obstacles, and the state space $\mathcal{X}$ consists of the free squares. The action set can be $\mathcal{U} = \{F, B, R, L\}$, representing go forward, go back, turn right, and turn left respectively. Depending on the current state (that is, the square currently occupied by the vehicle), some of these actions might not be permissible, due to the presence of obstacles. However, this situation can be tackled by assigning a large negative "reward" (that is, a large penalty) to an action that is not permitted. This approach provides a uniform set of actions for all states.

The most important concept in an MDP is that of a "policy," which is just a systematic way of choosing $U_t$ given $X_t$. One can make a distinction between deterministic and probabilistic policies. A deterministic policy is just a map from $\mathcal{X}$ to $\mathcal{U}$. A probabilistic policy is a map from $\mathcal{X}$ from the set of probability distributions on $\mathcal{U}$, denoted by $\mathbb{S}(\mathcal{U})$. Let $\Pi_d$, $\Pi_p$ denote respectively the set of deterministic, and the set of probabilistic, policies. Clearly the number of deterministic policies is $|\mathcal{U}|^{|\mathcal{X}|}$, while $\Pi_p$ is uncountable. Observe that a policy $\pi \in \Pi_d$ can be represented by a matrix $P_\pi$ of dimensions $|\mathcal{X}| \times |\mathcal{U}|$, where each row of $P_\pi$ contains a single one and the rest are zeros. Thus in $i$, the one is in column $\pi(x_i)$ and the rest are zeros. If $\pi \in \Pi_p$, then $P_\pi$ need not be binary, but $P_\pi$ must have only nonnegative elements, and the sum of each row must equal one.

Now we make an important observation. Whether a policy $\pi$ is deterministic or probabilistic, the resulting stochastic process $\{X_t\}$ is a Markov process with the state transition matrix denoted by $A^\pi$ determined as follows: If $\pi \in \Pi_d$, then

$$[A^\pi]_{ij} = \Pr\{X_{t+1} = x_j | X_t = x_i, \pi\} = a_{ij}^{\pi(x_i)}. \tag{5.2.2}$$

If $\pi \in \Pi_p$ and

$$\pi(x_i) = [\ \phi_{i1} \quad \cdots \quad \phi_{im}\ ], \tag{5.2.3}$$

where $m = |\mathcal{U}|$, then

$$[A^\pi]_{ij} = \Pr\{X_{t+1} = x_j | X_t = x_i, \pi\} = \sum_{k=1}^{m} \phi_{ik} a_{ij}^{u_k}. \tag{5.2.4}$$

Equation (5.2.4) contains (5.2.3) as a special case, by setting $\phi_{ij} = 1$ if $\pi(x_i) = u_j$, and zero otherwise.

In a similar manner, for every policy $\pi$, the reward function $R : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ can be converted into a reward map $R_\pi : \mathcal{X} \to \mathbb{R}$, or a reward vector $\mathbf{r}_\pi$, as follows: If $\pi \in \Pi_d$, then

$$R_\pi(x_i) = R(x_i, \pi(x_i)), \tag{5.2.5}$$

whereas if $\pi \in \Pi_p$, then

$$R_\pi(x_i) = \sum_{k=1}^{m} \phi_{ik} R(x_i, u_k). \tag{5.2.6}$$

Equations (5.2.4) and (5.2.6) can be put into "closed-form" using the notion of a Hadamard product. The standard definition of a Hadamard product is this: If $M, N$ are matrices of equal dimensions, then their Hadamard product $M \circ N$ has the same dimensions, and is defined by

$$[M \circ N]_{ij} = m_{ij} n_{ij},\ \forall i, j.$$

We now extend the definition as follows: Suppose $M$ is a matrix, and $N$ is a column vector, where both $M, N$ have the same number of rows. Then we define $M \circ N$ as a matrix that has the same dimensions as $M$, given by

$$[M \circ N]_{ij} = m_{ij} n_i,\ \forall i, j.$$

With this definition, we can write both $A_\pi$ and $\mathbf{r}_\pi$ as follows: Defiine the matrix $P_\pi \in [0,1]^{|\mathcal{X}| \times |\mathcal{U}|}$ associated with the policy $\pi$. Note that if $\pi \in \Pi_d$ is a deterministic policy, then $P_\pi \in \{0,1\}^{|\mathcal{X}| \times |\mathcal{U}|}$. Let $[P_\pi]_k$ denote

the $k$-th column of $P_\pi$. Then

$$A^\pi = \sum_{k=1}^{|\mathcal{U}|} A^{(u_k)} \circ [P_\pi]_k. \tag{5.2.7}$$

Next, write the Reward matrix $R$ as a matrix of dimensions $|\mathcal{X}| \times |\mathcal{U}|$, where

$$R_{ik} = R(x_i, u_k).$$

Then

$$\mathbf{r}_\pi = [R \circ P_\pi] \cdot \mathbf{1}_{|\mathcal{U}|} = \sum_{k=1}^{|\mathcal{U}|} R \circ [P_\pi]^k. \tag{5.2.8}$$

Suppose $|\mathcal{X}| = 3, |\mathcal{U}| = 2$. Thus there are three states and two actions. Suppose that the two state transition matrices are given by

$$A^{(u_1)} = \begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0.5 & 0.4 & 0.1 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}, A^{(u_2)} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.3 & 0.2 & 0.5 \\ 0.1 & 0.2 & 0.7 \end{bmatrix}.$$

As required both matrices are row-stochastic. Further, suppose that the associated reward matrix is given by

$$R = \begin{bmatrix} 1 & 6 \\ 4 & 3 \\ 2 & 5 \end{bmatrix}.$$

This means that the reward associated with the state $x_1$ and action $u_1$ is 1, and so on. The reward can represented conveniently in matrix form as above.

Now suppose we choose the deterministic policy $\pi_1$ as $\pi_1(x_1) = u_1, \pi_1(x_2) = u_2, \pi_1(x_3) = u_1$ . This means that when $X_t = x_1$, we choose the action $U_t = u_1$ etc., irrespective of the value of the time index $t$. Thus the policy matrix $P_{\pi_1}$ is given by

$$P_{\pi_1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Now we can apply (5.2.7) to deduce that

$$A_{(u_1)} \circ [P_{\pi_1}]_1 = \begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0 & 0 & 0 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}, A_{(u_2)} \circ [P_{\pi_1}]_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0.3 & 0.2 & 0.5 \\ 0 & 0 & 0 \end{bmatrix},$$

$$A_{\pi_1} = A_{(u_1)} \circ [P_{\pi_1}]_1 + A_{(u_2)} \circ [P_{\pi_1}]_2 = \begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0.3 & 0.2 & 0.5 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}.$$

Next, the reward vector $R_{\pi_1}$ can be computed using (5.2.8). It follows that

$$R \circ P_{\pi_1} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \\ 2 & 0 \end{bmatrix}, R_{\pi_1} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}.$$

### 5.2.2 Markov Decision Processes: Analysis

For a MDP, one can pose three questions:

1. **Policy evaluation:** For a given policy $\pi$, define $V_\pi(x_i)$ to be the "value" associated with the policy $\pi$ and initial state $x_i$, that is, the expected discounted future reward with $X_0 = x_i$. How can $V_\pi(x_i)$ be computed for each $x_i \in \mathcal{X}$?

2. **Optimal Value Determination:** For a specified initial state $x_i$, define

$$V^*(x_i) := \max_{\pi \in \Pi_p} V_\pi(x_i), \tag{5.2.9}$$

   to be the **optimal value** over all policies for that initial state. How can $V^*(x_i)$ be computed? Note that in (5.2.9), the optimum is taken over all *probabilistic* policies. It is shown in Theorem 5.9 in the sequel that the optimum can actually be achieved by a *deterministic* policy.

3. **Optimal Policy Determination:** Define the **optimal policy** map $\mathcal{X} \to \Pi_d$ via

$$\pi^*(x_i) := \arg\max_{\pi \in \Pi_d} V_\pi(x_i). \tag{5.2.10}$$

   How can the optimal policy map $\pi^*$ be determined? Note that in the previous item, we wish to find the optimal value associated with each state, whereas in this item, we wish to identify a policy that *achieves* the optimal value. It is possible restrict our search only to deterministic policies, because as stated above, the maximum over $\pi \in \Pi_p$ is not any larger. Moreover, it is again shown in Theorem 5.9 that there exists *one common optimal policy for all initial states.*

Next we present answers to the three questions above.

**Policy Evaluation:**

Suppose a policy $\pi \in \Pi_d$ is specified. Then the corresponding state transition matrix and reward are given by (5.2.2) and (5.2.5) respectively. Now suppose we define the vector $\mathbf{v}_\pi$ by

$$\mathbf{v}_\pi = [\ V_\pi(x_1) \quad \dots \quad V_\pi(x_n)\ ], \tag{5.2.11}$$

and the reward vector $\mathbf{r}_\pi$ by

$$\mathbf{r}_\pi = [\ R_\pi(x_1) \quad \dots \quad R_\pi(x_n)\ ], \tag{5.2.12}$$

where $R(x_i)$ is defined by (5.2.5) or (5.2.6) as appropriate. Then it readily follows from Theorem 5.1 that $\mathbf{v}_\pi$ satisfies an equation analogous to (5.1.6), namely

$$\mathbf{v}_\pi = \mathbf{r}_\pi + \gamma A^\pi \mathbf{v}_\pi. \tag{5.2.13}$$

As before, it is inadvisable to compute $\mathbf{v}_\pi$ via $\mathbf{v}_\pi = (I - \gamma A^\pi)^{-1}\mathbf{r}_\pi$. Instead, one should use value iteration to solve (5.2.13).

For future use we introduce another function $Q_\pi : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, known as the **action-value function**, which is defined as follows:

$$Q_\pi(x_i, u_k) := R(x_i, u_k) + E_\pi \left[ \sum_{t=1}^{\infty} \gamma^t R_\pi(X_t) | X_0 = x_i, U_0 = u_k \right]. \tag{5.2.14}$$

Apparently this function was first defined in [172]. Note that $Q_\pi$ is defined only for deterministic policies. In principle it is possible to define it for probabilistic policies, but this is not commonly done. In the above definition, the expectation $E_\pi$ is with respect to the evolution of the state $X_t$ under the policy $\pi$. When the

reward is a random function of $X_t$ and $U_t$, then inside the summation we would need to take the expected value of $R(X_t, \pi(X_t))$ for a deterministic policy.

The way in which a MDP is set up is that at time $t$, the Markov process reaches a state $X_t$, based on the previous state $X_{t-1}$ and the state transition matrix $A^\pi$ corresponding to the policy $\pi$. Once $X_t$ is known, the policy $\pi$ determines the action $U_t = \pi(X_t)$, and then the reward $R_\pi(X_t) = R(X_t, \pi(X_t))$ is generated. In particular, when defining the value function $V_\pi(x_i)$ corresponding to a policy $\pi$, we start off the MDP in the initial state $X_0 = x_i$, *and choose the action* $U_0 = \pi(x_i)$. However, in defining the action-value function $Q$, we do not feel compelled to set $U_0 = \pi(X_0) = \pi(x_i)$, and can choose an *arbitrary action* $u_k \in \mathcal{U}$. From $t = 1$ onwards however, the action $U_t$ is chosen as $U_t = \pi(X_t)$. This seemingly small change leads to some simpifications. Specifically, it will be seen in later chapters that it is often easier to approximate (or to "learn") the action-value function than it is to approximate the value function.

Just as we can interpret $V : \mathcal{X} \to \mathbb{R}$ as a $|\mathcal{X}|$-dimensional vector, we can interpret $Q : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ as an $|\mathcal{X}| \cdot |\mathcal{U}|$-dimensional vector, or as a matrix of dimension $|X| \times |\mathcal{U}|$. Consequently the $Q$-vector has higher dimension than the value vector.

