Gradient-based algorithms for zeroth-order optimization

Suggested Citation: Prashanth L. A. and Shalabh Bhatnagar (2024), "Gradient-based algorithms for zeroth-order optimization", : Vol. xx, No. xx, pp 1–16. DOI: 10.1561/XXXXXXXXX.

Prashanth L. A.
Department of Computer Science and Engineering,
Indian Institute of Technology Madras.
prashla@cse.iitm.ac.in

Shalabh Bhatnagar
Department of Computer Science and Automation,
Indian Institute of Science Bangalore.
shalabh@iisc.ac.in

now
the essence of knowledge
Boston — Delft

Contents

Gradient-based algorithms for zeroth-order optimization

Prashanth L. A.[1] and Shalabh Bhatnagar[2]

[1] *Indian Institute of Technology Madras; prashla@cse.iitm.ac.in*
[2] *Indian Institute of Science Bangalore; shalabh@iisc.ac.in*

ABSTRACT

This book deals with methods for stochastic or data-driven optimization. The overall goal in these methods is to minimize a certain parameter-dependent objective function that for any parameter value is an expectation of a noisy sample performance objective whose measurement can be made from a real system or a simulation device depending on the setting used. We present a class of model-free approaches based on stochastic approximation which involve random search procedures to efficiently make use of the noisy observations. The idea here is to simply estimate the minima of the expected objective via an incremental-update or recursive procedure and not to estimate the whole objective function itself. We provide both asymptotic as well as finite sample analyses of the procedures used for convex as well as non-convex objectives.

We present algorithms that either estimate the gradient in gradient-based schemes or estimate both the gradient and the Hessian in Newton-type procedures using random direction approaches involving noisy function measurements. Hence the class of approaches that we study fall under the broad category of zeroth order optimization methods. We provide both asymptotic convergence guarantees in the general setup as well as asymptotic normality results for various

Prashanth L. A. and Shalabh Bhatnagar (2024), "Gradient-based algorithms for zeroth-order optimization", : Vol. xx, No. xx, pp 1–16. DOI: 10.1561/XXXXXXXXX.

algorithms. We also provide an introduction to stochastic recursive inclusions as well as their asymptotic convergence analysis. This is necessitated because many of these settings involve set-valued maps for any given parameter. We also present a couple of interesting applications of these methods in the domain of reinforcement learning. Five appendices at the end quickly summarize the basic material for this text. A large portion of this work is driven by our own contributions to this area.

Preface

This monograph is written with the idea of providing a self-contained introduction to stochastic gradient algorithms for solving a zeroth-order optimization problem. Towards this goal, we have included a detailed introduction to stochastic approximation which provides the basic framework for the analysis of incremental update algorithms with noise, that indeed form the backbone of algorithms in areas such as reinforcement learning, and stochastic optimization with unbiased as well as biased gradient information. We provide a detailed coverage of zeroth-order gradient estimation procedures, including classic approaches such as simultaneous perturbation stochastic approximation (SPSA), smoothed functional (SF), as well as more recent approaches dealt with in the literature. The convergence analysis that we provide includes both asymptotic guarantees via the ordinary differential equation (ODE) and differential inclusion (DI) approaches, as well as non-asymptotic bounds. The convergence analyses should be of interest to students as well as researchers working in the broad area of stochastic optimization and machine learning.

Figure 1 provides a visual depiction of the dependencies between the individual chapters and appendices in the book.

We now provide a few guidelines on how to read this book.

- If you are an expert researcher well-versed in the field of stochastic approximation, then we suggest reading Chapters 3 to 5. These

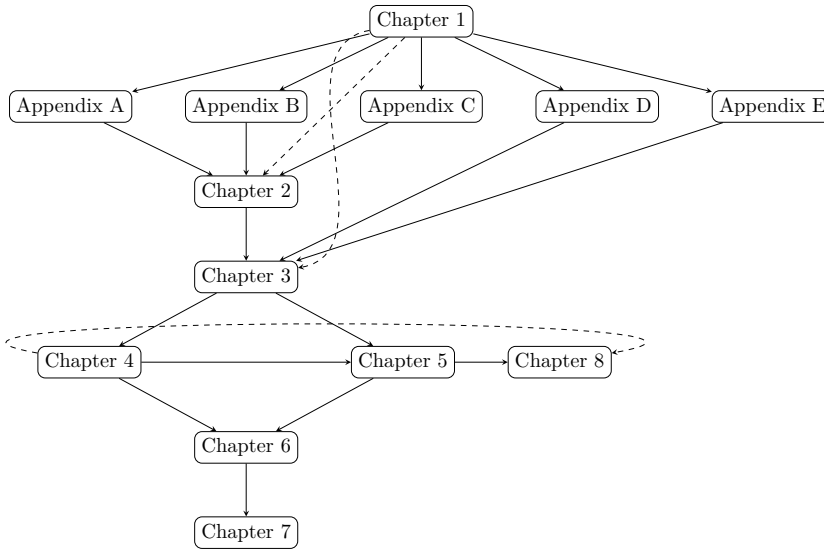


Figure 1: A schematic representation of the dependencies between the chapters and appendices in the book.

chapters cover (i) gradient estimation in a zeroth-order setting, where only noisy function measurements are available; and (ii) asymptotic as well as non-asymptotic analysis of stochastic gradient algorithms with zeroth-order gradient estimates. If you find the material in these chapters interesting, then you could go further to stochastic Newton algorithms with zeroth-order Hessian estimates. These topics are covered in Chapter 6. You could also check out Chapter 7, which describes variants of stochastic gradient/Newton algorithms designed to escape saddle points and converge to local optima.

- If you are student who has done a first course in probability, and someone who would like to conduct research in the area of zeroth-order optimization, then we suggest you pick up the background material covered in the appendices, in particular, ODEs and differential inclusions (Appendix A), conditional expectations and martingales (Appendix B) and smoothness/convexity (Appendix D). Thereafter, we recommend understanding stochastic

approximation, gradient estimation and analysis of stochastic gradient algorithms in that order from Chapters 2 to 4. Introduction to stochastic Newton methods and their analyses, which form the content of subsequent chapters, could be done after the zeroth-order gradient algorithms/analyses are covered.

- If you are also a reinforcement learning (RL) researcher, then the material covered in Chapter 8 could be of interest. In this chapter, we present zeroth-order variations of the well-known REINFORCE policy gradient method. In particular, we establish that such zeroth-order variants are competent and in many RL applications, REINFORCE style gradient estimation is not feasible, making zeroth-order schemes more amenable. One such setting that we cover is risk-sensitive RL, where the objective is not the usual value function, which is an expected value. Instead, we consider alternate functionals of the distribution and describe zeroth-order policy gradient algorithms for optimizing such functionals.

From a teaching viewpoint, the material in this book can be utilized for a semester-long course, with an optional followup course on the shorter side, say one-quarter. In the former course, the background material on ODEs and differential inclusions, conditional expectations and martingales and smoothness and convexity could be introduced first. These correspond to Appendices A, B and D. Next, the content in Chapters 1 to 5 on stochastic gradient algorithms/analyses could be covered. Sections 2.6 and 2.7 could be skipped in this course. The followup course could cover Chapters 6 to 8 on the stochastic Newton algorithms/analyses and RL applications as well as the skipped sections mentioned above.

We would like to thank Praneeth Netrapalli for useful inputs about perturbed gradient descent algorithm, and Aditya Mahajan for useful discussions on two timescale stochastic approximation. We would like to thank our students Soumen Pachal, Sumedh Gupte, Anmol Panda, Shaun Mathew and Ayman Akhter for pointing out typos and minor errors in the earlier versions of this manuscript. Part of this work was supported through a J. C. Bose Fellowship, Project No. DFTM/ 02/ 3125/M/04/AIR-04 from DRDO under DIA-RCOE, the Walmart

Center for Tech Excellence at IISc (CSR Grant WMGT-23-0001), and the RBCCPS, IISc. A portion of this book was written when the first author was visiting the Centre for Machine Intelligence and Data Sciences (C-MInDS) at the Indian Institute of Technology Bombay.

1

Introduction

1.1 Zeroth-order optimization

The underlying processes in many engineering systems can often be quantified by defining suitable objective functions. However, quite often, these functions are not analytically known but their noisy measurements or samples are available. Further, one is often interested in finding optima of such functions despite the challenge that the functions themselves are not known analytically. One may be tempted to try and estimate the whole function through multiple observations from the underlying process at different parameter values that would in turn reveal the function optima. However, such a function estimation scheme would in general be extremely computationally intensive, more so, since we are interested in obtaining the optima of objective functions over continuously valued sets.

Our primary objective here will be to find the minima of a performance objective whose analytical form is not known, however, noise-corrupted observations or samples from such a function are made available either through a simulation device or as 'real' data. The solution approaches that we present shall not aim at estimating the objective function itself but make use of the available 'noisy' data recursively

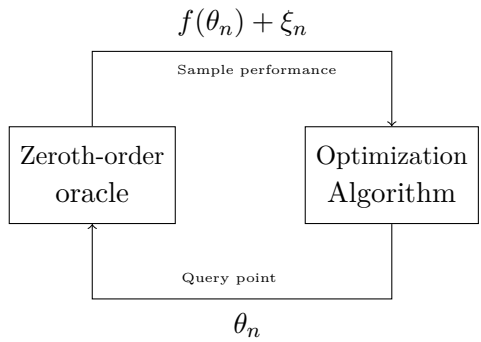


Figure 1.1: Model-free optimization framework

and converge thereby to the optima. Thus, in the end, even though we may still not know the precise nature of the performance objective, the scheme would nonetheless converge to an optimum of the unknown function.

To state it more formally, our goal here will be to find a parameter θ^* such that

$$\theta^* \in \arg\min_{\theta} f(\theta), \tag{1.1}$$

given noisy samples or observations of the performance objective f. As illustrated in Figure 1.1, an iterative optimization algorithm queries the zeroth-order oracle for the objective value at the parameter θ_n at time instant n, and receives the observation $f(\theta_n) + \xi_n$. Here ξ_n, $n \geq 1$ is a sequence of 'noise' random variables. For instance, as we consider in this book, this sequence could be a martingale difference sequence. It is important to note here that the noisy observations $f(\theta_n) + \xi_n$ above cannot be separated into the objective function value $f(\theta_n)$ and the noise component ξ_n to infer the form of the objective function directly from the given noise corrupted data. Thus, it is assumed that the noisy data samples are obtained either from a simulation device or a real system. The obtained data is then used by the optimization algorithm. Since we do not estimate the objective function f and yet run the optimization procedure using only noisy samples, we many times refer to techniques that solve such problems as model-free optimization

methods. On the contrary, approaches that are based on estimating the function f are called model-based optimization techniques. The performance value $f(\theta)$ and the sample performance $g(\theta, \xi) = f(\theta) + \xi$ are related as $f(\theta) = E[g(\theta, \xi)]$, where $E[\cdot]$ denotes the expectation w.r.t the distribution of ξ. It is assumed here that the noise random variable ξ has a mean of zero.

Note also that (1.1) contains '$\in$' instead of '$=$'. This is because the minimizer need not be unique and so $\arg\min_{\theta} f(\theta)$ would constitute the set of all parameters θ that attain the minimum. The set is a singleton if the minimizing parameter is unique. In general, finding one of the minimizers is sufficient in such problems. However, it is important to observe that finding a global minimum, in this setting, is far more computationally intensive than finding a local minimum. In this book, we shall focus on solution methods that aim at finding a local minimum. In most applications, the minima are also isolated in the sense that around any minimum, one can draw a ball of a small enough radius such that it contains only the given (and no other) minimum.

1.2 Applications

Several real-world systems in disciplines such as communication networks, healthcare, finance, are too complex to directly optimize among a set of choices. A viable alternative is to build a simulator for various components of the system, and then perform the optimization over decisions or choices via simulator access. Simulation optimization refers to this setting, where the goal is to find the optimum choice for a certain design parameter. For a given parametric description of the system, performance evaluations using the simulator are typically *noisy* (i.e., have a spread or distribution), and each simulation to obtain an evaluation is often computationally expensive. Thus, in addition to searching for optima, a good simulation optimization algorithm should ensure that the number of function evaluations is small.

Simulation optimization falls under the realm of zeroth-order optimization, and gradient-based algorithms are efficient solution alternatives for finding an optimum using observations from a simulator. The reader is referred to (Fu, 2015) for a detailed introduction to simulation

optimization. For a survey of simulation software catering to a variety of applications, see (Swain, 2017).

An area of practical interest for zeroth-order optimization algorithms is reinforcement learning (RL) (Sutton and Barto, 2018; Bertsekas and Tsitsiklis, 1996; Bertsekas, 2019; Meyn, 2022). In a typical RL setting, the goal is to maximize the cumulative reward over time by learning an optimal policy to choose actions. The underlying formalism is of a Markov decision process (MDP), where the algorithm interacts with the environment through actions, and as a response the environment changes its state and provides a reward. In an MDP, the next state depends on the current state and the chosen action.

Policy gradient methods (Sutton *et al.*, 1999; Konda and Tsitsiklis, 2003; Bhatnagar *et al.*, 2009) are a popular solution approach for such problems. The basis for such algorithms is the policy gradient theorem, which motivates the use of likelihood ratio based gradient estimates. While such an approach of obtaining unbiased gradient estimates works in a risk-neutral RL setting, the same is not true if one incorporates a risk measure in the problem framework. As an example, one could modify the problem to find a policy with the highest mean cumulative discounted reward, while imposing a constraint on the variance. In such a setting, it is difficult to employ the likelihood ratio method for estimating gradient, and simultaneous perturbation methods, which we discuss in detail in this book, are a viable alternative. In (Prashanth and Ghavamzadeh, 2016), the authors employ such an approach to find a risk-optimal policy, which handles a mean-variance tradeoff. Moreover, in (Vijayan and Prashanth, 2021), the authors show that a policy gradient algorithm employing the simultaneous perturbation method for gradient estimation performs on par with REINFORCE — an algorithm that uses the likelihood ratio method for gradient estimation.

More generally, zeroth order optimization approaches have been found useful in the context of simulation optimization under inequality constraints (Bhatnagar *et al.*, 2011a), actor-critic algorithms which are RL algorithms based on the policy iteration procedure (Bhatnagar and Kumar, 2004; Abdulla and Bhatnagar, 2007), simulation-based algorithms for finding optimal policies in finite horizon MDPs (Bhatnagar and Abdulla, 2008), RL algorithms for constrained MDPs (Bhatnagar,

2010; Bhatnagar and Lakshmanan, 2012), as well as discrete parameter simulation optimization (Gerencser *et al.*, 1999; Bhatnagar *et al.*, 2011b). In (Bhatnagar *et al.*, 2006), the problem of finding the optimal policy in an MDP setting conditioned on a rare event is considered and a zeroth order simulation optimization algorithm is presented and analysed. It is shown that the resulting scheme has close connections with risk sensitive MDPs with exponentiated costs. In most of the aforementioned settings, it is not easy to obtain likelihood ratio based sample gradient estimates, hence application of zeroth order methods becomes inevitable.

A more recent application of zeroth-order optimization algorithms, of the type discussed in this book, is in the context of large language models (LLMs), which are nearly ubiquitous, with widespread adoption across various disciplines. Traditional methods for LLM tuning involve high compute costs. To reduce the computational burden of LLM tuning, zeroth-order optimization methods have been explored recently, cf. (Malladi *et al.*, 2023). This approach is less compute-intensive compared to a traditional backpropagation scheme with the well-known ADAM step-size schedule.

Adversarial machine learning is another recent application, where zeroth-order optimization techniques have been applied successfully to construct black-box adversarial examples, cf. (Ilyas *et al.*, 2018; Chen *et al.*, 2017; Bhagoji *et al.*, 2018; Ilyas *et al.*, 2019; Alzantot *et al.*, 2019; Chen *et al.*, 2020a; Mukhoty *et al.*, 2023; Dong *et al.*, 2020). The idea here is to use zeroth-order gradient estimates, similar to SPSA discussed earlier, to approximate the gradient of a target neural network, and use this model to general adversarial images that lead to misclassification. Such adversarial examples are concerning from a security viewpoint, in a safety critical application such as autonomous driving. Zeroth-order gradient estimates have also been employed to make machine learning models robust during training, see (Zhang *et al.*, 2022).

1.3 Stochastic approximation algorithms

The algorithms that we shall present here are all going to be of the stochastic approximation type. The basic stochastic approximation scheme, also referred to as the Robbins-Monro algorithm, named after

its inventors, H.Robbins and S.Monro, see (Robbins and Monro, 1951), was designed to find the zeros of an unknown function $h : \mathbb{R}^d \to \mathbb{R}^d$. The algorithm tunes up the parameter values incrementally based on noisy observations of the function h obtained using the most recent parameter values as they become available. The basic stochastic approximation scheme has the following form:

$$\theta_{n+1} = \theta_n + a(n)(h(\theta_n) + \xi_n), \tag{1.2}$$

starting from an initial parameter estimate $\theta_0 \in \mathbb{R}^d$. Here, $a(n), n \geq 0$ is the step-size sequence of positive real numbers. Given the parameter update θ_n at the nth epoch, a noise-corrupted measurement $h(\theta_n) + \xi_n$ of the objective is obtained and used to update the parameter θ_n to obtain a new parameter θ_{n+1} according to (1.2). As can be seen, smaller step sizes while reducing the noise effects result in more graceful albeit slower convergence. On the other hand, larger step sizes result in faster tracking of the function's zeros though at the cost of higher variance in the iterates. A crucial aspect is one of ensuring convergence that would result in the desired outcome. This and other related aspects will be made more precise in later chapters.

Typical applications of stochastic approximation algorithms include finding the fixed points of a function whose noisy estimates alone are available, as well as finding a minimum of a function again under noisy observations. In the former case, $h(\theta)$ in (1.2) can have the form $h(\theta) = g(\theta) - \theta$ for some function $g : \mathbb{R}^d \to \mathbb{R}^d$, while in the latter, $h(\theta)$ can be of the form $h(\theta) = -\nabla f(\theta)$ for some function $f : \mathbb{R}^d \to \mathbb{R}$. The gradient form of the objective will be of interest to us here except that we will assume that just like the objective function, even the gradient is also not known analytically to us. Noisy function measurements will be used to estimate the gradient. We shall also present some recent Hessian estimation approaches in addition to gradient estimation procedures that will be used in noisy Newton-based schemes. We shall see that one may write the noisy gradient scheme involving gradient estimates as

$$\theta_{n+1} = \theta_n + a(n)(-\nabla f(\theta_n) + \xi_n + \eta_n). \tag{1.3}$$

Here $h(\theta_n)$ in (1.2) is replaced with $-\nabla f(\theta_n)$. However, the important difference is that there is an extra error term η_n in (1.3) that is however not present in (1.2). This error arises because of the gradient estimates obtained from noisy function measurements.

The original Robbins-Monro algorithm was aimed at solving the root finding problem under noisy observations of the function objective with the noise random variables assumed to be forming an independent and identically distributed (i.i.d) sequence. Under certain conditions, convergence was shown in (Robbins and Monro, 1951) to the root of the desired system of equations in the mean-squared sense. Kiefer and Wolfowitz developed a stochastic approximation algorithm to find the maximizer of a given objective function, see (Kiefer and Wolfowitz, 1952). We shall discuss this algorithm in more detail in the next section as indeed this was the first zeroth-order stochastic optimization algorithm and used a finite-difference gradient estimates derived from noisy function measurements. As with (Robbins and Monro, 1951), the objective function in (Kiefer and Wolfowitz, 1952) was considered to be a regression function. The iterate-sequence was shown to converge in probability to the optimum. In (J.R.Blum, 1954), weaker conditions were developed to ensure that both Robbins-Monro and Kiefer-Wolfowitz algorithms converge with probability one to the desired equilibria. In (A.Dvoretzky, 1956), a more general objective function was considered and under weaker conditions both mean-squared convergence and convergence with probability one were shown.

In another major development, the ordinary differential equation (ODE)-based analysis of stochastic approximation algorithms was introduced by Ljung, 1977 and Kushner and Clark, 1978. It was shown that under certain conditions, one may study the asymptotic behavior of a stochastic approximation algorithm by analyzing the same for an associated ODE. The ODE associated with (1.2) can be seen to correspond to

$$\dot{\theta}(t) = h(\theta(t)). \tag{1.4}$$

The main result of Ljung, 1977 and Kushner and Clark, 1978 would say the following:

Let θ^* denote a stable equilibrium of (1.4). Then, under certain conditions on the driving vector field $h(\cdot)$, noise sequence $\xi_n, n \geq 0$, learning rates $a(n), n \geq 0$, if the sequence θ_n governed by (1.2) enters infinitely often a compact subset of the domain of attraction of θ^*, then $\theta_n \to \theta^*$ almost surely.

The above corresponds to a strong notion of recurrence for the ODE, and may not be applicable in many situations. In (Benaïm, 1996), (Benaïm, 1999) and (Benaïm and Hirsch, 1996), the ODE based analysis of (Ljung, 1977) and (Kushner and Clark, 1978) has been extended to the setting where the asymptotic behavior of the algorithm is analyzed via a weaker notion of recurrence, namely *chain recurrence*, of the underlying ODE. Most of the modern ODE based analyses follow the latter approaches.

1.4 Zeroth-order stochastic gradient (SG) algorithm

Consider the following stochastic approximation scheme:

$$\theta_{n+1} = \theta_n + a(n)(-\widehat{\nabla} f(\theta_n)), \tag{1.5}$$

where $\widehat{\nabla} f(\theta_n)$ is a noisy estimate of the gradient of $f(\theta_n)$, with $f : \mathbb{R}^d \to \mathbb{R}$ being the objective function to be minimized. The Kiefer-Wolfowitz scheme, see (Kiefer and Wolfowitz, 1952), estimates the gradient $\nabla f(\theta)$ using the following estimator: For $i = 1, \ldots, d$,

$$\begin{aligned} \widehat{\nabla}_i f(\theta_n) &= \frac{1}{2\delta}\left(f(\theta_n + \delta e_i) + \xi_i^+(n) - f(\theta_n - \delta e_i) - \xi_i^-(n)\right), \\ &= \frac{1}{2\delta}\left((f(\theta_n + \delta e_i) - f(\theta_n - \delta e_i)) + \left(\xi_i^+(n) - \xi_i^-(n)\right)\right), \end{aligned} \tag{1.6}$$

where, $\widehat{\nabla}_i f(\theta_n)$ denotes the estimate of the ith partial derivative of $f(\theta_n)$. Further, $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ is the unit d-dimensional vector with 1 as the ith place and all other entries as 0. Further, $\xi_i^+(n)$ (resp. $\xi_i^-(n)$) is the noise associated with the estimate of the function f measured at the parameter value $(\theta_n + \delta e_i)$ (resp. $(\theta_n - \delta e_i)$).

Notice that in (1.6), assuming the function f to be sufficiently smooth, a first order Taylor's expansion would lead to

$$\frac{f(\theta_n + \delta e_i) - f(\theta_n - \delta e_i)}{2\delta} = \nabla_i f(\theta_n) + O(\delta^2).$$

This happens because the first and the third terms in the Taylor's expansion get canceled as a consequence of the balanced nature of the estimate. The term comprising $O(\delta^2)$ contributes to the bias in the gradient estimate. In relation to (1.3), if $\delta \to 0$ as $n \to \infty$ above, the analysis turns out to be a simple extension of the corresponding analysis for (1.2), see Chapter 2 of (Borkar, 2022). However, letting the δ-parameter approach zero results in constraining the choice of the step-size sequence $\{a(n)\}$. Nonetheless, the recursion in such a case can be shown to track the ODE

$$\dot{\theta}(t) = -\nabla f(\theta(t)). \tag{1.7}$$

For a fixed $\delta > 0$, on the other hand, it can be shown that for an algorithm as in (1.5) with say the Kiefer-Wolfowitz gradient estimator (1.6), given $\epsilon > 0$, $\exists \delta_0 > 0$, such that when the 'perturbation parameter' $\delta \in (0, \delta_0]$, the term η_n is $O(\epsilon)$. Analyses with a fixed δ can be carried out by viewing the resulting algorithm as one involving a set-valued map $H(\theta) = \nabla f(\theta) + \bar{B}(0, \epsilon)$, where $\bar{B}(0, \epsilon)$ is a closed ball of radius ϵ around the origin. The resulting scheme can then be analysed by viewing the limiting system as the Differential Inclusion (DI)

$$\dot{\theta}(t) \in -H(\theta(t)), \tag{1.8}$$

see, for instance, Ramaswamy and Bhatnagar, 2018.

A disadvantage with the gradient estimator defined above is that it requires $2d$ function measurements or simulations to run one update of the parameter according to (1.5). The amount of computation thus can be very high for a large value of d. In (Spall, 1992), the following

estimator of the gradient has been proposed that uses only two function measurements regardless of the value of d.

$$\widehat{\nabla}_i f(\theta_n) = \frac{f(\theta_n + \delta\Delta(n)) + \xi^+(n) - f(\theta_n - \delta\Delta(n)) - \xi^-(n)}{2\delta\Delta_i(n)}. \tag{1.9}$$

Here, $\Delta(n) = (\Delta_1(n), \ldots, \Delta_d(n))^T$ is a vector of i.i.d random variables $\Delta_j(n)$, $j = 1, \ldots, d, n \geq 0$ that are typically zero-mean with a finite inverse moment bound. Independent symmetric Bernoulli random variables such as $\Delta_j(n) = \pm 1$ w.p. $1/2$ are commonly used here. A Taylor's expansion as with the Kiefer-Wolfowitz estimator would give the following in this case:

$$\begin{aligned} &\frac{f(\theta_n + \delta\Delta(n)) - f(\theta_n - \delta\Delta(n))}{2\delta\Delta_i(n)} = \frac{\Delta(n)^T \nabla f(\theta_n)}{\Delta_i(n)} + O(\delta^2) \\ &= \nabla_i f(\theta_n) + \sum_{j \neq i} \frac{\Delta_j(n) \nabla_j f(\theta_n)}{\Delta_i(n)} + O(\delta^2). \end{aligned} \tag{1.10}$$

Note the presence of an extra (the second) term on the RHS that contributes to the bias. It may however be observed that

$$E\left[\sum_{j \neq i} \frac{\Delta_j(n) \nabla_j f(\theta_n)}{\Delta_i(n)} \mid \theta_n\right] = 0.$$

It can therefore be seen that

$$\left\| \mathbb{E}\left[\widehat{\nabla} f(\theta_n) \mid \theta_n\right] - \nabla f(\theta_n) \right\| \leq C\delta^2, \tag{1.11}$$

for some positive scalar C.

Since this estimate of ∇f is used in the recursion (1.5), a stochastic approximation scheme, one recovers the expectation in the asymptotic limit of the iterate sequence as the noise effects die down. A one-simulation estimator was proposed in (Spall, 1997) where the form of the estimator was simply

$$\widehat{\nabla}_i f(\theta_n) = \frac{f(\theta_n + \delta\Delta(n)) + \xi^+(n)}{\delta\Delta_i(n)}, \ i = 1, \ldots, d. \tag{1.12}$$

A Taylor's expansion on the function value without the noise term in (1.12) gives

$$\frac{f(\theta_n + \delta\Delta(n))}{\delta\Delta_i(n)} = \frac{f(\theta_n)}{\delta\Delta_i(n)} + \nabla_i f(\theta_n) + \sum_{j\neq i} \frac{\Delta_j(n)\nabla_j f(\theta_n)}{\Delta_i(n)} + O(\delta).$$

The third term on the RHS above is the same as a corresponding term that contributes to the bias in (1.10). However, there is an additional first term on the RHS that also has zero mean given the parameter update θ_n. The latter term, however, is primarily responsible for below par performance of this estimate because of the presence of δ, a typically small quantity, in the denominator. The aforementioned estimators are popularly referred to as two-measurement and one-measurement simultaneous perturbation stochastic approximation (SPSA) estimators.

Deterministic perturbation versions of the above algorithms have been proposed in (Bhatnagar *et al.*, 2003) and are seen to yield better performance particularly for the one-simulation estimators when compared with their random perturbation counterparts. This is because of a regular (cyclic) cancellation of the previously mentioned bias terms when deterministic perturbation schedules are used. In other work along similar lines, the smoothed functional estimators have been studied in (Rubinstein, 1981), (Katkovnik and Kulchitsky, 1972), (Bhatnagar and Borkar, 2003), (Bhatnagar, 2007), (Bhatnagar *et al.*, 2013), where the underlying perturbation distributions are primarily Gaussian, uniform and Cauchy. In (Ghoshdastidar *et al.*, 2014b; Ghoshdastidar *et al.*, 2014a), smoothed functional algorithms with q-Gaussian perturbations have been presented that are seen to significantly extend the class of perturbations and allowing for a continuum of distributions depending on the value of the q-parameter.

Random directions stochastic approximation (RDSA) algorithm has been presented in (Kushner and Clark, 1978) where the underlying distribution has been considered to be uniform on the surface of a sphere that is akin to the multivariate Gaussian distribution. In (Prashanth *et al.*, 2017), algorithms with i.i.d., uniformly distributed perturbations have been proposed. These perturbations lie within a d-dimensional cube. Further, in (Prashanth *et al.*, 2020), deterministic perturbation versions of these algorithms have been studied and analyzed. We shall

be discussing some of these algorithms in more detail in a later chapter.

1.5 Zeroth-order stochastic Newton (SN) algorithm

Recall that a SG algorithm involves the following recursion:

$$\theta_{n+1} = \theta_n - a(n)\widehat{\nabla} f(\theta_n), \tag{1.13}$$

where $\widehat{\nabla} f(\theta_n)$ is an estimate of the gradient $\nabla f(\theta_n)$.

There are three main shortcomings in employing a SG algorithm. First, from an asymptotic convergence rate analysis (cf. (Fabian, 1968)), it is apparent that the SG algorithm would achieve an order $O\left(\frac{1}{\sqrt{n}}\right)$ convergence when the stepsize is set using the curvature of f, i.e., $a_n = a_0/n$ with $a_0 > \delta/2\lambda_{\min}(\nabla^2 f(\theta^*))$. In practice, such curvature information is seldom available, and hence, it is problematic to assume such knowledge in setting the step-size for optimal convergence speed. Second, it is widely observed empirically that a SG algorithm declines fast initially, but slows down towards the end, i.e., when the SG iterate is near an optimum θ^*. Third, the update rule (1.13) is *not* scale-invariant, i.e., changing θ to $B\theta$ for some matrix B, would imply a change in the update (1.13). Finally, a SG algorithm may get stuck in traps or unstable equilibria such as local maxima and saddle points, while the goal is for it to converge to local minima (esp. since convexity is not assumed).

A second-order SN algorithm overcomes the shortcomings of a first-order SG algorithm mentioned above. A general gradient-search algorithm involves an update rule of the form:

$$\theta_{n+1} = \theta_n - a(n)B(\theta_n)^{-1}\nabla f(\theta_n), \tag{1.14}$$

where $B(\theta)$ for any $\theta \in \mathbb{R}^d$ is a $d \times d$ matrix. The following choices of the $B(\theta)$ matrix are widely popular (see (Bertsekas, 1999)):

(i) $B(\theta) = I$ (the identity matrix) for all θ: In this case, the algorithm (1.14) reduces to the first-order SG algorithm (1.13).

(ii) $B(\theta)$ is a diagonal matrix with diagonal entries being $\nabla^2_{i,i} f(\theta)$. This corresponds to the (second-order) Jacobi algorithm.

(iii) $B(\theta) = \nabla^2 f(\theta)$: This corresponds to the second-order SN algorithm.

In the following, we focus on the SN algorithm (corresponding to the full Hessian case). As illustrated in Figure 1.2, the update rule above then requires computation of the Hessian as well as the gradient estimate at any parameter update θ_n.

We elaborate on the advantages of such an algorithm over the first-order scheme in (1.13) (or alternatively the case of $B(\theta) = I$ in (1.14)). First, such algorithms achieve the optimum speed of convergence without the knowledge of $\lambda_{\min}(\nabla^2 f(\theta^*))$. Setting $a_0 = 1$ would suffice. Second, it is generally observed that second-order methods exhibit faster convergence in the final phase, i.e., when the iterates are close to the optima. This can be attributed to the fact that second-order methods minimize a quadratic model of f, while SG algorithm (1.13) uses a first-order Taylor's approximation. Third, second-order algorithms are scale-invariant, i.e., they auto-adjust to the scale of θ. Finally, second-order algorithms avoid traps naturally, since they factor in curvature information through the Hessian. On the flip side, second-order methods have a higher per-iteration cost than their first-order counterparts, as the Hessian matrix has to be inverted during each iteration.

In the zeroth-order optimization setting that we consider, we do not have direct access to the gradient and the Hessian of the objective function. Instead, as illustrated in Figure 1.2, both gradient and Hessian have to estimated from noisy function observations before performing a parameter update. In other words, letting $\widehat{\nabla} f(\theta_n)$ and $\overline{H}_n$ denote the gradient and Hessian estimates, we update the parameter as follows:

$$\theta_{n+1} = \theta_n - a(n) \left(\overline{H}_n\right)^{-1} \widehat{\nabla} f(\theta_n). \tag{1.15}$$

The topic of gradient estimation is handled in Chapter 3, while Chapter 6 focuses on Hessian estimation, and the convergence analysis of (1.15), where we use zeroth-order estimates of both the gradient and the Hessian.

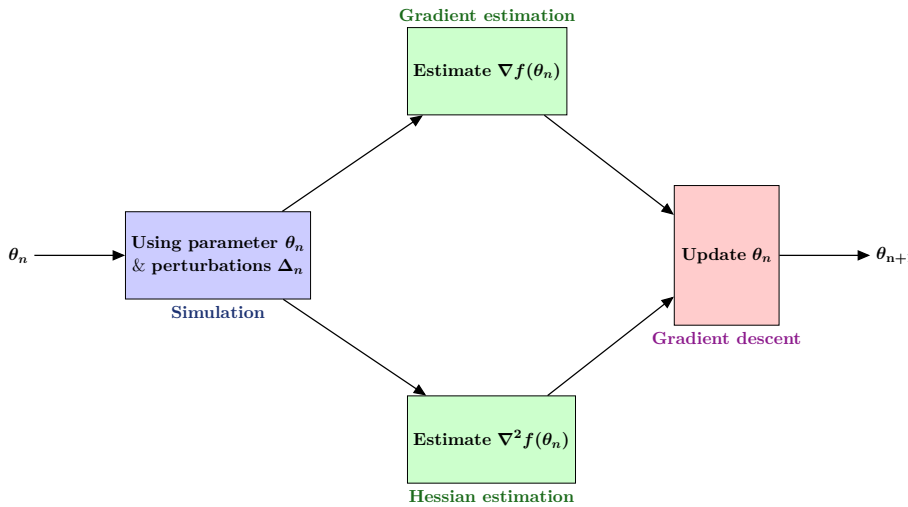


Figure 1.2: Overall flow of a second-order stochastic gradient algorithm

To understand the problem of Hessian estimation, we now discuss a finite difference approximation, which requires $O(d^2)$ function measurements. The simultaneous perturbation trick brings this number down to a small constant, regardless of the parameter dimension d. We shall discuss these schemes in detail in Chapter 6.

Consider a scalar variable θ. A finite difference approximation of the first derivative for this simple case of a scalar parameter θ is:

$$\frac{df(\theta)}{d\theta} \approx \left(\frac{f(\theta+\delta) - f(\theta-\delta)}{2\delta} \right). \tag{1.16}$$

Assuming the objective is smooth, and employing Taylor series expansions of $f(\theta+\delta)$ and $f(\theta-\delta)$ around θ, we obtain:

$$f(\theta \pm \delta) = f(\theta) \pm \delta \frac{df(\theta)}{d\theta} + \frac{\delta^2}{2} \frac{d^2 f(\theta)}{d\theta^2} + O(\delta^3),$$

$$\text{Thus, } \frac{f(\theta+\delta) - f(\theta-\delta)}{2\delta} = \frac{df(\theta)}{d\theta} + O(\delta^2).$$

From the above, it is easy to see that the estimate (1.16) converges to the true gradient $\dfrac{df(\theta)}{d\theta}$ in the limit as $\delta \to 0$.

This idea can be extended to estimate the second derivative by applying a finite difference approximation to the derivative in (1.16) as

follows:

$$\frac{d^2 f(\theta)}{d\theta^2} \approx \frac{\left(\frac{f(\theta+\delta+\delta)-f(\theta+\delta-\delta)}{2\delta}\right) - \left(\frac{f(\theta-\delta+\delta)-f(\theta-\delta-\delta)}{2\delta}\right)}{2\delta} \tag{1.17}$$

As before, using Taylor series expansions, it can be shown that the RHS above is a good approximation to the second derivative.

For the case of a vector parameter, one needs to perturb each coordinate separately, leading to the following scheme for estimating the Hessian $\nabla^2 f(\theta)$: For any $i, j \in \{1, \ldots, d\}$,

$$\nabla^2_{ij} f(\theta) \approx \frac{1}{4\delta^2}\Big(f(\theta+\delta e_i+\delta e_j) + f(\theta+\delta e_i-\delta e_j) - (f(\theta-\delta e_i+\delta e_j) - f(\theta-\delta e_i-\delta e_j))\Big). \tag{1.18}$$

Such an approach requires $4d^2$ number of function measurements to form the Hessian estimate. In the next section, we overcome this limitation by employing the simultaneous perturbation trick. Before that, we extend the estimate in (1.18) to the noisy case as follows: Suppose we have the following function measurements: For any $i, j \in \{1, \ldots, d\}$,

$$y_1 = f(\theta+\delta e_i+\delta e_j) + \xi_{1ij}, y_2 = f(\theta+\delta e_i-\delta e_j) + \xi_{2ij}, \tag{1.19}$$

$$y_3 = f(\theta-\delta e_i+\delta e_j) + \xi_{3ij} \text{ and } y_4 = f(\theta-\delta e_i-\delta e_j) + \xi_{4ij}. \tag{1.20}$$

Using these function measurements, we form the Hessian estimate $\widehat{H}$ as follows:

$$\widehat{H}_{ij} = \left(\frac{y_1 - y_2 - y_3 + y_4}{4\delta^2}\right), \forall i, j \tag{1.21}$$

Assuming the function is sufficiently smooth, as in the gradient case and the noise elements in the function measurements are zero mean, it can be shown through Taylor series expansions that

$$\mathbb{E}[\widehat{H}_{ij} \mid \theta] = \frac{1}{4\delta^2}\Big(f(\theta+\delta e_i+\delta e_j) + f(\theta+\delta e_i-\delta e_j)$$

$$- \left(f(\theta - \delta e_i + \delta e_j) - f(\theta - \delta e_i - \delta e_j)\right)\Big)$$
$$= \nabla^2_{ij} f(\theta) + O(\delta^2).$$

While the bias of the estimator is on the lower side, with explicit control via the δ parameter, the problem is in the number of function measurements. The latter number is $4d^2$, limiting the practical viability on high-dimensional problems. In Chapter 6, we discuss several alternative schemes using the simultaneous perturbation method for Hessian method. These schemes use a small (constant) number of function measurements (regardless of the parameter dimension d), while ensuring a bias of $O(\delta^2)$.

1.6 Organization of the book

We now describe the organization of the rest of the book.