**Theorem 5.5.** *The function $Q$ satisfies the recursive relationship*

$$Q_\pi(x_i, u_k) = R(x_i, u_k) + \gamma \sum_{j=1}^{n} a_{ij}^{u_k} Q_\pi(x_j, \pi(x_j)). \tag{5.2.15}$$

*Proof.* Observe that at time $t = 0$, the state transition matrix is $A^{u_k}$. So, given that $X_0 = x_i$ and $U_0 = u_k$, the next state $X_1$ has the distribution

$$X_1 \sim [a_{ij}^{u_k}, j = 1, \cdots, n].$$

Moreover, $U_1 = \pi(X_1)$ because the policy $\pi$ is implemented from time $t = 1$ onwards. Therefore

$$\begin{aligned} Q_\pi(x_i, u_k) &= R(x_i, u_k) + E_\pi \left[ \sum_{j=1}^{n} a_{ij}^{u_k} \left( \gamma R(x_j, \pi(x_j)) + \sum_{t=2}^{\infty} \gamma^t R_\pi(X_t) | X_1 = x_j, U_1 = \pi(x_j) \right) \right] \\ &= R(x_i, u_k) + E_\pi \left[ \gamma \sum_{j=1}^{n} a_{ij}^{u_k} \left( R(x_j, \pi(x_j)) + \sum_{t=1}^{\infty} \gamma^t R_\pi(X_t) | X_1 = x_j, U_1 = \pi(x_j) \right) \right] \\ &= R(x_i, u_k) + \gamma \sum_{j=1}^{n} a_{ij}^{u_k} Q(x_j, \pi(x_j)). \end{aligned}$$

This is the desired conclusion. □

**Theorem 5.6.** *The functions $V_\pi$ and $Q_\pi$ are related via*

$$V_\pi(x_i) = Q_\pi(x_i, \pi(x_i)). \tag{5.2.16}$$

*Proof.* If we choose $u_k = \pi(x_i)$ then (5.2.15) becomes

$$Q_\pi(x_i, \pi(x_i)) = R_\pi(x_i) + \gamma \sum_{j=1}^{n} a_{ij}^{\pi(x_j)} Q(x_j, \pi(x_j)).$$

This is the same as (5.2.1) written out componentwise. We know that (5.2.1) has a unique solution. This shows that (5.2.16) holds. □

In view of (5.2.16), the recursive equation for $Q_\pi$ can be rewritten as

$$Q_\pi(x_i, u_k) = R(x_i, u_k) + \gamma \sum_{j=1}^{n} a_{ij}^{u_k} V_\pi(x_j). \tag{5.2.17}$$

**Optimal Value Determination:**

For a policy $\pi \in \Pi_d$ or $\pi \in \Pi_p$, define the associated map $T_\pi : \mathbb{R}^n \to \mathbb{R}^n$ via

$$T_\pi \mathbf{v} = \mathbf{r}_\pi + \gamma A^\pi \mathbf{v}. \tag{5.2.18}$$

Then it follows from Theorem 5.2 that $T_\pi$ is monotone and is a contraction with respect to the $\ell_\infty$-norm, with contraction constant $\gamma$.

Now we introduce one of the key ideas in Markov Decision Processes. Define the **Bellman iteration map** $B : \mathbb{R}^n \to \mathbb{R}^n$ via

$$(B\mathbf{v})_i := \max_{u_k \in \mathcal{U}} \left[ R(x_i, u_k) + \gamma \sum_{j=1}^{n} a_{ij}^{u_k} v_j \right]. \tag{5.2.19}$$

**Theorem 5.7.** *The map $B$ is monotone and a contraction with respect to the $\ell_\infty$-norm.*

*Proof.* The theorem has two claims: The first claim is that the map $B$ is monotone, meaning that if $\mathbf{v}_1 \leq \mathbf{v}_2$ componentwise, then $B(\mathbf{v}_1) \leq B(\mathbf{v}_2)$ componentwise. The second claim is that $B$ is a contraction with respect to the $\ell_\infty$-norm. Note that, unlike the value iteration map $T_\pi$ defined in (5.2.18), the map $B$ is not affine.

Let us begin with the first claim. Suppose $\mathbf{v}_1 \leq \mathbf{v}_2$. Then

$$\begin{aligned}
(B(\mathbf{v}_1))_i &= \max_{u_k \in \mathcal{U}} \left[ R(x_i, u_k) + \gamma \sum_{j=1}^{n} a_{ij}^{u_k} v_{1j} \right] \\
&\leq \max_{u_k \in \mathcal{U}} \left[ R(x_i, u_k) + \gamma \sum_{j=1}^{n} a_{ij}^{u_k} v_{2j} \right] \\
&= (B(\mathbf{v}_2))_i.
\end{aligned}$$

Here we use the fact that $a_{ij}^{u_k} \geq 0$ for all $i, j$. This establishes that $B$ is monotone, which is the first claim.

The proof of the second claim is a bit more elaborate. We begin by establishing that

$$\left| \max_{u_k \in \mathcal{U}} g(x_i, u_k) - \max_{u_k \in \mathcal{U}} h(x_i, u_k) \right| \leq \max_{u_k \in \mathcal{U}} |g(x_i, u_k) - h(x_i, u_k)|, \ \forall x_i \in \mathcal{X}. \tag{5.2.20}$$

To prove (5.2.20), we begin with the obvious observation that, if $\alpha, \beta$ are real numbers, then

$$\alpha - \beta \leq |\alpha - \beta| \implies \alpha \leq |\alpha - \beta| + \beta.$$

Note that this inequality holds irrespective of the signs of $\alpha$ and $\beta$. Fix $x_i \in \mathcal{X}, u_k \in \mathcal{U}$ and apply the above inequality with $\alpha = g(x_i, u_k), \beta = h(x_i, u_k)$. This gives

$$g(x_i, u_k) \leq |g(x_i, u_k) - h(x_i, u_k)| + h(x_i, u_k).$$

Now take the maximum of both sides over $u_k \in \mathcal{U}$. This gives

$$\begin{aligned}
\max_{u_k \in \mathcal{U}} g(x_i, u_k) &\leq \max_{u_k \in \mathcal{U}} [|g(x_i, u_k) - h(x_i, u_k)| + h(x_i, u_k)] \\
&\leq \max_{u_k \in \mathcal{U}} |g(x_i, u_k) - h(x_i, u_k)| + \max_{u_k \in \mathcal{U}} h(x_i, u_k).
\end{aligned}$$

Rearranging gives

$$\max_{u_k \in \mathcal{U}} g(x_i, u_k) - \max_{u_k \in \mathcal{U}} h(x_i, u_k) \leq \max_{u_k \in \mathcal{U}} |g(x_i, u_k) - h(x_i, u_k)|.$$

By symmetry, we can interchange $g$ and $h$, which gives

$$\max_{u_k \in \mathcal{U}} h(x_i, u_k) - \max_{u_k \in \mathcal{U}} g(x_i, u_k) \leq \max_{u_k \in \mathcal{U}} |g(x_i, u_k) - h(x_i, u_k)|.$$

Combining these two inequalities gives (5.2.20).

Now we make use of (5.2.20) to show that $B$ is a contraction with respect to the $\ell_\infty$-norm. Let $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^n$ be arbitrary, and fix $x_i \in \mathcal{X}$. Then, by using the definition of $B$ and (5.2.20), we get

$$\begin{aligned}
|(B(\mathbf{v}_1))_i - (B(\mathbf{v}_2))_i| &= \left| \max_{u_k \in \mathcal{U}} \left[ R(x_i, u_k) + \gamma \sum_{j=1}^n a_{ij}^{u_k} v_{1j} \right] - \max_{u_k \in \mathcal{U}} \left[ R(x_i, u_k) + \gamma \sum_{j=1}^n a_{ij}^{u_k} v_{2j} \right] \right| \\
&\leq \max_{u_k \in \mathcal{U}} \left| R(x_i, u_k) + \gamma \sum_{j=1}^n a_{ij}^{u_k} v_{1j} - R(x_i, u_k) - \gamma \sum_{j=1}^n a_{ij}^{u_k} v_{2j} \right| \\
&= \max_{u_k \in \mathcal{U}} \gamma \left| \sum_{j=1}^n a_{ij}^{u_k} (v_{1j} - v_{2j}) \right| \leq \max_{u_k \in \mathcal{U}} \gamma \left| \sum_{j=1}^n a_{ij}^{u_k} |v_{1j} - v_{2j}| \right| \\
&\leq \gamma \|\mathbf{v}_1 - \mathbf{v}_2\|_\infty. \qquad (5.2.21)
\end{aligned}$$

Here we use the facts

$$|v_{1j} - v_{2j}| \leq \|\mathbf{v}_1 - \mathbf{v}_2\|_\infty \; \forall j, \sum_{j=1}^n a_{ij}^{u_k} = 1, \; \forall i, \; \forall u_k \in \mathcal{U}$$

Because the inequality (5.2.21) holds for *every* index $i$, it follows that

$$\|B(\mathbf{v}_1) - B(\mathbf{v}_2)\|_\infty \leq \gamma \|\mathbf{v}_1 - \mathbf{v}_2\|_\infty.$$

This shows that the map $B$ is a contraction with respect to the $\ell_\infty$-norm, which is the second claim. □

**Theorem 5.8.** *Define $\bar{\mathbf{v}} \in \mathbb{R}^n$ to be the unique fixed point of $B$, and define $\mathbf{v}^* \in \mathbb{R}^n$ to equal $[V^*(x_i), x_i \in \mathcal{X}]$, where $V^*(x_i)$ is defined in* (5.2.9)*. Then $\bar{\mathbf{v}} = \mathbf{v}^*$.*

*Proof.* By definition, for every $\pi \in \Pi_d$, we have that

$$\begin{aligned}
[T_\pi(\bar{\mathbf{v}})]_i &= R(x_i, \pi(x_i)) + \sum_{j=1}^n a_{ij}^{\pi(x_i)} \bar{V}_j \\
&\leq \max_{u_k \in \mathcal{U}} \left[ R(x_i, u_k) + \gamma \sum_{j=1}^n a_{ij}^{u_k} \bar{V}_j \right] = (B(\bar{\mathbf{v}}))_i = \bar{V}_i, \qquad (5.2.22)
\end{aligned}$$

because $\bar{\mathbf{v}}$ is a fixed point of the map $B$. If $\pi \in \Pi_p$, say

$$\pi(x_i) = [ \; \phi_{i1} \quad \cdots \quad \phi_{im} \; ] \in \mathbb{S}_m,$$

then

$$\begin{aligned}
[T_\pi(\mathbf{v})]_i &= \sum_{l=1}^l \phi_{il} \left[ R(x_i, u_l) + \sum_{j=1}^n a_{ij}^{u_l} \bar{V}_j \right] \\
&\leq \max_{u_k \in \mathcal{U}} \left[ R(x_i, u_k) + \sum_{j=1}^n a_{ij}^{u_k} \bar{V}_j \right] \\
&= (B(\bar{\mathbf{v}}))_i = \bar{V}_i. \qquad (5.2.23)
\end{aligned}$$

Because (5.2.22) and (5.2.23) hold for every index $i$, it follows that

$$T_\pi(\bar{\mathbf{v}}) \leq \bar{\mathbf{v}}.$$

Next, because $T_\pi$ is monotone as per Theorem 5.2, it follows that

$$T_\pi^2(\bar{\mathbf{v}}) = T_\pi(T_\pi(\bar{\mathbf{v}})) \leq T_\pi(\bar{\mathbf{v}}) \leq \bar{\mathbf{v}}.$$

The reasoning can be repeated to show that

$$T_\pi^l(\bar{\mathbf{v}}) \leq \bar{\mathbf{v}}, \ \forall l.$$

Now let $l \to \infty$. Then the left side approaches the fixed point of the map $T_\pi$, which is $\mathbf{v}_\pi$. Thus we conclude that, for all policies in $\Pi_d$ or $\Pi_p$, we have that

$$\mathbf{v}_\pi \leq \bar{\mathbf{v}}. \tag{5.2.24}$$

Therefore, for each $x_i \in \mathcal{X}$, we infer that

$$V^*(x_i) = \max_\pi V(x_i) \leq \bar{V}_i, \ \forall i, \text{ or } \mathbf{v}^* \leq \bar{\mathbf{v}}. \tag{5.2.25}$$

To show that $\bar{\mathbf{v}} \leq \mathbf{v}^*$, define a deterministic policy $\bar{\pi} \in \Pi_d$ by

$$\bar{\pi}(x_i) = \arg\max_{u_k \in \mathcal{U}} \left[ R(x_i, u_k) + \sum_{j=1}^n a_{ij}^{u_k} \bar{V}_j \right]. \tag{5.2.26}$$

In case of ties, choose any deterministic tie-breaking rule, e.g., choose the $u_k$ with the lowest index. Then, since the right side of (5.2.26) equals $(B(\bar{\mathbf{v}}))_i = \bar{V}_i$, we conclude that

$$\bar{V}_i = R(x_i, \bar{\pi}(x_i)) + \sum_{j=1}^n a_{ij}^{\bar{\pi}(x_i)} \bar{V}_j, \ \forall i. \tag{5.2.27}$$

Hence $T_{\bar{\pi}}(\bar{\mathbf{v}}) = \bar{\mathbf{v}}$. But since $T_{\bar{\pi}}$ is a contraction, it has a unique fixed point, which shows that $\bar{V}_i = V_{\bar{\pi}}(x_i)$ for all $i$. Therefore, for each index $i$, we have that

$$\bar{V}_i = V_{\bar{\pi}}(x_i) \leq V^*(x_i), \ \forall i, \text{ or } \bar{\mathbf{v}} \leq \mathbf{v}^*.$$

Taken together with (5.2.24), this shows that $\bar{\mathbf{v}} = \mathbf{v}^*$. □

By replacing $\bar{\mathbf{v}}$ in Theorem 5.8 by $\mathbf{v}^*$ (which equals $\bar{\mathbf{v}}$), we derive the following fundamental result for Markov Decision Processes.

**Theorem 5.9.** *Define the optimal value function $V^*(x_i)$ as in (5.2.9). Then*

1. *The optimal value function $V^* : \mathcal{X} \to \mathbb{R}$ is the unique solution of the following recursive relationship, known as the **Bellman Optimality Equation**:*

$$V^*(x_i) = \max_{u_k \in \mathcal{U}} \left[ R(x_i, u_k) + \gamma \sum_{j=1}^n a_{ij}^{u_k} V^*(x_j) \right]. \tag{5.2.28}$$

2. *There is at least one deterministic policy $\pi \in \Pi_d$ such that*

$$V_\pi(x_i) = V^*(x_i), \ \forall i \in \mathcal{X}. \tag{5.2.29}$$

*Specifically, the policy $\bar{\pi}$ defined by restating* (5.2.26) *with $\bar{V}_j$ replaced by $V_j^*$, namely*

$$\pi^*(x_i) = \arg\max_{u_k \in \mathcal{U}} \left[ R(x_i, u_k) + \sum_{j=1}^{n} a_{ij}^{u_k} V_j^* \right]. \tag{5.2.30}$$

*satisfies* (5.2.29) *and is thus an optimal policy.*

Note that Item 2 of the theorem states that enlarging the policy space to include probabilistic policies *does not* increase the maximum value. Also, there is *one common policy* that achieves the optimal value for *every* state $x_i$. Perhaps neither of these statements is obvious on the surface.

In analogy with the optimal value function, we can also define an optimal action-value function.

**Theorem 5.10.** *Define $Q^* : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ by*

$$Q^*(x_i, u_k) = R(x_i, u_k) + \gamma \sum_{j=1}^{n} a_{ij}^{u_k} V^*(x_j). \tag{5.2.31}$$

*Then $Q^*(\cdot, \cdot)$ satisfies the following relationships:*

$$Q^*(x_i, u_k) = R(x_i, u_k) + \gamma \sum_{j=1}^{n} a_{ij}^{u_k} \max_{w_l \in \mathcal{U}} Q^*(x_j, w_l). \tag{5.2.32}$$

$$V^*(x_i) = \max_{u_k \in \mathcal{U}} Q^*(x_i, u_k), \tag{5.2.33}$$

*Moreover, every policy $\pi \in \Pi_d$ such that*

$$\pi^*(x_i) = \arg\max_{u_k \in \mathcal{U}} Q^*(x_i, u_k) \tag{5.2.34}$$

*is optimal.*

*Proof.* Since $Q^*(\cdot, \cdot)$ is defined by (5.2.31), it follows that

$$\max_{u_k \in \mathcal{U}} Q^*(x_i, u_k) = \max_{u_k \in \mathcal{U}} \left[ R(x_i, u_k) + \gamma \sum_{j=1}^{n} a_{ij}^{u_k} V^*(x_j) \right] = V^*(x_i),$$

by (5.2.28). This establishes (5.2.33) and (5.2.34). Substituting from (5.2.33) into (5.2.31) gives (5.2.32). □

Now we define an iteration on action-functions that is analogous to (5.2.19) for value functions. As with the value function, the action-value function can either be viewed as a map $Q : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, or as a vector in $\mathbb{R}^{|\mathcal{X}| \cdot |\mathcal{U}|}$. Define $F : \mathbb{R}^{|\mathcal{X}| \times |\mathcal{U}|} \to \mathbb{R}^{|\mathcal{X}| \times |\mathcal{U}|}$ by

$$[F(Q)](x_i, u_k) := R(x_i, u_k) + \gamma \sum_{j=1}^{n} a_{ij}^{u_k} \max_{w_l \in \mathcal{U}} Q(x_j, w_l). \tag{5.2.35}$$

**Theorem 5.11.** *The map $F$ is monotone and is a contraction. Therefore for all $Q_0 : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, the sequence of iterations $\{F^t(Q_0)\}$ converges to $Q^*$ as $t \to \infty$.*

*Proof.* The proof is very similar to that of Theorem 5.9. Given a map $Q : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, define the associated map $\mathcal{M}(Q) : \mathcal{X} \to \mathbb{R}$ by

$$[\mathcal{M}(Q)](x_i) = \max_{u_k \in \mathcal{U}} Q(x_i, u_k),$$

and rewrite (5.2.35) as

$$[F(Q)](x_i, u_k) := R(x_i, u_k) + \gamma \sum_{j=1}^{n} a_{ij}^{u_k} [\mathcal{M}(Q)](x_j). \tag{5.2.36}$$

Also, if $Q, Q' : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, let $Q \leq Q'$ denote that $Q(x_i, u_k) \leq Q_i'(x_i, u_k)$ for all $x_i, u_k$. Then it is clear that if $Q \leq Q'$, then $\mathcal{M}(Q) \leq \mathcal{M}(Q')$. Because $a_{ij}^{u_k}$ is always nonnegative, it follows that the map $F$ is monotone.

Next, as in the proof of Theorem 5.7, for arbitrary maps $Q_1, Q_2 : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, we have

$$\begin{aligned} |[\mathcal{M}(Q_1)](x_i) - [\mathcal{M}(Q_2)](x_i)| &= \left| \max_{u_k \in \mathcal{U}} Q_1(x_i, u_k) - \max_{u_k \in \mathcal{U}} Q_2(x_i, u_k) \right| \\ &\leq \max_{u_k \in \mathcal{U}} |Q_1(x_i, u_k) - Q_2(x_i, u_k)|, \ \forall x_i \in \mathcal{X}. \end{aligned}$$

As a result

$$\|\mathcal{M}(Q_1) - \mathcal{M}(Q_2)\|_\infty \leq \|Q_1 - Q_2\|_\infty.$$

Substituting this into (5.2.36) gives

$$\|F(Q_1) - F(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty. \tag{5.2.37}$$

The desired conclusion now follows. □

If we were to rewrite (5.2.28) and (5.2.32) in terms of expected values, the advantages of the $Q$-function would become apparent. We can rewrite (5.2.28) as

$$V^*(X_t) = \max_{U_t \in \mathcal{U}} \{R(X_t, U_t) + \gamma E[V^*(X_{t+1}) | X_t]\}, \tag{5.2.38}$$

and (5.2.32) as

$$Q^*(X_t, U_t) = R(X_t, U_t) + \gamma E\left[\max_{U_{t+1} \in \mathcal{U}} Q^*(X_{t+1}, U_{t+1})\right]. \tag{5.2.39}$$

Thus in the Bellman formulation and iteration, the maximization occurs *outside* the expectation, whereas with the $Q$-formulation and $F$-iteration, the maximization occurs *inside* the expectation. As shown in later chapter, learning $Q^*$ is easier than learning $V^*$.

The idea of learning $Q^*$ instead of learning $V^*$ is introduced in [172].

**Optimal Policy Determination:**

Theorems 5.8 and 5.9 together show the following: Start with any initial guess $\mathbf{v}_0 \in \mathbb{R}^n$, and apply the Bellman iteration $B$ defined in (5.2.19). Then the sequence $\{\mathbf{v}_k\}$ with $\mathbf{v}_{k+1} = B\mathbf{v}_k$ converges monotonically to the optimal value $\mathbf{v}^*$. Once $\mathbf{v}^*$ is determined, then an optimal policy can be determined using (5.2.30). This approach to determining $\mathbf{v}^*$ is known as **value iteration**. While this is a useful result, a shortcoming is that the intermediate vectors $\mathbf{v}_k$ do not necessarily correspond to any policy. An easy remedy is to choose the starting point of the iterations $\mathbf{v}_0$ to be the value of some policy $\pi_0$. Then each successive iteration $\mathbf{v}_k$ also corresponds to a policy $\pi_k$. In this way, we generate a sequence of suboptimal policies $\pi_k$ with the property that the associated value vector $\mathbf{v}_k = \mathbf{v}_{\pi_k}$ converges to the optimal value. This approach is known as **policy iteration**. This is made precise as follows:

**Theorem 5.12.** *Choose an arbitrary policy $\pi_0 \in \Pi_d$, and compute the corresponding value $\mathbf{v}_{\pi_0}$. At the $k$-th iteration, choose an updated policy $\pi_{k+1} \in \Pi_d$ according to*

$$\pi_{k+1}(x_i) = \arg\max_{u_k \in \mathcal{U}} \left[ R(x_i, u_k) + \gamma \sum_{j=1}^{n} a_{ij}^{u_k} (\mathbf{v}_{\pi_k})_j \right]. \tag{5.2.40}$$

*Then*

1. $\mathbf{v}_{\pi_{k+1}} \geq \mathbf{v}_{\pi_k}$, *where the dominance is componentwise.*

2. $\{\mathbf{v}_{\pi_k}\} \uparrow \mathbf{v}^*$ *as* $k \to \infty$.

The proof is quite straightforward. The key step is to verify that if we define the updated policy $\pi_{k+1}$ according to (5.2.40), then the corresponding value $\mathbf{v}_{\pi_{k+1}}$ is just $B\mathbf{v}_{\pi_k}$; but this is obvious.

All of the material above provides a theoretical foundation for determining optimal values and optimal policies for MDPs. However, when the size of the state space is very large, as it often is, one is forced to resort to approximations to find "nearly optimal" values and/or policies. Approaches to do this are discussed in later chapters.

**Example 5.2.** Now we return to the game of Blackjack. A detailed discussion of the game is given in [145, Example 5.1]. To describe the original game briefly, it is played between a player and the "House." (It is possible to have more than one player playing against the House, but we don't study that problem in the interests of simplicity.) At each turn, the player and the House have the option of drawing a card ("hit") or not drawing ("stick"). Each card is counted as its face value, with picture cards counted as 10. An ace can count as either 1 of 11 at the player's preference. The objective of the player is to exceed the total of the House without going over 21.

From the description, it is obvious that if the player's current total is eleven or less, then the best strategy is to hit, because there is no chance of losing on the next draw. Hence the issue of what to do arises only when the player's total reaches 12 or higher. Indeed, if the target were to be changed to some number $N$, then it is clear that if the player's total is $N - 10$ or less, then the correct solution is to hit. It can also be assumed that the probability of any particular card being the next card drawn is the same, no matter what cards have been drawn until then (infinitely many card decks being used). In the original Blackjack game, only one card of the House is visible. In what follows, for the purposes of illustration, we eliminate all of these complications, and introduce a simplified game.

Suppose that, instead of drawing a card, the player rolls a fair four-sided die. Since there are only four possible outcomes, irrespective of what the target total might be, it is reasonable to suppose that the state $P_t$ of the player lies in the set $\{0, 1, 2, 3, W, L\}$, with 0 being the start state. It can be assumed that the *current* state is in $\{0, 1, 2, 3\}$, while $W$ and $L$ are terminal states. To simplify the problem further, suppose that the House adopts the strategy that it does not roll the die further once its state is in $\{1, 2, 3\}$ (i.e., it does not try for a win from any of these states). Therefore the state $H_t$ of the house lies in the set $\{1, 2, 3\}$. The overall state $(P_t, H_t)$ lies in the Cartesian product $\{0, 1, 2, 3, W, L\} \times \{1, 2, 3\}$. Out of these, there are twelve possible *current* states, namely $\{0, 1, 2, 3\} \times \{1, 2, 3\}$ where the first number is the state of the player and the second is the state of the House. If the player rolls the die, the possible *next* states are $\{1, 2, 3, W, L\} \times \{1, 2, 3\}$, or a total of fifteen states. In this game, as in the snakes and ladders game, the reward is random and is a function of the next state.