In Chapter 2, we provide an introduction to stochastic approximation algorithms, and outline a few popular applications such as mean estimation, gradient-type algorithms, fixed-point iterations, and quantile estimation. These algorithms are incremental update procedures that work with stochastic or noisy data as it becomes available and are model-free procedures. In Chapter 2, we provide a detailed introduction to stochastic approximation algorithms, provide motivating applications, and subsequently provide the main results on convergence of these schemes. It turns out that many of the stochastic optimization schemes require a treatment of algorithms with set-valued maps. We also present such algorithms in settings where data samples become available one at a time in real time, and so are Markovian. We therefore discuss the main convergence results in connection with these as well. In addition, Newton-based stochastic optimization schemes involve estimating the inverse of the Hessian of the objective. This cannot be done using the standard stochastic approximation template and we need such algorithms to perform updates using two-timescale procedures. We therefore also discuss two-timescale stochastic approximation algorithms (including those with set-valued maps) in this chapter.

In Chapter 3, we provide a variety of gradient estimators using the

simultaneous perturbation method. These include unified two-point as well as one-point gradient estimation schemes. The unified estimates feature abstract random perturbations that are required to satisfy certain conditions to ensure that the bias and variance of the estimates is manageable. Specializing these estimates with specific choice of random perturbations leads to several well-known simultaneous perturbation-based schemes such as the smoothed functional scheme (Katkovnik and Kulchitsky, 1972) with later refinements in (Polyak and Tsybakov, 1990; Dippon, 2003; Nesterov and Spokoiny, 2017), random direction stochastic approximation (RDSA) scheme proposed by (Kushner and Clark, 1978), and recently enhanced in (Prashanth *et al.*, 2017), and the popular simultaneous perturbation stochastic approximation (SPSA) scheme proposed by (Spall, 1992). While most estimators that we present require one or two function measurements in order to estimate the gradient, we also touch upon a recently developed class of generalized simultaneous perturbation gradient estimators that provide estimators requiring a number of function measurements that depends on the bias in the gradient estimator. We analyze the bias and variance of the aforementioned estimators in the convex as well as the non-convex regimes. In either case, the analysis requires the objective to be smooth.

In Chapter 4, we present a detailed mathematical treatment of a stochastic gradient algorithm that employs simultaneous perturbation-based gradient estimates. In particular, we cover asymptotic convergence of the stochastic gradient scheme using the popular ordinary differential equation (or ODE) method. It turns out that in many of these algorithms, it makes sense to hold the sensitivity parameter in the gradient estimation procedure fixed and not push it to zero in order that the estimator variance does not blow up. In such a case, we observe that the resulting scheme can be viewed as a stochastic recursive inclusion, i.e., one involving set-valued maps. Thus, we use here the theory of differential inclusions to establish that the stochastic gradient algorithm converges to a chain-recurrent set of an underlying differential inclusion.

In Chapter 5, we present the non-asymptotic analysis for the zeroth-order SG (ZSG) algorithm. In the case of a non-convex objective, we bound the expected decrease in the objective function in each iteration using the bias and variance properties of the gradient estimators together

with a standard Taylor series argument. The expected decrease is used to provide an overall bound, which shows that the stochastic gradient algorithm converges to an approximate stationary point of the objective, with a rate $O(\frac{1}{\sqrt{N}})$, where N is the number of iterations. In this chapter, we also analyze the rate of convergence of ZSG algorithm when the underlying objective is either convex or else strongly-convex. In the former case, we bound the optimization error (difference in function value between that of the iterate and the optimum), while in the latter case, we bound the parameter error, which is the norm of the distance between ZSG iterate and the optimum. Strong convexity allows a bound on the parameter error, while in the case of a non-strongly convex function, only a bound on the difference in function value is feasible. This is true even in the deterministic optimization setting, though the rates are slower in the stochastic zeroth-order setting that we study in this book. In this chapter, we also present a minimax lower bound using information-theoretic arguments, and this bound shows that the upper bounds for the ZSG algorithm are optimal up to a constant factor for the convex/strongly-convex cases.

In Chapter 6, we cover Hessian estimation using simultaneous perturbation methods. In particular, we provide a theoretical introduction to second-order SPSA proposed in (Spall, 2000) as well as its later enhancements in (Bhatnagar, 2005; Bhatnagar and Prashanth, 2015). We also describe second-order smoothed functional (Bhatnagar, 2007) and second-order RDSA (Prashanth *et al.*, 2017) schemes. We analyze the bias in these Hessian estimates, and establish that each of these aforementioned schemes results in an asymptotically unbiased Hessian estimate. In this chapter, we also analyze a stochastic Newton algorithm using gradient/Hessian estimates based on the simultaneous perturbation method. As mentioned previously, these algorithms involve two-timescale stochastic approximation schemes. The theoretical guarantees that we provide include the asymptotic almost sure convergence of the stochastic Newton scheme, and an asymptotic normality result that can be used to bound the asymptotic covariance, which in turn helps one understand the mean-square error of the algorithm after a sufficiently large number of iterations. The latter analysis provides

a convergence rate for the stochastic Newton algorithm, albeit in an asymptotic sense.

In Chapter 7, we focus on the points of convergence of the stochastic approximation schemes. An important consideration such algorithms is to ensure that the stochastic algorithm converges to local minima and not to saddle points that while being stationary points of the system, are in fact, unstable equilibria of the underlying ODE. Two schemes to escape saddle points are presented. In the first scheme, additional assumptions on the richness of noise are provided in the case of a general zeroth order gradient estimation scheme that would ensure avoidance of saddle points. We review these conditions on the noise from (Pemantle, 1990) and provide the basic results. The second scheme deals with a cubic regularized Newton-based formulation from (Maniyar *et al.*, 2024) with gradient and Hessian estimates obtained using zeroth-order estimation procedures. Convergence to an ϵ-second order stationary point is then shown.

In Chapter 8, we provide applications of simultaneous perturbation methods in the reinforcement learning (RL) context. The first application involves a constrained discounted Markov decision process (MDP). In an RL setting, direct gradient measurements of the objective or value function are not available. Instead, one can estimate the value function using a Monte Carlo scheme, or the popular temporal difference (TD) learning algorithm. We consider the stochastic shortest path setting here. Assuming a smooth class of parameterized policies, we describe a policy gradient scheme that employs SPSA-based gradient estimates in conjunction with value function estimation using Monte Carlo samples as with the REINFORCE algorithm. We present a convergence analysis of our algorithm, which shows that the algorithm converges almost surely to local optima in the asymptotic limit. The second application considers a risk-sensitive RL problem, where the goal is to find a policy that maximizes the value function while satisfying a constraint that is formed using a risk measure. As in the first application, we describe a policy gradient algorithm for solving the risk-constrained MDP, and provide an asymptotic convergence analysis of this algorithm.

We also provide five appendices of useful background material. We outline the content of the appendices below.

Appendix A covers significant material on ODEs and differential inclusions, specifically from the viewpoint of stability, equilibria, attractors, as well as weaker notions of recurrence. These concepts are required for the asymptotic analysis of stochastic gradient and Newton algorithms in Chapters 2, 4 and 6, respectively.

Appendix B provides an introduction to selected topics in probability that are relevant to this book. In particular, we discuss various notions of convergence of random variables in this appendix. Next, we cover conditional expectation and provide a detailed introduction to martingales, including examples from stochastic approximation and asymptotic convergence results. The latter results on martingale convergence are useful in the analysis of stochastic approximation algorithms in general (see Chapter 2), and zeroth-order gradient-based algorithms in particular (see Chapters 4 and 6). To elaborate, stochastic gradient and Newton algorithms involve increments with a martingale difference noise sequence and it is important to understand when this sequence converges, so that an ODE or differential inclusions-based analysis of the aforementioned algorithms is feasible.

Appendix C provides an introduction to Markov chains in discrete time. This background is useful in understanding stochastic approximation algorithms with a Markovian noise component (see Section 2.6.

Appendix D provides foundational material on smooth optimization. In particular, first/second-order optimality conditions, smoothness and convexity are discussed in detail in this appendix.

Appendix E provides an introduction to information theoretic concepts such as entropy, and KL-divergence, followed by a statement with proof of a simpler version of the well-known Pinsker's inequality. This background is useful for understanding the minimax lower bounds derived for a gradient-based algorithm with zeroth-order information in Section 5.6.

1.7 Bibliographic remarks

Kiefer and Wolfowitz in (Kiefer and Wolfowitz, 1952) presented the first paper on stochastic gradient descent with zeroth order estimators

and analysed their algorithm using the approach in (Robbins and Monro, 1951). A comprehensive and detailed treatment of stochastic optimization including direct methods and evolutionary algorithms, in addition to zeroth order methods such as SPSA is available in (Spall, 2005). A detailed treatment of stochastic simulation of random variables and processes including those driven by stochastic differential equations that also contains stochastic optimization is given in (Asmussen and Glynn, 2007b). Another textbook primarily on stochastic simulation that also deals with Markov chain Monte Carlo and discrete event system simulation, in addition to stochastic optimization (specifically, smoothed functional approaches) is (Rubinstein, 1981).

A text that deals primarily with the theory of stochastic approximation is (Borkar, 2022) that however also has a chapter on stochastic zeroth order methods for gradient estimation where methods such as SPSA and SF are briefly surveyed. Discrete event system simulation and optimization has been well-studied and analysed using perturbation analysis based methods in (Cassandras and Lafortune, 2008). A text mainly dedicated to optimal control and reinforcement learning but which also delves a bit on zeroth order stochastic optimization is (Meyn, 2022). A recent text on stochastic optimization and reinforcement learning covering a wide range of topics in these domains is (Powell, 2021).

A textbook treatment of zeroth-order stochastic optimization approaches is available in (Bhatnagar *et al.*, 2013). The focus of the approaches presented in that text was to find the optimum parameter of an objective which in itself is a certain long-run average cost over noisy cost samples. A variety of methods for both unconstrained and constrained optimization including reinforcement learning are presented there. The resulting algorithms largely have a multi-timescale structure and the asymptotic convergence analysis of these algorithms is presented. In our current text, we primarily consider single-timescale stochastic optimization algorithms that estimate the gradient and (in some cases) the Hessian using zeroth order estimators though we also consider two-timescale algorithms for the latter case. We present newer and more general analyses of these algorithms and provide in detail both asymptotic as well as non-asymptotic convergence analyses of the

presented algorithms. The asymptotic analyses are shown using limiting arguments involving underlying ordinary differential equations (ODE) or differential inclusions (with set-valued maps) as the case may be. Our current text also covers many recent algorithms not contained in (Bhatnagar *et al.*, 2013).

2

Stochastic approximation

In this chapter, we provide an introduction to stochastic approximation algorithms, and outline a few popular applications such as mean estimation, gradient-type algorithms, fixed-point iterations, and quantile estimation. We provide the main asymptotic convergence results under two approaches, namely ODE and differential inclusions. The former approach is applicable to Lipschitz continuous objective functions, which allows viewing a linearly-interpolated stochastic approximation algorithm's sample path as approximating the trajectory of an ODE. Using this 'dynamical systems' viewpoint, we list the assumptions that ensure almost sure convergence of stochastic approximation iterates to the equilibria of the underlying ODE. The approach of recursive inclusions is useful for handling objective functions with discontinuities. As in the ODE case, the stochastic approximation algorithm's interpolated trajectory is seen as an approximation to that of the recursive inclusion, leading to an almost sure convergence result. In the context of this book, when the perturbation constant δ, which features in the simultaneous perturbation-based gradient estimator presented above, is taken to zero, the stochastic gradient algorithm's behavior can be analyzed using ODEs, while treatment of a constant δ requires one to consider a

differential inclusions-based analysis.

2.1 Introduction

The basic stochastic approximation recursion is of the following form:

$$\theta_{n+1} = \theta_n + a(n)(h(\theta_n) + M_{n+1}), \tag{2.1}$$

where $\theta_n \in \mathbb{R}^d$, $n \geq 0$, is the stochastic sequence of iterates that are updated according to (2.1), $h : \mathbb{R}^d \to \mathbb{R}^d$ is a point-to-point map, $M_{n+1}, n \geq 0$ is the associated noise sequence, and the multipliers $a(n), n \geq 0$ form a sequence of positive step sizes or learning rates.

Under certain conditions on the aforementioned quantities that we shall discuss in this chapter, one can show that the recursion (2.1) almost surely tracks asymptotically the limit sets of the ODE (2.2) in a manner that will be made precise later.

$$\dot{\theta}(t) = h(\theta(t)). \tag{2.2}$$

We shall also consider here generalizations of the scheme (2.1) via stochastic recursive inclusions as well as recursions with an additional Markov noise component. Stochastic recursive inclusions are algorithms as in (2.1) except that the function $h(\theta)$ is in general a set instead of a point for any given θ. Such a scheme in general will have the following form:

$$\theta_{n+1} = \theta_n + a(n)(y_n + M_{n+1}), \tag{2.3}$$

where $M_n, n \geq 0$ is the noise sequence as before and $y_n \in h(\theta_n)$, where it will be now assumed that $h : \mathbb{R}^d \to \mathbb{R}^d$ is a set-valued map. Under some assumptions, such recursions will also be seen to almost surely track asymptotically the underlying differential inclusion

$$\dot{\theta}(t) \in h(\theta(t)). \tag{2.4}$$

The reader is referred to Appendix A for an introduction to ODEs and differential inclusions.

2.2 Applications

We begin with a few well-known applications of stochastic approximation. These include minimizing a function given noisy function measurements, which forms the core content of this book, as well as estimation of various quantities, e.g., mean, fixed-point, quantile, from noisy observations.

2.2.1 Mean estimation

Consider a random variable (r.v.) $\mathcal{X}$ with mean μ and variance σ^2. Suppose we are given independent and identically distributed (i.i.d.) samples $X_1, \ldots, X_n$ from the distribution of $\mathcal{X}$. Let $\theta_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$ be the sample mean computed using these n samples. We now derive an iterative scheme for updating the sample mean.

$$\begin{aligned}\theta_{n+1} &= \frac{1}{n+1}\sum_{k=1}^{n+1} X_k = \frac{n}{n+1}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) + \frac{1}{n+1}X_{n+1} \\ &= \frac{n}{n+1}\theta_n + \frac{1}{n+1}X_{n+1} \\ \theta_{n+1} &= \theta_n + \frac{1}{n+1}\left(X_{n+1} - \theta_n\right). \end{aligned} \tag{2.5}$$

The update rule above is a stochastic approximation scheme with step size $a(n) = \dfrac{1}{n+1}$, $n \geq 0$.

By the strong law of large numbers, one obtains

$$\theta_n \to \mu \text{ a.s. as } n \to \infty.$$

One may instead use a more general step size sequence $a(n), n \geq 0$ and write the update rule (2.5) as

$$\begin{aligned}\theta_{n+1} &= \theta_n + a(n)\left(X_{n+1} - \theta_n\right) \\ &= \theta_n + a(n)\left[(\mu - \theta_n) + (X_{n+1} - \mu)\right]\end{aligned}$$

Letting $M_{n+1} = X_{n+1} - \mu$, it is easy to see that $M_n, n \geq 0$ is a martingale difference sequence[1] satisfying $\mathbb{E}M_n^2 < \infty$.

[1] The reader is referred to Appendix B for an introduction to martingales.

From an application of the Kushner-Clark lemma, to be presented later, it can again be shown that $\theta_n \to \mu$ almost surely as $n \to \infty$ and this happens for general step sizes that satisfy the following conditions (see Theorem 2.4(i)):

$$\sum_n a(n) = \infty \text{ and } \sum_n a(n)^2 < \infty. \tag{2.6}$$

Clearly $a(n) = 1/(n+1)$ is a special case of the above. This means that the above result using the strong law of large numbers continues to hold with more general step sizes. The Kushner-Clark lemma is the main tool to infer asymptotic convergence of stochastic approximation algorithms. We shall present a precise statement and a proof of this result in Section 2.3.

2.2.2 Stochastic gradient algorithm using unbiased gradient information

Consider the following problem: Find

$$\theta^* \in \arg\min_\theta f(\theta), \tag{2.7}$$

where f is a smooth function (see Appendix D for background material on smoothness).

A stochastic gradient algorithm for solving (2.7) would update as follows:

$$\theta_{n+1} = \theta_n - a(n)\widehat{\nabla} f(\theta_n). \tag{2.8}$$

In the above, $\widehat{\nabla} f(\theta_n)$ is an estimate of the gradient $\nabla f(\theta_n)$, and $\{a(n)\}$ are (pre-determined) step sizes satisfying standard stochastic approximation conditions (see (2.6) above).

Here we shall assume unbiased gradient information is available, i.e., $\mathbb{E}\left[\widehat{\nabla} f(\theta_n) \mid \theta_n\right] = \nabla f(\theta_n)$. In this case, the algorithm in (2.8) becomes an instance of the seminal stochastic approximation scheme proposed by Robbins and Monro in 1951. The latter algorithm was proposed to find the zeroes of a function, and in the case of (2.8), the function of interest is ∇f. If the gradient estimates $\widehat{\nabla} f(\theta_n)$ have bounded variance, then the algorithm in (2.8) can be shown to converge to the stationary points of f. We make this claim precise later in Section 4.1.

We now describe a popular optimization setting, where unbiased gradient information is available. Consider the following problem that is ubiquitous in machine learning applications involving training over a given dataset of m samples, say $\{(x_i, y_i), i = 1, \ldots, m\}$:

$$\min_{\theta} f(\theta) = \frac{1}{m} \sum_{i=1}^{m} f_i(\theta). \tag{2.9}$$

In the above, f_i denotes the loss associated with sample i. A simple example is the square-loss in a linear regression problem, where $f_i(\theta) = (y_i - \theta^{\mathsf{T}} x_i)^2$. It is common to assume that the loss functions $f_i, \forall i$ are smooth, and f is convex or strongly convex.

A batch gradient descent algorithm would solve the problem above using the following update iteration:

$$\theta_{n+1} = \theta_n - a(n) \left(\frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\theta_n) \right). \tag{2.10}$$

The above algorithm is a noise-less algorithm, and for large m, it is computationally expensive. In ML parlance, m is the number of training examples.

A computationally efficient alternative is stochastic gradient descent, popularly known as SGD. This algorithm involves picking a training sample uniformly at random, i.e., a r.v. i_n with the following distribution:

$$i_n = \begin{cases} 1 & \text{w.p. } \frac{1}{m} \\ . & \\ . & \\ m & \text{w.p. } \frac{1}{m}. \end{cases}$$

SGD would then update the iterate as follows:

$$\theta_{n+1} = \theta_n - a(n) \nabla f_{i_n}(\theta_n). \tag{2.11}$$

Rewriting the above update rule, we obtain

$$\theta_{n+1} = \theta_n - \alpha_n \left(\frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\theta_n) \right) - \alpha_n \left(\nabla f_{i_n}(\theta_n) - \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\theta_n) \right)$$

$$= \theta_n - \alpha_n \left(\frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\theta_n) + w_{n+1} \right),$$

where $\{w_{n+1} = \nabla f_{i_n}(\theta_n) - \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\theta_n)\}$ is a martingale difference sequence because $\mathbb{E}[w_{n+1}|\theta_1, \ldots \theta_n] = 0$.

Several applications involving learning and optimization involve martingale difference noise terms, and the convergence of the stochastic approximation algorithm is tied to whether the effect of underlying noise (martingale difference) can be ignored in the long run. For an introduction to martingales, the reader is referred to Appendix B.

2.2.3 Stochastic gradient algorithm using a zeroth-order oracle

In a zeroth-order setting, the gradient information is not directly available, and instead, the optimization algorithm has oracle access to noise-corrupted function measurements, as illustrated in the figure below.

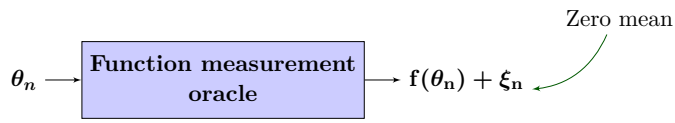


Figure 2.1: Simulation optimization

The stochastic gradient algorithm updates as follows:

$$\theta_{n+1} = \theta_n - a(n)\widehat{\nabla} f(\theta_n), \tag{2.12}$$

where $\widehat{\nabla} f(\theta_n)$ is formed from the function measurements. Two such gradient estimators, using two function measurements, were presented earlier in (1.6) and (1.9), respectively. Such estimates are not unbiased, but feature a parameter that can reduce the bias at the cost of variance. In the next chapter, we present the simultaneous perturbation trick that generalizes the example in (1.9).

Under suitable assumptions, $\theta_n, n \geq 0$, governed by (2.12) can be shown to converge almost surely to the set $\bar{H} = \{x \mid \nabla f(\theta) = 0\}$. We provide this result later in Section 4.1.

2.2.4 Stochastic fixed point iterations

Consider a function $f : \mathbb{R}^d \to \mathbb{R}^d$ that satisfies

$$\|f(x) - f(y)\| \leq \alpha \|x - y\|, \tag{2.13}$$

for any $x, y \in \mathbb{R}^d$. Here $\alpha \in (0, 1)$, and $\|\cdot\|$ is the ℓ_2-norm associated with $\mathbb{R}^d$. Such an f is called a contraction map. Since the underlying space is complete, by the Banach fixed point theorem, there exists a unique fixed point θ^* of the function f.

A first attempt at finding such a fixed point is via the following iterative scheme: start with some $\theta_0 \in \mathbb{R}^d$ and update as

$$\theta_{n+1} = f(\theta_n).$$

A smoothened variation to this update rule is given by

$$\theta_{n+1} = (1 - a(n))\theta_n + a(n)f(\theta_n),$$

where $a(n)$ is the step size. Note that if $\theta_n \to \theta^*$ and f is continuous at θ^*, then $f(\theta^*) = \theta^*$.

So far we have assumed that f is perfectly observable for any given input parameter. However, in many learning scenarios, e.g., reinforcement learning, this isn't the case. In particular, consider the setting where f is not precisely known, but we have black box access to f, as illustrated in Figure 2.1. The simplest noise model would correspond to i.i.d., e.g., $\mathcal{N}(0, 1)$, while a martingale difference noise structure is more general.

For this setting, a stochastic fixed point iteration would update as follows:

$$\theta_{n+1} = (1 - a(n))\theta_n + a(n)(f(\theta_n) + \xi_{n+1}), \tag{2.14}$$

where as described above, a simple setting is where $\{\xi_n\}$ is an i.i.d. sequence with $\mathbb{E}[\xi_n] = 0$, and $\mathbb{E}\|\xi_n\|^2 < \infty$, for all n. Now, it is desirable to have $\theta_n \to \theta^*$ almost surely as $n \to \infty$. From the convergence analysis of stochastic approximation algorithms, to be presented later, we shall see that $\theta_n \to \theta^*$ if (i) f is a contraction, see (2.13); (ii) step sizes satisfy standard stochastic approximation conditions, see (2.6); and (iii) noise ξ_n is a martingale difference sequence that has bounded variance, or satisfies a linear growth condition (see Assumption A2.8 below).

Remark 2.1. The stochastic fixed point iteration algorithm discussed above would not necessarily converge if the modulus of contraction $\alpha = 1$ in (2.13). In this case, a fixed point is not even guaranteed to exist, e.g., consider $f(\theta) = \theta + 1$. Alternatively, more than one fixed point may exist (e.g., $f(\theta) = \theta$), or only one fixed point exists (e.g. $f(\theta) = -\theta$). Under an additional assumption that at least one fixed point exists, the stochastic fixed point iteration (2.14) is guaranteed to converge almost surely to a sample path dependent fixed point solution.

Stochastic fixed point iterations are ubiquitous in the context of reinforcement learning. In particular, the well-known TD-learning and Q-learning algorithms are stochastic fixed-point iterations. The reader is referred to (Bertsekas, 2012; Sutton and Barto, 2018; Bertsekas and Tsitsiklis, 1996) for a detailed introduction to these algorithms.

2.2.5 Linear stochastic approximation

Consider the following stochastic approximation algorithm:

$$\theta_{n+1} = \theta_n + a(n)\left(A_{n+1}\theta_n + b_{n+1}\right),$$

where the step size $a(n)$ satisfies $\sum_n a(n) = \infty$, and $\sum_n a(n)^2 < \infty$. Further, A_n and b_n are matrices and vectors that satisfy

$$\mathbb{E}\left[A_{n+1} \mid \theta_1, \ldots, \theta_n\right] = A, E\left[b_{n+1} \mid \theta_1, \ldots, \theta_n\right] = b,$$

where A is a negative-definite matrix. Moreover, $\mathbb{E}\left[\|(A_n - A)\|^2\right] \leq C_1$ and $\mathbb{E}\left[\|b_n - b\|^2\right] \leq C_2$. In this setting, applying the Kushner-Clark lemma (to be presented later), it can be shown that

$$\theta_n \to \theta^* \text{ a.s. as } n \to \infty,$$

where the limit θ^* satisfies $A\theta^* + b = 0$.

A prominent LSA algorithm is TD-learning with linear function approximation, see (Tsitsiklis and Van Roy, 1997). Other examples include solving a linear regression problem using a stochastic gradient algorithm (Prashanth *et al.*, 2021; Mou *et al.*, 2020), and linear approximations to non-learning SA recursions (Chen *et al.*, 2020b).

2.2.6 Quantile estimation

Consider the following problem, which is a variant of mean estimation. For a continuous random variable (r.v.) X with cumulative distribution function F and for a given $\alpha \in (0,1)$, define

$$q_\alpha(X) = F^{-1}(\alpha).$$

Notice that $q_\alpha(X)$ is the median of the distribution of X when $\alpha = 0.5$.

Let $\{X_n\}_{n\geq 1}$ be a independent sequence of r.v.s with common distribution F. Notice that $F(q_\alpha(X)) = \mathbb{E}[\mathbb{I}\{X \leq q_\alpha(X)\}] = \alpha$. A stochastic approximation algorithm for estimating $q_\alpha(X)$ for a pre-specified α can be arrived at as follows: Let q_n denote an estimate of $q_\alpha(X)$ after observing samples $X_1, \ldots, X_n$. On observing X_{n+1}, q_n is updated as follows:

$$q_{n+1} = q_n + a(n)\left(\mathbb{I}\{X_{n+1} \leq q_n\} - \alpha\right), \tag{2.15}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function, i.e., $\mathbb{I}\{A\} = 1$ if A happens and 0 otherwise.

Notice that the update is iterative, i.e., given an estimate q_n at time instant n and a new sample X_{n+1}, the algorithm should perform an incremental update using q_n, X_{n+1} to arrive at q_{n+1}.

Consider the following alternative observation model: At time instant n, the stochastic approximation algorithm picks a threshold, say T, and the environment returns a Boolean that indicates whether $X_{n+1} < T$ or not. Quantile estimation in this threshold-based model would follow the same iterative scheme as (2.15). To see this, let

$$Y_{n+1} = \begin{cases} 1 & \text{if } X_{n+1} \leq q_n \\ 0 & \text{else.} \end{cases}.$$

Then, the update rule in (2.15) is equivalent to

$$q_{n+1} = q_n + a(n)\left(Y_{n+1} - \alpha\right). \tag{2.16}$$

Using a variant of Kushner Clark lemma, to be presented later, it is possible to establish almost sure convergence of q_n to $q_\alpha(X)$.

In finance literature, a risk measure closely related to quantiles is 'Value at Risk (VaR)'. For any random variable X, we define the VaR at level $\alpha \in (0,1)$ as

$$\text{VaR}_\alpha(X) = \inf\{\xi \mid \mathbb{P}(X \leq \xi) \geq \alpha\}.$$

If the distribution of X is continuous, then VaR is the lowest solution to $\mathbb{P}(X \leq \xi) = \alpha$. VaR as a risk measure has several drawbacks, which precludes using standard stochastic optimization methods. This motivated the definition of coherent risk measures in (Artzner *et al.*, 1999). A risk measure is coherent if it is convex, monotone, positive homogeneous and translation equi-variant. Conditional Value at Risk (CVaR) is a popular risk measure defined by

$$\text{CVaR}_\alpha(X) = \inf_\xi \left\{\xi + \frac{1}{(1-\alpha)}\mathbb{E}(X-\xi)_+\right\}, \tag{2.17}$$

where $(a)_+ = \max(a,0)$ denotes the positive part of a real number a. For a continuous random variable X, it can be shown that

$$\text{CVaR}_\alpha(X) := \mathbb{E}[X | X \geq \text{VaR}_\alpha(X)].$$

Unlike VaR, the above is a coherent risk measure.

A well-known result from (Rockafellar and Uryasev, 2000) is that both VaR and CVaR can be obtained from the solution of a certain convex optimization problem and we recall this result next.

Theorem 2.1. For any random variable X and a $\alpha \in (0,1)$, let

$$v(\xi, X) := \xi + \frac{1}{1-\alpha}(X-\xi)_+ \text{ and } V(\xi) = \mathbb{E}[v(\xi, X)]. \tag{2.18}$$

Then, $\text{VaR}_\alpha(X) = \left(\arg\min V := \{\xi \in \mathbb{R} \mid V'(\xi) = 0\}\right)$, where V' is the derivative of V w.r.t. ξ. Further, $\text{CVaR}_\alpha(X) = V(\text{VaR}_\alpha(X))$.

From the above, it is clear that in order to estimate VaR/CVaR, one needs to find a ξ that satisfies $V'(\xi) = 0$. Stochastic approximation (SA) is a natural tool to use in this situation. Recall that SA is used to solve the equation $h(\theta) = 0$ when analytical form of h is not known. However, noisy measurements $h(\theta_n) + \xi_n$ can be obtained, where $\theta_n, n \geq 0$ are

the input parameters and $\xi_n, n \geq 0$ are zero-mean random variables, that are not necessarily i.i.d.

Using the stochastic approximation principle and the result in Theorem 2.1, we have the following scheme to estimate the VaR/CVaR simultaneously from the samples $\{X_1, \ldots, X_n\}$:

$$\text{VaR: } q_{n+1} = q_n - a(n)(1 - \frac{1}{1-\alpha}\mathbb{I}\{X_{n+1} \geq q_n\}), \tag{2.19}$$

$$\text{CVaR: } \psi_{n+1} = \psi_n - \frac{1}{n+1}\left(\psi_n - v(q_n, X_{n+1})\right). \tag{2.20}$$

In the above, (2.19) can be seen as a gradient descent rule, while (2.20) can be seen as a plain averaging update. Since CVaR estimate depends on the VaR estimate, whereas the converse is not true, the update recursions (2.19)–(2.20) exhibit a one-way coupling, which implies the $1/(n+1)$ step size in (2.20) can be replaced by $a(n)$ for the sake of analysis.

An interesting question is whether the stochastic gradient-based estimation scheme in (2.19) converges faster than the root-finding estimation scheme in (2.15).

2.3 Convergence analysis using the ODE approach

So far, we have provided an introduction to stochastic approximation, and outlined a few popular applications. We now cover preliminary results on the convergence of stochastic approximation algorithms using the limit sets of the associated ordinary differential equation (ODE). In the next section, we provide convergence results with stochastic recursive inclusions, i.e., those algorithms that involve set-valued maps.

Consider now the following recursion:

$$\theta_{n+1} = \theta_n + a(n)(h(\theta_n) + \beta_n + \eta_n). \tag{2.21}$$

Definition 2.1. We denote by $L(\{\theta_n, n \geq 0\})$ the limit set of the sequence $\theta_n, n \geq 0$ obtained from (2.21). In other words, it is the set of all limit points of the sequence $\{\theta_n\}$ obtained from (2.21). Thus, consider all such subsequences $\{n_m\}$ of $\{n\}$ with $n_m \to \infty$ for

which $\theta_{n_m} \to \check{\theta}$ for some $\check{\theta} \in \mathbb{R}^d$. The collection of all these points $\check{\theta}$ obtained as limits of such subsequences $\{\theta_{n_m}\}$ of $\{\theta_n\}$ is defined as $L(\{\theta_n, n \geq 0\})$. Note also that $L(\{\theta_n, n \geq 0\})$ is a sample-path dependent set that can vary in general from one sample path to another.

Consider the following ODE associated with (2.21):

$$\dot{\theta}(t) = h(\theta(t)). \tag{2.22}$$

This is the same ODE as (2.2). Define a sequence $\{t(n), n \geq 0\}$ of time points as follows:

$$t(0) = 0, \; t(n) = \sum_{k=0}^{n-1} a(k), n \geq 1. \tag{2.23}$$

We now state the main result for the convergence of (2.21), see Theorem 1.2 of (Benaïm, 1996)), under the assumptions below.

A2.1. $h : \mathbb{R}^d \to \mathbb{R}^d$ is a Lipschitz continuous function with Lipschitz constant $L > 0$.

A2.2. $\lim_{n\to\infty} \beta_n = 0$ w.p.1.

A2.3. The step sizes satisfy $a(n) > 0$, $\forall n$, $a(n) \to 0$ as $n \to \infty$ and $\sum_n a(n) = \infty$.

A2.4. For each $T > 0$, $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\sup_{j\geq n} \max_{t\leq T} \| \sum_{i=m(jT)}^{m(jT+t)-1} a(i)\eta_i \| \geq \epsilon\right) = 0 \text{ w.p.1},$$

where

$$m(t) = \begin{cases} \max\{n | t(n) \leq t\}, & t \geq 0, \\ 0 & t < 0. \end{cases}$$

A2.5. $\sup_n \|\theta_n\| < \infty$ w.p.1.

A2.6. There exists a locally asymptotically stable attractor $\theta^* \in \mathbb{R}^d$ of the ODE (2.22) with domain of attraction $\breve{\Omega} \subset \mathbb{R}^d$.

We now discuss these assumptions. Assumption A2.1 ensures that the ODE (2.22) is well-posed. Assumption A2.2 ensures that the bias β_n vanishes asymptotically. We shall discuss Assumption A2.3 and Assumption A2.4 in detail below. Assumption A2.6 is satisfied for most gradient systems. This assumption can however be easily relaxed to the case where the attractor is a compact connected set of points instead of being 'isolated'. Theorem 2.2 however takes the form of Theorem 2.3 (a more general result) when one does not have an attractor in the underlying system.

The stability requirement in A2.5, while hard to ensure directly, is common to the analysis of stochastic approximation algorithms. A commonly used procedure to ensure stability is to employ a projection operator onto a large enough compact and convex constraint set that keeps the iterate sequence $\{\theta_n\}$ bounded. One then uses the following update rule in place of (4.7):

$$\theta_{n+1} = \Pi\left(\theta_n + a(n)(h(\theta_n) + \beta_n + \eta_n)\right), \tag{2.24}$$

where Π is a projection operator that keeps the iterates bounded within a compact and convex set, say $\Theta \subset \mathbb{R}^d$. For instance, a computationally inexpensive projection onto $\Theta \stackrel{\triangle}{=} \prod_{i=1}^{d} [\theta^i_{\min}, \theta^i_{\max}]$ can be realized by setting $\Pi_i(\theta) = \min(\max(\theta^i_{\min}, \theta^i), \theta^i_{\max}),\ i \in \{1 \ldots d\}$. If the projected region Θ contains all the attractors of the gradient ODE, then θ_n updated according to (2.24) would likely converge to such an attractor, except that the projection set boundary also introduces spurious attractors, see (Kushner and Yin, 2003). In the case when some of the attractors lie outside the constraint set, the iterate-sequence $\theta_n, n \geq 0$, may get stuck at the boundary of Θ, trying to push forward in the direction of the aforementioned attractors. To avoid the latter situation, one could gradually grow the region of projection as suggested in (Chen *et al.*, 1987), or perform projection infrequently as in (Dalal *et al.*, 2018). In (Yaji and Bhatnagar, 2019), the iterate sequence is reset to a compact set at increasingly sparse instants (in case it goes out of that set) if

the mean field has a globally attracting set. Such a scheme is shown to remain both stable and convergent in (Yaji and Bhatnagar, 2019) with the number of resets remaining finite.

The focus of this book is gradient estimation in a zeroth-order setting, and for the analysis, we assume that the iterates are stable. As discussed above, one could employ a projection operator, to workaround the stability issue, see Section 2.4 for further details. Also, independent of projection, certain verifiable sufficient conditions for stability of stochastic approximations in the literature, cf. (Borkar and Meyn, 2000) and (Abounadi *et al.*, 2002) for two such conditions, and (Ramaswamy and Bhatnagar, 2016a) and (Ramaswamy and Bhatnagar, 2021) for similar conditions in the context of set-valued stochastic approximation.

Motivation for step size assumptions: One can reason about the need for the step size conditions using a simpler noise setting as follows: Suppose $\beta_n = 0, \forall n$ and $\{\eta_n\}$ is an i.i.d. sequence with mean zero and variance σ^2. Then, variance of θ_{n+1} is

$$\begin{aligned}\text{Var}(\theta_{n+1}) &= \text{Var}\left[\theta_n + a(n)h(\theta_n)\right] + a(n)^2 Var(\eta_{n+1}) \\ &= \text{Var}\left[\theta_n + a(n)h(\theta_n)\right] + a(n)^2\sigma^2 \\ &\geq a(n)^2\sigma^2.\end{aligned}$$

If we choose a constant stepsize, i.e., $a(n) = a\ \forall n$, then, $Var(\theta_{n+1}) \geq a^2\sigma^2$. Thus, with a constant step size, $\theta_n \not\longrightarrow \theta^*$ almost surely, motivating the need for having a diminishing step size that vanishes asymptotically. However, such a step size cannot go down too fast, since

$$\theta_{m+1} = \theta_m + a(m)(h(\theta_m) + \eta_{m+1}),$$

$$\|\theta_m - \theta_0\| \leq \sum_{\tau=0}^{m-1} a(\tau)|h(\theta_\tau) + \eta_{\tau+1}|$$

If $|h(\theta_\tau) + \eta_{\tau+1}| \leq C_1$ and $\sum_{\tau=0}^{\infty} a(\tau) \leq C_2 < \infty$, then $\|\theta_m - \theta_0\|$ is bounded above. This implies that θ_m is forced to be within a ball of radius C_1C_2 around the initial point θ_0, for all m. This then puts an artificial constraint on $\{\theta_m\}$ as θ^* can always lie outside this ball. Thus, we need $\sum_{\tau} a(\tau) = \infty$.

Remark 2.2 (Only Diminishing vs. Square Summable Step-Sizes)**.** Assumption A2.3 is a condition on the step-size sequence $\{a(n)\}$ and is weaker than standard Robbins-Monro step-size requirements such as Assumption A2.7 that requires square summability of the step-sizes. However, as Theorems 2.2 and 2.3 suggest, if one makes Assumption A2.3 on the step-size sequence, then one needs to additionally make Assumption A2.4 on the noise sequence $\{\eta_n\}$. Verifying the latter independently may not be straightforward.

On the other hand, assuming the noise sequence $\eta_n, n \geq 0$ is a martingale difference satisfying Assumption A2.8, one can prove that Assumption A2.4 holds under Assumption A2.7 and Assumption A2.5. This is indeed shown in Remark 2.4. As mentioned, this will require the step-size sequence to be square summable, not just asymptotically diminishing. Theorem 2.4 is a variant of Theorem 2.3 that is based on Assumptions A2.7–A2.8 in place of Assumptions A2.3–A2.4, respectively, while continuing with the other assumptions.

The original convergence result of Kushner and Clark, see Theorem 2.3.1 of (Kushner and Clark, 1978), that establishes convergence of (2.21) is the following:

Theorem 2.2 (Kushner and Clark Theorem)**.** Under A2.1–A2.6, outside a set of zero probability, if there is a compact set $A \subset \breve{\Omega}$ such that $\{\theta_n\}$ given by (2.21) satisfies $\theta_n \in A$ infinitely often, then $\theta_n \to \theta^*$ as $n \to \infty$.

We briefly present a proof of this result which follows along the lines of Theorem 2.3.1 of (Kushner and Clark, 1978). A more generalized result is then provided as Theorem 2.3 which is from (Benaïm, 1996) [Theorem 1.2].

Proof. Recall the stochastic recursion (2.21):

$$\theta_{n+1} = \theta_n + a(n)(h(\theta_n) + \beta_n + \eta_n).$$

Let $\theta^0(t), t \geq 0$, denote a continuous linear interpolation of the θ_n

iterates obtained as follows: For $t \in [t(n), t(n+1)]$, $n \geq 0$,

$$\theta^0(t) = \frac{t(n+1)-t}{t(n+1)-t(n)}\theta_n + \frac{t-t(n)}{t(n+1)-t(n)}\theta_{n+1}.$$

Similarly, for t as above, let

$$\beta^0(t) = \frac{t(n+1)-t}{t(n+1)-t(n)}\left(\sum_{i=0}^{n-1} a(i)\beta_i\right) + \frac{t-t(n)}{t(n+1)-t(n)}\left(\sum_{i=0}^{n} a(i)\beta_i\right),$$

$$\eta^0(t) = \frac{t(n+1)-t}{t(n+1)-t(n)}\left(\sum_{i=0}^{n-1} a(i)\eta_i\right) + \frac{t-t(n)}{t(n+1)-t(n)}\left(\sum_{i=0}^{n} a(i)\eta_i\right),$$

respectively. We also define a piecewise constant interpolated process $\bar{\theta}^0(\cdot)$ according to

$$\bar{\theta}^0(t) = \theta_n, \ \theta \in [t(n), t(n+1)).$$

Then the recursion (2.21) can be written in continuous time as

$$\theta^0(t) = \theta^0(0) + \int_0^t h(\bar{\theta}^0(\tau))d\tau + \beta^0(t) + \eta^0(t), \ t \geq 0. \tag{2.25}$$

From these continuous-time functions, we define a sequence of left-shifted functions $\theta^n(\cdot), \beta^n(\cdot), \eta^n(\cdot)$ as follows: For $n \geq 0$,

$$\theta^n(t) = \begin{cases} \theta^0(t+t(n)), & t \geq -t(n) \\ \theta_0, & t \leq -t(n) \end{cases}$$

$$\eta^n(t) = \begin{cases} \eta^0(t+t(n)) - \eta^0(t(n)), & t \geq -t(n) \\ -\eta^0(t(n)), & t \leq -t(n) \end{cases}$$

$$\beta^n(t) = \begin{cases} \beta^0(t+t(n)) - \beta^0(t(n)), & t \geq -t(n) \\ -\beta^0(t(n)), & t \leq -t(n) \end{cases}$$

respectively.

Before proceeding further, we show that under Assumption A2.3 and Assumption A2.4, $\eta^0(\cdot)$ is uniformly continuous on $[0, \infty)$ almost surely. Further, for any $0 < T < \infty$,

$$\lim_{t\to\infty} \sup_{|s|\leq T} \|\eta^0(t+s) - \eta^0(t)\| = 0 \text{ w.p.1.}$$

By Assumption A2.4, given $\epsilon > 0$, there exists $n_k > 0$ such that

$$P\left(\sup_{j\geq n_k}\max_{t\leq T}\|\sum_{i=m(jT)}^{m(jT+t)-1}a(i)\eta_i\|\geq\epsilon\right)\leq\frac{1}{2^k}.$$

Thus,

$$\sum_k P\left(\sup_{j\geq n_k}\max_{t\leq T}\|\sum_{i=m(jT)}^{m(jT+t)-1}a(i)\eta_i\|\geq\epsilon\right)<\infty.$$

Thus, corresponding to $\{n_k\}$, we get a sequence of events $\{E_k\}$ where

$$E_k=\{\sup_{j\geq n_k}\max_{t\leq T}\|\sum_{i=m(jT)}^{m(jT+t)-1}a(i)\eta_i\|\geq\epsilon\}.$$

By the Borel-Cantelli lemma, $P(E_k \text{ infinitely often}) = 0$. Thus,

$$\sup_{\{|s|\leq T, t\geq n_k\}}\|\eta^0(t+s)-\eta^0(t)\|<\epsilon,$$

for all but finite number of n_k (integers) w.p.1. Since $\eta^0(\cdot)$ is continuous w.p.1 on $[0,\infty)$, the above implies that $\eta^0(\cdot)$ is also uniformly continuous w.p.1. Thus, $\{\eta^n(\cdot)$ is uniformly continuous on $\mathbb{R}$, bounded on compacts and $\eta^n(\cdot)\to 0$ w.p.1 uniformly on compacts in $\mathbb{R}$. Likewise, from Assumption A2.2, $\{\beta^n(\cdot)\}$ is uniformly continuous on $\mathbb{R}$, bounded on compacts and $\beta^n(\cdot)\to 0$ w.p.1 uniformly on compacts in $\mathbb{R}$.

Now, (2.25) can be equivalently written as follows: For $t\geq 0$,

$$\theta^n(t)=\theta^n(0)+\int_0^t h(\bar{\theta}^0(t(n)+\tau))d\tau+\beta^n(t)+\eta^n(t)$$

$$=\theta^n(0)+\int_0^t h(\theta^n(\tau))d\tau+\epsilon^n(t)+\beta^n(t)+\eta^n(t), \tag{2.26}$$

where

$$\epsilon^n(t)=\int_0^t h(\bar{\theta}^0(t(n)+\tau))d\tau-\int_0^t h(\theta^n(\tau))d\tau.$$

Note that by Lipschitz continuity of $h(\cdot)$ (cf. Assumption A2.1),

$$\|\epsilon^n(t)\|\leq L\int_0^t\|\bar{\theta}^0(t(n)+\tau))-\theta^n(\tau))\|d\tau, \tag{2.27}$$

where $L > 0$ is the Lipschitz constant of the function $h(\cdot)$. Now, observe that

$$\theta^n(t) = \theta^0(t+t(n)) = \bar{\theta}^0(t+t(n)) + \int_0^t h(\bar{\theta}^0(t(n)+\tau))d\tau + \beta^n(t) + \eta^n(t).$$

Thus,

$$\|\theta^n(t) - \theta^0(t+t(n))\| \leq \int_0^t \|h(\bar{\theta}^0(t(n)+\tau))\|d\tau + \|\beta^n(t)\| + \|\eta^n(t)\|. \tag{2.28}$$

Now, by Lipschitz continuity of $h(\cdot)$,

$$\begin{aligned}\|h(\bar{\theta}^0(t(n)+\tau))\| - \|h(0)\| &\leq \|h(\bar{\theta}^0(t(n)+\tau)) - h(0)\| \\ &\leq L\|\bar{\theta}^0(t(n)+\tau)\|.\end{aligned}$$

Thus, with $\check{L} = \max(L, \|h(0)\|)$, we get that

$$\|h(\bar{\theta}^0(t(n)+\tau))\| \leq \check{L}(1 + \|\bar{\theta}^0(t(n)+\tau)\|).$$

Since, outside a set of zero probability, $\exists \check{M} > 0$ such that $\|\bar{\theta}^0(t(n)+\tau)\| \leq \check{M}$. Thus,

$$\|h(\bar{\theta}^0(t(n)+\tau))\| \leq \check{K},$$

where $\check{K} \stackrel{\triangle}{=} \check{L}(1+\check{M}) > 0$. Thus, from (2.28), it follows that

$$\|\theta^n(t) - \theta^0(t+t(n))\| \leq a(n)\check{K} + \|\beta^n(t)\| + \|\eta^n(t)\|.$$

The RHS above $\to 0$ as $n \to \infty$ uniformly on compact intervals. Substituting the above inequality in (2.27), one obtains

$$\|\epsilon^n(t)\| \leq La(n)(a(n)\check{K} + \|\beta^n(t)\| + \|\eta^n(t)\|) \to 0,$$

as $n \to \infty$ uniformly on compact intervals. Thus, $(\epsilon^n(t) + \beta^n(t) + \eta^n(t)) \to 0$ as $n \to \infty$ uniformly on compact intervals. From Assumption A2.5, $\{X^n(\cdot)\}$ is bounded and further it is easy to observe that this sequence is equicontinuous. From the Arzela-Ascoli theorem, it then follows that $\{\Theta^n(\cdot)\}$ is relatively compact. Thus, there exists a convergent subsequence that we continue to call $\{\theta^n(\cdot)\}$ itself without loss of generality. Let $\theta(\cdot)$ be the limiting function of this sequence. Then $\theta(\cdot)$ can be seen to satisfy the limiting ODE (2.22) as

$$\theta(t) = \theta(0) + \int_0^t h(\theta(\tau))d\tau,$$

which is the integral form of the ODE (2.22).

Now note that under Assumption A2.6, $\theta^* \in \mathbb{R}^d$ is an attractor for the ODE (2.22). Let $\epsilon_1, \epsilon_2 > 0$ be two scalars with $\epsilon_1 < \epsilon_2$ with ϵ_1 being small in particular. Then the ϵ_1 and ϵ_2 neighborhoods of θ^* satisfy $N_{\epsilon_1}(\theta^*) \subset N_{\epsilon_2}(\theta^*)$ and let $N_{\epsilon_2}(\theta^*) \subset A$. Since $\theta_n \in A$ infinitely often, it follows that there exists a subsequence $\{n_m\}$ of $\{n\}$ such that $\theta_{n_m} \in A, \forall n_m$. Consider then the process $\theta^{n_m}(\cdot)$ which will have a subsequence (also indexed by $\{n_m\}$ for simplicity) that will converge to a limit $\hat{\theta}(\cdot)$ that in turn will satisfy the ODE (2.22). Since $\hat{\theta}(0) \in A$ and θ^* is asymptotically stable, $\hat{\theta}(t) \to \theta^*$ as $t \to \infty$.

Consider again the process $\theta^{n_m}(\cdot)$ formed from the stochastic iterates. Since $\theta^{n_m}(\cdot) \to \hat{\theta}(\cdot)$ uniformly on compacts and $\hat{\theta}(t) \to \theta^*$, it follows that there is a subsequence $\{\theta_{n_{m_j}}\}$ of $\{\theta_{n_m}\}$ that will be contained in $N_{\epsilon_1}(\theta^*)$. However, we know that $\{\theta_{n_m}\}$ is entirely contained in A. Suppose then that there is a subsequence $\{\theta_{n_{mk}}$ of $\{\theta_{n_m}\}$ that is entirely contained in $A \backslash N_{\epsilon_2}(\theta^*)$, i.e., $A \cap N^c_{\epsilon_2}(\theta^*)$. Then $\{\theta_{n_m}\}$ will move from $N_{\epsilon_1(\theta^*)}$ to $A \backslash N_{\epsilon_2}(\theta^*)$ and back infinitely often since there are an infinite number of points in each of these sets. Then there is a sequence of time points $\tau_1 < \bar{\tau}_1 < \tau_2 < \bar{\tau}_2 \cdots$ such that $\theta^0(\tau_j) \in \partial N_{\epsilon_1}(\theta^*)$ and $\theta^0(\bar{\tau}_j) \in \partial N_{\epsilon_2}(\theta^*)$, $\forall j$. Further, $\theta^0(t) \in \bar{N}_{\epsilon_2}(\theta^*) \backslash N_{\epsilon_1}(\theta^*)$, for $t \in (\tau_j, \bar{\tau}_j)$ for all j. Consider the $[\tau_j, \bar{\tau}_j]$ portions of the trajectory $\theta^0(\cdot)$. This sequence will have a convergent subsequence whose limit is say $\tilde{\theta}(\cdot)$ which again satisfies (2.22). Consider two cases: (i) There is a $T > 0$ such that along a subsequence $\bar{\tau}_j - \tau_j \to T$. Then, $\tilde{\theta}(0) \in \partial N_{\epsilon_1}(\theta^*)$ and $\tilde{\theta}(T) \in \partial N_{\epsilon_2}(\theta^*)$. This is not possible by asymptotic stability of θ^* since $\epsilon_1 > 0$ is small. (ii) Let $r_j - l_j \to \infty$. Then the set of $\{[l_j, \infty)\}$ segments of $\theta^0(\cdot)$ are bounded and equicontinuous. Again by the Arzela-Ascoli theorem, one can obtain a convergent subsequence with limit say $\breve{\theta}(\cdot)$ that again satisfies (2.22). Then $\breve{\theta}(0) \in \partial N_{\epsilon_1}(\theta^*)$ and $\breve{\theta}(t) \in \bar{N}_{\epsilon_2}(\theta^*) \backslash N_{\epsilon_1}(\theta^*)$. This contradicts that θ^* is asymptotically stable. The claim follows. □

Remark 2.3. A more formal argument on the tracking of the iterate sequence to the underlying ODE (2.22) is provided in Chapter 2 of Borkar, 2022. We briefly sketch that argument here for completeness.

Consider $\{t(n)\}$ as in (2.23) and let $T > 0$ be a given time element

and define a sequence of time points $\{T_n\}$ as follows: Let $T_0 = t(0) = 0$. Further, for $n \geq 1$, let

$$T_n = \min\{t(m) | t(m) \geq T_{n-1} + T\},$$

denote a sequence of time points. Let $\theta^{T_n}(t), t \geq T_n$ denote the solution to the ODE (2.22) with $\theta^{T_n}(T_n) = \theta^0(T_n)$ as the initial condition of the ODE. It is argued in Lemma 1, Chapter 2, of Borkar, 2022, using an application of the Gronwall's inequality (see Lemma A.1), that

$$\lim_{n\to\infty} \max_{t\in[T_n, T_{n+1}]} \|\theta^0(t) - \theta^{T_n}(t)\| = 0,$$

almost surely. In fact, the above holds for any time point $s \in \mathbb{R}$ (in positive and negative time), not just the time instants T_n above. Now if the ODE has a globally asymptotically stable attractor A, any trajectory of the ODE (2.22) will eventually converge to it, and so will the interpolated iterates $\theta^0(t)$, and thereby the iterate sequence $\theta_n, n \geq 0$. Figure 2.2 illustrates this iterate-tracking process.

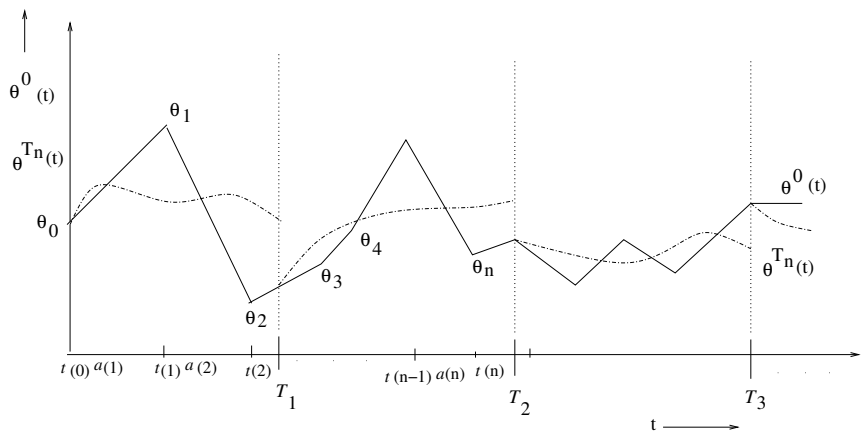


Figure 2.2: The continuously interpolated algorithm's trajectory $\theta^0(t)$ represented by the solid line asymptotically tracks the ODE's trajectory (the dashed-dotted line) $\theta^{T_n}(t)$ suitably reset to the algorithm's trajectory after every (regular) time interval approximately T instants long. On the X-axis are the instants $t(0), t(1), \cdots$, with $t(n) - t(n-1) = a(n)$, $\forall n$ with $t(0) = 0$. From the step size conditions, it follows that $t(n) \to \infty$ as $n \to \infty$. This ensures that the algorithm does not converge prematurely.

Theorem 2.3 (A More General Kushner and Clark Theorem)**.** Under A2.1–A2.5, $\{\theta_n\}$ governed according to (2.21) converges almost surely to $L(\{\theta_n, n \geq 0\})$ (see Definition 2.1). Further,$L(\{\theta_n, n \geq 0\})$ is a connected internally chain recurrent set[2] for the ODE (2.22).

This result is a generalization of the Kushner and Clark lemma (cf. (Kushner and Clark, 1978)) and is stated under the same assumptions as used in the aforementioned result.

We now state some alternative assumptions that in fact we shall use for our analysis.

A2.7. $a(n) > 0, \forall n, \ \sum_n a(n) = \infty$ and $\sum_n a(n)^2 < \infty$.

A2.8. $\{\eta_n\}$ is a square integrable martingale difference sequence with respect to the filtration $\{\mathcal{F}_n\}$, with $\mathcal{F}_n = \sigma(\theta_m, \beta_m, m \leq n, \eta_m, m < n)$, $n \geq 0$. Further,

$$\mathbb{E}[\|\eta_{n+1}\|^2 \mid \mathcal{F}_n] \leq C_0(1 + \|\theta_n\|^2), \ n \geq 0.$$

Remark 2.4. As discussed in Remark 2.2, Assumption A2.7 is stronger than Assumption A2.3. However, Assumptions A2.7 and A2.8, in addition to A2.5 turn out to be sufficient conditions for the verification of Assumption A2.4. This can be seen as follows: Let

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

Then, from Assumption A2.8, it will follow that $(\chi_n, \mathcal{F}_n), n \geq 0$ is a martingale sequence. Moreover,

$$E\left[\sum_n \|\chi_{n+1} - \chi_n\|^2 \mid \mathcal{F}_n\right]$$
$$= E\left[\sum_n a(n)^2\|\eta_n\|^2 \mid \mathcal{F}_n\right]$$

[2]See Appendix A for the definition of internally chain recurrent set of an ODE and other related concepts.

$$\leq \sum_n a(n)^2 C_0 \left(1 + \|\theta_n\|^2\right) \qquad \text{(by Assumption A2.8)}$$

$$< \infty \text{ a.s.} \qquad \text{(by Assumption A2.5.)}$$

Thus the quadratic variation process associated with the martingale $\{\chi_n\}$ is almost surely convergent. Hence, by the martingale convergence theorem for square integrable martingales, see Theorem B.7 in Appendix B, $\{\chi_n\}$ itself is almost surely convergent. Assumption A2.4 will thus follow.

We now present the generalized form of the Kushner-Clark theorem for convergence of algorithms of the form (2.21), where we also consider the case when $h(\theta) = -\nabla f(\theta)$, for a continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$. This result is stated under Assumptions A2.1, A2.2, A2.7, A2.5 and A2.8 and will be used for the analysis of our gradient search algorithms. In this case, the recursion (2.21) takes the form

$$\theta_{n+1} = \theta_n + a(n)(-\nabla f(\theta_n) + \beta_n + \eta_n), \ n \geq 0. \tag{2.29}$$

The ODE associated with (2.29) is the following:

$$\dot{\theta}(t) = -\nabla f(\theta(t)). \tag{2.30}$$

Theorem 2.4. (i) The recursion (2.21), under Assumptions A2.1, A2.2, A2.5, A2.7, A2.8, converges almost surely to $L(\{\theta_n, n \geq 0\})$, see Definition 2.1. Further, $L(\{\theta_n, n \geq 0\})$ is a connected internally chain recurrent set for the ODE (2.22).

(ii) Part (i) continues to hold for the case of (2.29) where $h(\theta) = -\nabla f(\theta)$ for a continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ and with the ODE (2.30) in place of (2.22). Further, $L(\{\theta_n, n \geq 0\}) \subset H \stackrel{\triangle}{=} \{\theta \mid \nabla f(\theta) = 0\}$.

Remark 2.5. (i) As mentioned previously, Theorem 2.4(i) is similar to Theorem 2.3 except that it is obtained under more directly verifiable noise condition in Assumption A2.8 as opposed to AssumptionA2.4 and under step-size Assumption A2.7 in place of Assumption A2.3.

(ii) In the case of stochastic gradient recursions as in (2.29), one can claim a stronger result, see Theorem 2.4(ii). In this case, $H = \{\theta | \nabla f(\theta) = 0\}$ denotes the set of all equilibria of the ODE (2.30), for which $V(\theta) = f(\theta)$ serves as a Lyapunov function since

$$\begin{aligned} \frac{dV(\theta)}{dt} &= \langle \nabla V(\theta), \dot{\theta} \rangle \\ &= \langle \nabla V(\theta), -\nabla V(\theta) \rangle \\ &\leq 0, \ \forall \theta \in \mathbb{R}^d. \end{aligned}$$

In particular, $\frac{dV(\theta)}{dt} < 0$, $\forall \theta \notin H$ and $\frac{dV(\theta)}{dt} = 0$ otherwise. Finally, Theorem 2.4(ii) is similar to Corollary 2.1 of (Borkar, 2022) even though the latter is stated for the case of a general recursion (not necessarily of the gradient type) but where a Lyapunov function exists for an ODE such as (2.22).

(iii) Note also that in the case of (2.29), if $L(\{\theta_n, n \geq 0\})$ comprises of only isolated limit points, then by Theorem 2.4(ii), these limit points of the algorithm constitute isolated equilibria of the ODE (2.30), and $\theta_n, n \geq 0$ will converge almost surely to a possibly sample path dependent equilibrium, see Corollary 2.2 of (Borkar, 2022).

Remark 2.6. Stability of stochastic approximation, i.e., Assumption A2.5, is one of the strongest requirements to ensure convergence of the stochastic iterates. Various sets of sufficient conditions to ensure stability of the stochastic iterates can be found in (Kushner and Yin, 2003; Borkar and Meyn, 2000; Abounadi *et al.*, 2002; Tsitsiklis, 1994) and other references.

2.4 Projected Stochastic Approximation

There are many practical situations where it is difficult to verify sufficient conditions for stability (as in Remark 2.6) of the stochastic recursions. In such scenarios, a popular approach is to enforce stability on the stochastic iterates by selecting a convex and compact set in which the parameter iterates can take values and thereafter projecting the iterates

to the aforementioned set whenever the iterates escape from the same. This approach also helps in situations where the parameter takes values only in a pre-specified compact set. Stability of the iterates is then enforced due to the projection.

We review here an important result originally due to Kushner and Clark (cf. Theorem 5.3.1 on pp. 191-196 of (Kushner and Clark, 1978)) that shows the convergence of projected stochastic approximations. While the result, as stated in (Kushner and Clark, 1978), is more generally applicable, we present its adaptation here that is relevant to the setting that we consider.

Let $C \subset \mathcal{R}^d$ be a compact and convex set and $\Gamma : \mathcal{R}^d \to C$ denote a projection operator that projects any $\theta = (\theta_1, \ldots, \theta_d)^T \in \mathcal{R}^d$ to its nearest point in C. Thus, if $\theta \in C$, then $\Gamma(\theta) \in C$ as well. For instance, if C is a d-dimensional rectangle having the form $C = \prod_{i=1}^{d} [a_{i,\min}, a_{i,\max}]$, where $-\infty < a_{i,\min} < a_{i,\max} < \infty$, $\forall i = 1, \ldots, d$, then a convenient way to identify $\Gamma(\theta)$ is according to $\Gamma(\theta) = (\Gamma_1(\theta_1), \ldots, \Gamma_N(\theta_d))^T$, where the individual operators $\Gamma_i : \mathcal{R} \to \mathcal{R}$ are defined by

$$\Gamma_i(\theta_i) = \min(a_{i,\max}, \max(a_{i,\min}, \theta_i)), \ i = 1, \ldots, d.$$

Let $\mathcal{C}(C)$ denote the space of all continuous functions from C to $\mathcal{R}^d$.

Consider the following d-dimensional stochastic recursion:

$$\theta_{n+1} = \Gamma(\theta_n + a(n)(h(\theta_n) + \xi_n + \beta_n)), \tag{2.31}$$

under the assumptions listed below.

Consider now the following ODE associated with (2.31):

$$\dot{\theta}(t) = \bar{\Gamma}(h(\theta(t))). \tag{2.32}$$

Here, $\bar{\Gamma} : \mathcal{C}(C) \to \mathcal{C}(\mathcal{R}^d)$ is defined according to

$$\bar{\Gamma}(v(\theta)) = \lim_{\eta \to 0} \left(\frac{\Gamma(\theta + \eta v(\theta)) - \theta}{\eta} \right), \tag{2.33}$$

for any continuous $v : C \to \mathcal{R}^d$. The limit in (2.33) exists and is unique since C is a convex set. In case C is not convex, the limit $\bar{\Gamma}(v(\theta))$ in (2.33) will not be unique in general for all θ and so $\bar{\Gamma}(h(\theta(t)))$ will be

a set of points for any $\theta(t)$, that is not necessarily a singleton, and so instead of the ODE (2.32), one may consider the following differential inclusion:

$$\dot{\theta}(t) \in \bar{\Gamma}(h(\theta(t))). \tag{2.34}$$

A similar result as below can then be seen to hold in this case. For simplicity, we shall restrict our attention to the case where C is a compact and convex set.

From the definition of $\bar{\Gamma}$ in (2.33), note that $\bar{\Gamma}(v(\theta)) = v(\theta)$ if $\theta \in C^o$ (the interior of C). This is because for such a θ, one can find $\eta > 0$ sufficiently small so that $\theta + \eta v(\theta) \in C^o$ as well and hence $\Gamma(\theta + \eta v(\theta)) = \theta + \eta v(\theta)$. On the other hand, if $\theta \in \partial C$ (the boundary of C) is such that $\theta + \eta v(\theta) \notin C$, for any small $\eta > 0$, then $\bar{\Gamma}(v(\theta))$ is the projection of $v(\theta)$ to the tangent space of ∂C at θ.

Consider now the assumptions listed below.

A2.9. The function $h : \mathcal{R}^d \to \mathcal{R}^d$ is continuous.

A2.10. The step sizes $a(n), n \geq 0$ satisfy

$$a(n) > 0 \ \forall n, \ \sum_n a(n) = \infty, \ \sum_n a(n)^2 < \infty.$$

A2.11. The sequence $\beta_n, n \geq 0$ is a bounded random sequence with $\beta_n \to 0$ almost surely as $n \to \infty$.

A2.12. $\{\eta_n\}$ is a square integrable martingale difference sequence with respect to the filtration $\{\mathcal{F}_n\}$, with $\mathcal{F}_n = \sigma(\theta_m, \beta_m, m \leq n, \eta_m, m < n)$, $n \geq 0$. Further,

$$\mathbb{E}[\|\eta_{n+1}\|^2 \mid \mathcal{F}_n] \leq C_0(1 + \|\theta_n\|^2), \ n \geq 0.$$

Let $K \subset \mathcal{R}^d$ denote the set of asymptotically stable attractors of (2.32). Then, (Kushner and Clark, 1978, Theorem 5.3.1 (pp. 191-196)) essentially says the following:

Theorem 2.5 (Kushner and Clark Theorem - Projected case)**.** Under

Assumptions A2.9–A2.12, almost surely, $X_n \to K$ as $n \to \infty$.

Remark 2.7. We wish to point out that the original theorem of Kushner and Clark (cited above) for the case of projected stochastic approximations is stated for the case of an analogous assumption as Assumption A2.4 in place of A2.12 and Assumption A2.3 in place of A2.10. As discussed in Remark 2.4, the noise assumption A2.12 in conjunction with A2.10 (and the fact that now the iterates are uniformly bounded throughout because of the projection), imply A2.4. Moreover, these assumptions are more easily verifiable in most applications.

2.5 Stochastic Recursive Inclusions

In many applications, one encounters set-valued maps $h(\theta)$, $\theta \in \mathbb{R}^d$ in place of point-to-point maps $h(\theta)$, for instance, resulting from dealing with partial observation settings. Let $h : \mathbb{R}^d \to \{\text{set of subsets of } \mathbb{R}^d\}$. A stochastic recursive inclusion has the following structure:

$$\theta_{n+1} - \theta_n - a(n)M_{n+1} \in a(n)h(\theta_n), \tag{2.35}$$

where $(M_n, \mathcal{F}_n), n \geq 0$, is a martingale difference sequence. Consider now the associated differential inclusion (DI):

$$\dot{\theta}(t) \in h(\theta(t)). \tag{2.36}$$

Let $t(n), n \geq 0$ be a sequence of time points defined as follows: $t(0) = 0$ and for $n \geq 1$, $t(n) = \sum_{k=0}^{n-1} a(k)$. Thus, $t(n+1) = t(n) + a(n)$. For any $t \geq 0$, let $m(t) \stackrel{\triangle}{=} \sup\{k \geq 0 \mid t \geq t(k)\}$. Define a continuous time affine interpolated process $W : [0, \infty) \to \mathbb{R}^d$ as follows:

$$W(t(n) + s) = \theta_n + s\left(\frac{\theta_{n+1} - \theta_n}{a(n)}\right), \ s \in [0, a(n)].$$

From the above, $W(t(n)) = \theta_n, \forall n$. Recall Definition A.12 for definition of a perturbed solution to a DI. The following result is from (Benaïm *et al.*, 2005, Proposition 1.3).

Proposition 2.1. Assume the following hold:

(i) For all $T > 0$,

$$\lim_{n\to\infty} \sup\{\| \sum_{k=n}^{l-1} a(k)M_{k+1}\| \mid k = n+1, \ldots, m(t(n)+T)\} = 0.$$

(ii) $\sup_n \|\theta_n\| < \infty$ almost surely.

Then the process $W(\cdot)$ is a perturbed solution of the DI (2.36).

Consider now the assumptions A2.7-A2.8 with $\eta_n = M_{n+1}, n \geq 0$ as the martingale difference sequence. Assume also the stability requirement on the iterates (2.35).

A2.13. The iterates (2.35) satisfy $\sup_n \|\theta_n\| < \infty$ almost surely.

Let $\zeta(n) = \sum_{m=0}^{n-1} a(m)M_{m+1}$, $n \geq 1$. Then $(\zeta(n), \mathcal{F}_n), n \geq 1$ can be seen to be a martingale sequence. From Assumptions A2.8 and A2.13, it can be seen that the quadratic variation process of the martingale $\{\zeta(n)\}$ converges almost surely, and by the martingale convergence theorem, the martingale itself converges almost surely. It is then clear that the requirement (i) in Proposition 2.1 is satisfied. Together with Assumption A2.13, it implies from Proposition 2.1 that the process $W(\cdot)$ is a bounded perturbed solution to the DI (2.36). Recall the definition of internally chain transitive sets of a DI (cf. Definition A.11). We have the following main result from (Benaïm *et al.*, 2005, Theorem 3.6).

Theorem 2.6. The limit set of $W(\cdot)$, the continuous time affine interpolated process obtained from the stochastic recursion (2.35) with $W(0) = z$, given by $L(z) = \bigcap_{t\geq 0} \overline{\{W[t, +\infty)\}}$, is internally chain transitive for the DI (2.36).

2.6 Stochastic Approximation with Markov Noise

An important setting not previously considered thus far in this text is of Markov noise in addition to the martingale difference noise sequence

when considering the stochastic iterates. Such a setting arises in the case of problems of optimization and control when data becomes available online one at a time in real time as well as in reinforcement learning with online updates. The results here are based on (Borkar, 2022; Ramaswamy and Bhatnagar, 2019). Consider the following update of the θ-parameter:

$$\theta_{n+1} = \theta_n + a(n)\left(h(\theta_n, X(n)) + M_{n+1}\right), \tag{2.37}$$

where $X(n), n \geq 0$ is the sequence of random variables characterizing Markov noise. Let $\check{S}$ denote the set of states for $\{X(n)\}$. Also, let $\mathcal{F}_n = \sigma(\theta(m), X(m), M_m, m \leq n)$, $n \geq 0$. We let

$$P(X(n+1) = j \mid \mathcal{F}_n) = p_{\theta_n}(X(n), j) \text{ a.s.},$$

where $p_{\theta_n}(\cdot, \cdot)$ are the transition probabilities that depend on the parameter iterates $\theta_n, n \geq 0$.

Consider now a sequence $\{t(n)\}$ of time points defined as before, i.e., $t(0) = 0$, $t(n) = \sum_{k=0}^{n-1} a(k)$, $n \geq 1$. Now define the algorithm's trajectory $\bar{\theta}(t)$ according to: $\bar{\theta}(t(n)) = \theta_n, \forall n$, and with $\bar{\theta}(t)$ defined as a continuous linear interpolation on each of the intervals $[t(n), t(n+1)]$.

Consider now the following assumptions:

A2.14. $h : \mathbb{R}^d \times \check{S} \to \mathbb{R}^d$ is Lipschitz continuous in the first argument, uniformly with respect to the second.

A2.15. For any given $\theta \in \mathbb{R}^d$, the set $D(\theta)$ of ergodic occupation measures of $\{X(n)\}$ is compact and convex.

A2.16. $\{M_n\}_{n \geq 0}$ is a square-integrable martingale difference sequence. Further, $\mathbb{E}\left[||M_{n+1}||^2 | \mathcal{F}_n\right] \leq K(1 + ||\theta_n||^2)$.

A2.17. The step size sequence $\{a(n)\}$ satisfies $a(n) > 0, \forall n$. Further, $\sum_{n=0}^{\infty} a(n) = \infty$ and $\sum_{n=0}^{\infty} a^2(n) < \infty$.

A2.18. Let $\tilde{h}(\theta, \nu) = \int h(\theta, x)\nu(dx)$, where $\nu \in D(\theta)$. Also, define a sequence of scaled functions $\tilde{h}_c(\theta, \nu) = \dfrac{\tilde{h}(c\theta, \nu(c\theta))}{c}$, $c \geq 1$.

(i) The limit $\tilde{h}_\infty(\theta,\nu) \stackrel{\triangle}{=} \lim_{c\to\infty} \tilde{h}_c(\theta,\nu)$ exists uniformly on compacts.

(ii) There exists an attracting set $\mathcal{A}$ associated with the DI $\dot{\theta}(t) \in H(\theta(t))$ where $H(\theta) = \bar{co}(\{\tilde{h}_\infty(\theta,\nu) : \nu \in D(\theta)\})$ such that $\sup_{u\in\mathcal{A}} ||u|| < 1$ and $\bar{B}_1(0) \stackrel{\triangle}{=} \{x \mid ||x|| \le 1\}$ is a fundamental neighborhood of $\mathcal{A}$.

Theorem 2.7. Under A2.14–A2.18, $\{\bar{\theta}(s+\cdot), s \ge 0\}$ remains uniformly bounded with probablity one and converges to an internally chain transitive invariant set of the DI

$$\dot{\theta}(t) \in \hat{h}(\theta(t)),$$

where $\hat{h}(\theta) = \{\tilde{h}(\theta,\nu) \mid \nu \in D(\theta)\}$. In particular, $\{\theta_t\}$ converges almost surely to such a set.

Example 2.1. We present here a simple example as an application to Theorem 2.7. The temporal difference (TD) learning algorithm in reinforcement learning (Sutton and Barto, 2018; Bertsekas and Tsitsiklis, 1996) has a similar structure as considered in this example. Consider a Markov chain $\{X(n)\}$ taking values in a set S (the state space) assumed finite for simplicity. Assume $\{X(n)\}$ is a given ergodic Markov process that does not depend on the parameter θ. Let ν denote the unique stationary distribution of $\{X(n)\}$. Consider now the following update of the parameter θ:

$$\theta_{n+1} = \theta_n + a(n)(A(X(n))\theta_n + b(X(n))), \tag{2.38}$$

where $A(X(n))$ for any $n \ge 0$ is a $d \times d$ matrix and $b(X(n)) \in \mathbb{R}^d$ is an d-dimensional vector. Further, suppose the step-size sequence $\{a(n)\}$ satisfies Assumption A2.17. Let

$$\bar{A} = \sum_{i\in S} A(i)\nu(i) \text{ and } \bar{b} = \sum_{i\in S} b(i)\nu(i).$$

Assume now that $\bar{A}$ is negative definite. In the setting of Theorem 2.7,

$$h(\theta, X) = A(X)\theta + b(X),$$

that is easily seen to satisfy Assumption A2.14. Since $\{X(n)\}$ is ergodic Markov, $D(\theta) = \{\nu\}$, a singleton set with ν independent of θ. Thus, Assumption A2.15 is trivially satisfied. Now note that in recursion (2.38), we do not have an explicit martingale difference noise term. Thus, one may let $M_{n+1} \equiv 0$ here for all n. Thus, Assumption A2.16 is trivially satisfied as well. We assume here that the step-sizes $\{a(n)\}$ above satisfy the standard Robbins-Monro conditions given in (A2.17). Now, as before, let

$$\tilde{h}(\theta, \nu) = \sum_i h(\theta, i)\nu(i) = \sum_i (A(i)\theta + b(i))\nu(i) = \bar{A}\theta + \bar{b}.$$

Again, let

$$\tilde{h}_c(\theta, \nu) = \frac{\tilde{h}(c\theta, \nu)}{c} = \bar{A}\theta + \frac{\bar{b}}{c}.$$

Now,

$$\tilde{h}_\infty(\theta, \nu) \stackrel{\triangle}{=} \lim_{c\to\infty} \tilde{h}_c(\theta, \nu) = \bar{A}\theta.$$

Note now that the set-valued map $H(\theta)$ in Theorem 2.7 takes the form $H(\theta) = \{\bar{A}\theta\}$, a singleton. Then the DI $\dot{\theta}(t) \in H(\theta(t))$ is actually the ODE $\dot{\theta}(t) = \bar{A}\theta(t)$. Let $V(\theta) = \frac{1}{2}\theta^T \bar{A}^T \bar{A}\theta$. It can be seen that $V(\theta)$ is a Lyapunov function for the above ODE since

$$\frac{dV(\theta)}{dt} = \nabla V(\theta)^T \dot{\theta} = \theta^T \bar{A}^T \bar{A}\bar{A}\theta = (\bar{A}\theta)^T \bar{A}(\bar{A}\theta)$$

Thus,

$$\frac{dV(\theta)}{dt} = \begin{cases} < 0 & \text{if } \theta \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The strict inequality above follows because $\bar{A}$ is negative definite and whereby $\bar{A}$ is also a full rank matrix. Thus, $\dot{\theta}(t) = \bar{A}\theta(t)$ has the origin as its unique globally asymptotically stable attractor with the unit ball $\bar{B}_1(0) = \{\theta | \|\theta\| \leq 1\}$ as the fundamental neighborhood of this attractor (i.e., the origin). Thus Assumption A2.18 holds as well.

Consider now the ODE

$$\dot{\theta}(t) = \bar{A}\theta + \bar{b}.$$

This ODE can be easily seen to have $\theta^* = -\bar{A}^{-1}\bar{b}$ as its unique globally asymptotically stable attractor where it is easy to verify (as before) that

$$W(\theta) = \frac{1}{2}(\bar{A}\theta + \bar{b})^T(\bar{A}\theta + \bar{b}),$$

serves as an associated Lyapunov function. The singleton set $\{\theta^*\}$ trivially serves as an internally chain transitive invariant set of the above ODE. Now from Theorem 2.7, $\{\theta_n\}$ remains uniformly bounded w.p.1. Moreover, it follows that $\theta_n \to \theta^*$ almost surely.

2.7 Two-timescale Stochastic Approximation

Many times, one is faced with the problem of optimizing parameters under a nested loop structure. The objective function to be optimized in such cases is obtained as a long-run average over other sample cost functions many times in non-i.i.d noise settings. The outer loop procedure in such a case would perform the optimization but the inner loop would perform the averaging corresponding to any given parameter value as determined by the outer-loop procedure and that in turn would have performed a parameter update using the averaged value provided by the inner-loop step in the previous round. Policy iteration in Markov decision processes to determine the optimal policy is an example of a numerical procedure where the policy evaluation step proceeds in the inner loop while policy improvement is conducted in the outer loop, cf. (Bertsekas and Tsitsiklis, 1996). In general, running a nested loop procedure, however, comes with the challenge of dealing with a potentially large computation time for the procedure.