As a part of the problem statement, we need to specify the dynamics of the Markov process. For the House, it does not play, so its state transition matrix is the $3 \times 3$ identity matrix, which ensures that $H_{t+1} = H_t$. As for the player's state $P_t$, if the action is to "stick," then the state transition matrix $A^S$ is the $5 \times 5$ identity matrix. If the action is to "hit," then the state transition matrix $A^H$ is given by

$$A^H = \begin{array}{c|cccccc} & 0 & 1 & 2 & 3 & W & L \\ \hline 0 & 0 & 0.25 & 0.25 & 0.25 & 0.25 & 0 \\ 1 & 0 & 0 & 0.25 & 0.25 & 0.25 & 0.25 \\ 2 & 0 & 0 & 0 & 0.25 & 0.25 & 0.50 \\ 3 & 0 & 0 & 0 & 0 & 0.25 & 0.75 \\ W & 0 & 0 & 0 & 0 & 1 & 0 \\ L & 0 & 0 & 0 & 0 & 0 & 1 \end{array}.$$

To complete the problem formulation, we need to specify the reward. Unlike the state transition matrix above, which is based on nothing more than the assumption that all four outcomes of the die are equally

likely, the reward is to some extent arbitrary. Let us assign the following rewards:

| | |
|---|---|
| $P_t > H_t$ | 2 |
| $P_t = H_t$ | 1 |
| $P_t < H_t$ | 0 |
| $P_t = W$ | 5 |
| $P_t = L$ | $-5$ |

With this problem specification, we should strive to find an optimal policy. Note that the action space $\mathcal{U} = \{H, S\}$ (for "hit" or "stick") has cardinality two. Hence the number of policies is $2^{12} = 4,096$, which is already large enough that simply enumerating all possibilities is not practicable.[2] Hence some kind of policy iteration is the only way.

For evaluating *a specific* policy, it can be noted that the duration of the game cannot exceed four time steps. This is because the player's position has to advance by at least one at each time step. So discount factors very close to 1 do not make sense. The discount $\gamma$ should be chosen much smaller, say 0.5.

**Problem 5.1.** Suppose that a Markov decision problem has four states and two actions. Suppose further that the two row-stochastic matrices corresponding to the two actions are as follows:

$$A^{u_1} = \begin{bmatrix} 0.1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.4 & 0.1 & 0.2 \\ 0 & 0.4 & 0.4 & 0.2 \\ 0.4 & 0.2 & 0.2 & 0.2 \end{bmatrix}, A^{u_2} = \begin{bmatrix} 0.3 & 0.2 & 0 & 0.5 \\ 0.1 & 0.1 & 0.2 & 0.6 \\ 0.2 & 0.5 & 0.1 & 0.2 \\ 0 & 0.1 & 0.5 & 0.4 \end{bmatrix}.$$

Suppose further that the reward map $R : \mathcal{X} \times \mathcal{U}$ is as follows (note that we write e.g., $(3,1)$ instead of $(x_3, u_1)$ to save space):

$$R = \begin{array}{|c|c|c|c|c|c|c|c|} \hline (1,1) & (1,2) & (2,1) & (2,2) & (3,1) & (3,2) & (4,1) & (4,2) \\ \hline 2 & 5 & -1 & 4 & 3 & 3 & 6 & -1 \\ \hline \end{array}.$$

- Suppose we define a deterministic policy $\pi$ by

$$\pi = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

  In other words, $\pi(x_1) = u_2, \pi(x_2) = u_1, \pi(x_3) = u_1, \pi(x_4) = u_2$. Compute the corresponding state transition matrix $A^\pi$ and reward map $R_\pi$.

- Suppose we define a probabilistic policy $\pi$ by

$$\pi = \begin{bmatrix} 0.3 & 0.4 & 0.2 & 0.6 \\ 0.7 & 0.6 & 0.8 & 0.4 \end{bmatrix}.$$

  Compute the corresponding state transition matrix $A^\pi$ and reward map $R_\pi$.

- How many deterministic policies can there be for this problem?

- With a discount factor of $\gamma = 0.9$, compute the optimal value and optimal policy using Theorem 5.12.

**Problem 5.2.** Prove Theorem 5.11.

**Problem 5.3.** Using the policy iteration method of Theorem 5.12, compute the optimal value function and optimal policy for the Markov decision process of Problem 5.1.

**Problem 5.4.** Show that, if $\pi^*$ is determined from (5.2.30), then $V_{\pi^*} = V^*$ as defined in (5.2.28).

[2] For the full Blackjack game, the number of policies is $2^{200}$ as shown in [145, Example 5.1].

## Notes and References

The material in this chapter is quite standard. A very old reference is [64]. A widely used reference is [119]. Some applications of MDPs can be found in [27].

# Chapter 6

# Reinforcement Learning

The contents of Chapter 5, specifically Section 5.2 are based on the assumption that the parameters of the Markov Decision Process are all known. In other words, the $|\mathcal{U}|$ possible state transition matrices $A^{u_k}$, as well as the reward map $R : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ (or its random version), are all available to the agent to aid in the choice of an optimal policy. One can say that the distinction between MDP theory and reinforcement learning (RL) theory is that in the latter, it is *not* assumed that the parameters of the MDP are known. Thus, in RL, one attempts to learn these parameters based on observations. At this point, one can make a distinction between "direct" methods and "indirect" methods.[1] In the "indirect" approach, one observes the trajectory of the "unknown" MDP, constructs a maximum likelihood estimate of the dynamics using the methods of Section 2.2.3, and then substitutes these estimates of the dynamics into the Bellman Optimality Equation (5.2.28), or the $F$-iteration (5.2.35). The logic is that, after a sufficiently long period of observation, the estimated parameters would be sufficiently close to the true but unknown parameters; as a result, the solutions of the fixed-point problems with estimated parameters would also be sufficiently close to the true fixed point. In the "direct" approach, one directly starts estimating the solution of the fixed-point problems on the basis of the available data. One hopes to prove mathematically that the "directly estimated" solutions would converge to the correct solution. In short, there is no attempt to estimate the unknown dynamics of the MDP.

In the RL literature, a couple of phrases are widely used without always being defined precisely. The first phrase is "tabular methods." As we will see, the methods presented in this chapter attempt to form estimates of the value function, which is a vector in $\mathbb{R}^n$, or the action-value function, which is a matrix of dimensions $n \times m$, for a specific policy. Recall that in many if not most MDPs (or RL problems), the number of actions $m$ is quite small. However, the number of states $n$ is often huge. Hence, instead of attempting to determine the $n$-dimensional value vector, it is often more convenient to find a *lower-dimensional approximation* of this vector. The phrase "tabular methods" thus refers to the situation where one attempts find the full $n$-dimensional vector without any reduction in dimension. The alternate is "value" determination *with function approximation.*

## 6.1 Value Determination Using Temporal Differences

In this section we present the "temporal difference method" for determining the value of a discounted Markov Reward Process *when the state transition matrix is unknown.* This method was pioneered by Sutton in [144]. Subsequently it was improved in various others papers, which are mentioned at the appropriate place.

The temporal difference method comes in two flavors: In the first, the unknown Markov Process is assumed to have a known set of absorbing states. Thus the state space $\mathcal{X}$ is partitioned as the union of transient states and absorbing states, and each set is known. However, the dynamics of the Markov process

[1] This terminology is quite common in the adaptive control theory literature, and less common in the RL literature.

are not assumed to be known. This is the version of TD learning studied in [144, 66]. In the other version, the Markov process is assumed to be either irreducible or both irreducible and aperiodic. This is the version studied in [172] which introduced TD-learning, as well as in [158] which made some fundamental contributions to the subject. In all of these references, TD-learning is "tabular," in the sense that the objective is to learn the true value of the Markov Reward Process. When the dimension $n$ of the state space is very high, it is common practice to *approximate* the value function by a function of a $d$-dimensional vector, where $d \ll n$. This is the version studied in [161].

### 6.1.1 $TD(\lambda)$-Learning Without Function Approximation

As in Section 5.1, let $\mathbf{r}$ denote the reward vector (assumed to be deterministic), $\gamma$ denote the discount factor, and $A$ the state transition matrix of the Markov process. In this case, the value vector $\mathbf{v}$ is specified by Theorem 5.1, specifically (5.1.6), as the unique solution of the equation

$$\mathbf{v} = \mathbf{r} + \gamma A\mathbf{v}, \tag{6.1.1}$$

Suppose however that $A$ is not known to the learner. Instead, a single sample path $\{X_t\}$ of the Markov process is available. With this information, the hope is to construct a sequence $\{\hat{\mathbf{v}}_t\}$ that converges almost surely to the true solution $\mathbf{v}$ of (6.1.1).

As in Chapter 5, it is convenient to view the value both as a *vector* $\mathbf{v}$ of dimension $n$, as well as a *read-out map* $V : \mathcal{X} \to \mathbb{R}$. Once the elements of $\mathcal{X}$ are ordered in some fashion as $\{x_1, \cdots, x_n\}$, the two interpretations are interchangeable. Hence we will use whichever is more convenient to the situation at hand.

The key to the Temporal Difference approach is the following result.

**Lemma 6.1.** *Suppose $\{X_t\}$ is a sample path of a Markov process with an unknown state transition matrix $A$, and that $\mathbf{v}$ is a given (known and deterministic) vector in $\mathbb{R}^n$. Then, for each time $t$, the component $V(X_{t+1})$ is an unbiased estimator of the $X_t$-th component of $A\mathbf{v}$, with conditional variance no larger than $4\|\mathbf{v}\|_\infty^2$.*

**Remarks:**

1. In (6.1.1), the discount factor $\gamma$ is usually chosen by the learner and is thus known. If the reward is deterministic, then once the sample path traverses every state at least once, the reward vector $\mathbf{r}$ is also known. Thus the only unknown is the state transition matrix $A$.

2. An "indirect" approach to solving (6.1.1) might go like this. After observing a "sufficiently long" sample path $\{X_t\}$, the learner can construct a maximum-likelihood estimate of $A$ using the approach from Section 2.2.3; call it $\hat{A}$. Then (6.1.1) can be solved with $\hat{A}$ in place of the unknown $A$. The temporal difference approach is "direct" in that it generates a sequence of estimates $\{\mathbf{v}_t\}$ which converges to the true value vector as $t \to \infty$. There is no attempt to generate estimates of $A$.

3. The time index $t$ plays no role in the lemma or corollary. If $\{X_t\}$ is a sample path of a Markov process with the (unknown) state transition matrix $A$, and if $X_t = x_i, X_{t+1} = x_j$, then $v_j$ is an unbiased estimate of the product $[A\mathbf{v}]_i$.

4. A key attribute of this lemma is that the bound on the conditional variance of the estimate is *independent* of the unknown matrix $A$.

5. If $\mathbf{v}$ is a fixed vector, then this lemma is not all that useful. However, the way in which the lemma is used is that $\mathbf{v}_t$ *is itself a function of time*, and at each step $t$, the lemma can be used to generate an unbiased estimate of *one component* of $A\mathbf{v}_t$.

*Proof.* Let $\{\mathcal{F}_t\}$ be the filtration generated by $\{X_t\}$. Suppose $X_t = x_i$, and define $\xi_{t+1,i}$ as the error in estimating $[A\mathbf{v}]_i$ at time $t+1$; that is

$$\xi_{t+1,i} = V(X_{t+1}) - [A\mathbf{v}]_i, \ \forall i \in [n].$$

Then

$$\Pr\{X_{t+1} = x_j | X_y = x_i\} = a_{ij},\ \forall j \in [n],$$
$$E[V(X_{t+1}|X_t = x_i] = \sum_{j=1}^{n} a_{ij} v_j = [A\mathbf{v}]_i.$$

Since this is true for each state $x_i$, it follows that $V(X_{t+1})$ is an unbiased estimate of $[A\mathbf{v}]_i$, where $X_t = x_i$.