To simplify such dual-loop computations, particularly in the model-free setting, one often resorts to stochastic approximation with two timescales. In these algorithms, the aforementioned nested loop structure is replaced with two recursions that perform updates simultaneously but using different step size schedules, both of which satisfy the usual Robbins-Monro step size conditions though one of these tends to zero at a rate faster than the other. The actor-critic algorithm, (Sutton and Barto, 2018; Konda and Tsitsiklis, 2003; Bhatnagar *et al.*, 2009), in reinforcement learning (that mimics policy iteration) or the simulation

optimization algorithm for optimizing long-run average cost objectives under Markov noise, see for instance (Bhatnagar and Borkar, 1998; Bhatnagar *et al.*, 2013).

Suppose θ_n, γ_n, $n \geq 0$ be two parameter sequences that are governed according to

$$\theta_{n+1} = \theta_n + \alpha(n)(f(\theta_n, \gamma_n) + N^1_{n+1}), \tag{2.39}$$
$$\gamma_{n+1} = \gamma_n + \beta(n)(g(\theta_n, \gamma_n) + N^2_{n+1}), \tag{2.40}$$

where $\theta_n \in \mathbb{R}^d$ and $\gamma_n \in \mathbb{R}^l$, $\forall n \geq 0$ under the following assumptions:

A2.19. The functions $f : \mathbb{R}^d \times \mathbb{R}^l \to \mathbb{R}^d$ and $g : \mathbb{R}^d \times \mathbb{R}^l \to \mathbb{R}^l$ are both Lipschitz continuous.

A2.20. The step size sequences $\{\alpha(n)\}$ and $\{\beta(n)\}$ satisfy $\alpha(n), \beta(n) > 0$, $\forall n$. In addition,

$$\sum_n \alpha(n) = \sum_n \beta(n) = \infty, \quad \sum_n \left(\alpha(n)^2 + \beta(n)^2\right) < \infty, \tag{2.41}$$

$$\lim_{n\to\infty} \frac{\beta(n)}{\alpha(n)} = 0. \tag{2.42}$$

A2.21. The noise sequences $\{N^1_n\} \subset \mathbb{R}^d$ and $\{N^2_n\} \subset \mathbb{R}^l$ are both martingale difference sequences w.r.t. the sequence of σ-fields $\bar{\mathcal{F}}_n = \sigma(\theta_m, \gamma_m, N^1_m, N^2_m, m \leq n)$, $n \geq 0$, and further satisfy

$$E[\| N^i_{n+1} \|^2 | \bar{\mathcal{F}}_n] \leq D(1 + \| \theta_n \|^2 + \| \gamma_n \|^2), \ i = 1, 2, \ n \geq 0,$$

for $i = 1, 2$ and some constant $D < \infty$.

A2.22. $\sup_n \| \theta_n \|, \sup_n \| \gamma_n \| < \infty$ almost surely.

In Assumption A2.20, (2.42) is an important requirement which results in the separation of timescales. As a consequence of (2.42), $\beta(n) \to 0$ faster than $\{\alpha(n)\}$. Consider now the system of ODEs:

$$\dot{\theta}(t) = f(\theta(t), \gamma(t)), \tag{2.43}$$

$$\dot{\gamma}(t) = 0. \tag{2.44}$$

As a consequence of (2.44), one can alternatively consider the ODE

$$\dot{\theta}(t) = f(\theta(t), \gamma) \tag{2.45}$$

in place of (2.43), where because of (2.44), $\gamma(t) \equiv \gamma$, a constant.

A2.23. The ODE (2.45) has a unique globally asymptotically stable equilibrium $\mu(\gamma)$ where $\mu : \mathbb{R}^l \to \mathbb{R}^d$ is a Lipschitz continuous function.

Consider also the ODE

$$\dot{\gamma}(t) = g(\mu(\gamma(t)), \gamma(t)). \tag{2.46}$$

A2.24. The ODE (2.46) has a unique globally asymptotically stable attractor $\gamma^\star$.

Define two real-valued sequences $\{r_n\}$ and $\{s_n\}$ as $r_n = \sum_{m=0}^{n-1} \alpha(m)$ and $s_n = \sum_{m=0}^{n-1} \beta(m)$, respectively, $n \geq 1$ and with $r_0 = s_0 = 0$. Define continuous time processes $\bar{\theta}(r)$, $\bar{\gamma}(r)$, $r \geq 0$ as follows:

$$\bar{\theta}(r) = \frac{r_{n+1} - r}{r_{n+1} - r_n}\theta_n + \frac{r - r_n}{r_{n+1} - r_n}\theta_{n+1}, \ r \in [r_n, r_{n+1}],$$

$$\bar{\gamma}(r) = \frac{r_{n+1} - r}{r_{n+1} - r_n}\gamma_n + \frac{r - r_n}{r_{n+1} - r_n}\gamma_{n+1}, \ r \in [r_n, r_{n+1}].$$

For $s \geq 0$, let $\theta^s(r)$, $\gamma^s(r)$, $r \geq s$ denote the trajectories of (2.43)-(2.44) with $\theta^s(s) = \bar{\theta}(s)$ and $\gamma^s(s) = \bar{\gamma}(s)$. Note that because of (2.44), $\gamma^s(r) = \bar{\gamma}(s)$ $\forall r \geq s$. Now (2.39)-(2.40) can be viewed as 'noisy' Euler discretizations of the ODEs (2.43)-(2.44) when the time discretization corresponds to $\{r_n\}$. This is because (2.40) can be written as

$$\gamma_{n+1} = \gamma_n + \alpha(n)\left(\frac{\beta(n)}{\alpha(n)}\left(g(\theta_n, \gamma_n) + N_{n+1}^2\right)\right),$$

and (2.42) implies that the term multiplying $\alpha(n)$ on the RHS above vanishes in the limit. One can now show, see (Borkar, 2022), using a

sequence of approximations involving the Gronwall inequality that for any given $T > 0$, with probability one, $\sup_{r\in[s,s+T]} \| \bar{\theta}(r) - \theta^s(r) \| \to 0$ as $s \to \infty$. The same is also true for $\sup_{r\in[s,s+T]} \| \bar{\gamma}(r) - \gamma^s(r) \|$ as well. Further, using the time discretization $\{s_t\}$ for the ODE (2.46), a similar conclusion with regards to iteration (2.40) (and ODE (2.46)) can be drawn following a continuous time trajectory that is obtained with the iterates in (2.40) interpolated along the time line $\{s_n\}$ according to

$$\check{\gamma}(s) = \frac{s_{n+1} - s}{s_{n+1} - s_n}\gamma_n + \frac{s - s_n}{s_{n+1} - s_n}\gamma_{n+1}, \ s \in [s_n, s_{n+1}].$$

The following is the main two-timescale convergence result (cf. (Borkar, 2022)).

Theorem 2.8. Under Assumptions A2.19–A2.24, with probability one, $(\theta_n, \gamma_n) \to (\mu(\gamma^\star), \gamma^\star)$ as $n \to \infty$.

Consider now the case that the (A2.24) is replaced by the more general assumption:

A2.25. The ODE (2.46) has a set A of isolated local attractors that are individually asymptotically stable.

Assumption A2.25 relaxes the requirement that the ODE (2.46) have a unique globally asymptotically stable attractor by allowing instead for a set A of isolated attractors. Theorem 2.8 in this case takes the form:

Theorem 2.9. Under Assumptions A2.19–A2.23 and A2.25, with probability one, $(\theta_n, \gamma_n) \to \{(\mu(\gamma^\star), \gamma^\star) | \gamma^* \in A\}$ as $n \to \infty$.

The proof of this result follows in the same manner as Theorem 2.8. The only difference is that since now one allows for multiple isolated attractors for the ODE (A2.25), $\{\gamma_n\}$ will converge to a possibly sample path dependent local attractor $\gamma^* \in A$, see (Borkar, 2022, Corollary 2.4). The claim in Theorem 2.9 will then follow. The above result will be generalized further in the next section.

2.8 Two-timescale Stochastic Recursive Inclusions

In this section, we present a generalization of the results in Section 2.7. Specifically, we consider two-timescale recursions with both recursions having set-valued maps. More importantly, we weaken the requirement of existence of unique globally asymptotically stable attractors in Assumptions A2.23–A2.24 corresponding to the ODEs (2.45)–(2.46), respectively. The results we present below are from (Ramaswamy and Bhatnagar, 2016b).

Consider the following two-timescale recursion:

$$\theta_{n+1} = \theta_n + \alpha(n)(\kappa_n + N^1_{n+1}), \tag{2.47}$$

$$\gamma_{n+1} = \gamma_n + \beta(n)(\zeta_n + N^2_{n+1}), \tag{2.48}$$

where $\kappa_n \in f(\theta_n, \gamma_n)$ and $\zeta_n \in g(\theta_n, \gamma_n)$, respectively, where $f(\theta_n, \gamma_n)$ and $g(\theta_n, \gamma_n)$ are set-valued maps with $f(\theta_n, \gamma_n) \subset \mathbb{R}^d$ and $g(\theta_n, \gamma_n) \subset \mathbb{R}^l$ respectively. Further, the parameters that are getting updated, viz., $\theta_n \in \mathbb{R}^d$ and $\gamma_n \in \mathbb{R}^l$, $\forall n \geq 0$ under the following assumptions:

A2.26. The set-valued maps f and g are Marchaud or Peano maps.

A2.27. The step-size sequences $\{\alpha(n)\}$ and $\{\beta(n)\}$ satisfy

$$\begin{aligned}
&\alpha(n), \beta(n) > 0, \ \forall n; \\
&\sum_n \alpha(n) = \sum_n \beta(n) = \infty; \\
&\sum_n (\alpha(n)^2 + \beta(n)^2) < \infty; \\
&\lim_{n\to\infty} \frac{\beta(n)}{\alpha(n)} = 0.
\end{aligned}$$

A2.28. The sequences $\{N^1_n\}$ and $\{N^2_n\}$ are square integrable martingle differences w.r.t. the common filtration $\mathcal{F}_n = \sigma(\theta_m, \gamma_m, N^1_m, N^2_m, m \leq n)$, $n \geq 0$. Further, for a given constant $M > 0$ and $\forall n$, we have

$$E[\|N^i_{n+1}\|^2 \mid \mathcal{F}_n] \leq M(1 + \|\theta_n\|^2 + \|\gamma_n\|^2, \ i = 1, 2.$$

A2.29. We have that $\sup_n \|\theta_n\| < \infty$ and $\sup_n \|\gamma_n\| < \infty$ almost surely.

A2.30. For each $\theta \in \mathbb{R}^d$, the DI

$$\dot{\theta}(t) \in f(\theta(t), \gamma)$$

has a globally attracting set B_γ that is also Lyapunov stable. Moreover, $\sup_{\theta \in A_\gamma} \|\theta\| \leq K(1 + \|\gamma\|)$. The set-valued map $\mu : \mathbb{R}^l \to \{\text{subsets of } \mathbb{R}^d\}$ defined by $\mu(\gamma) = B_\gamma$ is upper semi-continuous.

For each $\gamma \in \mathbb{R}^l$, let

$$G(\gamma) \stackrel{\triangle}{=} \bar{co}\left(\bigcup_{\theta \in \mu(\gamma)} g(\theta, \gamma)\right),$$

denote the closed convex hull of the set $\left(\bigcup_{\theta \in \mu(\gamma)} g(\theta, \gamma)\right)$.

A2.31. The DI $\dot{\gamma}(t) \in G(\gamma(t))$ has a globally attracting set $\check{B}$ that is also Lyapunov stable.

The main result is then the following, see (Ramaswamy and Bhatnagar, 2016b)(Theorem 3.10):

Theorem 2.10. Under Assumptions A2.26–A2.31, the set of accumulation points of the algorithm (2.47)-(2.48) is given by

$$\{(\theta, \gamma) | \liminf_{n \to \infty} d((\theta_n, \gamma_n), (\theta, \gamma)) = 0\} \subset \bigcup_{\gamma \in \check{B}} \{(\theta, \gamma) | \theta \in \mu(\gamma)\}. \tag{2.49}$$

Remark 2.8. In relation to Theorem 2.10, we note the following with regards the role played by Assumption A2.31:

(i) In the absence of Assumption A2.31 (assuming the other assumptions continue to hold), the RHS of (2.49) will get replaced by the much larger set $\bigcup_{\gamma \in \mathbb{R}^l} \{(\theta, \gamma) | \theta \in \mu(\gamma)\}$.

(ii) If Assumption A2.31 holds but $\check{B}$ is only a singleton, say $\{\gamma_0\}$, then the RHS of (2.49) is simply $\{(\theta, \gamma_0) | \theta \in \mu(\gamma_0)\}$.

(iii) In addition to (ii) above, if $\mu(\gamma_0)$ is the only point in the set, i.e., is a singleton, then the RHS of (2.49) is $(\mu(\gamma_0), \gamma_0)$.

2.9 Exercises

Exercise 1. Let $h(\theta^{(1)}, \theta^{(2)}) = \begin{bmatrix} 2\theta^{(1)} + 2\theta^{(2)} + 5 \\ 2\theta^{(1)} + 3\theta^{(2)} + 7 \end{bmatrix}$.

Answer the following:

(a) Find θ^* such that $h(\theta^*) = 0$.

(b) Consider a root-finding algorithm with the following update iteration:

$$\theta_{n+1} = \theta_n - a(n)h(\theta_n). \tag{2.50}$$

Specify a value for $a(n)$ that ensures $\theta_n \to \theta^*$ as $n \to \infty$. Justify your answer.

(c) Suppose h is not directly observable. Instead, for any θ, we have a noisy observation $\hat{h}(\theta)$ that satisfies

$$\mathbb{E}\left[\hat{h}(\theta) \mid x\right] = h(\theta) \text{ and } \mathbb{E}\left[\left\|\hat{h}(\theta)\right\|^2\right] \leq \sigma^2.$$

Specify a stochastic approximation variant of (2.50) and establish asymptotic convergence of the stochastic approximation iterate to θ^*.

Exercise 2. Recall the linear stochastic approximation algorithm from Section 2.2.5:

$$\theta_{n+1} = \theta_n + a(n)\left(A_{n+1}\theta_n + b_{n+1}\right),$$

where the step size $a(n)$ satisfies $\sum_n a(n) = \infty$, and $\sum_n a(n)^2 < \infty$. Further, A_n and b_n are matrices and vectors that satisfy

$$\mathbb{E}\left[A_{n+1} \mid \theta_1, \ldots, \theta_n\right] = A, E\left[b_{n+1} \mid \theta_1, \ldots, \theta_n\right] = b,$$

where A is a negative-definite matrix. Moreover, $\mathbb{E}\left[\|(A_n - A)\|^2\right] \leq C_1$ and $\mathbb{E}\left[\|b_n - b\|^2\right] \leq C_2$. Use the Kushner-Clark lemma to establish asymptotic convergence of θ_n?

Exercise 3. Consider the following update iteration:

$$\theta_{(n+1)L} = \theta_{nL} + \sum_{i=0}^{L-1} a(nL+i)\left(h(\theta_n) + M_{nL+i}\right), \tag{2.51}$$

where $a(n)$ is a random step size, $L > 1$ is a given integer, $\{M_{nL+i}, n \geq 0\}$ is a martingale difference sequence. Suppose the step sizes satisfy $\sum_n a(n) = \infty$ a.s. and $\sum_n a(n)^2 < \infty$ a.s.

Assume h is Lipschitz. Prove convergence of the stochastic approximation scheme given above, while making any additional assumptions as required.

2.10 Bibliographic remarks

Stochastic approximation has a long history, starting with the seminal work of Robbins and Monro (Robbins and Monro, 1951), who provided a stochastic root finding scheme. Subsequently, (Kiefer and Wolfowitz, 1952) analyzed a zeroth-order stochastic gradient scheme. For a textbook introduction, the reader is referred to (Borkar, 2022; Kushner and Yin, 2003). The main convergence result in Section 2.3 is the well-known Kushner Clark lemma, see (Kushner and Clark, 1978), while the Markov noise case is handled in (Borkar, 2022; Ramaswamy and Bhatnagar, 2019). The convergence result for two timescale stochastic approximation is based on Theorem 8.1 of (Borkar, 2022) and its generalization is based on (Ramaswamy and Bhatnagar, 2016b). The reader may refer to (Karmakar and Bhatnagar, 2018) for an analysis of two-timescale stochastic approximation with Markov noise. Finally, the reader is referred to (Benaïm *et al.*, 2005) for a detailed introduction to one-timescale stochastic recursive inclusions and their convergence analysis using differential inclusions. Finally, a detailed convergence analysis of two-timescale stochastic recursive inclusions with non-ergodic Markov noise appears in (Yaji and Bhatnagar, 2020).

On the applications side, reinforcement learning is popular and (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 2018; Bertsekas, 2019; Powell, 2021; Meyn, 2022) provide textbook introductions, see also (Bertsekas, 2012) for an extensive treatment on approximate dynamic programming, the backbone of modern RL.

Simultaneous perturbation based approaches in conjunction with reinforcement learning have been found to perform exceedingly well on several applications. For instance, (Bhatnagar and Kumar, 2004) presents and analyses an actor-critic algorithm with a temporal difference critic and an actor based on simultaneous perturbation gradient estimates. Further, an application on the available bit rate (ABR) service in asynchronous transfer mode (ATM) networks is studied. In (Bhatnagar and Babu, 2008) and (Bhatnagar and Lakshmanan, 2016), actor-critic style RL algorithms are developed to mimic q-learning but where the critic is updated on a slower timescale as compared to the actor. The algorithm in (Bhatnagar and Babu, 2008) is for the look-up table case while the algorithm in (Bhatnagar and Lakshmanan, 2016) caters to the case with function approximation. The actor recursion in each case involves SPSA based gradient estimates. These algorithms are also studied on problems of routing in communication networks. The algorithm in (Bhatnagar and Lakshmanan, 2016) has further been explored in (Prashanth *et al.*, 2014) for a problem of intrusion detection in adhoc wireless sensor networks. Further, in an application on vehicular traffic control, (Prashanth and Bhatnagar, 2012) incorporates Q-learning with a graded feedback control where the threshold levels are tuned using an SPSA based algorithm on a slower timescale.

For quantile estimation and CVaR estimation using stochastic approximation, see (Bardou *et al.*, 2009). Stochastic approximation is popular for estimating other risk measures, e.g., utility-based shortfall risk (Hegde *et al.*, 2021; Dunkel and Weber, 2010).

3

Gradient estimation

In this chapter, we introduce the simultaneous perturbation trick for gradient estimation, given noisy measurements from a zeroth-order oracle. These estimates are not unbiased, but feature a parameter that controls the bias, usually at the cost of variance. We discuss several popular gradient estimates in the literature, through a unified estimator. These estimates form the basis for a stochastic gradient algorithm, which is presented in Algorithm 1.

Algorithm 1: Zeroth-order stochastic gradient (ZSG) algorithm

Input: Initial point $\theta_0 \in \mathbb{R}^d$, iteration limit m, step sizes $\{a_k\}_{k \geq 1}$.

for $k = 1, \ldots, m$ **do**

Form the gradient estimate $\widehat{\nabla} f(\theta_k)$ using one or more function measurements;

Perform the following stochastic gradient descent update:

$$\theta_{k+1} = \theta_k - a_k \widehat{\nabla} f(\theta_k).$$

end for

Output: θ_m

In the following section, we present schemes for devising $\widehat{\nabla} f(\cdot)$ with an estimation error (bias) that can be made to vanish asymptotically. For the sake of analyzing the bias and variance properties of the gradient estimators in this chapter, we shall consider two classes of smooth functions, as given below. For a detailed introduction to smoothness, the reader is referred to Appendix D.

Definition 3.1. Consider a function $f : \mathbb{R}^d \to \mathbb{R}$.

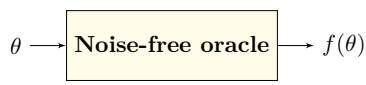
(i) f **is** L**-smooth** if for some constant $L > 0$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

(ii) $f \in \mathcal{C}^3$ if f is three times continuously differentiable with $\left|\nabla^3_{i_1 i_2 i_3} f(\theta)\right| < \alpha_0 < \infty$, for $i_1, i_2, i_3 = 1, \ldots, d$ and for all $\theta \in \mathbb{R}^d$. Here $\nabla^3 f(\theta) = \dfrac{\partial^3 f(\theta)}{\partial \theta^\mathsf{T} \partial \theta^\mathsf{T} \partial \theta^\mathsf{T}}$ denotes the third derivative of f at θ, and $\nabla^3_{i_1 i_2 i_3} f(\theta)$ denotes the $(i_1 i_2 i_3)$th entry of $\nabla^3 f(\theta)$, for $i_1, i_2, i_3 = 1, \ldots, d$.

3.1 Finite differences

As a gentle start, consider a noise-free zeroth-order oracle, as illustrated below.



In this setting, one could form an estimate $\widehat{\nabla} f(\theta)$ using $d + 1$ queries to the oracle above as follows:

$$\widehat{\nabla}_i f(\theta) = \frac{1}{\delta}\left(f(\theta + \delta e_i) - f(\theta)\right), \quad i = 1, \ldots, d. \tag{3.1}$$

How good an estimate is (3.1)? Assuming $f \in \mathcal{C}^3$, i.e., f is three-times continuously differentiable, we can employ Taylor series expansion of f as follows[1]:

$$f(\theta + \delta e_i) = f(\theta) + \delta \nabla f(\theta)^\top e_i + \frac{\delta^2}{2} e_i^\top \nabla^2 f(\theta) e_i + O(\delta^3),$$

[1] For the sake of simplicity, we have chosen to hide the constants through a $O(\delta^3)$ term. The latter constants can be made precise, as in Proposition 3.1 below.

leading to the estimation error:

$$\left\|\widehat{\nabla} f(\theta) - \nabla f(\theta)\right\| = O(\delta).$$

Using $2d$ queries to the oracle mentioned above, we define a two-sided variant of the estimate in (3.2) below.

$$\widehat{\nabla}_i f(\theta) = \frac{1}{2\delta}\left(f(\theta + \delta e_i) - f(\theta - \delta e_i)\right), \quad i = 1, \ldots, d. \tag{3.2}$$

Employing Taylor-series expansions as before, leads to the following bound on the estimation error:

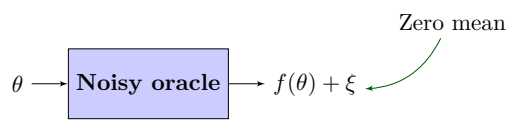
$$\left\|\widehat{\nabla} f(\theta) - \nabla f(\theta)\right\| = O(\delta^2).$$

Thus, using a two-sided estimate reduced the error to $O(\delta^2)$, while the number of sample measurements went up to $2d$ from $d + 1$.

The two estimates presented in (3.1) and (3.2) fall under the realm of finite difference stochastic approximation (FDSA), and such schemes can be extended to handle noise-corrupted function observations, as we show next. As an aside, a major disadvantage with FDSA estimates is the high measurement cost, since $O(d)$ calls to the oracle are needed to form an estimate.

FDSA with noisy measurements

We consider a zeroth-order oracle, which outputs noisy observations of the objective at any query point, as illustrated below.



Consider the following two-sided estimate, formed using noisy function measurements[2]:

$$\widehat{\nabla}_i f(\theta) = \frac{1}{2\delta}\left\{f(\theta + \delta e_i) + \xi_i^+ - (f(\theta - \delta e_i) + \xi_i^-)\right\}, \quad i = 1, \ldots, d.$$

[2]Here and in what follows, ξ^+ and ξ^- are real valued r.v.s and this notation is not be confused with positive and negative parts of a measurable function (Royden and Fitzpatrick, 2010).

Suppose that $\mathbb{E}\left[\xi^+ - \xi^-\right] = 0$ and also that $\mathbb{E}\left[\xi^{\pm 2}\right] \leq \sigma^2 < \infty$. Then, assuming $f \in \mathcal{C}^2$, one can establish the near-unbiasedness of the estimate above using Taylor-series expansions as follows:

$$\begin{aligned} & f(\theta \pm \delta e_i) = f(\theta) \pm \delta \nabla f(\theta)^\top e_i + \frac{\delta^2}{2} e_i^\top \nabla^2 f(\theta) e_i + O(\delta^3). \\ & \Rightarrow \mathbb{E}(\widehat{\nabla}_i f(\theta)) = \frac{1}{2\delta} \left(f(\theta + \delta e_i) - f(\theta - \delta e_i)\right) \\ & \Rightarrow \left\| \mathbb{E}\widehat{\nabla} f(\theta) - \nabla f(\theta) \right\| = O(\delta^2). \end{aligned}$$

With $2d$ queries, an FDSA estimate would be $O(\delta^2)$ from the true gradient, even in the case when function measurements are noisy.

Next, we will present a series of estimates that achieve the same level of accuracy as FDSA, but with only two measurements, irrespective of the dimension d.

3.2 Simultaneous perturbation method

FDSA perturbs co-ordinates one-at-a-time, leading to $2d$ queries to the oracle. The number of queries get reduced by randomly perturbing all co-ordinate directions simultaneously. This is the idea behind the SPSA scheme proposed by (Spall, 1992), which we describe below.

Let $y^+ = f(\theta + \delta\Delta) + \xi^+$ and $y^- = f(\theta - \delta\Delta) + \xi^-$, where $\Delta = (\Delta_1, \ldots, \Delta_d)^\top$ is a d-vector of independent, symmetric, ± 1-valued Bernoulli r.v.s, i.e., $\Delta_i = +1$ w.p. $1/2$ and -1 w.p. $1/2$, for $i = 1, \ldots, d$. It is important to mention that (Spall, 1992) provides general conditions on the perturbation distribution and symmetric Bernoulli is a popular special case that we also consider here for simplicity. The gradient estimate here is given as follows:

$$\widehat{\nabla}_i f(\theta) = \left[\frac{y^+ - y^-}{2\delta\Delta_i}\right], \; i = 1, \ldots, d. \tag{3.3}$$

In expectation, the estimate defined above is nearly unbiased, and this can be argued as follows: Assuming $\mathbb{E}\left[\xi^+ - \xi^-\right] = 0$,

$$\mathbb{E}\left[\widehat{\nabla}_i f(\theta)\right] = \mathbb{E}\left[\frac{f(\theta + \delta\Delta) - f(\theta - \delta\Delta)}{2\delta\Delta_i}\right]. \tag{3.4}$$

Here and in what follows, we assume that θ is given, and the expectation is over other random terms.

Next, assuming $f \in \mathcal{C}^3$, and employing Taylor series expansions, we obtain

$$f(\theta \pm \delta\Delta) = f(\theta) \pm \delta\Delta^{\mathrm{T}}\nabla f(\theta) + \frac{\delta^2}{2}\Delta^{\mathrm{T}}\nabla^2 f(\theta)\Delta + O(\delta^3). \tag{3.5}$$

From the above, it is easy to see that

$$\frac{f(\theta+\delta\Delta) - f(\theta-\delta\Delta)}{2\delta\Delta_i} - \nabla_i f(\theta) = \underbrace{\sum_{j=1, j\neq i}^{d} \frac{\Delta_j}{\Delta_i}\nabla_j f(\theta)}_{(I)} + O(\delta^2).$$

In expectation given θ, term (I) above is zero, since $\Delta_l, l = 1, \ldots, d$ are independent, symmetric, Bernoulli ± 1-valued r.v.s. Hence,

$$\mathbb{E}\left[\widehat{\nabla}_i f(\theta)\right] = \nabla_i f(\theta) + O(\delta^2).$$

From the above, it is easy to see that the expected value of the estimate (3.3) converges to the true gradient $\nabla f(\theta)$ in the limit as $\delta \to 0$. Thus, if one uses a gradient estimate as in (3.3) in a stochastic approximation algorithm, and lets $\delta \to 0$ slowly enough, the overall scheme will converge to local minima of the function f. This will be made precise in the next chapter.

We demonstrated the simultaneous perturbation trick through the SPSA scheme, which employed independent symmetric Bernoulli r.vs for random perturbations. As mentioned before, the trick is more generally valid and is not restricted to this choice of random perturbations alone. Furthermore, this trick can be used to estimate the Hessian, and not just the gradient, as we illustrate later.

In the next section, we present a unified gradient estimate that covers several schemes in the literature.

3.2.1 A unified estimate

Let $y^+ = f(\theta+\delta U)+\xi^+$, and $y^- = f(\theta-\delta U)+\xi^-$. Using these function values, we form the gradient estimate as follows:

$$\widehat{\nabla} f(\theta) = \left(\frac{y^+ - y^-}{2\delta} \right) V. \tag{3.6}$$

The estimate defined above can be specialized to cover several popular simultaneous perturbation-based gradient estimates, and we list some of these below.

- Setting $U \sim \mathcal{N}(0, I_d)$, where $\mathcal{N}(0, I_d)$ denotes the d-dimensional standard Gaussian vector, and $V = U$, we obtain the smoothed functional scheme proposed by (Styblinski and Tang, 1990) (see also (Katkovnik and Kulchitsky, 1972) for a one-sided variant). The latter scheme has been refined by (Polyak and Tsybakov, 1990), and also studied by (Dippon, 2003; Bhatnagar and Borkar, 2003; Bhatnagar, 2007; Nesterov and Spokoiny, 2017).

- Setting U_i to be symmetric ± 1-valued Bernoulli r.v.s and $V = U$, we obtain the SPSA gradient estimate, which was defined earlier in (3.3).

- $U \sim \text{Unif}(\mathbb{S}_N)$, i.e., U is chosen uniformly at random on the surface of a d-dimensional unit sphere, and with $V = dU$, we obtain the random direction stochastic approximation (RDSA) scheme proposed by (Kushner and Clark, 1978). A variant of the RDSA scheme with other choices for random perturbations is discussed next.

- Setting U_i to be a uniformly distributed r.v. in $[-\eta, \eta]$, and $V = \frac{3}{\eta^2} U_i$, leads to the 1RDSA-Unif variant of (Prashanth *et al.*, 2017). On the other hand, setting U_i to be an asymmetric Bernoulli r.v., i.e., taking values -1 and $1+\varepsilon$ with probabilities $\frac{1+\varepsilon}{2+\varepsilon}$ and $\frac{1}{2+\varepsilon}$, respectively, and $V_i = \frac{1}{1+\varepsilon} U_i$ leads to the 1RDSA-Asymber variant of (Prashanth *et al.*, 2017). Here $\varepsilon > 0$ is a constant, usually set to a small value.

We make the following assumptions for analyzing the unified estimator presented above:

A3.1. Let U, V be random d-vectors satisfying $\mathbb{E}\left[VU^\top\right] = I$ and $\mathbb{E}\left[\|V\|^2\right] < \infty$.

A3.2. The noise factors $\xi^\pm$ in (3.6) satisfy

$$\mathbb{E}[\xi^+ - \xi^- \,|\, U, V] = 0, \quad \text{and} \quad \mathbb{E}[(\xi^+ - \xi^-)^2 \,|\, U, V] \leq \sigma^2 < \infty\,. \tag{3.7}$$

A3.3. The objective f satisfies

$$\sup_{\theta \in \mathbb{R}^d} \mathbb{E}[f(\theta \pm \delta U)^2] \leq B < \infty\,. \tag{3.8}$$

The result below provide bias and variance bounds for the unified estimate presented above.

Proposition 3.1. Assume A3.1–A3.3, $\mathbb{E}\left[\|V\|\,\|U\|^3\right] < \infty$, and also that $f \in \mathcal{C}^3$, with $\left|\nabla^3_{i_1 i_2 i_3} f(\theta)\right| < \tilde{B} < \infty$, for $i_1, i_2, i_3 = 1, \ldots, d$ and for all $\theta \in \mathbb{R}^d$. Then, the gradient estimate defined in (3.6) satisfies the following bounds for any given θ:

$$\left\|\mathbb{E}\left[\widehat{\nabla} f(\theta)\right] - \nabla f(\theta)\right\| \leq C_1 \delta^2, \text{ and}$$
$$\mathbb{E}\left[\left\|\widehat{\nabla} f(\theta) - \mathbb{E}\left[\widehat{\nabla} f(\theta)\right]\right\|^2\right] \leq \frac{C_2}{\delta^2},$$

where $C_1 = \dfrac{\tilde{B}\mathbb{E}\left[\|V\|\,\|U\|^3\right]}{6}$, and $C_2 = \mathbb{E}\left[\|V\|^2\right]\left(\sigma^2 + B^2\right)$.

From the result above, it is apparent that the sensitivity parameter δ controls the bias-variance tradeoff, i.e., small values of δ imply a low bias and high variance, while large values of δ implies high bias and low variance in the gradient estimate.

Proof. Notice that

$$\mathbb{E}[\widehat{\nabla} f(\theta)] = \mathbb{E}\left[V\,\frac{f(\theta + \delta U) - f(\theta - \delta U)}{2\delta}\right],$$

since $\mathbb{E}\left[V\left(\dfrac{\xi^+ - \xi^-}{2\delta}\right)\right] = 0$ from A3.2.

Since $f \in C^3$, we have the following Taylor series expansion of f around θ:

$$f(\theta \pm \delta U) = f(\theta) \pm \delta\, U^{\mathsf{T}} \nabla f(\theta) + \frac{\delta^2}{2} U^{\mathsf{T}} \nabla^2 f(\theta) U \pm \frac{\delta^3}{6} \nabla^3 f(\tilde{\theta}^{\pm})(U \otimes U \otimes U), \tag{3.9}$$

where $\otimes$ denotes the Kronecker product and $\tilde{\theta}^{+}$ (resp. $\tilde{\theta}^{-}$) is on the line segment between θ and $(\theta + \delta U)$ (resp. $(\theta - \delta U)$).

Now,

$$V \frac{f(\theta + \delta U) - f(\theta - \delta U)}{2\delta} = V U^{\mathsf{T}} \nabla f(\theta) + \frac{\delta^2}{12} V \left(\nabla^3 f(\tilde{\theta}^{+}) + \nabla^3 f(\tilde{\theta}^{-}) \right) (U \otimes U \otimes U). \tag{3.10}$$

Taking expectations of both sides above, using $\mathbb{E}\left[VU^{\mathsf{T}}\right] = I$, $|\nabla^3 f(\tilde{\theta}^{\pm})| < \tilde{B}$, and $|\nabla^3 f(\bar{\theta})(U \otimes U \otimes U)| \leq \tilde{B} \|U\|^3$ for any $\bar{\theta}$, we obtain

$$\left\| \mathbb{E}\left[\widehat{\nabla} f(\theta) \right] - \nabla f(\theta) \right\| \leq C_1\, \delta^2\,, \text{ where } C_1 = \frac{\tilde{B} \mathbb{E}\left[\|V\| \, \|U\|^3 \right]}{6}.$$

Next, we prove the second claim concerning the variance of $\widehat{\nabla} f(\theta)$. Notice that

$$\begin{aligned}
&\mathbb{E} \left\| \widehat{\nabla} f(\theta) - \mathbb{E}\left[\widehat{\nabla} f(\theta) \right] \right\|^2 \leq \mathbb{E} \left\| \widehat{\nabla} f(\theta) \right\|^2 \\
&= \mathbb{E} \left(\|V\|^2 \left(\left(\frac{\xi^{+} - \xi^{-}}{2\delta} \right)^2 + 2 \left(\frac{\xi^{+} - \xi^{-}}{2\delta} \right) \left(\frac{f(\theta + \delta U) - f(\theta - \delta U)}{2\delta} \right) \right. \right. \\
&\qquad \left. \left. + \left(\frac{f(\theta + \delta U) - f(\theta - \delta U)}{2\delta} \right)^2 \right) \right) \\
&= \mathbb{E} \left(\|V\|^2 \left(\frac{\xi^{+} - \xi^{-}}{2\delta} \right)^2 \right) + 4\mathbb{E} \left(\|V\|^2 \left(\frac{f(\theta + \delta U) - f(\theta - \delta U)}{2\delta} \right)^2 \right) \qquad (3.11) \\
&\leq \frac{C_2}{\delta^2}\,,
\end{aligned}$$

where $C_2 = \mathbb{E}\left[\|V\|^2 \right] \left(\sigma^2 + B^2 \right)$. The equality in (3.11) follows from $\mathbb{E}\left[\xi^{+} - \xi^{-} \,|\, U, V \right] = 0$. □

3.2.2 The convex case

We now analyze the bias and variance properties of the estimator in (3.6) under a convex objective f. In this case, we do not require higher-order smoothness, and instead it is enough to assume first-order smoothness.

Proposition 3.2. Assume A3.1–A3.3, $\mathbb{E}[\|V\| \, \|U\|^2] < \infty$, and also that the function f is convex and L-smooth, as specified in Definition 3.1. Then the gradient estimate defined in (3.6) satisfies the following bounds for any given θ:

$$\left\| \mathbb{E}\left[\widehat{\nabla} f(\theta)\right] - \nabla f(\theta) \right\| \leq C_1 \delta, \text{ and } \mathbb{E}\left[\left\| \widehat{\nabla} f(\theta) - \mathbb{E}\left[\widehat{\nabla} f(\theta)\right] \right\|^2 \right] \leq \frac{C_2}{\delta^2},$$

where $C_1 \triangleq \frac{L}{2}\mathbb{E}[\|V\| \, \|U\|^2]$ and C_2 is as specified in Proposition 3.1.

Proof. For any convex function f with an L-Lipschitz gradient, for any $\delta > 0$, it holds that

$$\frac{\langle \nabla f(\theta), \delta u \rangle}{2\delta} \leq \frac{f(\theta + \delta u) - f(\theta)}{2\delta} \leq \frac{\langle \nabla f(\theta), \delta u \rangle + (L/2) \, \|\delta u\|^2}{2\delta}.$$

Using similar inequalities for $f(\theta - \delta u)$, we obtain

$$\langle \nabla f(\theta), u \rangle - \frac{L\delta \, \|u\|^2}{2} \leq \frac{f(\theta + \delta u) - f(\theta - \delta u)}{2\delta} \leq \langle \nabla f(\theta), u \rangle + \frac{L\delta \, \|u\|^2}{2}.$$

Letting $\phi(\theta, \delta, u) := \frac{1}{\delta}\left(\frac{f(\theta + \delta u) - f(\theta - \delta u)}{2\delta} - \langle \nabla f(\theta), u \rangle \right)$, we get

$$|\phi(\theta, \delta, u)| \leq \frac{L}{2} \|u\|^2 .$$

Using $\mathbb{E}\left[VU^\top\right] = I$ and A3.2, we obtain

$$\begin{aligned} \mathbb{E}[\widehat{\nabla} f(\theta)] &= \mathbb{E}\left[V \left(\frac{f(\theta + \delta U) - f(\theta - \delta U)}{2\delta} \right) \right] \\ &= \mathbb{E}\left[VU^\top \nabla f(\theta) + \delta \phi(\theta, \delta, U) V \right] \\ &= \nabla f(\theta) + \delta \widehat{\phi}(\theta, \delta), \end{aligned}$$

where $\widehat{\phi}(\theta, \delta)$ satisfies $\left\|\widehat{\phi}(\theta, \delta)\right\| \leq C_1 = \frac{L}{2}\mathbb{E}[\|V\| \, \|U\|^2]$. The first claim concerning the bias of the gradient estimate follows.