Next, conditioned on the event $X_t = x_i$, we have that, with probability $a_{ij}$,

$$\xi_{t+1,i} = v_j - \mathbf{a}^i\mathbf{v} = \mathbf{e}_j^\top \mathbf{v} - \mathbf{a}^i\mathbf{v},$$

where $\mathbf{e}_j$ is the $j$-elementary unit vector and $\mathbf{a}^i$ is the $i$-row of $A$. Note that $\|\mathbf{e}\|_1 = \|\mathbf{a}^i\|_1 = 1$. Hence, by Hölder's inequality

$$|\mathbf{e}_j^\top \mathbf{v}| \le \|\mathbf{v}\|_\infty, |\mathbf{a}^i\mathbf{v}| \le \|\mathbf{v}\|_\infty, |\mathbf{e}_j^\top \mathbf{v} - \mathbf{a}^i\mathbf{v}| \le 2\|\mathbf{v}\|_\infty,$$

$$\begin{aligned} CV_t(\boldsymbol{\xi}_{t+1}) &= \sum_{i=1}^{n} E[\xi_{t+1,i}^2 | X_t = x_i] \cdot \Pr\{X_t = x_i\} \\ &\le \sum_{i=1}^{n} 4\|\mathbf{v}\|_\infty^2 \cdot \Pr\{X_t = x_i\} = 4\|\mathbf{v}\|_\infty^2. \end{aligned}$$

□

**Corollary 6.1.** *Suppose $\{X_t\}$ is a sample path of a Markov process with an unknown state transition matrix $A$, and that $\mathbf{v}$ is a given (known and deterministic) vector in $\mathbb{R}^n$. Then, for each time $t$ and each time $\tau \ge 1$, the component $V(X_{t+\tau})$ is an unbiased estimator of the $X_t$-th component of $A^\tau\mathbf{v}$, with conditional variance no larger than $4\|\mathbf{v}\|_\infty^2$.*

*Proof.* The proof is the same as that of Lemma 6.1, after observing that

$$\Pr\{X_{t+\tau} = x_j | X_y = x_i\} = [A^\tau]_{ij},$$

□

Papers by Sutton [144], Tsitsiklis [158] and Jaakkola et al. [66].

## 6.2 $TD(\lambda)$-Learning With Function Approximation

Papers by Tsitsiklis and Van Roy [160, 161, 162, 163].

### 6.2.1 Discounted Reward Processes

### 6.2.2 Average Reward Processes

## 6.3 Simultaneous Value and Policy Approximation

### 6.3.1 Two Time-Scale Stochastic Approximation: Reprise

Papers by Borkar [21], and by Lakshminarayanan and Bhatnagar [88]

### 6.3.2 Average Reward Processes: Reprise

### 6.3.3 Policy Gradient Theorem

[137, 146, 98, 15, 149].

### 6.3.4 Actor-Critic Methods

[81, 82, 100, 80]
[89, 13, 14, 88, 150, 151, 152, 153]

## 6.4 $Q$-Learning

[147, 46, 132, 170, 171, 172]

## 6.5 Zap $Q$-Learning

### 6.5.1 Stochastic Newton-Raphson Approximation

[128, 116].

### 6.5.2 Zap $Q$-Learning

[42, 40, 41, 104]

## Notes and References

This part will be written after the contents of the chapter are fleshed out. It will discuss each of the references above, and their role in the overall RL scene.

# Chapter 7

# Background Material

The objective of this chapter is to collect in one place the background material required to understand the remainder of the notes. While much of the chapter consists of standard material that is found elsewhere, some parts of the chapter are not "background," because many if not most standard texts do not contain the material. An example is the material on stopping times and maximum likelihood estimation of Markov processes in Section 2.2. Where the material is genuinely background in nature and adequate references are found elsewhere, the treatment here is rigorous but cursory, and several references are given throughout, In such a case these notes are not, by themselves, sufficient to gain a mastery over these topics. A reader who is encountering these background topics for the first time is strongly encouraged to consult the various references in order to understand the topics more thoroughly.

## 7.1 Contraction Mapping Theorem

In this section we introduce a very powerful theorem known as the contraction mapping theorem (also known as the Banach fixed point theorem), which provides an iterative technique for solving noninear equations. It holds in extremely general settings. We present a version that is sufficient for the present purposes.

**Theorem 7.1.** *Suppose* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ *and that there exists a constant* $\rho < 1$ *such that*

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| \leq \rho \|\mathbf{x} - \mathbf{y}\|, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \tag{7.1.1}$$

*where* $\|\cdot\|$ *is some norm on* $\mathbb{R}^n$*. Then there is a unique* $\mathbf{x}^* \in \mathbb{R}^n$ *such that*

$$\mathbf{f}(\mathbf{x}^*) = \mathbf{x}^*. \tag{7.1.2}$$

*To find* $\mathbf{x}^*$*, choose an arbitrary* $\mathbf{x}_0 \in \mathbb{R}^n$ *and define* $\mathbf{x}_{l+1} = \mathbf{f}(\mathbf{x}_l)$*. Then* $\{\mathbf{x}_l\} \to x^*$ *as* $l \to \infty$*. Moreover, we have the explicit estimate*

$$\|\mathbf{x}^* - \mathbf{x}_l\| \leq \frac{\rho^l}{1-\rho} \|\mathbf{x}_1 - \mathbf{x}_0\|. \tag{7.1.3}$$

*Proof.* By definition, we have that

$$\|\mathbf{x}_{l+1} - \mathbf{x}_l\| \leq \rho \|\mathbf{x}_l - \mathbf{x}_{l-1}\| \leq \cdots \leq \rho^l \|\mathbf{x}_1 - \mathbf{x}_0\|. \tag{7.1.4}$$

Suppose $m > l$, say $m = l + r$ with $r > 0$. Then

$$\begin{aligned} \|\mathbf{x}_m - \mathbf{x}_l\| &= \|\mathbf{x}_{l+r} - \mathbf{x}_l\| \leq \sum_{i=0}^{r-1} \|\mathbf{x}_{l+i+1} - \mathbf{x}_{l+i}\| \\ &\leq \sum_{i=0}^{r-1} \rho^{l+i} \|\mathbf{x}_1 - \mathbf{x}_0\| \leq \sum_{i=0}^{\infty} \rho^{l+i} \|\mathbf{x}_1 - \mathbf{x}_0\| = \frac{\rho^l}{1-\rho} \|\mathbf{x}_1 - \mathbf{x}_0\|. \end{aligned} \tag{7.1.5}$$

Therefore $\|\mathbf{x}_m - \mathbf{x}_l\| \to 0$ as $\min\{m, l\} \to \infty$. Such a sequence is called a **Cauchy sequence**. In $\mathbb{R}^n$, a Cauchy sequence always converges to a limit. Denote this limit by $\mathbf{x}^*$. Then $\mathbf{x}^* = \lim_{l\to\infty} \mathbf{x}_l$. Now (7.1.1) makes it clear that the function $\mathbf{f}$ is continuous. Therefore

$$\mathbf{f}(\mathbf{x}^*) = \lim_{l\to\infty} \mathbf{f}(\mathbf{x}_l) = \lim_{l\to\infty} \mathbf{x}_{l+1} = \mathbf{x}^*.$$

Therefore $\mathbf{x}^*$ satisfies (7.1.2). To show that $\mathbf{x}^*$ is unique, suppose $\mathbf{f}(\mathbf{y}^*) = \mathbf{y}^*$. Then it follows from (7.1.1) that

$$\|\mathbf{x}^* - \mathbf{y}^*\| = \|\mathbf{f}(\mathbf{x}^*) - \mathbf{f}(\mathbf{y}^*)\| \leq \rho\|\mathbf{x}^* - \mathbf{y}^*\|.$$

Since $\rho < 1$, the only way in which the above inequality can hold is if $\|\mathbf{x}^* - \mathbf{y}^*\| = 0$, i.e., if $\mathbf{x}^* = \mathbf{y}^*$. Finally, let $m \to \infty$ in (7.1.5) so that $\mathbf{x}_m \to \mathbf{x}^*$ and $\|\mathbf{x}_m - \mathbf{x}_l\| \to \|\mathbf{x}^* - \mathbf{x}_l\|$. Then (7.1.5) becomes (7.1.3). □

The bound (7.1.3) is extremely useful. Note that $\|\mathbf{x}_1 - \mathbf{x}_0\| = \|\mathbf{f}(\mathbf{x}_0) - \mathbf{x}_0\|$. Therefore $\|\mathbf{x}_1 - \mathbf{x}_0\|$ is a measure of how far off the initial guess $\mathbf{x}_0$ is from being a fixed point of $\mathbf{f}$. Then (7.1.3) gives an *explicit* estimate of how far $\mathbf{x}_l$ is from $\mathbf{x}^*$, for each iteration $\mathbf{x}_l$. Note that the bound on the right side of (7.1.3) decreases by a factor of $\rho$ at each iteration.

## 7.2 Some Elements of Lyapunov Stability Theory

The study of nonlinear differential equations (ODEs) is a centuries-old and well-established subject. In the context of Reinforcement Learning, nonlinear ODEs arise when studying the convergence properties of the Stochastic Approximation (SA) algorithm in its various formulations; see Chapter 2. Therefore the present section presents a tiny slice of this very rich subject, just enough to serve our rather narrow objective. Authoritative treatments of nonlinear ODEs can be found [58, 166, 76]. Where required, more specific citations are given.

Throughout, we study ODEs of the form

$$\dot{\boldsymbol{\theta}}(t) = \mathbf{f}(\boldsymbol{\theta}(t)), \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0, \tag{7.2.1}$$

where $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$. In some situations, we study a *linear* ODE of the form

$$\dot{\boldsymbol{\theta}}(t) = A\boldsymbol{\theta}(t), \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0, \tag{7.2.2}$$

where $A \in \mathbb{R}^{d\times d}$. The linear ODE (7.2.2) always has a unique solution corresponding to each initial condition $\boldsymbol{\theta}_0$. It is given by

$$\boldsymbol{\theta}(t) = e^{At}\boldsymbol{\theta}_0, \text{ where } e^{At} = \sum_{k=0}^{\infty} \frac{A^k t^k}{k!}. \tag{7.2.3}$$

The summation in (7.2.3) is well-defined for all $t$, and is called the **matrix exponential**. However, unless the function $\mathbf{f}(\cdot)$ in (7.2.1) satisfies some assumptions, there is no guarantee that (7.2.1) has a unique solution. One of the most widely used sufficient conditions is presented next.

**Definition 7.1.** A function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$ is said to be **globally Lipschitz continuous** with constant $L$ if

$$\|\mathbf{f}(\boldsymbol{\theta}) - \mathbf{f}(\boldsymbol{\phi})\| \leq L\|\boldsymbol{\theta} - \boldsymbol{\phi}\|, \ \forall \boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^d. \tag{7.2.4}$$

Note that we have not specified which norm is used in (7.2.4). Since all norms on $\mathbb{R}^d$ are equivalent, the Lipschitz continuity (or the lack of it) of a function $\mathbf{f}(\cdot)$ does not depend on the norm used in (7.2.4). However, the *value of the Lipschitz constant* $L$ could depend on the norm used. In RL, the most commonly used norms are $\|\cdot\|_2$ and $\|\cdot\|_\infty$.

**Theorem 7.2.** *(See Theorems 2.4.25 and 2.4.57 of [166].) Suppose the function* $\mathbf{f}(\cdot)$ *satisfies* (7.2.4) *for some finite constant* $L$*. Then, for each* $\boldsymbol{\theta}_0 \in \mathbb{R}^d$*, there exists a unique solution* $\mathbf{s}(\cdot, \boldsymbol{\theta}_0)$ *that satisfies* (7.2.1)*. Further, given any* $\epsilon > 0$*, and* $T < \infty$*, there exists a* $\delta = \delta(\epsilon, T)$ *such that*

$$\|\mathbf{s}(t, \boldsymbol{\theta}_0) - \mathbf{s}(t, \boldsymbol{\phi}_0)\|_2 \leq \epsilon \; \forall t \in [0, T], \text{ if } \|\boldsymbol{\theta}_0 - \boldsymbol{\phi}_0\|_2 \leq \delta. \tag{7.2.5}$$

The next set of definitions are from the stability theory of ODEs. The interested reader may consult [166, Chapter 5] for more details.

**Definition 7.2.** A vector $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is said to be an **equilibrium** of (7.2.1) if $\mathbf{f}(\boldsymbol{\theta}^*) = \mathbf{0}$.

Note that, if $\boldsymbol{\theta}^*$ is an equilibrium of (7.2.1), then the solution trajectory $\mathbf{s}(t, \boldsymbol{\theta}^*) = \boldsymbol{\theta}^*$ for all $t \geq 0$. There are various types of equilibria. The next definition introduces some types that are most relevant to RL.

**Definition 7.3.** We present several notions of stability for an equilibrium.

1. An equilibrium $\boldsymbol{\theta}^*$ of (7.2.1) is said to be **stable** if, for every $\epsilon > 0$, there exists a $\delta > 0$ such that
$$\|\mathbf{s}(t, \boldsymbol{\theta}_0)\|_2 \leq \epsilon \; \forall t \geq 0, \text{ if } \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2 \leq \delta. \tag{7.2.6}$$
2. An equilibrium $\boldsymbol{\theta}^*$ of (7.2.1) is said to be **globally attractive** if
$$\mathbf{s}(t, \boldsymbol{\theta}_0) \to \boldsymbol{\theta}^* \text{ as } t \to \infty, \; \forall \boldsymbol{\theta}_0 \in \mathbb{R}^d. \tag{7.2.7}$$
3. An equilibrium $\boldsymbol{\theta}^*$ of (7.2.1) is said to be **globally asymptotically stable (GAS)** if it is both stable and globally attractive.
4. An equilibrium $\boldsymbol{\theta}^*$ of (7.2.1) is said to be **globally exponentially stable (GES)** if there exist constants $\mu < \infty$ and $\kappa > 0$ such that
$$\|\mathbf{s}(t, \boldsymbol{\theta}_0)\|_2 \leq \mu \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2 \exp(-\kappa t), \; \forall t \geq 0, \; \forall \boldsymbol{\theta}_0 \in \mathbb{R}^d. \tag{7.2.8}$$

**Remark:**

1. The above definition contains a bare minimum from a very rich set of concepts from nonlinear stability theory. Thorough treatments can be found in [58, 166, 76].
2. The concept of stability becomes clear if one were to compare (7.2.5) and (7.2.6). Equation (7.2.5) holds for every *finite* $T$, and it is possible that $\delta(\epsilon, T) \to 0$ as $T \to \infty$. In contrast, (7.2.6) implies (7.2.5) with the *uniform bound* $\delta(\epsilon, T) = \delta(\epsilon)$.
3. It is possible for an equilibrium to be globally attractive without being stable. An example, originally due to Vinogradov, is reproduced in [58, Section 40] and again in [166, Example 5.1.32].
4. If $\boldsymbol{\theta}^*$ is GAS or GES, then $\boldsymbol{\theta}^*$ is the only solution of $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{0}$.

Sufficient conditions for GAS and GES are given in terms of the existence of a "Lyapunov function" $V$ that satisfies appropriate conditions. Suppose $V : \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{C}^1$ (continuously differentiable). Then the function $\dot{V} : \mathbb{R}^d \to \mathbb{R}$ associated with $V$ and the ODE (7.2.1) is defined by

$$\dot{V}(\boldsymbol{\theta}) := \langle \nabla V(\boldsymbol{\theta}), \mathbf{f}(\boldsymbol{\theta}) \rangle. \tag{7.2.9}$$

Note that the same function $V$ associated with a different ODE could have a different $\dot{V}$. The rationale behind the definition of $\dot{V}$ is that, along the solution trajectories of (7.2.1), we have

$$\frac{d}{dt} V(\boldsymbol{\theta}(t)) = \langle \nabla V(\boldsymbol{\theta}), \dot{\boldsymbol{\theta}}(t) \rangle = \langle \nabla V(\boldsymbol{\theta}), \mathbf{f}(\boldsymbol{\theta}) \rangle = \dot{V}(\boldsymbol{\theta}(t)). \tag{7.2.10}$$

**Definition 7.4.** A function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is said to **belong to class** $\mathcal{K}$, denoted by $\phi \in \mathcal{K}$, if $\phi(0) = 0$, and $\phi(\cdot)$ is strictly increasing. A function $\phi \in \mathcal{K}$ is said to **belong to class** $\mathcal{K}R$, denoted by $\phi \in \mathcal{K}R$, if in addition, $\phi(r) \to \infty$ as $r \to \infty$.

**Definition 7.5.** Suppose $\boldsymbol{\theta}^*$ is the unique equilibrium of (7.2.1), and that $V : \mathbb{R}^d \to \mathbb{R}$ is continuous. Then

1. The function $V$ is said to be **positive definite** at $\boldsymbol{\theta}^*$ if there exists a function $\phi$ of Class $\mathcal{K}$ such that
$$V(\boldsymbol{\theta}) \geq \phi(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2. \tag{7.2.11}$$
$V$ is said to be **negative definite** if $-V$ is positive definite.

2. The function $V$ is said to be **positive definite and radially unbounded** if there exists a function $\phi$ of Class $\mathcal{K}R$ such that (7.2.11) holds.

Now we reproduce some classical results from [166]. Since we deal with time-invariant systems, the condition in [166] that $V$ should be decrescent is automatically satisfied.

**Theorem 7.3.** *(See [166, Theorem 5.3.56].) Suppose* $\mathbf{f}(\cdot)$ *in* (7.2.1) *is globally Lipschitz-continuous. and that* $\boldsymbol{\theta}^*$ *is the unique equilibrium of* (7.2.1). *Then* $\boldsymbol{\theta}^*$ *is globally asymptotically stable if there exists a* $\mathcal{C}^1$ *function* $V : \mathbb{R}^d \to \mathbb{R}_+$ *such that* $V$ *is positive definite and radially unbounded, and* $\dot{V}$ *is negative definite.*

**Theorem 7.4.** *(See [166, Theorem 5.3.62].) Suppose* $\mathbf{f}(\cdot)$ *in* (7.2.1) *is globally Lipschitz-continuous. and that* $\boldsymbol{\theta}^*$ *is the unique equilibrium of* (7.2.1). *Then the equilibrium* $\boldsymbol{\theta}^*$ *of* (7.2.1) *is globally exponentially stable if there exists a* $\mathcal{C}^1$ *function* $V : \mathbb{R}^d \to \mathbb{R}_+$ *and constants* $a, b, c > 0$ *such that*

$$a\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \leq V(\boldsymbol{\theta}) \leq b\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2, \ \forall \boldsymbol{\theta} \in \mathbb{R}^d, \tag{7.2.12}$$

$$\dot{V}(\boldsymbol{\theta}) \leq -c\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2, \ \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{7.2.13}$$

*Proof.* Let $\boldsymbol{\theta}(t)$ denote the solution of (7.2.1) with the initial condition $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$. (Thus $\boldsymbol{\theta}(t)$ is shorthand for $\mathbf{s}(t, \boldsymbol{\theta}_0)$.) Recall that

$$\dot{V}(\boldsymbol{\theta}(t)) = \frac{d}{dt} V(\boldsymbol{\theta}(t)).$$

Now (7.2.11) and (7.2.12) together imply that

$$\frac{d}{dt} V(\boldsymbol{\theta}(t)) \leq -c\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \leq -\frac{c}{a} V(\boldsymbol{\theta}(t)).$$

In other words,

$$V(\boldsymbol{\theta}(t)) \leq V(\boldsymbol{\theta}_0) \exp(-(c/a)t), \ \forall t \geq 0.$$

Now we again use (7.2.11) to turn this into a bound for $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*\|_2^2$.

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*\|_2^2 \leq \frac{V(\boldsymbol{\theta}(t))}{a} \leq \frac{1}{a} V(\boldsymbol{\theta}_0) \exp(-(c/a)t) \leq \frac{b}{a} \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2^2 \exp(-(c/a)t, \ \forall t \geq 0.$$

This bound can be readily recast in the form (7.2.8). □

Next, we present an improvement of Theorem 7.3. Unlike Theorems 7.3 and 7.4, which are classical and of long-standing, Theorem 7.5 below is of quite recent origin; thus, strictly speaking, it does belong under "Background." Nevertheless, it is included here to maintain the flow of ideas. This material is taken from [168].

In order to state this theorem, we introduce the concept of a function of Class $\mathcal{B}$. It is introduced in [52] but without giving it a name. The formal definition is given in [168, Definition 1].

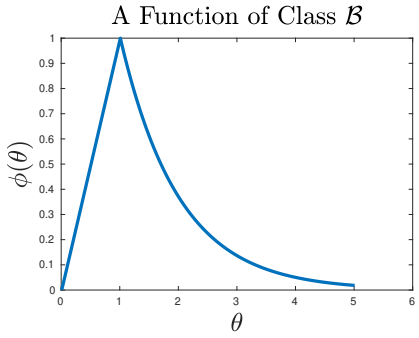



Figure 7.1: An illustration of a function in Class $\mathcal{B}$

**Definition 7.6.** A function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is said to **belong to Class** $\mathcal{B}$ if $\phi(0) = 0$, and in addition, for arbitrary real numbers $0 < \epsilon \leq M$, it is true that

$$\inf_{\epsilon \leq r \leq M} \phi(r) > 0.$$

Note that $\phi(\cdot)$ is *not* assumed to be monotonic, or even to be continuous. However, if $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous, then $\phi(\cdot)$ belongs to Class $\mathcal{B}$ if and only if (i) $\phi(0) = 0$, and (ii) $\phi(r) > 0$ for all $r > 0$. Such a function is called a "class P function" in [55]. Thus a Class $\mathcal{B}$ function is slightly more general than a function of Class $P$.

As example of a function of Class $\mathcal{B}$ is given next:

**Example 7.1.** Define a function $f : \mathbb{R}_+ \to \mathbb{R}_+$ by

$$\phi(\theta) = \begin{cases} \theta, & \text{if } \theta \in [0, 1], \\ e^{-(\theta-1)}, & \text{if } \theta > 1. \end{cases}$$

Then $\phi$ belongs to Class $\mathcal{B}$. A sketch of the function $\phi(\cdot)$ is given in Figure 7.1. Note that, if we were to change the definition to:

$$\phi(\theta) = \begin{cases} \theta, & \text{if } \theta \in [0, 1], \\ 2e^{-(\theta-1)}, & \text{if } \theta > 1, \end{cases}$$

then $\phi(\cdot)$ would be discontinuous at $\theta = 1$, but it would still belong to Class $\mathcal{B}$. Thus a function need not be continuous to belong to Class $\mathcal{B}$.

**Theorem 7.5.** *Suppose* $\mathbf{f}(\cdot)$ *in* (7.2.1) *is globally Lipschitz-continuous. and that* $\boldsymbol{\theta}^*$ *is the unique equilibrium of* (7.2.1). *Further, suppose that there exists a function* $V : \mathbb{R}^d \to \mathbb{R}_+$ *and functions* $\eta, \psi \in \mathcal{KR}, \phi \in \mathcal{B}$ *such that*

$$\eta(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2) \leq V(\boldsymbol{\theta}) \leq \psi(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2), \ \forall \boldsymbol{\theta} \in \mathbb{R}^d, \tag{7.2.14}$$

$$\dot{V}(\boldsymbol{\theta}) \leq -\phi(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2), \ \forall \boldsymbol{\theta} \in \mathbb{R}^d, \tag{7.2.15}$$

*Then* $\boldsymbol{\theta}^*$ *is a globally asymptotically stable equilibrium of the ODE* (7.2.1).

*Proof.* Let $\boldsymbol{\theta}(\cdot)$ denote a solution trajectory of the ODE (7.2.1). Then (7.2.15) implies that $V(\boldsymbol{\theta}(t))$ is a nonincreasing function of $t$, and therefore has a limit as $t \to \infty$. Since $\dot{V}(\boldsymbol{\theta}(t)) = (d/dt)(V(\boldsymbol{\theta}(t))$, this implies that $\dot{V}(\boldsymbol{\theta}(t)) \to 0$ as $t \to \infty$, as shown next. Suppose that $V(\infty) =: V_\infty > 0$. Then the right-side bound in (7.2.14) implies that

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \geq \psi^{-1}(V_\infty) > 0, \ \forall t.$$

In turn, this implies that

$$\dot{V}(\boldsymbol{\theta}(t)) \leq -\phi(\psi^{-1}(V_\infty)) < 0, \ \forall t.$$

This contradicts the fact that $\dot{V}(\boldsymbol{\theta}(t)) \to 0$ as $t \to \infty$. Therefore $V(\boldsymbol{\theta}(t)) \to 0$ as $t \to \infty$. Now the left inequality in (7.2.14) shows that $\|\boldsymbol{\theta}(t)\|_2 \to 0$ as $t \to \infty$. □

We conclude this section with a discussion of linear ODEs. Such ODEs arise naturally in studying the convergence of RL algorithms, specifically several variants of the Stochastic Approximation algorithm, as shown in Chapter 3 If the object of study is a *linear* ODE of the form (7.2.2), then the situation is simpler. The stability of linear ODEs is found in [166, Sec. 5.4]. The relevant results are summarized below.