The bound on the variance of the gradient estimate in (3.6) follows in a similar manner to the proof of Proposition 3.1. □

3.3 Variants

3.3.1 One-point gradient estimate

The gradient estimate presented earlier required two function evaluations. In this section, we describe a variant that requires only one function evaluation. Let $y = f(\theta + \delta U) + \xi$. Using this function value, we form a gradient estimate as follows:

$$\widehat{\nabla} f(\theta) = \frac{y}{\delta} V, \tag{3.12}$$

where U, V are random perturbations as in the case of two-point estimate (3.6), and ξ is a zero-mean noise r.v., i.e., satisfying $\mathbb{E}[\xi|V] = 0$.

Proposition 3.3. Assume A3.1, A3.3, $\mathbb{E}[V] = 0$, and $\mathbb{E}[\xi|V] = 0$. Further, assume that U is symmetrically distributed, and V is an odd function of U. Then, for $f \in \mathcal{C}^3$, the gradient estimate defined in (3.12) satisfies

$$\left\|\mathbb{E}\left[\widehat{\nabla} f(\theta)\right] - \nabla f(\theta)\right\| \leq C_1 \delta^2, \text{ and } \mathbb{E}\left[\left\|\widehat{\nabla} f(\theta) - \mathbb{E}\left[\widehat{\nabla} f(\theta)\right]\right\|^2\right] \leq \frac{C_2}{\delta^2}.$$

The $O(\delta^2)$ bound on the bias above is comparable to the one obtained for the two-point estimate (3.6) in Proposition 3.1. However, a closer inspection of the proof reveals that the first and second term in the Taylor expansion (see (3.9)) cancel out in the case of the two-point estimate, while no such cancellation occurs in the one-point case. Instead, in the latter case, the corresponding Taylor terms turn out to be mean zero (see (3.13) in the proof below). Hence, the two-point estimate is preferable. Moreover, empirically the two-point estimate usually outperforms its one-point counterpart, as noted in (Spall, 1997).

Proof. Using $\mathbb{E}[\xi|V] = 0$, we have

$$\mathbb{E}[\widehat{\nabla} f(\theta)] = \mathbb{E}\left[V\left(\frac{f(\theta+\delta U)}{\delta}\right)\right].$$

By Taylor's expansion in (3.9), we obtain

$$\begin{aligned}
&\mathbb{E}\left[V\frac{f(\theta+\delta U)}{\delta}\right] \\
&= \mathbb{E}\left[V\frac{f(\theta)}{\delta}\right] + \mathbb{E}\left[VU^{\mathsf{T}}\nabla f(\theta)\right] + \mathbb{E}\left[\frac{\delta}{2}VU^{\mathsf{T}}\nabla^2 f(\theta)U\right] \\
&\quad + \mathbb{E}\left[\frac{\delta^2}{6}V\nabla^3 f(\tilde{\theta}^+)(U\otimes U\otimes U)\right] \\
&= \nabla f(\theta) + \mathbb{E}\left[\frac{\delta^2}{6}V\nabla^3 f(\tilde{\theta}^+)(U\otimes U\otimes U)\right]. \qquad (3.13)
\end{aligned}$$

The final equality above follows from the facts that $\mathbb{E}[V] = 0$, $\mathbb{E}[VU^{\mathsf{T}}] = I$ and for any $i, j = 1, \ldots, d$, $E[V_i U_j^2] = 0$ since V is a deterministic odd function of U, with U having a symmetric distribution. Using the fact that $|\nabla^3 f(\tilde{\theta}^+)(U\otimes U\otimes U)| \leq \tilde{B}\,\|U\|^3$, we obtain

$$\left\|\mathbb{E}\left[\widehat{\nabla} f(\theta)\right] - \nabla f(\theta)\right\| \leq C_1\,\delta^2, \text{ where } C_1 = \frac{B_3\mathbb{E}\left[\|V\|\,\|U\|^3\right]}{6}.$$

The proof of the second claim concerning the variance of the estimate $\widehat{\nabla} f(\theta)$ follows using arguments similar to those used in the proof of Proposition 3.1. □

3.3.2 Deterministic perturbations

So far, we have shown that one can use random perturbations to construct a gradient estimate with controllable bias. In this section, we show that one can achieve similar bias control through a deterministic perturbation sequence. To illustrate, we demonstrate (i) a permutation matrix-based perturbation sequence in the context of an RDSA scheme; and (ii) a Hadamard matrix-based perturbation sequence in an SPSA-type gradient estimate.

Permutation matrices for RDSA

The analysis of the biasedness of the unified estimator in (3.6) relied on suitable Taylor's expansions to arrive at the following:

$$V\left[\frac{f(\theta+\delta U)-f(\theta-\delta U)}{2\delta}\right]=VU^{\intercal}\nabla f(\theta)+O(\delta^2).$$

The random perturbations U, V satisfying $\mathbb{E}VU^{\intercal}=\mathbb{I}_d$ resulted in a nearly unbiased estimator (see Proposition 3.1). Now, if U, V are chosen in a deterministic fashion, such that $VU^{\intercal}$ sums to identity over a loop, i.e., $\sum_{m=0}^{\tau} V_m U_m^{\intercal}=\mathbb{I}_d$ for some τ, then $\widehat{\nabla} f(\theta)$ would be nearly unbiased, in the spirit of the guarantees in Proposition 3.1. We present below a deterministic perturbation scheme, where we loop through the rows of a permutation matrix.

A permutation matrix is a matrix whose rows are the rows of an identity matrix in some order. For instance, the permutation matrices in two dimension are

$$\begin{bmatrix}1&0\\0&1\end{bmatrix}\text{ and }\begin{bmatrix}0&1\\1&0\end{bmatrix}.$$

In three dimensions, there are 6 permutation matrices. In general, there are $d!$ permutation matrices in dimension d.

We now present an RDSA-style gradient estimate using permutation matrix-based deterministic perturbations below.

$$\widehat{\nabla} f(\theta)=\sum_{m=0}^{d-1}\Delta_m\left[\frac{y_m^+-y_m^-}{2\delta_m}\right]. \tag{3.14}$$

In the above, $y_m^+=f(\theta+\delta_m\Delta_m)+\xi_m^+$ and $y_m^-=f(\theta-\delta_m\Delta_m)+\xi_m^-$, where $\xi_m^{\pm}$ is the measurement noise. Further, Δ_m is the mth row of the d-dimensional permutation matrix. Table 3.1 illustrates the perturbations d_m used in (3.14), for $d=2$ and $d=3$. In a nutshell, the sequence shown in Table 3.1 loops through the rows of the identity matrix in some order.

Table 3.1: Illustration of the permutation matrix-based deterministic perturbation sequence construction for two-dimensional and three-dimensional settings.

(a) Case $d = 2$

Inner loop counter m	D_2^1	D_2^2
0	1	0
1	0	1

(b) Case $d = 3$

Inner loop counter m	D_3^1	D_3^2	D_3^3
0	0	1	0
1	0	0	1
2	1	0	0

Hadamard matrices for SPSA

A Hadamard matrix is a square matrix with entries ± 1 that satisfies $H^\intercal H = mI_m$, where I_m denotes the $m \times m$ identity matrix. Further, a Hadamard matrix is said to be normalized if all the elements of its first row and column are 1. A simple and systematic way of constructing normalized Hadamard matrices of order $m = 2^k$ is as follows:

For $k = 1$,

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

and for general $k > 1$,

$$H_{2^k} = \begin{bmatrix} H_{2^{k-1}} & H_{2^{k-1}} \\ H_{2^{k-1}} & -H_{2^{k-1}} \end{bmatrix}.$$

Let $P = 2^{\lceil log_2(d+1) \rceil}$, where, as mentioned before, d is the parameter dimension. This implies $P \geq d + 1$. Now construct a normalized Hadamard matrix H_P of order P using the above procedure. Let $h(1), \ldots, h(d)$ be any d columns other than the first column of H_P. The first column is not considered because all elements in the first column are 1, while all the other columns have an equal number of $+1$ and -1 elements. The latter property aids in cancellation of some of the bias terms. Now form a new matrix $\widetilde{H}_P$ of order $P \times d$ with $h(1), \ldots, h(d)$ as its columns. Let $\widetilde{\triangle}(k), k = 1, \ldots, P$ denote the rows of $\widetilde{H}_P$. The perturbation sequence $\{\triangle(m)\}$ is now generated by cycling through the rows of $\widetilde{H}_P$, i.e.,

$$\triangle(n) = \widetilde{\triangle}(n \bmod P + 1), \forall n \geq 0.$$

Remark 3.1. Under assumptions similar to those used in Proposition 3.1, it can be shown that the gradient estimate formed using either permutation matrices for RDSA or Hadamard matrices for SPSA satisfies the following inequality:

$$\left\|\mathbb{E}\left[\widehat{\nabla} f(\theta)\right]-\nabla f(\theta)\right\| \leq C_1\delta^2, \text{ and } \mathbb{E}\left[\left\|\widehat{\nabla} f(\theta)-\mathbb{E}\left[\widehat{\nabla} f(\theta)\right]\right\|^2\right] \leq \frac{C_2}{\delta^2}.$$

3.3.3 Gaussian smoothing

In this section, we analyze the estimation error of a special case of the unified estimate with Gaussian perturbations, using the technique from (Nesterov and Spokoiny, 2017).

Let $y^+ = f(\theta+\delta\Delta)+\xi^+$ and $y = f(\theta)+\xi^-$, where Δ is a d-dimensional Gaussian vector composed of standard normal r.v.s., i.e., $\Delta \sim N(0, I_d)$, and ξ^+, ξ^- are noise factors. Then, the "Gaussian smoothing" gradient estimate is formed as follows:

$$\widehat{\nabla} f(\theta) = \Delta\left[\frac{y^+ - y}{\delta}\right], \tag{3.15}$$

where Δ is a d-dimensional Gaussian vector composed of standard normal r.v.s., i.e., $\Delta \sim N(0, I_d)$.

Proposition 3.4. Assume A3.2, A3.3 and that f is L-smooth (see 3.1). The estimate defined in (3.15) satisfies

$$\left\|\mathbb{E}\left[\widehat{\nabla} f(\theta)\right]-\nabla f(\theta)\right\| \leq C_1\delta, \text{ and} \tag{3.16}$$

$$\mathbb{E}\left[\left\|\widehat{\nabla} f(\theta)-\mathbb{E}\left[\widehat{\nabla} f(\theta)\right]\right\|^2\right] \leq \frac{C_2}{\delta^2}, \tag{3.17}$$

for some constants $C_1, C_2 > 0$.

Proof. For any $\theta \in \mathbb{R}^d$, define

$$\begin{aligned} f_\delta(\theta) &= \frac{1}{(2\pi)^{\frac{d}{2}}}\int_{-\infty}^{\infty} f(\theta+\delta u)\exp\left(-\frac{\|u\|^2}{2}\right)du \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}\delta^d}\int_{-\infty}^{\infty} f(y)\exp\left(-\frac{\|y-\theta\|^2}{2\delta^2}\right)dy. \end{aligned}$$

The function f_δ denotes the smoothed version of the objective f, and is obtained by a convolution of f with Gaussian density. Notice that

$$\begin{aligned}\nabla f_\delta(\theta) &= \frac{1}{(2\pi)^{\frac{d}{2}}\delta^{d+2}} \int_{-\infty}^{\infty} f(y) \exp\left(-\frac{\|y-\theta\|^2}{2\delta^2}\right)(y-\theta)dy \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}\delta} \int_{-\infty}^{\infty} f(\theta+\delta u) \exp\left(-\frac{\|u\|^2}{2}\right) u\, du \\ &\qquad\qquad \text{(Letting } \delta u = y-\theta\text{)} \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{-\infty}^{\infty} \left(\frac{f(\theta+\delta u)-f(\theta)}{\delta}\right) \exp\left(-\frac{\|u\|^2}{2}\right) u\, du, \quad (3.18)\end{aligned}$$

where the final equality follows by using $\int_{-\infty}^{\infty} \exp\left(-\frac{\|u\|^2}{2}\right) u\, du = 0$.

Also,

$$\nabla f_\delta(\theta) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{-\infty}^{\infty} \frac{f(\theta)-f(\theta-\delta u)}{\delta} \exp\left(-\frac{\|u\|^2}{2}\right) u\, du. \quad (3.19)$$

Using (3.18) and (3.19), we obtain

$$\nabla f_\delta(\theta) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{-\infty}^{\infty} \frac{f(\theta+\delta u)-f(\theta-\delta u)}{2\delta} \exp\left(-\frac{\|u\|^2}{2}\right) u\, du.$$

Notice that

$$\begin{aligned}&\frac{1}{(2\pi)^{\frac{d}{2}}} \int_{-\infty}^{\infty} \langle \nabla f(\theta), u\rangle \exp\left(-\frac{\|u\|^2}{2}\right) u\, du \\ &= \sum_{i=1}^{d} \nabla_i f(\theta) \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{-\infty}^{\infty} u_i \exp\left(-\frac{\|u\|^2}{2}\right) u\, du \\ &= \sum_{i=1}^{d} \nabla_i f(\theta) \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{-\infty}^{\infty} \left(u_1 u_i, \ldots, u_{i-1}u_i, u_i^2, u_{i+1}u_i, \ldots, u_i u_d\right) \\ &\qquad\qquad \times \exp\left(-\frac{\|u\|^2}{2}\right) du \\ &= \sum_{i=1}^{d} \nabla_i f(\theta) \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{-\infty}^{\infty} u_i^2 \exp\left(-\frac{\|u\|^2}{2}\right) du \\ &= \sum_{i=1}^{d} \nabla_i f(\theta) \frac{1}{(2\pi)^{\frac{d}{2}}} \left(\prod_{j\neq i} \int_{-\infty}^{\infty} \exp\left(-\frac{u_j^2}{2}\right) du_j\right) \int_{-\infty}^{\infty} u_i^2 \exp\left(-\frac{u_j^2}{2}\right) du_i\end{aligned}$$

$$= \nabla f(\theta), \tag{3.20}$$

where the penultimate equality uses $\int_{-\infty}^{\infty} u_i \, u_j \exp\left(-\frac{\|u\|^2}{2}\right) du = 0$ for $i \neq j$, which holds owing to the symmetry of the Gaussian distribution.

Using (3.20), we obtain

$$\begin{aligned}
&\|\nabla f_\delta(\theta) - \nabla f(\theta)\| \\
&\leq \frac{1}{(2\pi)^{\frac{d}{2}}\delta} \int_{-\infty}^{\infty} |f(\theta + \delta u) - f(\theta) - \delta\langle \nabla f(\theta)), u\rangle| \, \|u\| \exp\left(-\frac{\|u\|^2}{2}\right) du \\
&\leq \frac{1}{(2\pi)^{\frac{d}{2}}} \frac{\delta L}{2} \int_{-\infty}^{\infty} \|u\|^3 \exp\left(-\frac{\|u\|^2}{2}\right) du \\
&\leq \frac{\delta L (d+3)^{\frac{3}{2}}}{2},
\end{aligned} \tag{3.21}$$

where the penultimate inequality follows by using the following inequality

$$|f(y) - f(\theta) - \langle \nabla f(\theta)), y - \theta\rangle| \leq \frac{1}{2} L \, \|\theta - y\|^2 ,$$

whereas the last inequality is a straightforward moment calculation for a multivariate Gaussian, cf. (Nesterov and Spokoiny, 2017, Lemma 1).

The claim in (3.16) concerning the bias of the Gaussian smoothing estimator now follows by combining (3.21) with A3.2.

The claim in (3.17) follows in a similar manner as in the proof of Proposition 3.1. □

We collect a few useful facts about Gaussian smoothing in the following lemma. These facts are extracted from the proof of Proposition 3.4 above.

Lemma 3.1. Suppose f is L-smooth. Let $f_\delta(\theta)$ denote the smoothed functional of f, which is defined as follows: For any $\theta \in \mathbb{R}^d$,

$$f_\delta(\theta) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{-\infty}^{\infty} f(\theta + \delta\Delta) \exp\left(-\frac{\|\Delta\|^2}{2}\right) d\Delta,$$

where Δ denotes a standard Gaussian vector.

The gradient of the $f_\delta(\cdot)$ is given by

$$\nabla f_\delta(\theta) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{-\infty}^{\infty} \frac{f(\theta + \delta\Delta) - f(\theta - \delta\Delta)}{2\delta} \exp\left(-\frac{\|\Delta\|^2}{2}\right) \Delta \, d\Delta.$$

Further, the smoothed functional f_δ is L-smooth and satisfies

$$\|\nabla f_\delta(\theta) - \nabla f(\theta)\| \leq \frac{\delta L (d+3)^{\frac{3}{2}}}{2}.$$

3.3.4 Common random numbers

Consider the classic simulation optimization setting, where the objective is $f(\theta) = \mathbb{E}(F(\theta, \psi))$, with ψ denoting the noise element, and $F(\cdot, \cdot)$ the sample performance. Notice that the observation noise is $\xi = F(\theta, \psi) - f(\theta)$, and one usually assumes that ξ is zero-mean, and i.i.d. when one obtains multiple function measurements.

In this section, we consider a special case where ψ can be kept fixed across function measurements. For instance, one could obtain function measurements $F(\theta_1, \psi)$ and $F(\theta_2, \psi)$. More precisely,

$$f(\theta) = \int F(\theta, \psi) P_\psi(d\psi)\,, \tag{3.22}$$

where $\psi \in \mathbb{R}$ is chosen by the algorithm. To reiterate, the algorithm can call the zeroth-order oracle by selecting both the input parameter θ and noise element ψ. In simulation optimization problems, where the function measurements are obtained from a computer simulation, and the source of randomness is common random numbers, one has the luxury of controlling the noise by initializing the seed. Thus, setting the same seed for two different input parameters would amount to having the same set of random numbers across simulations.

In this specialized setting, we now construct a two-point gradient estimate with the same noise element in both function measurements. Let $y^+ = F(\theta + \delta U, \psi)$, and $y^- = F(\theta - \delta U, \psi)$. Using these function values, we form the gradient estimate as follows:

$$\widehat{\nabla} f(\theta) = \left(\frac{y^+ - y^-}{2\delta} \right) V. \tag{3.23}$$

We shall establish now that the additional 'common random noise' structure allows the algorithm to reduce the variance of the gradient estimates, under the following additional smoothness assumption:

A3.4. The function F has a L-Lipschitz continuous gradient a.s. for any ψ, i.e.,

$$\|\nabla F(x, \psi) - \nabla F(y, \psi)\| \leq L \|x - y\| \text{ a.s.}$$

Proposition 3.5. Assume A3.1, A3.3, A3.4, and also that the function f is convex. Then the gradient estimate defined in (3.23) satisfies the following bounds for any given θ:

$$\left\| \mathbb{E}\left[\widehat{\nabla} f(\theta) \right] - \nabla f(\theta) \right\| \leq C_1 \delta, \text{ and} \tag{3.24}$$

$$\mathbb{E}\left[\left\| \widehat{\nabla} f(\theta) - \mathbb{E}\left[\widehat{\nabla} f(\theta) \right] \right\|^2 \right] \leq C_2 + C_3 \delta^2. \tag{3.25}$$

Proof. As in the proof of Proposition 3.2, for any convex function h with an L-Lipschitz gradient, for any $\delta > 0$, we have

$$\frac{\langle \nabla h(\theta), \delta u \rangle}{2\delta} \leq \frac{h(\theta + \delta u) - h(\theta)}{2\delta} \leq \frac{\langle \nabla h(\theta), \delta u \rangle + (L/2) \|\delta u\|^2}{2\delta}.$$

Using similar inequalities for $h(\theta - \delta u)$, we obtain

$$\langle \nabla h(\theta), u \rangle - \frac{L\delta \|u\|^2}{2} \leq \frac{h(\theta + \delta u) - h(\theta - \delta u)}{2\delta} \leq \langle \nabla h(\theta), u \rangle + \frac{L\delta \|u\|^2}{2}.$$

Letting $\phi(\theta, \delta, u) := \frac{1}{\delta} \left(\frac{h(\theta + \delta u) - h(\theta - \delta u)}{2\delta} - \langle \nabla h(\theta), u \rangle \right)$, we get

$$|\phi(\theta, \delta, u)| \leq \frac{L}{2} \|u\|^2 .$$

Using $\mathbb{E}\left[V U^\top \right] = I$, we obtain

$$\mathbb{E}\left[V \left(\frac{h(\theta + \delta U) - h(\theta - \delta U)}{2\delta} \right) \right] = \mathbb{E}\left[V U^\top \nabla h(\theta) + \delta \phi(\theta, \delta, U) V \right]$$

$$= \nabla h(\theta) + \delta \widehat{\phi}(\theta, \delta),$$

where $\widehat{\phi}(\theta, \delta)$ satisfies $\left\| \widehat{\phi}(\theta, \delta) \right\| \leq \frac{L}{2} \mathbb{E}[\|V\| \, \|U\|^2]$.

Applying the above expression to $F(\cdot, \psi)$ and using (3.23), we have

$$\mathbb{E}\left[\widehat{\nabla} f(\theta) \right] = \nabla F(\theta, \psi) + \delta \widehat{\phi}(\theta, \delta) \text{ a.s.},$$

where $\widehat{\phi}(\theta, \delta)$ satisfies $\left\| \widehat{\phi}(\theta, \delta) \right\| \leq \frac{L}{2} \mathbb{E}[\|V\| \, \|U\|^2]$.

A3.4 together with dominated convergence theorem leads to $E[\nabla F(\theta, \psi)] = \nabla f(\theta)$. Using this fact, we obtain

$$\begin{aligned}
&\left\| \mathbb{E}\left[\widehat{\nabla} f(\theta) \right] - \nabla f(\theta) \right\| \\
&= \left\| \mathbb{E}\left[V \left(\frac{f(\theta + \delta U) - f(\theta - \delta U)}{2\delta} \right) - V U^\top \nabla f(\theta) \right] \right\| \\
&\leq \delta \left\| \mathbb{E}[V \phi(\theta, \delta, U)] \right\| \\
&\leq \frac{\delta L}{2} \mathbb{E}[\|V\| \, \|U\|^2],
\end{aligned}$$

and the claim for the bias follows by setting $C_1 = \frac{L}{2} \mathbb{E}[\|V\| \, \|U\|^2]$.

We now bound $\mathbb{E}\left[\left\| \widehat{\nabla} f(\theta) \right\|^2 \right]$ as follows:

$$\begin{aligned}
\mathbb{E}\left\| \widehat{\nabla} f(\theta) \right\|^2 &= \mathbb{E}\left\| V \left(\delta \phi(\theta, \delta, U) + U^\top \nabla f(\theta) \right) \right\|^2 \\
&\leq \mathbb{E}\left[\left(\|V U^\top \nabla f(\theta)\| + \frac{\delta L}{2} \|V\| \, \|U\|^2 \right)^2 \right] \\
&\leq 2\mathbb{E}\left[\|V U^\top \nabla f(\theta)\|^2 \right] + \frac{\delta^2 L^2}{2} \mathbb{E}\left[\|V\|^2 \, \|U\|^4 \right],
\end{aligned}$$

and the claim for the variance follows by setting $C_2 = 2B_1^2 + \frac{L^2}{2} \mathbb{E}\left[\|V\|^2 \, \|U\|^4 \right]$ with $B_1 = \sup_\theta \|\nabla f(\theta)\|$. □

3.3.5 Gradient estimation with truncated Cauchy distribution

One can also use the truncated Cauchy distribution as the smoothing density in smoothed functional algorithms as shown and analyzed recently in (Mondal *et al.*, 2024). We first describe the truncated Cauchy distribution below.

Definition 3.2. A random variable u is said to follow the truncated (to the δ-sphere) Cauchy distribution with mean vector zero and covariance matrix $\Sigma = \delta^2 \mathbb{I}_{d\times d}$ if u has the following PDF:

$$h_\delta(u) = \frac{\Gamma(\frac{d+1}{2})}{\pi^{\frac{d+1}{2}} c_1 \delta^d (1 + \frac{\|u\|^2}{\delta^2})^{\frac{d+1}{2}}} \quad \text{for } \|u\| \leq \delta, \tag{3.26}$$

with $h_\delta(u) = 0$ for $\|u\| > \delta$. In the above, $c_1 > 0$ is a normalizing constant.

We define the smoothed version $g_\delta : \mathbb{R}^d \to \mathbb{R}$ of the given objective function $f : \mathbb{R}^d \to \mathbb{R}$ as follows:

$$g_\delta(\theta) \triangleq \mathbb{E}_{h_\delta(u)}[f(\theta + u)], \tag{3.27}$$

where $h_\delta(\theta)$ is the aforementioned smoothing kernel. One may also define another smoothed function $f_\delta : \mathbb{R}^d \to \mathbb{R}$ based on difference of objectives as follows:

$$\begin{aligned} f_\delta(\theta) &= \mathbb{E}_{h_\delta(u)}(f(\theta + u) - f(\vartheta)) \\ &= \mathbb{E}_{h_\delta(\theta-u)}[f(u) - f(\theta - u)]. \end{aligned} \tag{3.28}$$

The following result provides expressions for the gradient of the smoothed functions g_δ and f_δ, respectively. These expressions can be seen to help derive the one-measurement and two-measurement forms for the gradient estimators. We however mention that only the two-measurement form of the gradient estimator in (3.30) below has been studied in (Mondal *et al.*, 2024) as it provides lower bias than the one-measurement form. The reader is referred to (Mondal *et al.*, 2024) for a proof of Proposition 3.6.

Proposition 3.6. We have

$$\nabla g_\delta(\theta) = \frac{1}{\delta}\mathbb{E}_u \left[f(\theta + \delta u) \frac{(d+1)u}{(1 + \|u\|^2)} \right], \tag{3.29}$$

$$\nabla f_\delta(\theta) = \frac{1}{\delta}\mathbb{E}_u \left[(f(\theta + \delta u) - f(\theta)) \frac{(d+1)u}{(1 + \|u\|^2)} \right]. \tag{3.30}$$

A two-sample gradient estimate $G(\theta, \xi^+, \xi, u, \delta)$ of $\nabla f(\theta)$ is formed as follows:

$$G(\theta, \xi^+, \xi, u, \delta)$$

$$= \left(\frac{F(\theta + \delta u, \xi^+) - F(\theta, \xi)}{\delta} \right) \frac{(d+1)u}{(1 + \|u\|^2)}, \tag{3.31}$$

where ξ^+ and ξ are independent, zero-mean noise random vectors constituting measurement noise in the two function measurements. The function measurements themselves are represented using the function $F(\cdot)$. Further, the perturbation random variable $u \sim h_\delta$, the truncated Cauchy PDF.

A3.5. The function f is three-times continuously differentiable with $\|\nabla f(\theta)\| \leq B < \infty$ and $\|\nabla^3_{i_1,i_2,i_3} f(\theta)\| \leq B_1$ for all $\theta \in \mathbb{R}^d$ and for all $i_1, i_2, i_3 = 1, \ldots, d$.

Lemma 3.2 (Bias Lemma)**.** Under Assumption A3.5, we have a.s.

$$\mathbb{E}[G(\theta, \xi^+, \xi, u, \delta) | \theta, u] = c_2 \nabla f(\theta) + \delta w, \tag{3.32}$$

where $c_2 = \mathbb{E}_u \left[\frac{(d+1)(u^1)^2}{1 + \|u\|^2} \right] > 0$, with u^1 denoting the first component of the random vector u, and $w = \mathbb{E}\left[\left(\frac{u^T \nabla^2 f(\bar{\theta}^+) u}{2} \right) \frac{(d+1)u}{1 + \|u\|^2} | \theta, u \right]$ with $\bar{\theta}^+$ being a suitable point on the line segment joining θ and $\theta + \delta u$.

Proof. See (Mondal *et al.*, 2024, Lemma 1). □

Remark 3.2. Note here that the bias lemma in this case has a different form than corresponding results in other cases such as the one-sided Gaussian SF, cf. Proposition 3.4. In particular the conditional expectation of $G(\theta, \xi^+, \xi, u, \delta)$ given θ, u has $O(\delta)$ bias though in comparison with $c_2 \nabla f(\theta)$ instead of $\nabla f(\theta)$ with $c_2 > 0$ as the multiplying factor. We explain in Remark 4.1 about the impact on convergence of the resulting scheme due to this additional factor.

3.3.6 Generalized simultaneous perturbation method

A recently proposed approach in the class of random difference methods is Generalized SPSA (Bhatnagar and Prashanth, 2023; Pachal *et al.*, 2023). The idea is to use a multi-variate Taylor's expansion of the

objective function at a perturbed parameter and thereafter terminate the expansion after a certain number of terms. The larger the number of terms used in the expansion, smaller is the bias. Thus, the approach allows one to construct finite difference estimators of $\nabla f(\theta)$ for any given order of the bias. Chapter VII.1a of (Asmussen and Glynn, 2007a) explores this idea in the context of scalar functions $f : \mathbb{R} \to \mathbb{R}$. For the case of vector-valued parameters and for functions $f : \mathbb{R}^d \to \mathbb{R}$, this idea has been presented in (Bhatnagar and Prashanth, 2023; Pachal *et al.*, 2023).

Let $\mathcal{D}^\beta f(\theta) = \dfrac{\partial^{|\beta|} f(\theta)}{\partial\theta_1^{\beta_1} \cdots \partial\theta_d^{\beta_d}}$ with $|\beta| = \beta_1 + \cdots + \beta_d$ and $\theta^\beta = \theta_1^{\beta_1} \cdots \theta_d^{\beta_d}$. Further, $\beta! = \beta_1!\beta_2! \cdots \beta_d!$. The multi-variate Taylor's expansion has the following form:

$$f(\theta+\delta\Delta) = \sum_{|\beta|=0}^{\infty} \frac{\mathcal{D}^\beta f(\theta)}{\beta!}(\delta\Delta)^\beta = \sum_{|\beta|=0}^{\infty} \left(\frac{(\delta\Delta\mathcal{D})^\beta}{\beta!}\right) f(\theta) = \exp(\delta\Delta\mathcal{D}) f(\theta), \tag{3.33}$$

assuming f is infinitely many times continuously differentiable. Let $\tau_{\delta\Delta} f(\theta) \equiv f(\theta + \delta\Delta)$, with $\tau_{\delta\Delta} = \exp(\delta\Delta\mathcal{D})$ as the associated shift operator. Thus,

$$\mathcal{D} = \frac{1}{\delta\Delta} \log(\tau_{\delta\Delta}),$$

where, $\dfrac{1}{\delta\Delta} \triangleq \left(\dfrac{1}{\delta\Delta_1}, \ldots, \dfrac{1}{\delta\Delta_d}\right)^T$. An expansion of the log function gives

$$\mathcal{D} = \frac{1}{\delta\Delta} \sum_{j=1}^{\infty} \frac{(\tau_{\delta\Delta} - \mathcal{I})^j}{j} (-1)^{j+1},$$

where $\mathcal{I}$ denotes the identity operator and moreover, $\tau_{\delta\Delta}^k = \tau_{k\delta\Delta}$. The generalized gradient operator can then be viewed as follows: Let $\mathcal{D} = (\mathcal{D}_i, i = 1, \ldots, d)^T$, where for $i = 1, \ldots, d$,

$$\mathcal{D}_i = \frac{1}{\delta\Delta_i} \sum_{j=1}^{\infty} \frac{(\tau_{\delta\Delta} - \mathcal{I})^j}{j} (-1)^{j+1}. \tag{3.34}$$

From the above, one can obtain an estimator of order k by taking just the sum of the first k terms above. This will then require that the function f be only k times continuously differentiable and not infinite times continuously differentiable as in the beginning of this section.

Two-measurements (unbalanced) SPSA

The two-measurements version of SPSA here when using the GSPSA estimator (3.34) will correspond to just taking the first term in the summation above. Then, we will get

$$\mathcal{D}_i^1 f(\theta) \stackrel{\triangle}{=} \left(\frac{\tau_{\delta\Delta} - \mathcal{I}}{\delta\Delta_i}\right) f(\theta) = \frac{f(\theta + \delta\Delta) - f(\theta)}{\delta\Delta_i},$$

where $\mathcal{D}^1 = (\mathcal{D}_i^1, i = 1, \ldots, d)^T$ denotes the first order approximation operator. When acting on $f(\theta)$, it gives the one-sided (unbalanced) version of SPSA, see (Chen *et al.*, 1999; Bhatnagar *et al.*, 2013). Observe that a Taylor's expansion of $f(\theta + \delta\Delta)$ around θ gives

$$\mathcal{D}_i^1 f(\theta) = \frac{f(\theta + \delta\Delta) - f(\theta)}{\delta\Delta_i} = \frac{\Delta^T \nabla f(\theta)}{\Delta_i} + O(\delta). \tag{3.35}$$

Three-measurements SPSA

This estimator is obtained from (3.34) by truncating the series at $j = 2$.

$$\begin{aligned}
\mathcal{D}_i^2 f(\theta) &\stackrel{\triangle}{=} \left[\left(\frac{\tau_{\delta\Delta} - \mathcal{I}}{\delta\Delta_i}\right) - \frac{(\tau_{\delta\Delta} - \mathcal{I})^2}{2\delta\Delta_i}\right] f(\theta) \\
&= \left[\left(\frac{\tau_{\delta\Delta} - \mathcal{I}}{\delta\Delta_i}\right) - \left(\frac{\tau_{2\delta\Delta} + \mathcal{I} - 2\tau_{\delta\Delta}}{2\delta\Delta_i}\right)\right] f(\theta) \\
&= \left(\frac{f(\theta + \delta\Delta) - f(\theta)}{\delta\Delta_i}\right) \\
&\quad - \left(\frac{f(\theta + 2\delta\Delta) + f(\theta) - 2f(\theta + \delta\Delta)}{2\delta\Delta_i}\right) \\
&= \left(\frac{4f(\theta + \delta\Delta) - 3f(\theta) - f(\theta + 2\delta\Delta)}{2\delta\Delta_i}\right).
\end{aligned}$$

As before, $\mathcal{D}^2$ indicates the second order approximation operator. This is a new gradient SPSA estimator that has previously not been proposed. Through suitable Taylor's expansions, one obtains

$$\mathcal{D}_i^2 f(\theta) = \frac{\Delta^T \nabla f(\theta)}{\Delta_i} + O(\delta^2). \tag{3.36}$$

The first term in the expansion in (3.36) is the same as the first term in (3.35). However, the second term in (3.36) is $O(\delta^2)$ as opposed to $O(\delta)$ in (3.35).

Four-measurements SPSA

This estimator is obtained from (3.34) by truncating the series at $j = 3$.

$$\begin{aligned}
&\mathcal{D}_i^3 f(\theta) \\
&= \left[\left(\frac{\tau_{\delta\Delta} - \mathcal{I}}{\delta\Delta_i}\right) - \frac{(\tau_{\delta\Delta} - \mathcal{I})^2}{2\delta\Delta_i} + \frac{(\tau_{\delta\Delta} - \mathcal{I})^3}{3\delta\Delta_i}\right] f(\theta) \\
&= \left[\left(\frac{\tau_{\delta\Delta} - \mathcal{I}}{\delta\Delta_i}\right) - \left(\frac{\tau_{2\delta\Delta} + \mathcal{I} - 2\tau_{\delta\Delta}}{2\delta\Delta_i}\right)\right. \\
&\qquad \left. + \left(\frac{\tau_{3\delta\Delta} - 3\tau_{2\delta\Delta} + 3\tau_{\delta\Delta} - \mathcal{I}}{3\delta\Delta_i}\right)\right] f(\theta) \\
&= \frac{2f(\theta + 3\delta\Delta) - 9f(\theta + 2\delta\Delta) + 18f(\theta + \delta\Delta) - 11f(\theta)}{6\delta\Delta_i}.
\end{aligned}$$

Note that this estimator requires four function measurements at the parameter values θ, $\theta + \delta\Delta$, $\theta + 2\delta\Delta$ and $\theta + 3\delta\Delta$, respectively. The RHS above is obtained upon simplification and Taylor's expansions as before give us in this case

$$\mathcal{D}_i^3 f(\theta) = \frac{\Delta^T \nabla f(\theta)}{\Delta_i} + O(\delta^3). \tag{3.37}$$

The zeroth order as well as second and third order terms turn out to be zero due to cancellations of the various terms in the expansions resulting in (3.37).

Generalized $(k+1)$-measurements SPSA

Proceeding in a similar manner, one can obtain the kth order estimator by truncating the series in (3.34) at the kth term in the summation. Thus, one has in this (general) case

$$\mathcal{D}_i^k = \frac{1}{\delta\Delta_i} \sum_{j=1}^{k} \frac{(\tau_{\delta\Delta} - \mathcal{I})^j}{j} (-1)^{j+1} = \frac{1}{\delta\Delta_i} \sum_{l=0}^{k} \frac{(-1)^{1-l} (\tau_{\delta\Delta})^l}{l!} C_l,$$

where $C_l = \frac{1}{l}\prod_{j=0}^{l-1}(k-j)$. Then,

$$\begin{aligned}\mathcal{D}_i^k f(\theta) &= \left[\frac{1}{\delta\Delta_i}\sum_{l=0}^{k}\frac{(-1)^{1-l}C_l\tau_{l\delta\Delta}}{l!}\right] f(\theta) \\ &= \frac{1}{\delta\Delta_i}\sum_{l=0}^{k}\frac{(-1)^{1-l}C_l f(\theta+l\delta\Delta)}{l!}.\end{aligned}$$

This gradient estimator requires $(k+1)$ function measurements at the parameter values $\theta + l\delta\Delta$, $l = 0, 1, \ldots, k$. It has been shown in (Bhatnagar and Prashanth, 2023; Pachal *et al.*, 2023) that the order k generalized SPSA algorithm satisfies

$$\mathcal{D}_i^k f(\theta) = \frac{\Delta^T \nabla f(\theta)}{\Delta_i} + O(\delta^k). \tag{3.38}$$

Remark 3.3. Several remarks are in order.

1. As the Taylor's expansions of the various higher order generalized SPSA algorithms demonstrate, generalized SPSA of order k has a bias of order $O(\delta^k)$. Thus, an advantage with this class of algorithms is that given an accuracy level $O(\delta^k)$, one can find a gradient estimator within this class that provides this level of accuracy. Such guarantees are not available in the classes of estimators seen previously.

2. Note that here, one need not restrict oneself to only generalized SPSA estimators but in fact, the same can be done for smoothed functional and RDSA based perturbations, see Pachal *et al.*, 2023.

3. Finally, note that the estimators presented above are not balanced like two-simulation SPSA. In (Pachal *et al.*, 2023)[Section IV], balanced estimators are also derived by noting that

$$\tau_{\delta\Delta} f(\theta) - \tau_{-\delta\Delta} f(\theta) = f(\theta+\delta\Delta) - f(\theta-\delta\Delta).$$

   A similar calculation as before shows that

$$\tau_{\delta\Delta} - \tau_{\delta\Delta} = \exp(\delta\Delta\mathcal{D}) - \exp(-\delta\Delta\mathcal{D}) = 2\sinh 2\delta\mathcal{D}.$$

Thus,

$$\mathcal{D} = \frac{1}{\delta\Delta} \sinh^{-1}\left(\frac{\tau_{\delta\Delta} - \tau_{-\delta\Delta}}{2}\right).$$

Balanced estimators requiring even number of function measurements are then presented by terminating the infinite series of the $\sinh^{-1}(\cdot)$ function after varying number of steps.

4. We refer the reader to Pachal *et al.*, 2023 for detailed proofs of non-asymptotic and asymptotic convergence of the algorithms derived using the above gradient estimators.