**Theorem 7.6.** *(See [166, Theorem 5.4.29].) The equilibrium* $\mathbf{0}$ *of the linear ODE* (7.2.2) *is globally exponentially stable if and only all eigenvalues of* $A$ *have negative real parts.*

Note that a matrix whose eigenvalues all have negative real parts is called a **Hurwitz** matrix.

For linear sytems, a natural choice for a Lyapunov function is *quadratic*, in the form $V(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top P \boldsymbol{\theta}$. Note that it can be assumed without loss of generality that $P$ is symmetric. Then $V(\cdot)$ is a positive definite function if and only if $P$ is a positive definite matrix, that is, all of its eigenvalues are positive. Next, the function $\dot{V}(\cdot)$ is also quadratic, and equals $-\boldsymbol{\theta}^\top Q \boldsymbol{\theta}$, where $Q$ satisfies the **Lyapunov Matrix Equation**

$$A^\top P + PA = -Q. \tag{7.2.16}$$

Note that we have written $\dot{V}(\boldsymbol{\theta})$ as $-\boldsymbol{\theta}^\top Q \boldsymbol{\theta}$, with the hope that $Q$ would be positive definite, in which case $\dot{V}(\cdot)$ would be a negative definite function.

**Theorem 7.7.** *(See [166, Theorem 5.4.42].) Given a matrix* $A \in \mathbb{R}^{d \times d}$*, the following statements are equivalent:*

1. *$A$ is a Hurwitz matrix.*
2. *There exists a positive definite matrix $Q$ such that* (7.2.16) *has a unique solution for $P$, and that solution is positive definite.*
3. *For every positive definite matrix $Q$,* (7.2.16) *has a unique solution for $P$, and that solution is positive definite.*

## Notes and References

The material in this chapter is quite standard. Out of many possible sources, one can consult [166] because it contains proofs of both the contraction mapping theorem and all the elements of Lyapunov stability theory used here.

# Bibliography


[1] P. Abbeel, A. Coates, M. Quigley, and A. Y. Ng. An application of reinforcement learning to aerobatic helicopter flight. In *Advances in neural information processing systems*, pages 1–8, 2007.

[2] V. Apidopoulos, N. Ginatta, and S. Villa. Convergence rates for the heavy-ball continuous dynamics for non-convex optimization, under polyak–lojasiewicz condition. *Journal of Global Optimization*, 84:563–589, 2022.

[3] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus. Discrete-time controlled Markov processes with average cost criterion: A survey. *SIAM Journal of Control and Optimization*, 31(2):282–344, 1993.

[4] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1–2):165–214, 2023.

[5] J.-F. Aujol, C. Dossal, and A. Rondepierre. Optimal convergence rates for nesterov acceleration. *SIAM Journal on Optimization*, 29(4):3131–3153, 2019.

[6] J.-F. Aujol, C. Dossal, and A. Rondepierre. Convergence rates of the heavy-ball method under the lojasiewicz property. *Mathematical Programming*, 198(1):198–254, 2023.

[7] M. Benaim. *Dynamics of stochastic approximation algorithms*, volume 1709 of *Springer Lecture Notes in Mathematics*. Springer Verlag, 1999.

[8] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu. Advances in optimizing recurrent networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8624–8628, 2013.

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

[10] S. K. Berbarian. *Measure and Integration*. Chelsea, 1965.

[11] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, 1979.

[12] D. P. Bertsekas and J. N. Tsitsiklis. Global convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

[13] J. Bhandari, D. Russo, and R. Singal. A finite time analysis of temporal difference learning with linear function approximation. *Proceedings of Machine Learning Research*, 75(1–2):1691–1692, 2018.

[14] J. Bhandari, D. Russo, and R. Singal. A finite time analysis of temporal difference learning with linear function approximation. *Operations Research*, 69(3):950–973, May-June 2018.

[15] S. Bhatnagar, R. S. Sutton, M. Ghavamzadeh, and M. Lee. Natural actor–critic algorithms. *Automatica*, 45:2471–2482, 2009.

[16] P. Billingsley. *Probability and Measure (Third Edition)*. John Wiley, 1995.

[17] J. R. Blum. Multivariable stochastic approximation methods. *Annals of Mathematical Statistics*, 25(4):737–744, 1954.

[18] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter. Characterizations of lojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

[19] V. Borkar, S. Chen, A. Devraj, I. Kontoyiannis, and S. Meyn. The ode method for asymptotic statistics in stochastic approximation and reinforcement learning. *The Annals of Applied Probability*, 35(2):936–982, 2025.

[20] V. S. Borkar. *Probability Theory: An Advanced Course*. Springer-Verlag, 1995.

[21] V. S. Borkar. Stochastic approximation in two time scales. *Systems & Control Letters*, 29(5):291–294, February 1997.

[22] V. S. Borkar. Asynchronous stochastic approximations. *SIAM Journal on Control and Optimization*, 36(3):840–851, 1998.

[23] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[24] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint (Second Edition)*. Cambridge University Press, 2022.

[25] V. S. Borkar and S. P. Meyn. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38:447–469, 2000.

[26] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[27] R. J. Boucherie and N. M. van Dijk, editors. *Markov Decision Processes in Practice*. Springer Nature, 2017.

[28] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[29] L. Breiman. *Probability*. SIAM: Society for Industrial and Applied Mathematics, 1992.

[30] H. Cai, Y. Lou, D. Mckenzie, and W. Yin. A zeroth-order block coordinate descent algorithm for huge-scale black-box optimization. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1193–1203. PMLR, 18–24 Jul 2021.

[31] H. F. Chen, T. E. Duncan, and B. Pasik-Duncan. A kiefer–wolfowitz algorithm with randomized differences. *IEEE Transactions on Automatic Control*, 44(3):442–453, March 1999.

[32] Z. Chen, S. T. Maguluri, S. Shakkottai, and K. Shanmugam. A lyapunov theory for finite-sample guarantees of asynchronous q-learning and td-learning variants. arxiv:2102.01567v3, February 2021.

[33] Z. Chen, S. T. Maguluri, and M. Zubeldia. Concentration of contractive stochastic approximation: Additive and multiplicative noise. arxiv:2303.15740, March 2023.

[34] M. Corless and L. Glielmo. New converse Lyapunov theorems and related results on exponential stability. *Mathematics of Control, Signals, and Systems*, 11:79–100, 1998.

[35] B. D. Craven. Invex functions and constrained local minima. *Bulletin of the Australian Mathematical Society*, 24:357–366, 1981.

[36] B. D. Craven and B. M. Glover. Invex functions and duality. *Journal of the Australian Mathematical Society (Series A)*, 39:1–20, 1985.

[37] Deep Mind. Alphazero: Shedding new light on chess, shogi, and go. https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go.

[38] D. P. Derevitskii and A. L. Fradkov. Two models for analyzing the dynamics of adaptation algorithms. *Automation and Remote Control*, 35:59–67, 1974.

[39] C. Derman and J. Sacks. On Dvoretzky's stochastic approximation theorem. *Annals of Mathematical Statistics*, 30(2):601–606, 1959.

[40] A. M. Devraj, A. Busić, and S. Meyn. Zap Q-learning– a user's guide. In *Proceedings of the 2019 Fifth Indian Control Conference (ICC)*, pages 10–15, 2019.

[41] A. M. Devraj, A. Bušić, and S. P. Meyn. Optimal matrix momentum stochastic approximation and applications to q-learning. arxiv:1809.06277, September 2018.

[42] A. M. Devraj and S. Meyn. Zap Q-learning. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pages 2235–2244, 2017.

[43] R. M. Dudley. *Real Analysis and Probability.* Cambridge University Press, 2002.

[44] R. Durrett. *Probability: Theory and Examples (5th Edition).* Cambridge University Press, 2019.

[45] A. Dvoretzky. On stochastic approximation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statististics and Probability*, volume 1, pages 39–56. University of California Press, 1956.

[46] E. Even-Dar and Y. Mansour. Learning rates for q-learning. *Journal of machine learning Research*, 5:1–25, December 2003.

[47] R. Fletcher. *Practical Methods of Optimization, Volume 1: Unconstrained Optimization.* John Wiley, 1980.

[48] B. Franci and S. Grammatico. Convergence of sequences: A survey. *Annual Reviews in Control*, 53:1–26, 2022.

[49] S. Gadat, F. Panloup, and S. Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12:461–529, 2018.

[50] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson. Global convergence of the heavy-ball method for convex optimization. In *Proceedings of the 2015 European Control Conference*, pages 311–316, 2015.

[51] I. Ghosh. Introduction to mathematical optimization with python. https://indrag49.github.io/Numerical-Optimization/, 2021.

[52] E. G. Gladyshev. On stochastic approximation. *Theory of Probability and Its Applications*, X(2):275–278, 1965.

[53] O. N. Granichin. Randomized algorithms for stochastic approximation under arbitrary disturbances. *Automation and Remote Control*, 63(2):209–219, 2002.

[54] M. Gromniak and J. Stenzel. Deep reinforcement learning for mobile robot navigation. In *2019 4th Asia-Pacific Conference on Intelligent Robot Systems*, pages 68–73, 2019.

[55] L. Grüne and C. M. Kellett. ISS-Lyapunov Functions for Discontinuous Discrete-Time Systems. *IEEE Transactions on Automatic Control*, 59(11):3098–3103, November 2014.

[56] H. Gupta, R. Srikant, and L. Ying. Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. *Advances in Neural Information Processing Systems*, pages 4706–4715, 2019.

[57] H. Gupta, R. Srikant, and L. Ying. Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. arxiv:1907.0290v1, July 2019.

[58] W. Hahn. *Stability of Motion*. Springer-Verlag, 1967.

[59] M. A. Hanson. On sufficiency of kuhn-tucker conditions. *Journal of Mathematical Analysis and Applications*, 80:545–550, 1981.

[60] W. B. Haskell, R. Jain, and D. Kalathil. Empirical dynamic programming. *Mathematics of Operations Research*, 41(2):402–429, 2016.

[61] W. B. Haskell, R. Jain, H. Sharma, and P. Yu. A universal empirical dynamic programming algorithm for continuous state MDPs. *IEEE Transactions on Automatic Control*, 65(1):115–129, January 2020.

[62] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer-Verlag, Berlin and Heidelberg, 2001.

[63] R. A. Horn and C. R. Johnson. *Matrix Analysis (Second Edition)*. Cambridge University Press, 2013.

[64] R. Howard. *Dynamic Programming and Markov Decision Processes*. MIT Press, 1960.

[65] ImageNet. http://image-net.org/about-stats, 2010.

[66] T. Jaakkola, M. I. Jordan, and S. P. Singh. Convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6):1185–1201, November 1994.

[67] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is q-learning provably efficient? In *Proceedings of Advances in Neural Information Processing Systems 2018*, 2018.

[68] A. Juditsky, A. Nazin, A. Tsybakov, and N. Vayatis. Recursive aggregation of estimators by the mirror descent algorithm with averaging. *Problems of Information Transmission*, 41:368–384, 2005.

[69] R. L. Karandikar, B. V. Rao, and M. Vidyasagar. Revisiting stochastic approximation and stochastic gradient descent. https://arxiv.org/pdf/2505.11343, May 2025.

[70] R. L. Karandikar and M. Vidyasagar. Convergence rates for stochastic approximation: Biased noise with unbounded variance, and applications. https://arxiv.org/pdf/2312.02828v3.pdf, May 2024.

[71] R. L. Karandikar and M. Vidyasagar. Convergence rates for stochastic approximation: Biased noise with unbounded variance, and applications. *Journal of Optimization Theory and Applications*, 203:2412–2450, October 2024.

[72] R. L. Karandikar and M. Vidyasagar. Recent advances in stochastic approximation with applications to nonconvex optimization and fixed point problems. *Communications in Optimization Theory*, (to appear), 2024.

[73] R. L. Karandikar and M. Vidyasagar. Recent advances in stochastic approximation with applications to nonconvex optimization and fixed point problems. https://arxiv.org/pdf/2109.03445v6.pdf, February 2024.

[74] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak- lojasiewicz condition. *Lecture Notes in Computer Science*, 9851:795–811, 2016.