3.4 Summary

Property → **Gradient estimate** ↓	**Bias**	**Variance**
Two-point estimate (3.6), $f \in \mathcal{C}^3$	$C_1\delta^2$	$\frac{C_2}{\delta^2}$
Two-point estimate (3.6) f convex+smooth	$C_1\delta$	$\frac{C_2}{\delta^2}$
One-point estimate (3.12), $f \in \mathcal{C}^3$	$C_1\delta^2$	$\frac{C_2}{\delta^2}$
One-point estimate (3.12) f convex+smooth	$C_1\delta^2$	$\frac{C_2}{\delta^2}$
Gaussian smoothing (3.15), $f \in \mathcal{C}^1$	$C_1\delta$	$\frac{C_2}{\delta^2}$
Gaussian smoothing with L-smooth F, common random noise	$C_1\delta$	$C_2 + C_3\delta^2$

3.5 Bibliographic remarks

The idea of simultaneous perturbation dates back to (Katkovnik and Kulchitsky, 1972), where the authors proposed the smoothed functional scheme for gradient estimation. A closely related estimation scheme

is RDSA, proposed by (Kushner and Clark, 1978), where the random perturbation are chosen uniformly on the surface of a d-dimensional sphere. This idea is equivalent to using d-dimensional standard Gaussian vector for the random perturbations — a choice studied in (Polyak and Tsybakov, 1990; Dippon, 2003; Bhatnagar and Borkar, 2003; Bhatnagar, 2007; Nesterov and Spokoiny, 2017). The asymptotic convergence of a zeroth-order algorithm with Gaussian smoothing where the gradient is estimated using a single measurement $y^+ = f(\theta+\delta\Delta)+\xi^+$ alone is shown in (Bhatnagar and Borkar, 2003). The same with a balanced estimator with two measurements $y^+ = f(\theta+\delta\Delta)+\xi^+$ and $y^- = f(\theta-\delta\Delta)+\xi^-$ is shown in (Bhatnagar, 2007). The latter reference also proposes one and two measurement Newton algorithms where both the gradient and Hessian are estimated using y^+ and y^- respectively. In (Rubinstein, 1981), conditions on perturbation distributions needed to construct zeroth-order gradient estimators have been presented. It is also shown that the uniform, Cauchy and Gaussian distributions satisfy these properties. Simultaneous perturbation gradient search algorithms with q-Gaussian smoothed functionals have been proposed in (Ghoshdastidar *et al.*, 2014b) for a wide range of the q-value parameter, for which the aforementioned distributions namely uniform, Cauchy and Gaussian emerge as special cases for certain values of q. It is shown that Variants of RDSA, employing uniform and asymmetric Bernoulli distributed random perturbations, have been proposed recently in (Prashanth *et al.*, 2017). SPSA, proposed by (Spall, 1992), is a very popular simultaneous perturbation method, which also exhibits the lowest asymptotic mean-square error (cf. (Chin, 1997; Prashanth *et al.*, 2017)). Deterministic perturbation variants of SPSA have been proposed and analyzed in (Bhatnagar *et al.*, 2003), while the corresponding deterministic variation for RDSA has been proposed recently in (Prashanth *et al.*, 2020). A comprehensive text-book reference on simultaneous perturbation methods is (Bhatnagar *et al.*, 2013). The latter reference contains a rigorous treatment of SPSA/SF methods, and includes both first as well as second-order schemes.

We now briefly survey other recent work on stochastic optimization. In (Berahas *et al.*, 2022), the authors assume the measurement noise is bounded a.s. and analyze simultaneous perturbation-based gradient

estimators under this condition. In particular, they establish bounds on the bias and variance of the gradient estimators and also conduct detailed numerical experiments comparing the performance of finite difference-based estimators with those employing simultaneous perturbation on a synthetic setup. In (Gasnikov *et al.*, 2022), zeroth-order stochastic optimization algorithms for non-smooth convex optimization problems are presented with perturbations distributed uniform on the surface of a unit sphere. Bounds on the number of iterations needed as well as the complexity of the estimator are provided. In (Kozak *et al.*, 2023; Rando *et al.*, 2023; Rando *et al.*, 2024), one-sided zeroth-order gradient estimation algorithms for both convex objectives as well as non-convex objectives, involving perturbation matrices with orthogonal random directions are presented. At each iterate, a random matrix $\mathbb{P}$ of size $d \times l$ is obtained with in general, fewer columns than rows and satisfying the conditions (i) $\mathbb{P}^{\mathsf{T}}\mathbb{P} = (d/l)\mathbb{I}$ and (ii) $\mathbb{E}[\mathbb{P}\mathbb{P}^{\mathsf{T}}] = \mathbb{I}$ (the identity matrix). A total of l zeroth order gradient estimates are then obtained and summed with each column of the $\mathbb{P}$ matrix. This is then used in the gradient update procedure. Various cases such as coordinate descent, spherical smoothing etc., are then considered, and rate bounds on the algorithm in both non-convex and convex cases are obtained. Almost sure convergence of the iterates in the convex case is also shown in (Rando *et al.*, 2024).

Another recent work along these lines in (Wang and Feng, 2024). Measurement noise is not considered in the system observations and the only noise that is present is in the gradient search directions. The convergence rates of such algorithms for Lojasiewicz functions which are generalizations of the Polyak-Lojasiewicz (PL) functions are obtained. Assuming existence of an almost sure limit point of the parameter sequence $\theta_n, n \geq 0$, the rate of convergence of $\{f(\theta_n)\}$ and $\{\theta_n\}$ is obtained. For a class of smooth as well as convex and non-smooth Lojasiewicz functions, the convergence is shown to be faster than standard zeroth-order gradient search. The work of (Kornowski and Shamir, 2024) provides a zeroth-order stochastic optimization scheme that produces the complexity of obtaining a (δ, ϵ)-stationary point of a possibly non-smooth and non-convex Lipschitz objective. Their algorithm incorporates a two-measurement gradient estimator using a common random

noise sequence and with perturbations that are distributed uniform on the unit sphere. The proposed algorithm requires $O(d\delta^{-1}\epsilon^{-3})$ function evaluations which the authors argue is the best complexity obtained so far.

There is also work on algorithms that provide better bounds due to reduced variance in the iterates. In (Duchi *et al.*, 2012), a stochastic optimization algorithm for a smoothed convex function that works with sub-differentials is presented that is however not a zeroth-order stochastic optimization scheme. The authors consider sample average of the estimates and show that the same has a better (finite-time) convergence rate due to the resulting lower variance in the iterates with extra averaging. Zeroth-order stochastic gradient search algorithms with variance reduction have been presented in (Ji *et al.*, 2019). In (Huang *et al.*, 2020), a class of Franke-Wolfe methods using zeroth-order stochastic gradient estimation approaches involving an accelerated scheme with reduced variance are presented. Their approach shows improved function query complexity for finding an approximate stationary point.

4

Asymptotic analysis of stochastic gradient algorithms

Consider the following stochastic gradient algorithm for solving $\theta^* = \arg\min_{\theta \in \Theta} f(\theta)$, given noisy sample access to f:

$$\theta_{n+1} = \theta_n - a(n)\widehat{\nabla} f(\theta_n), n \geq 0. \tag{4.1}$$

In Chapter 3, we learned how to form $\widehat{\nabla} f(\theta_n)$ from function samples so that $\widehat{\nabla} f(\theta_n) \approx \nabla f(\theta_n)$. Recall that these estimators incorporate search directions based on randomly perturbed parameters. The question of the error in the simultaneous perturbation-based estimate was also handled in the earlier chapter. In this chapter, we shall be concerned with whether θ_n governed by (4.1) converges to a local optimum θ^* or a neighborhood of it, when the underlying gradient estimates are biased.

The update in (4.1) is equivalent to

$$\theta_{n+1} = \theta_n - a(n)\Big(\nabla f(\theta_n) + \beta_n + \eta_n\Big), \tag{4.2}$$

where $\eta_n = \widehat{\nabla} f(\theta_n) - \mathbb{E}\left[\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n\right]$ is a martingale difference term, and $\beta_n = \mathbb{E}\left[\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n\right] - \nabla f(\theta_n)$ is the error in the gradient estimate. Recall that the latter is of the order $O(\delta^2)$.

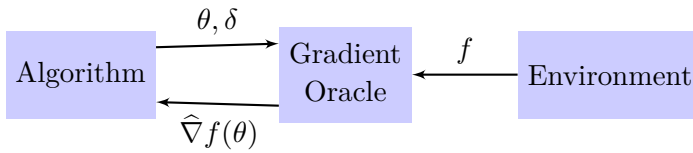


Figure 4.1: The interaction of the algorithms with a stochastic zeroth-order oracle that provides a gradient at the input point θ, with perturbation constant δ.

We analyse both cases of direct gradient measurements where information on sample performance (noisy though unbiased) gradients is available and zeroth order methods where one has access to only noisy function observations and not sample performance gradients. In the second case, we further consider the following sub-cases: (i) where the sensitivity parameter $\delta \equiv \delta_n \downarrow 0$ as $n \uparrow \infty$ and (ii) where the parameter $\delta > 0$ is held fixed in the algorithms. In the sub-case (ii), one can argue that there exists $\epsilon > 0$ such that $\beta_n \in \overline{B}_\epsilon(0)$ (the closed ball of radius ϵ centred at the origin) for all $n \geq 0$. In the above, $\mathcal{F}_n$ keeps a record of observations until time n. For instance, in the case of SPSA, one may let $\mathcal{F}_n = \sigma(\theta_m, m \leq n, \Delta_m, m < n), n \geq 1$, and $\mathcal{F}_0 = \sigma(\theta_0)$ as the sequence of sigma algebras generated by the associated quantities. This choice of $\mathcal{F}_n$ would ensure Δ_n is independent of $\mathcal{F}_n$, for all n.

Map of the results Table 4.1 provides a summary of the main convergence results for the stochastic gradient algorithm 4.1 with gradient estimates constructed using measurements from a zeroth-order oracle. The analysis of the previous chapter can be encapsulated into a biased gradient oracle, as illustrated in Figure 4.1. For a given input parameter θ and perturbation constant δ, one could use the schemes outlined in the previous chapter to obtain a gradient estimate $\widehat{\nabla} f(\theta)$ that satisfies

$$\left\| \mathbb{E}\left[\widehat{\nabla} f(\theta)\right] - \nabla f(\theta) \right\| \leq C_1 \delta^2, \text{ and } \mathbb{E}\left[\left\| \widehat{\nabla} f(\theta) - \mathbb{E}\left[\widehat{\nabla} f(\theta)\right] \right\|^2 \right] \leq \frac{C_2}{\delta^2}, \tag{4.3}$$

for given θ and some constants C_1 and C_2.

As mentioned before, we consider both (a) the case when noisy gradient-based though unbiased estimates are available and (b) the

gradient-free case where only noisy function measurements are available. In the second case (i.e., case (b)), we further consider two sub-cases for analysis. First, the gradient estimates at the nth update in (4.1) are obtained with input parameter θ_n and perturbation constant δ_n. The sequence $\{\delta_n\}$ is assumed to vanish asymptotically. Even though there is bias in the gradient estimates in this setting, the same asymptotically vanishes, as a result of which this setting allows analysis using the ODE approach for stochastic approximation. The unbiased (gradient-based) setting as well as the first case in the second setting (of asymptotically vanishing bias terms) form the content of Section 4.1.

The second sub-case above pertaining to zeroth-order gradient estimation, where the bias terms do not asymptotically vanish because the δ-parameter is kept constant requires a separate, more-detailed, analysis. Here, in the nth iteration of the stochastic gradient algorithm (4.1), the input parameter considered is θ_n while the perturbation parameter $\delta_n \equiv \delta > 0$ is a constant (i.e., is iteration-invariant). The analysis in this setting (with a constant δ) requires more sophisticated arguments as compared to the vanishing $\delta \equiv \delta_n$ case, and involves the theory of differential inclusions (DIs). Section 4.2 provides the DI analysis. Note that these algorithms are gradient-based algorithms where convergence can typically be claimed only to stationary points. In Section 7.2, however, we review work and provide some sufficient conditions under which one can avoid saddle points that form unstable equilibria of the associated ODE.

While this chapter focuses on the asymptotic convergence analysis, in the next chapter, we provide non-asymptotic bounds for the iterate sequence governed by (4.1).

4.1 Asymptotic convergence: An ODE approach

In this section, we analyse the asymptotic convergence of the algorithm (4.1) for two specific cases: (i) when direct (noisy and unbiased) gradient estimates are available, and (ii) when direct gradient estimates are not available but instead one has access to an oracle from where noisy objective function measurements at randomly perturbed parameter updates can be obtained and biased gradient estimates constructed

Table 4.1: Summary of the convergence results for the algorithm governed by (4.1)

Result type	Perturbation constant	Main result	Remark
Asymptotic convergence	diminishing	Theorem 4.4	Analysis via ODE limit
Asymptotic convergence	constant	Theorem 4.7	Analysis via DI limit
Non-asymptotic bound	constant	Theorem 5.2	Bound on iterate sequence

from these. For the latter case, we assume however, in this section, that the sensitivity parameter δ is diminishing. In other words, $\delta \equiv \delta_n \downarrow 0$ as $n \to \infty$, as a result of which we also show that the bias terms asymptotically vanish. This section thus treats the cases when either unbiased gradient estimates or gradient estimates that become asymptotically unbiased are available.

4.1.1 A variant of Kushner-Clark lemma for gradient systems

In this section, we provide a convergence result for a stochastic gradient algorithm with possibly biased gradient estimates. We apply this result to prove Theorem 4.4 for the case when unbiased gradient information is available. Subsequently, we analyze the stochastic gradient algorithm with biased gradient information, and use the aforementioned result in the latter setting to establish asymptotic convergence.

Consider a general stochastic gradient scheme as described in (1.3), involving the update rule below and under assumptions A2.1–A2.5.

$$\theta_{n+1} = \theta_n + a(n)(-\nabla f(\theta_n) + \beta_n + \eta_n). \tag{4.4}$$

The ODE associated with this scheme would be

$$\dot{\theta} = h(\theta) = -\nabla f(\theta). \tag{4.5}$$

For this ODE, $V(\theta) = f(\theta)$ serves as a Lyapunov function. Further, $\nabla V(\theta)^T h(\theta) \leq 0, \forall \theta$. One may now apply Lasalle's invariance principle, see Theorem A.7–Lemma A.9 to obtain the following:

Lemma 4.1. Any trajectory $\theta(\cdot)$ of (4.5) must converge to the largest invariant set that is a subset of $H \stackrel{\triangle}{=} \{\theta \mid \nabla f(\theta) = 0\}$

In the setting of gradient-based algorithms such as (1.3), we now have the following result that is easily obtained by combining Theorem 2.3 and Lemma 4.1.

Theorem 4.2. Under A2.1–A2.5, $\{\theta_n\}$ given by (4.4) satisfies $\theta_n \to \bar{H}$, where $\bar{H}$ denotes the largest invariant set contained in H.

In the case when the equilibrium points contained in $\bar{H}$ are isolated, we have the following result, see Corollary 3.3 of (Benaïm, 1996).

Corollary 4.3. Let the set H above comprise isolated equilibrium points. Then, under conditions of Theorem 4.2, $\{\theta_n\}$ given by (4.4) satisfies $\theta_n \to \theta^*$ for some (possibly sample path dependent) limit point $\theta^* \in \bar{H}$.

Corollary 4.3 is useful in most practical situations where the equilibrium points of the ODE (4.5) are isolated. Theorem 4.2 will be used in the analysis of algorithms that we shall present in later chapters. For this we shall assume that $\delta \to 0$ as $n \to \infty$. We shall also subsequently consider the case where the sensitivity parameter δ is held fixed to a small positive value and provide an asymptotic analysis where we show that the limiting dynamics of the recursion tracks a differential inclusion instead of an ODE.

If the set H specified in Theorem 4.2 consists of a single point, then the convergence would be to that point. Otherwise, the meaning of convergence to a set is depicted by two graphs in Figure 4.2. If all the elements in the set are disconnected, then convergence would be to a single point in the set, with the specific point to which the algorithm converges depending on the initial condition, the step size sequence, and the noise, as illustrated in the left graph of Figure 4.2, which contains two local minima and one local maximum. If some of the points are connected, then the algorithm could "bounce" between such points and not converge to a single point, as illustrated in the right graph of Figure 4.2, which contains a flat local minimal region and a saddle point.

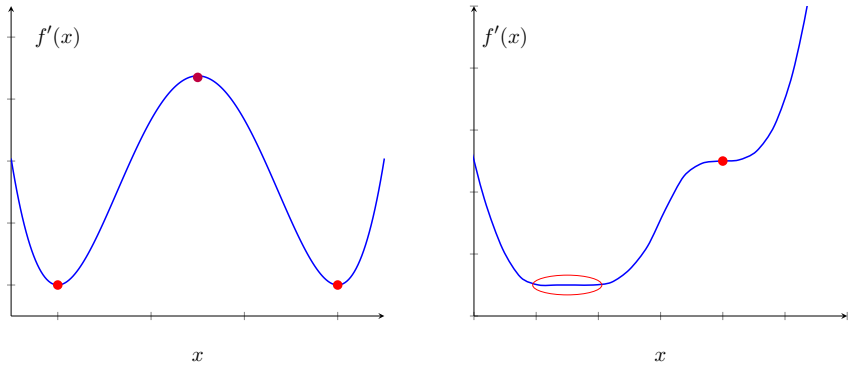


Figure 4.2: Two graphs illustrating the types of convergence for a stochastic gradient (SG) algorithm. In the left graph, an SG algorithm for minimization would converge to one of the two local minima or the local maximum indicated by the filled (red) circles, where which one it reaches depends on the starting point and the noise. In the right graph, the SG algorithm could converge to the saddle point indicated by the filled (red) circle or would eventually bounce between points in the circled (in red) interval unless the noise goes to zero. As long as the gradient estimate remains appropriately noisy, the SA algorithm would eventually move away from the local maximum in the left graph and away from the saddle point in the right graph.

"Unstable" points such as local maxima (in minimization problems) and saddle points can be avoided by ensuring that the gradient estimate is suitably noisy, to be described in more detail below.

Since the ODE tracked by the iteration (4.4) is $\dot{\theta} = -\nabla f(\theta)$, we know that its stationary points will be local maxima or minima, saddle points, or points of inflection. If these points are isolated, then the algorithm (4.4) will a.s. converge to a (possibly) sample path-dependent stationary point. Under additional assumptions, one can ensure convergence to a local minimum, thereby avoiding convergence to local maxima or saddle points. One such assumption is that the stationary points are hyperbolic, i.e., the Hessian $\nabla^2 f$ does not have eigenvalues on the imaginary axis. Then locally, it has a 'stable manifold' of dimension equal to the number of eigenvalues in the left half plane and an unstable manifold with the complementary dimension. A trajectory on the former converges to the stationary point along the stable manifold, whereas one on the latter moves away from it on the unstable manifold. A trajectory initiated anywhere else also eventually moves away. Thus, if there is

at least one unstable eigenvalue, the trajectories move away from the stationary point except on the stable manifold, a set of zero Lebesgue measure. Hence, if the noise is omnidirectional, i.e., rich in all directions in a certain precise sense, the iterations will be pushed away from the stable manifold often enough for the iterates to move away from the stationary point for good, a.s. Then the iterates will a.s. converge to a local minimum, where there are no unstable directions. In case the conditions on noise cannot be verified for the problem at hand, one can possibly add extraneous i.i.d. zero mean noise and have an SA update iteration of the form

$$\theta_{n+1} = \theta_n - a(n)(\widehat{\nabla} f(\theta_n) + \varphi_n), \tag{4.6}$$

where φ_n is extraneous noise added to ensure that the algorithm avoids saddle points/local maxima. A simple choice is to sample φ_n from the d-dimensional unit sphere uniformly. In practice, it may not be necessary to add such a noise factor extraneously, since the algorithm has an inherent noise component in the gradient estimates. We discuss escaping saddle points in more detail in Chapter 7.

4.1.2 Stochastic gradient algorithm using unbiased (direct) gradient estimates

We begin by considering the case when unbiased direct (noisy) gradient measurements are available. This would correspond to the setting of infinitesimal perturbation analysis (IPA) based estimators where information on direct sample performance gradients is available and one does not resort to zeroth-order gradient estimation methods.

To solve (1.1), a stochastic gradient algorithm would update as follows:

$$\theta_{n+1} = \theta_n - a(n)\widehat{\nabla} f(\theta_n), \tag{4.7}$$

where $\widehat{\nabla} f(\theta_n)$ is an estimate of the gradient $\nabla f(\theta_n)$, and $\{a(n)\}$ are (pre-determined) step-sizes satisfying standard Robbins-Monro step-size conditions (see A4.3 below).

In a zeroth-order setting, the gradient information is not directly available, and instead, the optimization algorithm has oracle access to

noise-corrupted function measurements. We also present in the latter case, an analysis of the resulting stochastic approximation scheme with gradient estimates obtained from zeroth-order information. Such estimates are not unbiased, but feature a parameter that can reduce the bias at the cost of variance. As mentioned, before getting to zeroth-order gradient estimation, we shall cover a simpler setting where unbiased gradient information is indeed available, i.e., $\mathbb{E}(\widehat{\nabla} f(\theta_n)) = \nabla f(\theta_n)$. In this case, the algorithm in (4.7) becomes an instance of the seminal stochastic approximation scheme proposed by Robbins and Monro in 1951. The latter algorithm was proposed to find the zeroes of a function, and in the case of (4.7), the function of interest is ∇f.

The algorithm in (4.7) can be shown to converge to local optima of f, and we make this claim precise, by starting with the necessary assumptions below.

A4.1. ∇f is a Lipschitz continuous $\mathbb{R}^d$-valued function.

A4.2. $\widehat{\nabla} f(\theta_n)$ is an unbiased estimate of the gradient $\nabla f(\theta_n)$, i.e., $\mathbb{E}\left[\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n\right] = \nabla f(\theta_n)$, where $\mathcal{F}_n = \sigma(\theta_m, m \leq n)$ denotes the underlying sigma-field. Further, there exists $\sigma > 0$ such that

$$\mathbb{E}\left[\left\|\widehat{\nabla} f(\theta_n) - \mathbb{E}\left[\widehat{\nabla} f(\theta_n)\middle| \mathcal{F}_n\right]\right\|^2\right] \leq \sigma^2 < \infty. \tag{4.8}$$

A4.3. The step-sizes satisfy $\sum_n a(n) = \infty$ and $\sum_n a(n)^2 < \infty$.

A4.4. The iterates $\{\theta_n, n \geq 0\}$ are stable, i.e., $\sup_n \|\theta_n\| < \infty$, a.s.

Theorem 4.4. Assume A4.1–A4.4. Let $\bar{H}$ denote the largest invariant set contained in $\{\theta \mid \nabla f(\theta) = 0\}$. Then, the sequence of iterates θ_n, $n \geq 0$, obtained from (4.7), satisfy

$$\theta_n \to \bar{H} \text{ a.s. as } n \to \infty.$$

Before presenting a proof of this result, we discuss below the assumptions made. First, the continuity requirement on the objective function $h(\theta)$ in A4.1 is standard to the analysis of stochastic approximation algorithms.

Indeed for the setting considered here, $h(\theta) = -\nabla f(\theta)$. Second, the unbiasedness condition in A4.2 is not satisfied in a zeroth-order optimization setting, where the gradient information is directly unavailable, and instead, one needs to infer this through measurements of the objective function at any query point. In the following section, we shall discuss the simultaneous perturbation trick, leading to asymptotically-unbiased gradient estimates, in place of A4.2. Third, the condition on step-sizes in A4.3 are standard requirements in stochastic approximation, and the reader is referred to the next chapter for a brief motivation (or Chapter 2 of (Borkar, 2022) for a detailed description). Fourth, the stability requirement in A4.4 is standard in the analysis of stochastic approximation algorithms, and this assumption was discussed in detail in the previous chapter, see Section 2.3.

Proof of Theorem 4.4

For proving Theorem 4.4, we shall invoke Theorem 4.2.

Proof. The update in (4.7) is equivalent to

$$\theta_{n+1} = \theta_n - a(n)\Big(\nabla f(\theta_n) + \eta_n\Big), \tag{4.9}$$

where $\eta_n = \widehat{\nabla} f(\theta_n) - \mathbb{E}\left[\widehat{\nabla} f(\theta_n)\middle| \mathcal{F}_n\right]$ is a martingale difference term. The equivalent update rule above used the fact that $\mathbb{E}\left[\widehat{\nabla} f(\theta_n)\middle| \mathcal{F}_n\right] = \nabla f(\theta_n)$, which holds by assumption A4.2.

The mean ODE underlying (4.1) is

$$\dot{\theta} = -\nabla f(\theta), \tag{4.10}$$

with limit set $H = \{\theta : \nabla f(\theta)) = 0\}$.

To apply Theorem 4.2, we verify a few conditions below.

1. A4.1 implies A2.1.
2. Since $\beta_n = 0,\ \forall n$, A2.2 is trivially satisfied.
3. A4.3 implies A2.3.

4. To verify A2.4, we first recall a martingale inequality attributed to Doob (also given as (2.1.7) on pp. 27 of (Kushner and Clark, 1978)):

$$\mathbb{P}\left(\sup_{m\geq 0}\|W_m\|\geq\epsilon\right)\leq\frac{1}{\epsilon^2}\lim_{m\to\infty}\mathbb{E}\,\|W_m\|^2\,. \tag{4.11}$$

Applying the inequality above to $W_m \triangleq \sum_{i=n}^{m} a(i)\eta_i$, $m \geq n$ and $n \geq 1$, we obtain

$$P\left(\sup_{m\geq n}\left\|\sum_{i=n}^{m}a(i)\eta_i\right\|\geq\epsilon\right)\leq\frac{1}{\epsilon^2}\mathbb{E}\left\|\sum_{i=n}^{\infty}a(i)\eta_i\right\|^2=\frac{1}{\epsilon^2}\sum_{i=n}^{\infty}a(i)^2\mathbb{E}\,\|\eta_i\|^2\,. \tag{4.12}$$

The last equality above follows by observing that, for $k < l$, $\mathbb{E}\,[\eta_k^{\mathrm{T}}\eta_l] = \mathbb{E}\,[\eta_k^{\mathrm{T}}\mathbb{E}\,[\eta_l|\,\mathcal{F}_k]] = 0$.

Now, using the square-summability of the stepsize in A4.3 and (4.8) in A4.2, we have

$$P\left(\sup_{m\geq n}\left\|\sum_{i=n}^{m}a(i)\eta_i\right\|\geq\epsilon\right)\leq\frac{1}{\epsilon^2}\sum_{i=n}^{\infty}a(i)^2\mathbb{E}\,\|\eta_i\|^2\leq\frac{\sigma^2}{\epsilon^2}\lim_{n\to\infty}\sum_{i=n}^{\infty}a(i)^2$$
$$\to 0 \text{ as } n\to\infty.$$

Thus, $\{\theta_n\}$ converges a.s. to the set $\bar{H}$ by an application of Theorem 4.2. □

4.1.3 Stochastic gradient algorithm using (zeroth-order) biased gradient estimates

We now consider the case where we have zeroth-order gradient estimates constructed from (noisy) function measurements obtained from an oracle. The bias in these gradient estimates is seen to vanish asymptotically as we allow the sensitivity parameter δ to tend to zero.

We analyze the following stochastic gradient algorithm:

$$\theta_{n+1} = \theta_n - a(n)\widehat{\nabla} f(\theta_n), \tag{4.13}$$

where $\widehat{\nabla} f(\theta_n)$ is formed using the unified estimate from the previous chapter, which is recalled below.

$$\widehat{\nabla} f(\theta_n) = \left(\frac{y_n^+ - y_n^-}{2\delta_n} \right) V(n), \tag{4.14}$$

where $y_n^+ = f(\theta_n + \delta_n U(n)) + \xi_n^+$, and $y_n^- = f(\theta_n - \delta_n U(n)) + \xi_n^-$. The reader is referred to Chapter 3 for a variety of choices for the random vectors $U(n), V(n)$.

For the analysis of this algorithm, we require the following assumptions in addition to A4.4 listed earlier: Let $\mathcal{F}_n = \sigma(\theta_i, i \le n, U(i), V(i), i < n, \xi_i^{\pm}, i < n)$, $n \ge 1$ denote a sequence of sigma fields.

A4.5. The noise factors $\xi^{\pm}$ in (4.14) satisfy

$$\mathbb{E}[\xi_n^+ - \xi_n^- | \mathcal{F}_n] = 0, \quad \text{and} \quad \mathbb{E}[(\xi_n^+ - \xi_n^-)^2 | \mathcal{F}_n] \le \sigma^2 < \infty, \forall n \ge 1. \tag{4.15}$$

A4.6. The objective function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$\mathbb{E}[f(\theta_n \pm \delta_n U(n))^2 \mid \mathcal{F}_n] \le B < \infty, \forall n. \tag{4.16}$$

A4.7. The step-sizes $a(n)$ and perturbation constants δ_n are positive, for all n and satisfy

$$a(n), \delta_n \to 0 \text{ as } n \to \infty, \sum_n a(n) = \infty \text{ and } \sum_n \left(\frac{a(n)}{\delta_n} \right)^2 < \infty.$$

Assuming $f \in \mathbb{C}^3$, and using Assumptions A4.5 to A4.6, it is possible to infer the following bias and variance bounds on the gradient estimator (4.14):

$$\begin{aligned} \forall n \ge 1, \left\| \mathbb{E}\left[\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n \right] - \nabla f(\theta_n) \right\| \le C_1 \delta_n^2, \text{ and} \\ \mathbb{E}\left[\left\| \widehat{\nabla} f(\theta_n) - \mathbb{E}\left[\widehat{\nabla} f(\theta_n) \middle| \mathcal{F}_n \right] \right\|^2 \right] \le \frac{C_2}{\delta_n^2}, \end{aligned} \tag{4.17}$$

for some constants C_1 and C_2. A straightforward adaptation of the proof of Proposition 3.1 leads to the bound in Equation (4.17).

The result below establishes asymptotic convergence of (4.13) to stationary points of f and the bounds in (4.17) is a crucial ingredient in the proof.

Theorem 4.5. Assume A4.5–A4.7, A4.4, and that f is L-smooth as well as three times continuously differentiable with bounded third derivative, i.e., $f \in \mathbb{C}^3$. Let $\bar{H}$ denote the largest invariant set contained in $\{\theta \mid \nabla f(\theta) = 0\}$. Then, the iterates θ_n, $n \geq 1$, updated according to (4.13), satisfy

$$\theta_n \to \bar{H} \text{ a.s. as } n \to \infty.$$

Proof. We first rewrite the update rule (4.13) as follows:

$$\theta_{n+1} = \theta_n - a(n)(\nabla f(\theta_n) + \eta_n + \beta_n), \tag{4.18}$$

where $\eta_n = \widehat{\nabla} f(\theta_n) - \mathbb{E}\left[\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n\right]$ is a martingale difference term, and $\beta_n = \mathbb{E}\left[\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n\right] - \nabla f(\theta_n)$ is the bias in the gradient estimate.

Convergence of (4.13) can be inferred from Theorem 4.2, provided we verify the necessary assumptions, and we do this verification below.

- f is L-smooth implies A2.1.
- From (4.17), we have $\beta_n = O(\delta_n^2)$. In conjunction with A4.7, we have $\beta_n \to 0$, verifying A2.2.
- Applying Doob's martingale inequality, $W_m = \sum_{i=n}^{m} a(i)\eta_i$, $m \geq n$ and $n \geq 1$, we obtain

$$\begin{aligned} \mathbb{P}\left(\sup_{m \geq n} \left\| \sum_{i=n}^{m} a(i)\eta_i \right\| \geq \epsilon\right) &\leq \frac{1}{\epsilon^2} \mathbb{E} \left\| \sum_{i=n}^{\infty} a(i)\eta_i \right\|^2 \\ &= \frac{1}{\epsilon^2} \sum_{i=n}^{\infty} a(i)^2 \mathbb{E} \left\| \eta_i \right\|^2, \end{aligned} \tag{4.19}$$

  where, as in the proof of Theorem 4.4, the last equality used $\mathbb{E}\left[\eta_k^{\mathsf{T}} \eta_l\right] = 0$ for $k < l$. This verifies A2.4.

Using (4.17), we have

$$\mathbb{E}\left\|\eta_n\right\|^2 \leq \frac{C_2}{\delta_n^2}. \tag{4.20}$$

Now, substituting the bound in (4.20) into (4.19), we obtain

$$\lim_{n\to\infty} P\left(\sup_{m\geq n}\left\|\sum_{i=n}^{m} a(i)\eta_i\right\| \geq \epsilon\right) \leq \frac{C_2}{\epsilon^2}\lim_{n\to\infty}\sum_{i=n}^{\infty}\frac{a(i)^2}{\delta_i^2} = 0.$$

The equality above follows from A4.7, as a consequence of $\sum_n \left(\frac{a(n)}{\delta_n}\right)^2 < \infty$.

The main claim now follows by an application of Theorem 4.2. □

Remark 4.1. The above result shows that the ODE tracked by (4.13) is (4.5). Except for one gradient estimation scheme, all the other schemes that we consider (see Chapter 3) track the ODE (4.5). However, for the case when the truncated Cauchy smoothed functional (TCSF) gradient estimator (3.31) is used, it can be seen that the ODE tracked is the following:

$$\dot{\theta}(t) = -c_2\nabla f(\theta), \tag{4.21}$$

with $c_2 > 0$. While the asymptotic convergence in this case is also to the stable fixed points of the ODE that is qualitatively the same as (4.5), the effect of the multiplicative constant $c_2 > 0$ manifests in the speed of convergence of the ODE's trajectories to the ODE's stable fixed points. In particular, $c_2 > 1$ would result in faster convergence of the trajectories of (4.21) as compared to that of the ODE (4.5). As mentioned, the latter ODE is the one tracked by all the other algorithms studied so far.

4.2 Asymptotic convergence: A differential inclusions approach

We now consider the case when gradient estimators such as (4.14) are considered but where $\delta > 0$ is held constant. This ensures that there is a bias in the gradient estimates that however does not asymptotically vanish as with the previous case.

4.2.1 Assumptions

We make the following assumptions:

A4.8. $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable. Furthermore, $\|\nabla f(\theta)\| \leq \tilde{K}(1 + \|\theta\|)$ for all $\theta \in \mathbb{R}^d$, for some $\tilde{K} > 0$.

A4.9. $\{\eta_n\}$ is a square-integrable martingale difference sequence w.r.t. the filtration $\{\mathcal{F}_n\}$, where $\mathcal{F}_n = \sigma(\theta_m, m \leq n, \eta_m, m < n), n \geq 0$. Further,

$$\mathbb{E}[\|\eta_n\|^2 \mid \mathcal{F}_n] \leq K_1(1 + \|\theta_n\|^2),$$

for some constant $K_1 > 0$.

A4.10. $a(n) > 0, \forall n$. Further, $\sum_n a(n) = \infty$ and $\sum_n a(n)^2 < \infty$.

A4.11. $\sup_n \|\theta_n\| < \infty$ w.p. 1.

A sufficient condition for the second part of Assumption A4.8 (in addition to f being continuously differentiable) is that the function ∇f is a Lipschitz continuous function of θ. This is because in such a case

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\| \leq Q \|\theta_1 - \theta_2\|,$$

for some constant $Q > 0$ and for any $\theta_1, \theta_2 \in \mathbb{R}^d$. Then by letting $\theta_1 = \theta$ and $\theta_2 = 0$, we get

$$\|\nabla f(\theta)\| - \|\nabla f(0)\| \leq \|\nabla f(\theta) - \nabla f(0)\| \leq Q \|\theta\|,$$

implying $\|\nabla f(\theta)\| \leq \tilde{K}(1 + \|\theta\|)$ with $\tilde{K} = \max(Q, \|\nabla f(0)\|)$.

Assumption A4.9 is on the noise sequence $\{\eta_n\}$. From the manner in which it is defined, viz., $\eta_n = \widehat{\nabla} f(\theta_n) - \mathbb{E}\left[\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n\right]$ and the various forms of the gradient estimators $\widehat{\nabla} f(\theta_n)$ discussed previously and the assumptions on the measurement noise there, it can be easily seen that this condition will be satisfied.

Assumption A4.10 is on the step size sequence and is a standard requirement in stochastic approximation schemes. The condition on

non-summability of the step size is needed to track the asymptotic behaviour of the limiting differential equation or inclusion as the case may be. The second condition ensures, in particular, that the errors due to noise asymptotically vanish.

Finally, assumption A4.11 is necessary to establish convergence of gradient-descent scheme but is a non-trivial requirement. Certain sufficient conditions for stability of stochastic approximation schemes that rely mainly on the underlying ODE and a certain scaling limit of the same are given in (Borkar and Meyn, 1999). For the case of stochastic recursive inclusions (SRI), i.e., stochastic approximations with set-valued maps, similar conditions have recently been provided in (Ramaswamy and Bhatnagar, 2016a; Ramaswamy and Bhatnagar, 2018). In particular, (Ramaswamy and Bhatnagar, 2018) considers a gradient recursion with errors in the setting of SRI and provides sufficient conditions for stability of the scheme. We present these conditions from (Ramaswamy and Bhatnagar, 2018) in the subsection following the convergence proof. Prior work, for instance, (Benaïm, 1996; Kushner and Clark, 1978; Kushner and Yin, 2003) show convergence of stochastic approximation assuming stability of the stochastic iterates. Further, (Benaïm *et al.*, 2005) proves the almost sure convergence of SRI again assuming stability of the iterates. As mentioned earlier, if one is unable to ensure stability of the stochastic iterates, a common approach is to project these to a large enough compact set that would ensure boundedness of the iterates. This however comes at the cost of introducing spurious fixed points on the projection set boundary to which the recursion might converge as well, see (Kushner and Clark, 1978; Kushner and Yin, 2003) for detailed analyses of projected stochastic approximations.

4.2.2 Proof of Convergence

Let $G(\theta) = \nabla f(\theta) + \overline{B}_\epsilon(0)$, where $\overline{B}_\epsilon(0)$ is a closed ball of radius $\epsilon > 0$ around the origin. In other words, $G(\theta) = \overline{B}_\epsilon(\nabla f(\theta))$ is a closed ball of radius $\epsilon > 0$ around $\nabla f(\theta)$.

Lemma 4.6. The set-valued map G is a Peano map.

Proof. Recall Definition A.6 for definition of Peano map. We shall verify the three conditions (i)-(iii) of Definition A.6. As noted earlier, for any $\theta \in \mathbb{R}^d$, $G(\theta)$ is a closed ball in $\mathbb{R}^d$ of radius ϵ centred at $\nabla f(\theta)$. Thus, it is clearly convex and compact. Now for any $y \in G(\theta)$,

$$\begin{aligned} \|y\| &\leq \|\nabla f(\theta)\| + \|y - \nabla f(\theta)\| \\ &\leq \tilde{K}(1 + \|\theta\|) + \epsilon \\ &\leq \bar{K}(1 + \|\theta\|), \end{aligned}$$

where $\bar{K} = \tilde{K} + \epsilon$. The second inequality above follows from the *smoothness* assumption A4.8. Since y above is arbitrary, it follows that

$$\sup_{y \in G(\theta)} \|y\| \leq \bar{K}(1 + \|\theta\|).$$

Thus $G(\theta)$ is pointwise bounded.

Finally, consider a sequence $\theta_n, n \geq 0$ of parameters and another sequence $y_n, n \geq 0$ of points such that $y_n \in G(\theta_n)$, $\forall n$. Further, let $\theta_n \to \theta$ and $y_n \to y$ as $n \to \infty$. Now given $\delta > 0$ small, let N be large enough so that $\|y_n - y\| < \delta/2$ and similarly $\|\nabla f(\theta_n) - \nabla f(\theta)\| < \delta/2$, respectively, $\forall n > N$. Then,

$$\begin{aligned} \|y - \nabla f(\theta)\| &\leq \|y - y_n\| + \|y_n - \nabla f(\theta_n)\| \\ &\quad + \|\nabla f(\theta_n) - \nabla f(\theta)\| \\ &\leq \epsilon + \delta. \end{aligned}$$

Since $\delta > 0$ is arbitrary, let $\delta \to 0$. It then follows that $\|y - \nabla f(\theta)\| \leq \epsilon$, implying that $y \in G(\theta)$. Thus G is also upper-semicontinuous and the claim follows. □

Consider now the Differential Inclusion (DI):

$$\dot{\theta}(t) \in -G(\theta(t)). \tag{4.22}$$

Here $-G(\theta(t))$ is used to denote the set $\{-g \mid g \in G(\theta(t))\}$. The next result follows directly from (Benaïm *et al.*, 2005).

Theorem 4.7. The iterates (4.2) converge to a closed connected internally chain transitive and invariant set of the DI (4.22).

Proof. The claim follows from Theorem 3.6 and Lemma 3.8 of (Benaïm *et al.*, 2005). □

Consider also the associated ODE that would result from the case of $\epsilon = 0$:

$$\dot{\theta}_t = -\nabla f(\theta_t). \tag{4.23}$$

This will be the case when either the information on the gradient $\nabla f(\theta)$ is fully known for all θ and a (true) gradient scheme with noise is used or else the sensitivity parameter δ is replaced by a slowly decreasing $\delta_n \to 0$. Both of these cases have been analysed for their convergence in Section 4.1.

As seen in Section 4.1, in the second case above, the square summability requirement of the step size sequence $\{a(n)\}$ is considerably tightened. More specifically, the condition $\sum_n a(n)^2 < \infty$ in A4.10 is replaced by the more stringent requirement $\sum_n \left(\frac{a(n)}{\delta_n}\right)^2 < \infty$ in A4.7. The latter has the effect of significantly constraining the learning rates in the update recursion.

Let $\mathcal{M}$ denote the minimum set of f and suppose the regular values of f, i.e., θ for which $\nabla f(\theta) \neq 0$ are dense in $\mathbb{R}^d$, then the chain recurrent set of f is a subset of it's minimum set, see *Proposition 4* of Hurley (Hurley, 1995). As shown earlier, the gradient descent scheme without errors (i.e., with $\epsilon = 0$), will converge to $\mathcal{M}$ almost surely.

We now state *Theorem 3.1* of (Benaïm *et al.*, 2012) adapted to the setting considered here..

Theorem 4.8. Given $\delta > 0$, $\exists \epsilon(\delta) > 0$ such that the chain recurrent set of (4.22) is within the δ-open neighborhood of the chain recurrent set of (4.23) for all $\epsilon \leq \epsilon(\delta)$.

It follows as a consequence of Theorem 4.7 and Theorem 4.8 that (4.2) with $\epsilon < \epsilon(\delta)$ (cf. Theorem 4.8) converges almost surely to $N^{\delta}(\mathcal{M})$.

4.2.3 A Set of Stability Conditions for Stochastic Recursive Inclusions

We now present a set of conditions from (Ramaswamy and Bhatnagar, 2016a; Ramaswamy and Bhatnagar, 2018) that ensure that the stochastic recursive inclusion (4.2) remains stable, i.e., that $\sup_n \|\theta_n\| < \infty$ a.s., that was the last assumption for our analysis of the recursion (4.2). The conditions that we present are a generalization of stability conditions for stochastic approximation presented in (Borkar and Meyn, 1999).

Recall from Lemma 4.6 that G is a Peano or Marchaud map. For each integer $c \geq 1$, let

$$G_c(\theta) := \left\{ \frac{y}{c} \mid y \in G(c\theta) \right\}.$$

Let

$$G_\infty(\theta) := \overline{co}(\text{Limsup}_{c\to\infty} G_c(\theta)),$$

where

$$\text{Limsup}_{x_n \to x} J(x_n) = \{y \in \mathbb{R}^d \mid \liminf_{x_n \to x} d(y, J(x_n)) = 0\},$$

see Definition A.7. Given $A \subseteq \mathbb{R}^d$, the convex closure of A, denoted by $\overline{co}(A)$, is the closure of the convex hull of A. It is worth noting that $Limsup_{c\to\infty} G_c(\theta)$ is non-empty for every $\theta \in \mathbb{R}^d$. It is also shown in Lemma 1 of (Ramaswamy and Bhatnagar, 2018) that G_∞ is Marchaud. Thus, from (Aubin and Cellina, 1984), the DI $\dot{\theta}(t) \in -G_\infty(\theta(t))$ has at least one solution that is absolutely continuous.

We make the following additional assumptions:

A4.12. $\dot{\theta}(t) \in -G_\infty(x(t))$ has an attractor set $\mathcal{A}$ such that $\mathcal{A} \subseteq B_a(0)$ for some $a > 0$ and $\overline{B}_a(0)$ is a fundamental neighborhood of $\mathcal{A}$.

Since $\mathcal{A} \subseteq B_a(0)$ is compact, we have that $\sup_{\theta\in\mathcal{A}} \|\theta\| < a$.

A4.13. Let $c_n \geq 1$ be an increasing sequence of integers such that $c_n \uparrow \infty$ as $n \to \infty$. Further, let $\theta_n \to \theta$ and $y_n \to y$ as $n \to \infty$, such that $y_n \in G_{c_n}(\theta_n)$, $\forall n$, then $y \in G_\infty(\theta)$.

It can be shown that the existence of a global Lyapunov function for $\dot{\theta}(t) \in -G_\infty(\theta(t))$ is sufficient to guarantee that A4.12 holds. Further, A4.13 is satisfied when ∇f is Lipschitz continuous.

Theorem 4.9. Under A4.8–A4.10 and A4.12-A4.13, the stochastic update sequence given by (4.2) remains stable, i.e., $\sup_n \|\theta_n\| < \infty$ almost surely.

A detailed proof of this result is given in Theorem 1 of (Ramaswamy and Bhatnagar, 2018). What is important to note here as also with the original result of (Borkar and Meyn, 1999) (that was for the case of stochastic updates involving single-valued functions as opposed to set-valued maps as considered above), both the additional assumptions A4.12 and A4.13 involve only deterministic systems, more precisely scaled Differential Inclusions. Asymptotic stability properties of these systems and in particular the limiting system are enough to guarantee stability of the original stochastic recursions.

4.3 Bibliographic remarks

The steps involved in establishing the convergence analysis of stochastic approximation algorithms with gradient estimators mirrors largely similar analysis for broader stochastic approximation algorithms dealt with in Chapter 2, see (Benaïm, 1996; Borkar, 2022). Convergence analyses of zeroth-order stochastic gradient algorithms with diminishing step-sizes are for instance, available in (Spall, 1992; Spall, 1997) for the case of two and one-sided SPSA, in (Prashanth *et al.*, 2017) for the case of RDSA, as well as (Prashanth *et al.*, 2020) for deterministic perturbation RDSA.

Stochastic recursive inclusions or stochastic approximation with set-valued maps have been analysed for the first time in (Benaïm *et al.*,

2005). The recursion there involves a set-valued map with a martingale difference noise sequence and assumes stability of the stochastic iterates. The works in (Ramaswamy and Bhatnagar, 2016a; Ramaswamy and Bhatnagar, 2021) provide the first and only available sets of stability conditions for such recursions. Stability conditions for stochastic recursive inclusions with non-ergodic Markov noise are available in (Ramaswamy and Bhatnagar, 2019).

The works in (Ramaswamy and Bhatnagar, 2016b; Yaji and Bhatnagar, 2020) present the first convergence analyses of two-timescale stochastic recursive inclusions with set-valued maps on both timescales. In (Yaji and Bhatnagar, 2018), the analysis of stochastic recursive inclusions with set-valued maps and Markov noise in addition, has been conducted for the first time and in (Yaji and Bhatnagar, 2018), the same in the two-timescale case is conducted in (Yaji and Bhatnagar, 2020). The Markov noise is assumed to be dependent on the parameter sequence and in addition, depends on an additional control-valued sequence, and furthermore is assumed to have multiple stationary distributions. This combination makes it the hardest so far case of stochastic inclusions that has been analysed in the literature. Such algorithms are however seen to have applications in stochastic optimization as well as reinforcement learning. For instance, an application of (Karmakar and Bhatnagar, 2018) on two-timescale stochastic approximation with Markov noise was studied on an application of off-policy gradient temporal difference learning algorithms (Sutton *et al.*, 2009). Finally, the material on convergence of a zeroth-order stochastic gradient algorithm for a fixed δ-parameter is based on (Ramaswamy and Bhatnagar, 2018), where a corresponding set-valued map is obtained and analysed.

5

Non-asymptotic analysis of stochastic gradient algorithms

We consider a SG algorithm for solving (1.1), with an update iteration of the form:

$$\theta_{n+1} = \theta_n - a(n)\widehat{\nabla} f(\theta_n), n \geq 0. \tag{5.1}$$

We analyze the algorithm above with inputs from either an unbiased gradient oracle or a biased one, i.e., corresponding to the cases where $\mathbb{E}\left[\widehat{\nabla} f(\theta) \mid \theta\right] = \nabla f(\theta)$ and $\mathbb{E}\left[\widehat{\nabla} f(\theta) \mid \theta\right] = \nabla f(\theta) + O(\delta^2)$, with δ denoting the perturbation constant (see Chapter 3), respectively. The analysis in the former case serves as a useful contrast to the biased case, since the proof technique is similar, while there is a loss in convergence rate when one moves from an unbiased to a biased gradient oracle.

We consider an SG algorithm that runs for N iterations, and outputs a (possibly random) point θ_R, that could be chosen based on the iterates $\theta_1, \ldots, \theta_N$. For a general SG algorithm, we consider different performance metrics based on the nature of the underlying objective. More precisely, we consider the following cases:
(i) convex; (ii) strongly convex; and (ii) non-convex.
In case (i), we provide bounds on the optimization error, i.e., $\mathbb{E}\left(f(\theta_R) - f(\theta^*)\right)$, where θ^* is a minimum of f, whereas in case (ii), we establish bounds on the parameter error $\mathbb{E}\left\|\theta_R - \theta^*\right\|^2$. On the other

hand, in case (iii), i.e., when the objective is non-convex, it is difficult to bound the optimization/parameter errors. A popular alternative is to establish local convergence. i.e., to a point where the gradient of the objective is small (cf. (Ghadimi and Lan, 2013; Bottou *et al.*, 2018)). The following definition makes the optimization objectives apparent in all the cases studied in this chapter.

Definition 5.1. Let $\theta_R \in \mathbb{R}^d$ be the output of a SG algorithm and $\epsilon > 0$ be a target accuracy, then:

1. If f is non-convex, θ_R is called an ϵ-stationary point of (1.1), if $\mathbb{E}\left\|\nabla f\left(\theta_R\right)\right\|^2 \leq \epsilon$;
2. If f is convex, θ_R is called an ϵ-optimal point of (1.1), if $\mathbb{E}[f\left(\theta_R\right)] - f(\theta^*) \leq \epsilon$, where θ^* is a minimizer of f.
3. If f is strongly convex, θ_R is called an ϵ-optimal point of (1.1), if $\mathbb{E}\left[\left\|\theta_R - \theta^*\right\|^2\right] \leq \epsilon$, where θ^* is the unique minimizer of f.

The SG algorithms are judged using the iteration complexity, which is defined below.

Definition 5.2. For a given $\epsilon > 0$, the iteration complexity of an algorithm $\mathcal{A}$ is the number of iterations of $\mathcal{A}$ before finding an ϵ-stationary (resp. ϵ-optimal) point for a non-convex (resp. convex/strongly-convex) objective function.

For a gradient descent type algorithm, results from deterministic optimization lead to complexity bounds listed in Table 5.1, cf. (Wright and Recht, 2022, Chapter 3). The bounds in Table 5.1 are useful to compare against the corresponding cases in the stochastic case that we consider in this chapter. Moreover, as we shall see later, the case of biased gradient oracle results in bounds that are weaker than the unbiased counterpart.

For the bounds in this chapter, we consider a variant of SG algorithm, namely randomized stochastic gradient (RSG), which was proposed in (Ghadimi and Lan, 2013). This is a well-known scheme that provides a

Function Type	Condition	Iteration Complexity
Non-convex	$\|\|\nabla f(\theta^*)\|\| \leq \epsilon$	$n \geq \frac{2L}{\epsilon^2}[f(\theta_0) - f(\theta^*)]$
Convex	$\|f(\theta) - f(\theta^*)\| \leq \epsilon$	$n \geq \frac{L}{2\epsilon}\|\|\theta_0 - \theta^*\|\|^2$
Strongly Convex	$\|f(\theta) - f(\theta^*)\| \leq \epsilon$	$n \geq \frac{L}{\mu}\log\left(\frac{f(\theta_0) - f(\theta^*)}{\epsilon}\right)$

Table 5.1: Summary of iteration complexities of a gradient descent algorithm for deterministic smooth optimization. Here iteration complexity is the number of iterations n required to satisfy the condition specified in the second column. Here θ^* denotes an optimum of f, θ_0 is the starting point of the gradient descent algorithm, μ is the strong-convexity parameter, and L is the smoothness constant.

non-asymptotic bound on a random iterate visited by a SG algorithm. More precisely, suppose $\theta_1, \ldots, \theta_m$ be the iterates visited along a sample path of a SG algorithm that is run for m iterations. Then, the RSG algorithm would return an iterate θ_R that is picked randomly from the set $\{\theta_1, \ldots, \theta_m\}$. For the case where θ_R is picked uniformly at random from the set mentioned above, the RSG scheme for picking the aforementioned random iterate resembles the well-known Polyak-Ruppert iterate averaging scheme (Polyak and Juditsky, 1992; Ruppert, 1985) for stochastic approximation. The latter scheme performs averaging of all the iterates $\{\theta_i,\ i = 1, \ldots, m\}$, while RSG achieves the same effect, except that the averaging happens in expectation. Algorithm 2 presents the pseudocode of RSG algorithm that takes as input the probability mass function $P_R(\cdot)$ for picking a random variable from the set $\{1, \ldots, m\}$. The bounds we present are for the special case where P_R is the discrete uniform distribution over the aforementioned set.

In this chapter, we provide non-asymptotic bounds for Algorithm 2 with unbiased and biased gradient information, respectively, for three different assumptions on the underlying objective, namely convex, strongly convex and non-convex. In a zeroth-order setting, the RSG algorithm is provided gradient estimates formed using the simultaneous perturbation

Algorithm 2: RSG algorithm

Input: Initial point $\theta_1 \in \mathbb{R}^d$, iteration limit m, step sizes $\{a(k)\}_{k\geq 1}$ and probability mass function $P_R(\cdot)$ of a random variable R supported on $\{1, \ldots, m\}$.

for $k = 1, \ldots, m$ **do**

Form the gradient estimate $\widehat{\nabla} f(\theta_k)$ using one or more function measurements;

Perform the following stochastic gradient descent update:

$$\theta_{k+1} = \theta_k - a(k)\widehat{\nabla} f(\theta_k).$$

end for

Output: θ_R

method described in Chapter 3.

The rest of this chapter is organized as follows: Sections 5.1–5.3 present the non-asymptotic bounds with proofs for non-convex, convex, and strongly convex functions, respectively. In Section 5.4, we present two settings where the non-asymptotic bounds for the RSG algorithm features improved dimension dependence, as compared to those in Sections 5.1–5.3. In Section 5.5, we discuss a zeroth-order model variant, where the function measurements are biased. In Section 5.6, we present a minimax lower bound for an algorithm that has access to gradient estimates that satisfy a bias-variance tradeoff (e.g., see (4.3)). In Section 5.7, we outline the connection between smooth optimization in a zeroth-order setting and bandit convex optimization.

5.1 The non-convex case

We begin by considering the case of a non-convex objective function $f : \mathbb{R}^d \to \mathbb{R}$.

5.1.1 RSG with an unbiased gradient oracle

As a gentle start, first, we provide bounds for the simple "unbiased gradient" model, and subsequently analyze the other challenging model involving biased gradients.

In this model, we assume access to a stochastic first-order oracle, which for a given θ_k outputs a random estimate $\widehat{\nabla} f(\theta_k)$ of the gradient of f. We assume that the gradient estimate $\widehat{\nabla} f(\theta_k)$ satisfies the following assumption:

A5.1. Let $\mathcal{F}_k = \sigma(\theta_i, i \leq k)$. Recall $\mathbb{E}_k$ denotes the expectation w.r.t. $\mathcal{F}_k$. For any $k \geq 1$, we have

1. $\mathbb{E}_k \left[\widehat{\nabla} f(\theta_k) \right] = \nabla f\left(\theta_k\right),$
2. $\mathbb{E}_k \left[\left\| \widehat{\nabla} f(\theta_k) - \nabla f\left(\theta_k\right) \right\|^2 \right] \leq \sigma^2$, for some parameter $\sigma \geq 0$.

From the above, it is apparent that $\widehat{\nabla} f(\theta_k)$ is an unbiased estimate of $\nabla f(\theta_k)$ with bounded variance.

The results provide a bound on the gradient norm after m iterations of RSG. As mentioned earlier, under a non-convex objective, bounding the optimization error, i.e., $f(\theta_R) - f(\theta^*)$ is difficult, where θ^* is a local optima. However, a popular alternative is to show that the RSG algorithm converges to a point, where the gradient of the objective is small (quantified by a bound on the squared norm of the gradient) (cf. (Ghadimi and Lan, 2013; Bottou *et al.*, 2018)).

Theorem 5.1. (**Unbiased gradients: Non-convex case**) Suppose f is L-smooth and satisfies A5.1. Suppose that the RSG algorithm is run with the stepsize sequence set as

$$a(k) = a, \forall k \text{ with } a = \min\left\{\frac{1}{L}, \frac{c}{\sqrt{m}}\right\}, \tag{5.2}$$

for some constant $c > 0$. Then, for any $m \geq 1$, we have

$$\mathbb{E}\left[\left\|\nabla f\left(\theta_R\right)\right\|^2\right] \leq \frac{2LD_f}{m} + \frac{1}{\sqrt{m}}\left[\frac{2D_f}{c} + L\sigma^2 c\right],$$

where R is uniformly distributed over $\{1, \ldots, m\}$, θ^* is an optimal

solution to (1.1), and

$$D_f = f(\theta_1) - f(\theta^*). \tag{5.3}$$

Proof. Since f is L-smooth, we have

$$\begin{aligned} f\left(\theta_{k+1}\right) &\leq f\left(\theta_k\right) + \left\langle \nabla f\left(\theta_k\right), \theta_{k+1} - \theta_k \right\rangle + \frac{L}{2}\left\|\theta_{k+1} - \theta_k\right\|^2 \\ &= f\left(\theta_k\right) - a(k)\left\langle \nabla f\left(\theta_k\right), \widehat{\nabla} f(\theta_k) \right\rangle + \frac{L}{2}a(k)^2\left\|\widehat{\nabla} f(\theta_k)\right\|^2 \end{aligned}$$

Using $\mathbb{E}_k\left[\widehat{\nabla} f(\theta_k)\right] = \nabla f\left(\theta_k\right)$, and the following inequality[1]:

$$\mathbb{E}_k\left[\left\|\widehat{\nabla} f(\theta_k)\right\|^2\right] \leq \left\|\mathbb{E}_k\left[\widehat{\nabla} f(\theta_k)\right]\right\|^2 + \sigma^2,$$

we obtain

$$\begin{aligned} &\mathbb{E}_k[f\left(\theta_{k+1}\right)] \\ &\leq f\left(\theta_k\right) - a(k)\left\|\nabla f\left(\theta_k\right)\right\|^2 + \frac{L}{2}a(k)^2\left[\left\|\nabla f\left(\theta_k\right)\right\|^2 + \sigma^2\right] \\ &= f\left(\theta_k\right) - \left(a(k) - \frac{L}{2}a(k)^2\right)\left\|\nabla f\left(\theta_k\right)\right\|^2 + \frac{L}{2}a(k)^2\sigma^2. \end{aligned} \tag{5.4}$$

Re-arranging the terms above and setting $a(k) = a, \forall k \geq 1$, we obtain

$$a\left\|\nabla f\left(\theta_k\right)\right\|^2 \leq \frac{2\left[f\left(\theta_k\right) - \mathbb{E}_k[f\left(\theta_{k+1}\right)]\right]}{(2 - La)} + \frac{La^2\sigma^2}{(2 - La)}$$

Now, summing up the above inequality for $k = 1$ to m, we obtain

$$a\sum_{k=1}^{m}\left\|\nabla f\left(\theta_k\right)\right\|^2 \leq 2\sum_{k=1}^{m}\frac{\left[f\left(\theta_k\right) - \mathbb{E}_k[f\left(\theta_{k+1}\right)]\right]}{(2 - La)} + \frac{mL\sigma^2a^2}{(2 - La)}.$$

Taking total expectations on both sides of above equation, and using $\mathbb{E}\left[f\left(\theta_k\right)\right] \geq f(\theta^*)$, for all $k \geq 1$, we obtain

$$a\sum_{k=1}^{m}\mathbb{E}\left\|\nabla f\left(\theta_k\right)\right\|^2 \leq \frac{2\left(f(\theta_1) - f(\theta^*)\right)}{(2 - La)} + \frac{mL\sigma^2a^2}{(2 - La)}.$$

[1] When $\|\cdot\|$ is defined from an inner product, we have $\mathbb{E}\left[\|X - \mathbb{E}\left[X\right]\|^2\right] = \mathbb{E}\left[\|X\|^2\right] - \|\mathbb{E}\left[X\right]\|^2$.

Since θ_R is picked uniformly at random from $\{\theta_1, \ldots, \theta_m\}$ and $a \leq 1/L$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f(\theta_R)\|^2\right] &= \frac{1}{m}\sum_{k=1}^{m}\mathbb{E}\|\nabla f(\theta_k)\|^2 \\
&\leq \frac{1}{ma}\left[\frac{2D_f}{(2-La)} + L\sigma^2 m\frac{a^2}{(2-La)}\right] \\
&\leq \frac{1}{ma}\left[2D_f + L\sigma^2 m a^2\right] \\
&= \frac{2D_f}{ma} + L\sigma^2 a \\
&\leq \frac{2D_f}{m} max\left\{L, \frac{\sqrt{m}}{c}\right\} + L\sigma^2\frac{c}{\sqrt{m}} \\
&\leq \frac{2LD_f}{m} + \frac{2D_f}{c\sqrt{m}} + L\sigma^2\frac{c}{\sqrt{m}} \\
&= \frac{2LD_f}{m} + \frac{1}{\sqrt{m}}\left[\frac{2D_f}{c} + L\sigma^2 c\right].
\end{aligned}$$

The claim follows. □

5.1.2 RSG with a biased gradient oracle

We make the following assumptions for the non-asymptotic analysis of RSG algorithm in the zeroth-order setting:

A5.2. There exists a constant $B > 0$ such that $\|\nabla f(x)\|_1 \leq B, \forall x \in \mathbb{R}^d$.

A5.3. The gradient estimate $\widehat{\nabla} f(\theta_k)$ satisfies the following inequalities for all $k \geq 1$:

$$\left\|\mathbb{E}_k\left[\widehat{\nabla} f(\theta_k)\right] - \nabla f(\theta_k)\right\| \leq c_1\delta^2, \tag{5.5}$$

and

$$\mathbb{E}_k\left[\left\|\widehat{\nabla} f(\theta_k)\right\|^2\right] \leq \left\|\mathbb{E}_k\left[\widehat{\nabla} f(\theta_k)\right]\right\|^2 + \frac{c_2}{\delta^2}. \tag{5.6}$$

In the above, $\mathbb{E}_k$ is shorthand for $\mathbb{E}(\cdot \mid \mathcal{F}_k)$, with $\mathcal{F}_k$ denoting the sigma-field $\sigma(\theta_i, i \leq k)$.

As mentioned before, in the non-convex case, the gradient norm is a standard benchmark for quantifying the convergence rate of stochastic gradient algorithms. The main result concerning RSG's non-asymptotic performance is presented below.

Theorem 5.2.
Suppose the objective function f is L-smooth (see Definition 3.1), and assumptions A5.2–A5.3 hold. Suppose that the RSG algorithm is run with the stepsize $a(k) = a$ and perturbation constant $\delta(k) = \delta$ for each $k = 1, \ldots, m$, where

$$a = min\left\{\frac{1}{L}, \frac{1}{m^{2/3}}\right\}, \ \delta = \frac{1}{m^{1/6}}, \ \forall k \geq 1. \tag{5.7}$$

Then, choosing θ_R uniformly at random from $\{\theta_1, \ldots, \theta_m\}$, we have

$$\mathbb{E}\left\|\nabla f\left(\theta_R\right)\right\|^2 \leq \frac{2L(f(\theta_1) - f(\theta^*))}{m} + \frac{\mathcal{K}_1}{m^{1/3}}, \tag{5.8}$$

where $\mathcal{K}_1 = 2D_f d^{4/3} + \frac{4Bc_1}{d^{5/3}} + \frac{Lc_1^2}{d^{11/3}m} + Lc_2 d^{1/3}$, constants c_1, c_2 are defined in A5.3, B is as defined in A5.2,

$$D_f = f(\theta_1) - f(\theta^*), \tag{5.9}$$

and θ^* is a global optima of f.

Remark 5.1. From the bound in the result above, it is easy to see that an order $\mathcal{O}\left(\frac{1}{\epsilon^3}\right)$ iterations of the RSG algorithm are enough to find a point θ_R that satisfies $\mathbb{E}\left\|\nabla f\left(\theta_R\right)\right\|^2 \leq \epsilon$.

Remark 5.2. In comparison to the unbiased gradient information case handled in the previous section, the $\mathcal{O}\left(\frac{1}{m^{1/3}}\right)$ bound obtained here is weaker. This drop in rate is owing to the bias-variance tradeoff in the gradient estimates, i.e., choosing a very small perturbation constant δ improve the accuracy of the gradient estimate at the cost of increased variance, see Assumption A5.3. However, with additional structure, the rate can be improved to $\mathcal{O}\left(\frac{1}{m^{1/2}}\right)$. We mention two such settings

next. First, in the case of common random noise, discussed earlier in Section 3.3.4, we have $f(\theta) = \mathbb{E}_\xi(F(\theta, \xi))$. Assuming F is smooth in θ, for any given ξ, it is possible to establish a $\mathcal{O}\left(\sqrt{\frac{d}{m}}\right)$ bound on the gradient norm square, i.e., $\mathbb{E}\left\|\nabla f\left(\theta_R\right)\right\|^2$. Notice that this bound improves both the dependency on dimension d as well number of iterations m. The proof for such a result is analogous to the proof of Theorem 5.2 given below, and we leave it as an exercise. The second setting with improved bounds is that of sparse optimization. Assuming the gradient is s-sparse, i.e., $\|\nabla f(x)\|_0 \leq s, \forall x$, it is possible to establish a $\mathcal{O}\left(\sqrt{\frac{\log d}{m}}\right)$ bound on the gradient norm square. In comparison to the first setting with smooth sample performance, this bound has a better dependence on the dimension, and this is due to the sparsity assumption.

Proof. ***(Theorem 5.2)***
Since f is L-smooth, we have

$$\begin{aligned} f\left(\theta_{k+1}\right) &\leq f\left(\theta_k\right) + \left\langle \nabla f\left(\theta_k\right), \theta_{k+1} - \theta_k \right\rangle + \frac{L}{2}\left\|\theta_{k+1} - \theta_k\right\|^2 \\ &\leq f\left(\theta_k\right) - a\left\langle \nabla f\left(\theta_k\right), \widehat{\nabla} f(\theta_k) \right\rangle + \frac{L}{2} a^2 \left\|\widehat{\nabla} f(\theta_k)\right\|^2. \end{aligned} \tag{5.10}$$

Taking expectations with respect to the sigma field $\mathcal{F}_k$ on both sides of (5.10), and using (5.5) and (5.6) from A5.3, we obtain

$$\begin{aligned} &\mathbb{E}_k\left[f\left(\theta_{k+1}\right)\right] \\ &\leq \mathbb{E}_k\left[f\left(\theta_k\right)\right] - a\left\langle \nabla f\left(\theta_k\right), \nabla f\left(\theta_k\right) + c_1\delta^2 \mathbf{1}_{d\times 1} \right\rangle \\ &\quad + \frac{L}{2} a^2 \left[\left\|\mathbb{E}_k\left[\widehat{\nabla} f(\theta_k)\right]\right\|^2 + \frac{c_2}{\delta^2}\right] \\ &\leq f\left(\theta_k\right) - a\left\|\nabla f\left(\theta_k\right)\right\|^2 + c_1\delta^2 a \mathbb{E}_k \|\nabla f\left(\theta_k\right)\|_1 \\ &\quad + \frac{L}{2} a^2 \left[\left\|\nabla f\left(\theta_k\right)\right\|^2 + 2c_1\delta^2 \mathbb{E}_k \|\nabla f\left(\theta_k\right)\|_1 + dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right] \qquad (5.11) \\ &\leq f\left(\theta_k\right) - \left(a - \frac{L}{2}a^2\right)\left\|\nabla f\left(\theta_k\right)\right\|^2 + c_1\delta^2 B\left(a + La^2\right) \qquad (5.12) \\ &\quad + \frac{L}{2} a^2 \left[dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right], \end{aligned}$$

where we have used the fact that $-\|y\|_1 \leq \sum_{i=1}^{N} y_i$ for any vector N-vector y, in arriving at the inequality (5.11). The last inequality follows from the fact that $\|\nabla f(\theta_k)\|_1 \leq B$ by assumption A5.2. Re-arranging the terms, we obtain

$$a\|\nabla f(\theta_k)\|^2 \leq \frac{2}{(2-La)}\Big[f(\theta_k) - \mathbb{E}_k f(\theta_{k+1}) + c_1\delta^2\left(a+La^2\right)B\Big] + \frac{La^2}{(2-La)}\left[dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right].$$

Now, summing up the inequality above for $k = 1$ to m, and taking expectations, we obtain

$$\begin{aligned}
&\sum_{k=1}^{m} a\mathbb{E}_m\|\nabla f(\theta_k)\|^2 \\
&\leq 2\sum_{k=1}^{m}\frac{(\mathbb{E}_m f(\theta_k) - \mathbb{E}_m f(\theta_{k+1}))}{(2-La)} + 2mc_1\delta^2 B\left(\frac{a+La^2}{2-La}\right) \\
&\quad + Lm\frac{a^2}{(2-La)}\left[dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right] \\
&= 2\left[\frac{f(\theta_1)}{(2-La)} - \frac{\mathbb{E}_m[f(\theta_{m+1})]}{(2-La(m))}\right] \\
&\quad + 2mc_1\delta^2 B\left(\frac{a+La^2}{2-La}\right) + Lm\frac{a^2}{(2-La)}\left[dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right].
\end{aligned}$$

Using $\mathbb{E}_m[f(\theta_k)] \geq f(\theta^*)$, we obtain

$$\begin{aligned}
\sum_{k=1}^{m} a\mathbb{E}_m\|\nabla f(\theta_k)\|^2 &\leq \frac{2(f(\theta_1) - f(\theta^*))}{(2-La)} + 2mc_1\delta^2 B\left(\frac{a+La^2}{2-La}\right) \\
&\quad + Lm\frac{a^2}{(2-La)}\left[dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right].
\end{aligned}$$

Using the fact that θ_R is picked uniformly at random from $\{\theta_1, \ldots, \theta_m\}$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f(\theta_R)\|^2\right] \leq \frac{1}{ma}\Bigg[&\frac{2D_f}{(2-La)} + 2Bmc_1\delta^2\left(\frac{a+La^2}{2-La}\right) \\
&+ Lm\frac{a^2}{(2-La)}\left[dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right]\Bigg]. \qquad (5.13)
\end{aligned}$$

Next, we simplify the bound obtained above by substituting the step-size and perturbation constant values specified in (5.7) as follows:

$$\mathbb{E}\left[\|\nabla f(\theta_R)\|^2\right] \leq \frac{1}{ma}\left[2D_f + 4maBc_1\delta^2 + Lma^2\left[dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right]\right] \tag{5.14}$$

$$\leq \frac{2D_f}{m} max\left\{L, m^{2/3}\right\} + 4B\left(\frac{c_1}{m^{1/3}}\right) + L\left[\frac{dc_1^2}{m^{2/3}} + \frac{c_2}{m^{-1/3}}\right]\frac{1}{m^{2/3}}. \tag{5.15}$$

In the above, the inequality (5.14) follows by using the fact that $a \leq 1/L$, while the inequality (5.15) uses the choice of δ in (5.7). The main claim follows follows by rearranging terms in (5.15). □

5.2 The convex case

We now study the non-asymptotic performance of the RSG algorithm presented earlier, assuming that the objective is convex and smooth. The main result that provides a non-asymptotic bound for RSG algorithm with gradient estimates satisfying A5.3 is given below.

Theorem 5.3.
Suppose the objective function f is L-smooth (see Definition 3.1), and convex. Assume A5.3 holds. Suppose that the RSG algorithm is run for m iterations with stepsize a, perturbation constant δ set as defined in (5.7). Let θ_R be chosen uniformly at random from $\{\theta_1, \ldots, \theta_m\}$. Then, for any $m \geq 1$, we have

$$\mathbb{E}\left[f(\theta_R)\right] - f(\theta^*) \leq \frac{LD^2}{m} + \frac{\mathcal{K}_1}{m^{1/3}},$$

where $\mathcal{K}_1 = D^2 + 4\sqrt{d}Dc_1\delta^2 + \dfrac{dc_1^2\delta^4}{m} + c_2$, constants c_1 and c_2 are specified in A5.3, and

$$D = \|\theta_1 - \theta^*\|, \tag{5.16}$$

with θ^* denoting a global optima of f.

The case of unbiased gradient information leads to a $O(1/\sqrt{m})$ bound and the proof is a complete parallel argument to the one employed for the biased case in the result above, and we omit the details.

Remark 5.3. From the result above, it is apparent that an $\mathcal{O}\left(\frac{1}{\epsilon^3}\right)$ number of iterations is necessary to find a point that satisfies $\mathbb{E}\left[f\left(\theta_R\right)\right] - f(\theta^*) \leq \epsilon$. Moreover, this rate is not improvable in a minimax sense for a gradient-based algorithm with inputs from a biased gradient oracle, which we formalize in the next section.

Remark 5.4. For the special case of noise originating from a common random number sequence that was discussed earlier in Section 3.3.4, it is possible to obtain an improved bound of the order $\mathcal{O}\left(\sqrt{\frac{d}{m}}\right)$. This improvement comes from the fact that the gradient estimate variance does not blow up as the perturbation constant δ goes to zero, see Proposition 3.5. The proof of this improved bound follows arguments similar to those employed in the proof of Theorem 5.3. We omit the details.

Remark 5.5. The bound in Theorem 5.3 above is for a random iterate θ_R. Using a different step size choice that decays in a geometric fashion, and a radically different proof technique, it is possible to infer a bound of the same order, i.e., $O\left(m^{-1/3}\right)$ for the last iterate θ_m. The reader is referred to Section IV-B of (Bhavsar and Prashanth, 2022) for the details. Note that the last iterate is preferred over a random iterate in practice, and hence, it is desirable to obtain bounds for the last iterate. For the non-convex case, to the best of our knowledge, there are no bounds available for a stochastic gradient algorithm with inputs from a biased gradient oracle.

Proof. ***(Theorem 5.3)***
Let $\Delta_k = \widehat{\nabla} f(\theta_k) - \nabla f(\theta_k)$ and $\omega_k = \|\theta_k - \theta^*\|, \forall k \geq 1$. Then for any $k = 1, \ldots, m$, we have

$$\omega_{k+1}^2 = \|\theta_{k+1} - \theta^*\|^2$$

$$= \|\theta_k - a\widehat{\nabla} f(\theta_k) - \theta^*\|^2$$
$$= \omega_k^2 - 2a\left\langle \widehat{\nabla} f(\theta_k), \theta_k - \theta^* \right\rangle + a^2 \left\| \widehat{\nabla} f(\theta_k) \right\|^2. \tag{5.17}$$

Taking expectations with respect to the sigma field $\mathcal{F}_k$ on both sides of (5.17), and using (5.5), (5.6), we obtain

$$\begin{aligned}\mathbb{E}[\omega_{k+1}^2] &\leq \mathbb{E}[\omega_k^2] - 2a\left\langle \nabla f(\theta_k), \theta_k - \theta^* \right\rangle - 2a\mathbb{E}\left[\langle \Delta_k, \theta_k - \theta^* \rangle\right] \\ &\quad + a^2\left[\left\| \mathbb{E}_k\left[\widehat{\nabla} f(\theta_k)\right] \right\|^2 + \frac{c_2}{\delta^2} \right] \\ &\leq \mathbb{E}[\omega_k^2] - 2a\left\langle \nabla f\left(\theta_k\right), \theta_k - \theta^* \right\rangle + 2ac_1\delta^2\|\theta_k - \theta^*\|_1 \\ &\quad + a^2\left[\|\nabla f\left(\theta_k\right)\|^2 + 2\sqrt{d}c_1\delta^2\|\nabla f\left(\theta_k\right)\| + dc_1^2\delta^4 + \frac{c_2}{\delta^2} \right],\end{aligned} \tag{5.18}$$

where the last inequality follows from the fact that $-\sum_{i=1}^{N} \theta_i \leq \|X\|_1$ for any vector X. Now, using the fact that f is convex, we have

$$\|\nabla f\left(\theta_k\right)\|^2 \leq L\left\langle \nabla f\left(\theta_k\right), \theta_k - \theta^* \right\rangle.$$

Further, since f is L-smooth, $\|\nabla f(\theta_k)\| \leq L\|\theta_k - \theta^*\|$. Plugging these inequalities in (5.18), we obtain

$$\begin{aligned}\mathbb{E}[\omega_{k+1}^2] &\leq \mathbb{E}[\omega_k^2] - 2a\left\langle \nabla f\left(\theta_k\right), \theta_k - \theta^* \right\rangle + 2ac_1\delta^2\|\theta_k - \theta^*\|_1 \\ &\quad + a^2\Bigg[L\left\langle \nabla f\left(\theta_k\right), \theta_k - \theta^* \right\rangle + 2\sqrt{d}c_1\delta^2 L\|\theta_k - \theta^*\| \\ &\qquad + dc_1^2\delta^4 + \frac{c_2}{\delta^2} \Bigg] \\ &\leq \mathbb{E}[\omega_k^2] - \left(2a_k - La^2\right)\left[f\left(\theta_k\right) - f(\theta^*)\right] \\ &\quad + 2\sqrt{d}\omega_k c_1\delta^2 a + La^2) + a^2\left[dc_1^2\delta^4 + \frac{c_2}{\delta^2} \right],\end{aligned}$$

where the last inequality follows from the fact that $f(\cdot)$ is convex along with $\|X\|_1 \leq \sqrt{d}\|X\|$ for any vector X. Re-arranging the terms, we obtain

$$a\left[f\left(\theta_k\right) - f(\theta^*)\right]$$

$$\leq \frac{1}{(2-La)}\left[\omega_k^2 - \mathbb{E}[\omega_{k+1}^2] + 2\sqrt{d}\omega c_1\delta^2(a+La^2) + a^2\left(dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right)\right].$$

Now summing up the inequality above from $k=1$ to m and taking expectations, we obtain

$$\begin{aligned}
\sum_{k=1}^{m} a\mathbb{E}_m\left[f\left(\theta_k\right) - f(\theta^*)\right] &\leq \sum_{k=1}^{m} \frac{\mathbb{E}_m[\omega_k^2] - \mathbb{E}_m[\omega_{k+1}^2]}{(2-La)} \\
&\quad + 2\sqrt{d}\sum_{k=1}^{m}\mathbb{E}_m\left[\omega_k\right]c_1\delta^2\frac{a+La^2)}{(2-La)} \\
&\quad + \sum_{k=1}^{m}\frac{a^2}{(2-La)}\left(dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right) \\
&= \frac{\omega_1^2}{(2-La)} - \frac{\mathbb{E}_m\left[\omega_{m+1}^2\right]}{(2-La)} \\
&\quad + 2\sqrt{d}\sum_{k=1}^{m}\mathbb{E}_m\left[\omega_k\right]c_1\delta^2\frac{(a+La^2)}{(2-La)} \\
&\quad + \sum_{k=1}^{m}\frac{a^2}{(2-La)}\left(dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right) \\
&\leq \frac{D^2}{(2-La)} + 2\sqrt{d}D\sum_{k=1}^{m}c_1\delta^2\frac{(a+La^2)}{(2-La)} \\
&\quad + \sum_{k=1}^{m}\frac{a^2}{(2-La)}\left(dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right)
\end{aligned}$$

where the last inequality follows by using (5.16), i.e., $\mathbb{E}_m\left[\omega_k\right] \leq D$. Combining the above result with the fact that θ_R is picked uniformly at random from $\{\theta_1, \ldots, \theta_m\}$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[f\left(\theta_R\right)\right] - f(\theta^*) \\
&\leq \frac{1}{ma}\left[\frac{D^2}{(2-La)} + 2\sqrt{d}D\sum_{k=1}^{m}c_1\delta^2\frac{(a+La^2)}{(2-La)}\right. \\
&\qquad \left. + \sum_{k=1}^{m}\frac{a^2}{(2-La)}\left(dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right)\right], \qquad (5.19)
\end{aligned}$$

Using (5.7) in (5.19), we obtain

$$\mathbb{E}\left[f\left(\theta_R\right)\right] - f(\theta^*)$$

$$\leq \frac{1}{ma}\left[\frac{D^2}{(2-La)} + 2\sqrt{d}D\sum_{k=1}^{m} c_1\delta^2\frac{a+La^2)}{(2-La)} + \sum_{k=1}^{m}\frac{a^2}{(2-La)}\left(dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right)\right]$$

$$\leq \frac{1}{ma}\left[D^2 + 4\sqrt{d}Dmac_1\delta^2 + ma^2\left(dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right)\right], \tag{5.20}$$

where the final inequality follows by using the fact that $a \leq 1/L$. The main claim follows by using the definition of a, δ given in (5.7) followed by simple algebraic manipulations. □

5.3 The strongly-convex case

In this section, we present non-asymptotic analysis for the SG algorithm (5.1) under a strongly convex objective, which is made precise in the definition below.