[75] A. Khaled and P. Richtárik. Better Theory for SGD in the Nonconvex World. arXiv:2002.03329, February 2020.

[76] H. K. Khalil. *Nonlinear Systems (Third Edition)*. Prentice Hall, 2002.

[77] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[78] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. A. Sallab, S. Yogamani, and P. Pérez. Deep reinforcement learning for autonomous driving: A survey. https://arxiv.org/pdf/2002.00444.pdf, February 2020.

[79] A. N. Kolmogorov. *Foundations of Probability (Second English Edition)*. Chelsea, 1950. (English translation by Kai Lai Chung).

[80] V. Konda and J. Tsitsiklis. On actor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003.

[81] V. R. Konda and V. S. Borkar. Actor-critic learning algorithms for Markov decision processes. *SIAM Journal on Control and Optimization*, 38(1):94–123, 1999.

[82] V. R. Konda and J. N. Tsitsiklis. Actor-critic algorithms. In *Neural Information Processing Systems (NIPS1999)*, pages 1008–1014, 1999.

[83] K. Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de L'Institut Fourier*, 48:769–783, 1998.

[84] H. J. Kushner. General convergence results for stochastic approximations via weak convergence theory. *Journal of Mathematical Analysis and Applications*, 61(2):490–503, 1977.

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

[86] H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, 1997.

[87] T. L. Lai. Stochastic approximation (invited paper). *The Annals of Statistics*, 31(2):391–406, 2003.

[88] C. Lakshminarayanan and S. Bhatnagar. A stability criterion for two timescale stochastic approximation schemes. *Automatica*, 79:108–114, 2017.

[89] C. Lakshminarayanan and C. Szepesvári. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *Proceedings of the AISTATS*, 2018.

[90] C. K. Lauand and S. Meyn. Markovian foundations for quasi-stochastic approximation in two timescales: Extended version. arxiv:2409.07842, September 2024.

[91] S. Li, Y. Xia, and Z. Xu. Simultaneous perturbation stochastic approximation: towards one-measurement per iteration. arxiv:2203.03075, March 2022.

[92] J. Liu, D. Xu, Y. Lu, J. Kong, and D. P. Mandic. Last-iterate convergence analysis of stochastic momentum methods for neural networks. *Neurocomputing*, 527:27–35, 2023.

[93] J. Liu and Y. Yuan. On almost sure convergence rates of stochastic gradient methods. In P.-L. Loh and M. Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2963–2983. PMLR, 02–05 Jul 2022.

[94] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22(6):551–575, 1977.

[95] L. Ljung. Strong convergence of a stochastic approximation algorithm. *Annals of Statistics*, 6:680–696, 1978.

[96] S. Lojasiewicz. *Une propriété topologique des sous-ensembles analytiques réels*, pages 87–89. Editions du centre National de la Recherche Scientifique, 1963.

[97] Z. Lu and L. Xiao. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152(1-2):615–642, Aug. 2014.

[98] H. R. Maei, C. Szepesvari, S. Bhatnagar, D. Precup, D. Silver, and R. S. Sutton. Convergent temporal-difference learning with arbitrary smooth function approximation. In *Neural Information Processing Systems (NIPS2009)*, pages 1–9, 2009.

[99] A. Mahajan, S.-I. Niculescu, and M. Vidyasagar. A vector almost-supermartingale convergence theorem and its applications. In *Proceedings of the 2024 Conference on Decision and Control*, pages 3877–3882, December 2024.

[100] P. Marbach and J. N. Tsitsiklis. Simulation-based optimization of markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, February 2001.

[101] S. M. Meerkov. On simplifying the description of slow Markov walks, I. *Automation and Remote Control*, 33(3):404–414, 1972.

[102] S. M. Meerkov. On simplifying the description of slow Markov walks, II. *Automation and Remote Control*, 33(5):761–764, 1972.

[103] M. Métivier and P. Priouret. Applications of kushner and clark lemma to general classes of stochastic algorithms. *IEEE Transactions on Information Theory*, IT-30(2):140–151, March 1984.

[104] S. Meyn. *Control Systems and Reinforcement Learning*. Cambridge University Press, 2022.

[105] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. In *NIPS Deep Learning Workshop*, 2013.

[106] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[107] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$n (in russian). *Soviet Mathematics Doklady*, 269:543–547, 1983.

[108] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Scientific+Business Media, 2004.

[109] Y. Nesterov. *Lectures on Convex Optimization (Second Edition)*. Springer Nature, 2018.

[110] Y. Nesterov and V. Spokoiny. Random Gradient-Free Minimization of Convex Functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[111] J. R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[112] S. Pachal, S. Bhatnagar, and L. A. Prashanth. Generalized simultaneous perturbation-based gradient search with reduced estimator bias. arxiv:2212.10477v1, December 2022.

[113] B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[114] B. T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[115] B. T. Polyak. *Introduction to optimization*. Optimization Software, Inc, 1987.

[116] B. T. Polyak. New method of stochastic approximation type (in russian). *Automation and Remote Control*, 51(7):937–946, July 1990.

[117] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal of Control and Optimization*, 30(4):838–855, July 1992.

[118] B. T. Polyak and Y. Z. Tsypkin. Pseudogradient adaptation and training algorithms. *Automation and Remote Control*, 34(3):377–397, 1973.

[119] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley, 2005.

[120] G. Qu and A. Wierman. Finite-time analysis of asynchronous stochastic approximation and q-learning. *Proceedings of Machine Learning Research*, 125:1–21, 2020.

[121] T. U. K. Reddy and M. Vidyasagar. Convergence of momentum-based heavy ball method with approximate gradients and/or block updating. https://arxiv.org/pdf/2303.16241v4.pdf, April 2025.

[122] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, Dec. 2012.

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

[124] H. Robbins and D. Siegmund. *A convergence theorem for non negative almost supermartingales and some applications*, pages 233–257. Elsevier, 1971.

[125] R. T. Rockafellar. *Convexity*. Princeton University Press, Princeton, NJ, 1970.

[126] R. T. Rockafellar and R. J.-B. Wets. On the interchange of subdifferentiation and conditional expectation for convex functionals. *Stochastics*, 7:173–182, 1982.

[127] H. L. Royden and P. M. Fitzpatrick. *Real Analysis (Fourth Edition*. Pearson Education, 2010.

[128] D. Ruppert. A newton-raphson version of the multivariate robbins-monro procedure. *Annals of Statistics*, 13(1):236–245, 1985.

[129] P. Sadegh and J. C. Spall. Optimal random perturbations for stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 43(10):1480–1484, 1998.

[130] O. Sebbouh, R. M. Gower, and A. Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Proceedings of Thirty Fourth Conference on Learning Theory, PMLR*, volume 134, pages 3935–3971, 2021.

[131] E. Seneta. *Non-negative Matrices and Markov Processes (Second Edition)*. Springer-Verlag, 1981.

[132] D. Shah and Q. Xie. Q-learning with nearest neighbors. In *Advances in Neural Information Processing Systems*, pages 3111–3121, 2018.

[133] C. E. Shannon. Programming a computer for playing chess. *Philosophical Magazine, Ser.7*, 41(314), March 1950.

[134] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

[135] L. Shen, C. Chen, F. Zou, Z. Jie, J. Sun, and W. Liu. A unified analysis of adagrad with weighted aggregation and momentum acceleration. arxiv:1808.03408, August 2018.

[136] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, December 2018.

[137] S. P. Singh and R. S. Sutton. Reinforcement learning with replacing eligibility traces. *Machine Learning*, 22(1–3):123–158, 1996.

[138] J. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

[139] M. W. Spong, S. R. Hutchinson, and M. Vidyasagar. *Robot Modeling and Control (Second Edition)*. John Wiley, 2020.

[140] R. Srikant. Rates of convergence in the central limit theorem for markov chains, with an application to td learning. *Mathematics of Operations Research*, (to appear), 2025.

[141] R. Srikant and L. Ying. Finite-time error bounds for linear stochastic approximation and td learning. arxiv:1902.00923v3, March 2019.

[142] W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17:1–43, 2016.

[143] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th international conference on machine learning (ICML-13)*, pages 1139–1147, 2013.

[144] R. S. Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

[145] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction (Second Edition)*. MIT Press, 2018.

[146] R. S. Sutton, D. McAllester, S. Singh, and Y.Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12 (Proceedings of the 1999 conference)*, pages 1057–1063. MIT Press, 2000.

[147] C. Szepesvári. The asymptotic convergence-rate of q-learning. In *Advances in Neural Information Processing Systems*, 1998.

[148] C. Szepesvári. *Algorithms for Reinforcement Learning*. Morgan and Claypool, 2010.

[149] V. Tadic. On the convergence of temporal-difference learning with linear function approximation. *Machine Learning*, 42:241–267, 2001.

[150] V. B. Tadić. Convergence of stochastic approximation under general noise and stability conditions. In *Proceedings of the IEEE Conference on Decision and Control*, volume 3, pages 2281–2286, 1997.

[151] V. B. Tadić. Asymptotic analysis of stochastic approximation algorithms under violated kushner-clark conditions with applications. In *Proceedings of the IEEE Conference on Decision and Control*, volume 3, pages 2875–2880, 2000.

[152] V. B. Tadić. Almost sure convergence of two time-scale stochastic approximation algorithms. In *Proceedings of the American Control Conference*, volume 4, pages 3802–3807, 2004.

[153] V. B. Tadić. On the robustness of two time-scale stochastic approximation algorithms. In *Proceedings of the IEEE Conference on Decision and Control*, volume 5, pages 5334–5339, 2004.

[154] G. Tesauro. Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6(2):215–219, 1994.

[155] G. Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.

[156] G. Tesauro. Programming backgammon using self-teaching neural nets. *Artificial Intelligence*, 134(1–2):181–199, 2002.

[157] G. Tesauro. Programming backgammon using self-teaching neural nets. *Artificial Intelligence*, 134(1-2):181–199, 2002.

[158] J. N. Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine Learning*, 16:185–202, 1994.

[159] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, September 1986.

[160] J. N. Tsitsiklis and B. V. Roy. Feature-based methods for large-scale dynamic programming. *Machine Learning*, 22:59–94, 1996.

[161] J. N. Tsitsiklis and B. V. Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, May 1997.

[162] J. N. Tsitsiklis and B. V. Roy. Average cost temporal-difference learning. *Automatica*, 35:1799–1808, 1999.

[163] J. N. Tsitsiklis and B. V. Roy. On average versus discounted reward temporal-difference learning. *Machine Learning*, 49:179–191, 2002.

[164] H. van Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the 2016 Association for the Advancement of Artificial Intelligence*, 2016.

[165] S. Vaswani, F. Bach, and M. Schmidt. Fast and Faster Convergence of SGD for Over-Parameterized Models (and an Accelerated Perceptron). In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1–10, 2019.

[166] M. Vidyasagar. *Nonlinear Systems Analysis (SIAM Classics Series)*. Society for Industrial and Applied Mathematics (SIAM), 2002.

[167] M. Vidyasagar. *Hidden Markov Processes: Theory and Applications to Biology*. Princeton University Press, 2014.

[168] M. Vidyasagar. Convergence of stochastic approximation via martingale and converse Lyapunov methods. *Mathematics of Controls Signals and Systems*, 35:351–374, 2023.

[169] M. Vidyasagar. Convergence of momentum-based optimization algorithms with time-varying parameters. *Pure and Applied Functional Analysis*, pages 1–32, 2025 (to appear).

[170] M. J. Wainwright. Stochastic approximation with cone-contractive operators: Sharp $\ell_\infty$-bounds for q-learning. arXiv:1905.06265, 2019.

[171] M. J. Wainwright. Variance-reduced q-learning is minimax optimal. arXiv:1906.04697, June 2019.

[172] C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

[173] D. Williams. *Probability with Martingagles*. Cambridge University Press, 1991.

[174] J. Wolfowitz. On the stochastic approximation method of Robbins and Monro. *Annals of Mathematical Statistics*, 23(3):457–461, 1952.

[175] S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

[176] Y. Xu and W. Yin. Block stochastic gradient iteration for convex and nonconvex optimization. *SIAM Journal on Optimization*, 25(3):1686–1716, 2015.

[177] Y. Yan, T. Yang, Z. Li, Q. Lin, and Y. Yang. A unified analysis of stochastic momentum methods for deep learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 2955–2961. AAAI Press, 2018.