Definition 5.3. A continuously differentiable function f is μ-strongly convex if the following condition holds for any θ, θ':

$$f(\theta') \geq f(\theta) + \nabla f(\theta)^T(\theta' - \theta) + \frac{\mu}{2}\left\|\theta' - \theta\right\|^2.$$

For a brief introduction to strong-convexity, the reader is referred to Appendix D. As in the previous sections, we consider unbiased as well as biased gradient information. We begin with the unbiased gradient case in the next section.

5.3.1 SG with unbiased gradient information

We consider the following update iteration:

$$\theta_{k+1} = \theta_k - a(k)\widehat{\nabla} f(\theta_k). \tag{5.21}$$

We first state and prove a result for the case of a constant step size.

Theorem 5.4. Let f be a μ-strongly convex function. Assume A5.1. Then, the SG algorithm governed by (5.21) and with $a(k) = a$ s.t.

$0 < aL < 1$, satisfies

$$\mathbb{E}[f(\theta_m) - f(\theta^*)] \leq \frac{aL\sigma^2}{2\mu} + (1 - a\mu)^{m-1}\left(f(\theta_1) - f(\theta^*) - \frac{aL\sigma^2}{2\mu}\right). \tag{5.22}$$

Proof. From the initial passage in the proof of Theorem 5.1, we have

$$\mathbb{E}_k[f(\theta_{k+1})] - f(\theta_k) \leq -a(k)(1 - \frac{1}{2}a(k)L)\|\nabla f(\theta_k)\|_2^2 + \frac{1}{2}a(k)^2 L\sigma^2.$$

Since $a(k) = a$ and $0 < aL < 1$, we have

$$\mathbb{E}_k[f(\theta_{k+1})] - f(\theta_k) \leq -\frac{1}{2}a\|\nabla f(\theta_k)\|_2^2 + \frac{1}{2}a^2 L\sigma^2. \tag{5.23}$$

Since f is μ-strongly convex, the following inequality, which is well-known as the **Polyak-Lojasiewicz (PL) condition** holds[2]:

$$f(\theta) - f(\theta^*) \leq \frac{1}{2\mu}\|\nabla f(\theta)\|_2^2, \ \forall \theta.$$

Using the above inequality in (5.23), we obtain

$$\mathbb{E}_k[f(\theta_{k+1})] - f(\theta_k) \leq -\mu a\left(f(\theta_k) - f(\theta^*)\right) + \frac{1}{2}a^2 L\sigma^2. \tag{5.24}$$

Subtracting $f(\theta^*)$ on both sides and re-arranging, we obtain

$$\mathbb{E}_k[f(\theta_{k+1}) - f(\theta^*)] \leq (1 - a\mu)[f(\theta_k) - f(\theta^*)] + \frac{1}{2}a^2 L\sigma^2. \tag{5.25}$$

[2]This inequality can be inferred as follows: Using strong convexity,

$$f(y) - f(x) \geq \nabla f(x)^{\mathsf{T}}(y - x) + \frac{\mu}{2}\|y - x\|^2.$$

Taking minimum over y on both sides, we have

$$\min_y \left(f(y) - f(x)\right) \geq \min_y \left(\nabla f(x)^{\mathsf{T}}(y - x) + \frac{\mu}{2}\|y - x\|^2\right).$$

The minimum on the RHS above is obtained for $y^* = -\frac{1}{\mu}\nabla f(x) + x$. Substituting this value on the RHS, we obtain

$$f(x^*) - f(x) \geq -\frac{1}{\mu}\|\nabla f(x)\|^2 + \frac{1}{2\mu}\|\nabla f(x)\|^2.$$

Re-arranging leads to PL-condition.

Taking expectations followed by straightforward simplifications, we obtain

$$\begin{aligned}
&\mathbb{E}[f(\theta_{k+1}) - f(\theta^*)] - \frac{aL\sigma^2}{2\mu} \\
&\leq (1 - a\mu)\mathbb{E}[f(\theta_k) - f(\theta^*)] + \frac{a^2 L\sigma^2}{2} - \frac{aL\sigma^2}{2\mu} \\
&= (1 - a\mu)\left(\mathbb{E}[f(\theta_k) - f(\theta^*)] - \frac{aL\sigma^2}{2\mu}\right). \qquad (5.26)
\end{aligned}$$

Using $a < 1/L$ by assumption, and $\mu \leq L$, we have[3]

$$a\mu < \frac{\mu}{L} \leq 1.$$

A repeated application of the above inequality leads to the following bound:

$$\mathbb{E}[f(\theta_m) - f(\theta^*)] \leq \frac{aL\sigma^2}{2\mu} + (1 - a\mu)^{m-1}\left(f(\theta_1) - f(\theta^*) - \frac{aL\sigma^2}{2\mu}\right). \qquad (5.27)$$

The claim follows. □

Remark 5.6. Using the bound on the optimization error (or the difference in function values) in the result above, we can establish a bound on the parameter error as follows: From μ-strong convexity of f, we have

$$f(\theta) + \nabla f(\theta)^\intercal(\tilde{\theta} - \theta) + \frac{\mu}{2}\left\|\tilde{\theta} - \theta\right\|^2 \leq f(\tilde{\theta}).$$

At $\theta = \theta^*$, $\nabla f(\theta^*) = 0$, implying

$$\left\|\tilde{\theta} - \theta\right\|^2 \leq \frac{2}{\mu}\left(f(\tilde{\theta}) - f(\theta^*)\right).$$

Thus, a bound on the difference in function values implies a bound on the parameter error.

[3]The second inequality can be inferred as follows: Using μ-strong convexity and L-smoothness of f, for any $\theta, \tilde{\theta}$, we have

$$f(\theta) + \nabla f(\theta)^\intercal(\tilde{\theta} - \theta) + \frac{\mu}{2}\left\|\tilde{\theta} - \theta\right\|^2 \leq f(\tilde{\theta}) \leq f(\theta) + \nabla f(\theta)^\intercal(\tilde{\theta} - \theta) + \frac{L}{2}\left\|\tilde{\theta} - \theta\right\|^2.$$

Thus, $\mu \leq L$.

Remark 5.7. Taking limits as $n \to \infty$, the bound in (5.22) converges to $\frac{aL\sigma^2}{2\mu}$. This observation implies that a constant step size stochastic gradient algorithm does not converge to the optima, and instead gets to within a ball around the optima.

Next, we consider the case of a diminishing step size.

Theorem 5.5. Let f be a μ-strongly convex function. Assume A5.1. Then, the SG algorithm governed by (5.21) and with $a(k) = \frac{c}{k+1}$ s.t. $\frac{1}{\mu} < c \leq L$, satisfies

$$\mathbb{E}\left[f(\theta_m) - f(\theta^*)\right] \leq \frac{1}{m+1} \max\left\{\frac{c^2 L\sigma^2}{2(c\mu - 1)}, 2\left(f(\theta_1) - f(\theta^*)\right)\right\}. \tag{5.28}$$

Proof. We prove by induction. The base case holds trivially. Assuming the claim holds for m, we show that it holds for $m+1$.

From (5.4), we have

$$\begin{aligned}
\mathbb{E}_k[f(\theta_{k+1})] - f(\theta_k)] &\leq -a(k)(1 - \frac{1}{2}a(k)L)\|\nabla f(\theta_k)\|_2^2 + \frac{1}{2}a(k)^2 L\sigma^2 \\
&\leq -a(k)\|\nabla f(\theta_k)\|_2^2 + \frac{1}{2}a(k)^2 L\sigma^2 \\
&\qquad \text{(Since } a(k)L \leq 1\text{)} \\
&\leq -a(k)\mu\left(f(\theta_k) - f(\theta^*)\right) + \frac{1}{2}a(k)^2 L\sigma^2 \\
&\qquad \text{(PL-condition)}
\end{aligned}$$

Thus,

$$\mathbb{E}[f(\theta_{k+1})] - f(\theta^*)] \leq (1 - a(k)\mu)\mathbb{E}[f(\theta_k) - f(\theta^*)] + \frac{1}{2}a(k)^2 L\sigma^2.$$

Using the induction hypothesis, the form of the step size $a(k)$ and letting $K = \max\left\{\frac{c^2 L\sigma^2}{2(c\mu - 1)}, 2\left(f(\theta_1) - f(\theta^*)\right)\right\}$, we obtain

$$\mathbb{E}[f(\theta_{m+1})] - f(\theta^*)] \leq \left(1 - \frac{c\mu}{m+1}\right)\frac{K}{m+1} + \frac{c^2 L\sigma^2}{2(m+1)^2}$$

$$= \frac{Km}{(m+1)^2} - \frac{(c\mu - 1)K}{(m+1)^2} + \frac{c^2 L\sigma^2}{2(m+1)^2}$$
$$\leq \frac{K}{m+2},$$

where the final inequality used the following fact:

$$-\frac{(c\mu - 1)K}{(m+1)^2} + \frac{c^2 L\sigma^2}{2(m+1)^2} \leq 0.$$

The inequality above holds by the definition of K, and simple algebra to infer $\frac{Km}{(m+1)^2} \leq \frac{K}{m+2}$.

The claim follows. □

Remark 5.8. In contrast to the constant step size case handled previously, with a diminishing step size, we have a bound that vanishes as $m \to \infty$. However, the step size choice requires the knowledge of the strong convexity parameter μ, while the constant step size case in Theorem 5.4 did not assume such information. On a related note, it is possible to obtain a bound of $O\left(1/\sqrt{m}\right)$ with a step size choice that does not require the knowledge of μ, and more importantly, with a bound that does not scale inversely with μ. Such a bound may be preferable for ill-conditioned problems, where μ is very small. The reader is referred to (Nemirovski *et al.*, 2009) for the details.

5.3.2 SG with biased gradient information

As before, we consider the update iteration in (5.21). Unlike the previous section, where we assumed unbiased gradient estimates (i.e., the condition A5.1 holds), here the estimate $\widehat{\nabla} f(\theta_k)$ is a biased approximation to the gradient of the objective function f at θ_k.

As in the asymptotic analysis in Section 4.1.3, the biased gradient estimate $\widehat{\nabla} f(\theta_k)$ can be decomposed as follows:

$$\widehat{\nabla} f(\theta_k) = \nabla f(\theta_k) + \beta_k + \eta_k, \text{ where} \tag{5.29}$$
$$\beta_k = E\left[\widehat{\nabla} f(\theta_k) \mid \mathcal{F}_k\right] - \nabla f(\theta_k),$$
$$\eta_k = \widehat{\nabla} f(\theta_k) - E\left[\widehat{\nabla} f(\theta_k) \mid \mathcal{F}_k\right],$$

where $\mathcal{F}_k$ is a σ-field generated by $\{\theta_i, i \leq k\}$. In the above, β_k is the bias in the gradient estimate and η_k, $n \geq 0$, is a martingale difference sequence.

Using a simultaneous perturbation-based gradient estimate implies $\beta_k = O(\delta_k^2)$, where δ_k is the perturbation parameter used in forming the estimate (see Chapter 3 for several examples). While the bias goes down as δ_k^2, the variance of the gradient estimate scales inversely with δ_k^2. This has been formalized earlier in assumptions A5.2–A5.3.

We now present a non-asymptotic bound in expectation for the SG algorithm (5.21) with inputs from a biased gradient oracle that satisfies the aforementioned assumptions.

Proposition 5.1. Suppose the objective function f is L-smooth (see Definition 3.1), and assumptions A5.2–A5.3 hold. Then, we have

$$\begin{aligned}\mathbb{E}\left\|\theta_{m+1}-\theta^*\right\|^2 \leq & \underbrace{2\exp(-2\mu\Gamma(m))\left\|\theta_0-\theta^*\right\|^2}_{\textbf{initial error}} \\ & \underbrace{+2\sum_{k=1}^{n} a_k^2\exp(-2\mu(\Gamma(m)-\Gamma_k))c_1^2\delta_k^4+}_{\textbf{bias error}} \\ & \underbrace{\sum_{k=1}^{n} a_k^2\exp(-2\mu(\Gamma(m)-\Gamma_k))c_2\delta_k^{-2}}_{\textbf{sampling error}}, \end{aligned} \tag{5.30}$$

where $\Gamma(k) := \sum_{i=1}^{k} a_i$.

Proof. Let $z_m = \theta_m - \theta^*$ denote the error at time instant n of the algorithm (5.21). Using $\nabla f(\theta^*) = 0$, we have

$$\left(\int_0^1 \nabla^2 f(\theta^* + \lambda(\theta_m - \theta^*))d\lambda\right) z_m = \nabla f(\theta_m).$$

Using the fact above, we arrive at a recursion for z_m from (5.29). Letting $J_m := \int_0^1 \nabla^2 f(\theta^* + \lambda(\theta_m - \theta^*)d\lambda$, we have

$$z_{m+1} = (I - a(m)J_m)z_m - a(m)\left(\beta_m + \eta_m\right)$$

$$=\Pi_m z_0 - \sum_{k=1}^{n} a(k)\Pi_m \Pi_k^{-1}(\beta_k + \eta_k),$$

where $\Pi_m := \prod_{k=1}^{n} (I - a(k)J_k)$.

By the conditional Jensen's inequality, we obtain

$$\begin{aligned}
&(\mathbb{E}_m \|z_{m+1}\|)^2 \leq \mathbb{E}_m(\langle z_m, z_m\rangle) \\
&= \mathbb{E}_m \left(\|\Pi_m z_0\|^2 + \left\|\sum_{k=1}^{n} a(k)\Pi_m\Pi_k^{-1}\beta_k\right\|^2 + \left\|\sum_{k=1}^{n} a(k)\Pi_m\Pi_k^{-1}\eta_k\right\|^2 \right. \\
&\quad - \left\langle \Pi_m z_0, \sum_{k=1}^{n} a(k)\Pi_m\Pi_k^{-1}\beta_k \right\rangle - \left\langle \Pi_m z_0, \sum_{k=1}^{n} a(k)\Pi_m\Pi_k^{-1}\eta_k \right\rangle \\
&\quad \left. - \left\langle \sum_{k=1}^{n} a(k)\Pi_m\Pi_k^{-1}\beta_k, \sum_{k=1}^{n} a(k)\Pi_m\Pi_k^{-1}\eta_k \right\rangle \right) \qquad (5.31) \\
&\leq 2\|\Pi_m z_0\|^2 + 2\sum_{k=1}^{n} a(k)^2 \left\|\Pi_m\Pi_k^{-1}\right\|^2 c_1^2\delta_k^4 \\
&\quad + \sum_{k=1}^{n} a(k)^2 \left\|\Pi_m\Pi_k^{-1}\right\|^2 \mathbb{E}\|\eta_k\|^2 . \qquad (5.32)
\end{aligned}$$

For the last inequality, we have used the following facts: (i) η_k is a martingale difference implying the last two cross terms in (5.31) are zero; (ii) $\beta_k \leq c_1\delta_k^2$ from A5.3; and (iii) Cauchy-Schwarz inequality for the first cross term in (5.31).

Now, we bound each of the square terms in (5.32) separately. Since the objective is strongly convex, we have that $\|I - a(m)J_m\| \leq \exp(-\mu a(m))$. Hence,

$$\begin{aligned}
\left\|\Pi_m\Pi_k^{-1}\right\|_2 &= \left\|\prod_{j=k+1}^{n} (I - a_j J_j)\right\|_2 \\
&\leq \prod_{j=k+1}^{n} \|(1 - a_j\mu)I - a_j(J_j - \mu I)\|_2 \\
&\leq \prod_{j=k+1}^{n} \|(1 - a_j\mu)I\|_2 \leq \prod_{j=k+1}^{n} (1 - a_j\mu) \\
&\leq \exp\left(-\mu(\Gamma(m) - \Gamma(k))\right). \qquad (5.33)
\end{aligned}$$

From A5.3, we can infer that the second moment of the martingale difference is bounded above by c_2/δ_k^2. The main claim now follows by plugging the bound on η_m and (5.33) into (5.32). □

By specializing the result in the proposition above, we derive a non-asymptotic bound of the order $O(1/\sqrt{m})$.

Theorem 5.6. (**Biased gradients and strongly convex objective**) Let $a(k) = c/k$ and $\delta_k = \delta_0/k^\delta$. Then,

$$\mathbb{E}\,\|\theta_m - \theta^*\| \leq \frac{\sqrt{2}\,\|\theta_0 - \theta^*\|}{m^{\mu c}} + \frac{\sqrt{2}cc_1\delta_0^2}{\sqrt{2\mu c - 4\delta - 1}} m^{-\frac{1}{2}-2\delta} + \frac{\sqrt{c_2}c}{\delta_0\sqrt{2\mu c + 2\delta - 1}} m^{\delta-\frac{1}{2}}.$$

Remark 5.9. Choosing $\delta = 0$, one can obtain a bound of the order $O\left(m^{-1/2}\right)$ for simultaneous perturbation schemes that lead to biased gradient estimates, and this bound matches the corresponding bound with unbiased gradient information up to constant factors. Contrast this with the difference in rates between biased and unbiased gradient information for the non-convex and convex cases in the previous sections.

Remark 5.10. Using L-smoothness of f and $\nabla f(\theta^*) = 0$, we have

$$\mathbb{E}\left[f\left(\theta_m\right)\right] - f(\theta^*) \leq \frac{L}{2}\mathbb{E}\,\|\theta_m - \theta^*\|^2 = O\left(\frac{1}{m}\right).$$

Proof. Bounding a sum by an integral, we obtain

$$\exp(-\mu\Gamma(m)) \leq \exp(-\mu c \ln m) = m^{-\mu c}.$$

Plugging $a(k) = c/k$ and $\delta_k = \delta_0/k^\delta$ into the bias error term in (5.30), we obtain

$$\begin{aligned}\sum_{k=1}^{m} a(k)^2 \exp(-2\mu(\Gamma(m) - \Gamma_k))c_1^2\delta_k^4 &\leq \sum_{k=1}^{m} \frac{c^2}{k^2} n^{-2\mu c} k^{2\mu c} c_1^2 \frac{\delta_0^4}{m^{4\delta}} \\ &\leq c^2 n^{-2\mu c} c_1^2 \delta_0^4 \sum_{k=1}^{m} k^{2\mu c - 4\delta - 2}\end{aligned}$$

$$\leq \frac{c^2 c_1^2 \delta_0^4}{(2\mu c - 4\delta - 1)} m^{-1-4\delta}.$$

Along similar lines, the sampling error term in (5.30) can be upper-bounded as follows:

$$\sum_{k=1}^{m} a(k)^2 \exp(-2\mu(\Gamma(m) - \Gamma_k)) \frac{c_2}{\delta_k^2} \leq \frac{c^2 c_2}{\delta_0^2 (2\mu c - 4\delta - 1)} m^{-1+2\delta}.$$

□

5.4 Bounds with improved dimension dependence

For a non-convex objective, under a smoothness assumption on the objective function, we presented $O(1/m^{1/3})$ bounds on the gradient norm square, see Theorem 5.2. Here m denotes the number of iterations of the RSG algorithm, and the gradient estimates had the usual bias-variance tradeoff (see Assumption A5.3). However, this bound has two shortcomings. First, the dependence on m is weaker as compared to the case where unbiased gradient information is available. We show in Section 5.6 that the $1/m^{1/3}$ dependence on m is unimprovable in the minimax sense, when the underlying gradient estimates exhibit a bias-variance tradeoff. Second, the dimension dependence in the $O(1/m^{1/3})$ bound mentioned above is not encouraging, since this dependence is not sub-linear in d.

In this section, we present two settings, where we exhibit sub-linear dependence on d and a $\frac{1}{\sqrt{m}}$ dependence on the number of iterations m, under additional assumptions on the system model. In the first setting we assume that the noisy observations are smooth, while the second setting considers the sparse objective gradient case.

5.4.1 Smooth sample performance

We obtain $O(1/\sqrt{m})$ bounds for an objective function of the form $f(\theta) = \mathbb{E}_\xi [F(\theta, \xi)]$ if the sample performance F is L-smooth, i.e., satisfying the following assumption:

A5.4. The sample performance F is such that (i) $\nabla f(\theta) = \mathbb{E}_\xi [\nabla F(\theta, \xi)]$; (ii) The gradient of F is Lipschitz continuous almost surely, for any ξ, i.e.,

$$\left\|\nabla F(\theta, \xi) - \nabla F(\tilde{\theta}, \xi)\right\| \leq L \left\|\theta - \tilde{\theta}\right\|, \forall \theta, \tilde{\theta} \in \mathbb{R}^d,$$

for some $L > 0$; and (iii) There exists a $\sigma > 0$ such that the following inequality holds for any θ:

$$\mathbb{E}\left[\|\nabla F(\theta, \xi) - \nabla f(\theta)\|^2\right] \leq \sigma^2.$$

Assumption A5.4 implies f is L-smooth. This can be seen as follows: For any $\theta, \tilde{\theta} \in \mathbb{R}^d$,

$$\begin{aligned}\left\|\nabla f(\theta) - \nabla f(\tilde{\theta})\right\| &\leq \left\|\nabla[\mathbb{E}_\xi \left(F(\theta, \xi) - F(\tilde{\theta}, \xi)\right)]\right\| \\ &\leq \mathbb{E}_\xi \left\|\nabla F(\theta, \xi) - \nabla F(\tilde{\theta}, \xi)\right\| \\ &\leq L \left\|\theta - \tilde{\theta}\right\|.\end{aligned}$$

Using such a smooth sample F, it is possible to construct a gradient estimate that satisfies the following conditions:

$$\left\|\mathbb{E}\left[\widehat{\nabla} f(\theta)\right] - \nabla f(\theta)\right\| \leq c_1 \delta, \tag{5.34}$$

$$\mathbb{E}\left[\left\|\widehat{\nabla} f(\theta) - \mathbb{E}\left[\widehat{\nabla} f(\theta)\right]\right\|^2\right] \leq c_2 \delta^2 + \widetilde{c}_2. \tag{5.35}$$

One way to construct a gradient estimate satisfying the conditions defined above is to employ the Gaussian smoothing approach, discussed earlier in Section 3.3.3. For ease of readability, we recall this estimator below.

$$\widehat{\nabla} f(\theta) = \Delta \left[\frac{F(\theta + \delta\Delta, \xi) - F(\theta, \xi)}{\delta}\right], \tag{5.36}$$

where Δ is a d-dimensional standard Gaussian vector. An important observation regarding the estimator above is that the noise ξ is common to the two function measurements. In practical settings, where the noise is generated using common random numbers, it is possible to keep the noise factor ξ same across function measurements — a setting discussed earlier in Section 3.3.4.

In Proposition 3.4, we established a $c_1\delta$ bias bound for the estimator in (5.36) without assuming that F is L-smooth and instead working

with only smoothness of the objective f. This bound is good enough to obtain the bias guarantee in (5.34). On the other hand, the variance bound in Proposition 3.4 is c_2/δ^2, which precludes choosing a very small δ in the gradient estimate. However, using a different proof technique, it is possible to obtain the variance bound $c_2\delta^2 + \widetilde{c_2}$ in (5.35). This proof would exploit the fact that F is L-smooth (see Assumption A5.4) in conjunction with the common noise in the gradient estimate. We present such a result below. On a related note, we established bounds similar to those in (5.34)–(5.35) for the case where F is L-smooth and in addition, f is convex, see Proposition 3.5.

Proposition 5.2. Assume A3.2, A3.3, A5.2, and A5.4. Then, the gradient estimate defined in (5.36), with Δ distributed as a multivariate standard Gaussian, satisfies the following bounds for any given θ:

$$\left\|\mathbb{E}\left[\widehat{\nabla} f(\theta)\right] - \nabla f(\theta)\right\| \leq C_1\delta, \text{ and} \tag{5.37}$$

$$\mathbb{E}\left[\left\|\widehat{\nabla} f(\theta) - \mathbb{E}\left[\widehat{\nabla} f(\theta)\right]\right\|^2\right] \leq \tilde{C}_2 + C_2\delta^2, \tag{5.38}$$

where C_1 is defined in Proposition 3.4, $C_2 = \dfrac{L^2(d+6)^3}{2}$, and $\tilde{C}_2 = 2(d+4)(B^2+\sigma^2)$, with σ^2 denoting a bound on the variance of $F(\theta,\xi)$ (see Assumption A5.4).

Proof. The first claim can be inferred from Proposition 3.4. We prove the second claim here.

For a L-smooth function h, define

$$g(x;\delta) = \frac{h(\theta+\delta\Delta) - h(\theta)}{\delta}\Delta.$$

where δ is the perturbation/smoothing constant and $\Delta \sim \mathcal{N}(0, I_d)$. Notice that

$$\mathbb{E}_\Delta\left[\|g(\theta,\delta)\|^2\right] = \frac{1}{\delta^2}\mathbb{E}_\Delta\left[(h(\theta+\delta\Delta) - h(\theta))^2\|\Delta\|^2\right].$$

Now,

$$\begin{aligned}(h(\theta+\delta\Delta) - h(\theta))^2 &\leq \|h(\theta+\delta\Delta) - h(\theta) - \delta\nabla h(\theta)^T\Delta\|^2 + \delta^2(\nabla h(\theta)^\intercal\Delta)^2 \\ &\leq 2\left(\frac{L\delta^2}{2}\|\Delta\|^2\right) + 2\delta^2(\nabla h(\theta)^\intercal\Delta)^2,\end{aligned}$$

where we used the fact that $|h(y)-h(x)-\nabla h(x)^{\mathsf{T}}(y-x)| \leq \frac{L}{2}\|y-x\|^2$. Thus,

$$\begin{aligned}\mathbb{E}_{\Delta}(||g(\theta,\delta)||^2) &\leq \frac{1}{\delta^2}\left[\frac{2\delta^4}{4}L^2\mathbb{E}_{\Delta}[||\Delta||^6] + 2\delta^2\mathbb{E}_{\Delta}((\nabla h(\theta)^{\mathsf{T}}\Delta)^2||\Delta||^2)\right] \\ &\leq \frac{\delta^2}{2}L^2(d+6)^3 + 2(d+4)||\nabla h(\theta)||^2, \end{aligned} \tag{5.39}$$

where we used the fact that $\mathbb{E}_{\Delta}[||\Delta||^k] \leq (d+k)^{k/2}$ for a standard Gaussian vector Δ, and $\mathbb{E}_{\Delta}\left((\nabla h(\theta)^{\mathsf{T}}\Delta)^2||\Delta||^2\right) \leq (d+4)||\nabla h(\theta)||^2$ (cf. Theorem 3 of (Nesterov and Spokoiny, 2017) for a proof).

Applying (5.39) for $h = F$, after noting that F is a L-smooth function (see the lemma above), we bound the second moment of the gradient estimate in (5.36) as follows:

$$\begin{aligned}\mathbb{E}(||\widehat{\nabla} f(\theta)||^2) &= \mathbb{E}\left(\left\|\Delta\left[\frac{F(\theta+\delta\Delta,\xi)-F(\theta,\xi)}{\delta}\right]\right\|^2\right) \\ &\leq \frac{\delta^2}{2}L^2(d+6)^3 + 2(d+4)\left[\frac{1}{2}\mathbb{E}||\nabla F(\theta,\xi)||^2 + \frac{\delta^2}{4}\right] \\ &\leq \frac{\delta^2}{2}L^2(d+6)^3 + 2(d+4)\left(||\nabla f(\theta)||^2 + \sigma^2\right), \end{aligned} \tag{5.40}$$

where the final inequality used the variance bound from Assumption A5.4. The bound in (5.38) follows by using (5.40) in conjunction with $\mathbb{E}\left\|\widehat{\nabla} f(\theta) - \mathbb{E}\left[\widehat{\nabla} f(\theta)\right]\right\|^2 \leq \mathbb{E}\left\|\widehat{\nabla} f(\theta)\right\|^2$. □

Using the gradient estimate in (5.36) in a stochastic gradient algorithm along the lines discussed in Section 5.1, it is possible to obtain an order $\mathcal{O}\left(\sqrt{\frac{d}{m}}\right)$ bound. This is an improvement over the $\mathcal{O}(m^{-1/3})$ derived for a smooth f in Section 5.1, see Theorem 5.2. The improvement is w.r.t. the number of iterations m as well as dimension d. The result below makes this claim precise.

Theorem 5.7.
Assume the conditions of Proposition 5.2 hold. Suppose the RSG algorithm is run with $a(k) = a$ and perturbation constant $\delta(k) = \delta$ for each $k = 1, \ldots, m$, where

$$a = \min\left\{\frac{1}{L}, \frac{1}{\sqrt{dm}}\right\}, \quad \delta = \frac{1}{d\sqrt{m}}. \tag{5.41}$$

Then, choosing θ_R uniformly at random from $\{\theta_1, \ldots, \theta_m\}$, we have

$$\mathbb{E}\left\|\nabla f\left(\theta_R\right)\right\|^2 \leq \frac{2LD_f}{m} + \frac{\mathcal{Z}_5}{\sqrt{m}},$$

where $\mathcal{Z}_5 = 2\sqrt{d}D_f + 4B\mathcal{K}_3 + L\left(\frac{\sqrt{d}\mathcal{K}_3^2}{m} + \frac{C_2}{md^{5/2}} + \frac{\tilde{C}_2}{\sqrt{d}}\right)$, $\mathcal{K}_3 = C_1 d^{-1}$, constants C_1, C_2, $\tilde{C}_2$ are as defined in Proposition 5.2, B is as defined in A5.2, and D_f is as defined in (5.3).

The bound above is $O\left(\sqrt{\frac{d}{m}}\right)$ if $m > d$. As mentioned before, a result in similar spirit can be claimed for the convex case, by using Proposition 3.5 in place of Proposition 5.2 and we omit the details as the proof is a completely parallel argument to Theorem 5.3.

Proof. Following the proof in a similar manner as that of Theorem 5.2, we obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla f\left(\theta_R\right)\right\|^2\right] \\ &\leq \frac{1}{ma}\left[\frac{2\left(f(\theta_1) - f(\theta^*)\right)}{(2 - La)} + 2mC_1\delta\left(\frac{a + La^2}{2 - La}\right)B\right. \\ &\qquad \left. + Lm\frac{a^2}{(2 - La)}\left[dC_1^2\delta^2 + C_2\delta^2 + \tilde{C}_2\right]\right].\end{aligned} \tag{5.42}$$

The main claim follows by plugging values of a and δ, defined in the theorem statement, in the inequality above. □

5.4.2 The sparse case

The zeroth norm $\|\theta\|_0$ of a vector θ is the number of non-zero entries in θ, i.e., $\|\theta\|_0 = \sum_{i=1}^{d} \mathbb{I}\left\{\theta^i \neq 0\right\}$, where θ^i denotes the ith coordinate of the vector θ. In this section, we make the following sparsity assumption on ∇f:

A5.5. For any $\theta \in \mathbb{R}^d$, the gradient of f is s-sparse, i.e.,

$$\|\nabla f(\theta)\|_0 \leq s,$$

where $s \ll d$.

The assumption above implies

$$\|\nabla f(\theta)\|_2 \leq \sqrt{s}\|\nabla f(\theta)\|_\infty \quad \text{and} \quad \|\nabla f(\theta)\|_1 \leq s\|\nabla f(\theta)\|_\infty.$$

Additionally, it follows that

$$\|\nabla f_\delta(\theta)\|_0 \leq s \quad \text{for all } \theta \in \mathbb{R}^d,$$

where $\nabla f_\delta(\theta) = \mathbb{E}_\Delta[\nabla f(\theta + \delta\Delta)]$. For the analysis in the sparse case, we make the following assumption that is a variant of Assumption A5.4:

A5.6. The sample performance F is such that (i) $\nabla f(\theta) = \mathbb{E}_\xi\left[\nabla F(\theta, \xi)\right]$; (ii) the gradient of F is Lipschitz continuous almost surely, for any ξ, i.e.,

$$\left\|\nabla F(\theta, \xi) - \nabla F(\tilde{\theta}, \xi)\right\|_1 \leq L\left\|\theta - \tilde{\theta}\right\|_\infty, \forall \theta, \tilde{\theta} \in \mathbb{R}^d,$$

for some $L > 0$; and (iii) there exists a $\sigma > 0$ such that the following inequality holds for any θ:

$$\mathbb{E}\left[\|\nabla F(\theta, \xi) - \nabla f(\theta)\|_1^2\right] \leq \sigma^2.$$

Next, we make the following assumption on the support of the gradient vector.

A5.7. There exists a set $S \subset \{1, \ldots, d\}$ s.t. $\nabla_i f(\theta) = 0$ if $i \notin S$, and non-zero if $i \in S$.

Since $\text{supp}(\nabla f(\theta)) = S$, it follows that $\text{supp}(\nabla f_\delta(\theta)) = S$. Thus,

$$\|\nabla f_\delta(\theta) - \nabla f(\theta)\|_2 = \left(\sum_{i\in S}(\nabla_i f_\delta(\theta) - \nabla_i f(\theta))^2\right)^{\frac{1}{2}},$$
$$\|\nabla f_\delta(\theta) - \nabla f(\theta)\|_2 \leq \sqrt{s}\|\nabla f_\delta(\theta) - \nabla f(\theta)\|_\infty. \tag{5.43}$$

The sparsity assumption A5.5 has been made earlier in (Balasubramanian and Ghadimi, 2022a). However, the non-asymptotic bound derived there is incorrect, as discussed in (Cai *et al.*, 2022). Motivated by the discussion in the aforementioned reference, we include the support assumption A5.7 as a fix to the non-asymptotic analysis of RSG in the sparse setting that we consider. While the support assumption in Assumption A5.7 is restrictive, the analysis goes through under any weaker assumption that ensures the condition in (5.43) holds.

Before presenting the main result, we provide a bound on the ℓ_∞-norm of the gradient estimate below.

Proposition 5.3. Suppose assumptions A5.5, A5.6, A5.7 hold. Then, the gradient estimator (5.36) satisfies

$$\mathbb{E}[\|\widehat{\nabla} f(\theta)\|_\infty^2] \leq C_a + C_b\|\nabla f(\theta)\|_1^2, \tag{5.44}$$

where $C_a = 4L^2\delta^2 C(\log(d))^3$, $C_b = 8C(\log(d))^2$.

In addition, with $f_\delta(\theta) = \mathbb{E}_\Delta(f(\theta + \delta\Delta)$ denoting the Gaussian smoothed functional, we have the following bound for any $\theta \in \mathbb{R}^d$:

$$\|\nabla f_\delta(\theta) - \nabla f(\theta)\|_2 \leq C\delta L\sqrt{2s}(\log(d))^{\frac{3}{2}}. \tag{5.45}$$

Proof. Notice that

$$\mathbb{E}\left[\|\widehat{\nabla} f(\theta)\|_\infty^2\right] = \mathbb{E}\left[\frac{(F(\theta+\delta\Delta,\xi) - F(\theta,\xi))^2}{\delta^2}\|\Delta\|_\infty^2\right] \tag{5.46}$$
$$= \mathbb{E}\left[\frac{(F(\theta+\delta\Delta,\xi) - F(\theta,\xi) - \delta\langle\nabla F(\theta,\xi),\Delta\rangle + \delta\langle\nabla F(\theta,\xi),\Delta\rangle)^2}{\delta^2} \times\|\Delta\|_\infty^2\right]$$
$$\leq \mathbb{E}\left[\frac{2\,(F(\theta+\delta\Delta,\xi) - F(\theta,\xi) - \delta\langle\nabla F(\theta,\xi),\Delta\rangle)^2}{\delta^2}\|\Delta\|_\infty^2\right]$$

$$
\begin{aligned}
&+ \mathbb{E}\left[\frac{2\left(\delta\langle\nabla F(\theta,\xi),\Delta\rangle\right)^2}{\delta^2}\|\Delta\|_\infty^2\right] \\
&\leq \mathbb{E}\left[\frac{2\left(\frac{L}{2}\delta^2\|\Delta\|_\infty^2\right)^2 + 2\left(\delta\langle\nabla F(\theta,\xi),\Delta\rangle\right)^2}{\delta^2}\|\Delta\|_\infty^2\right] \qquad (5.47)\\
&\leq \frac{L^2}{2}\delta^2\mathbb{E}\left[\|\Delta\|_\infty^6\right] + 2\mathbb{E}\left[\|\nabla F(\theta,\xi)\|_1^2\right]\mathbb{E}\left[\|\Delta\|_\infty^4\right] \\
&\leq 4L^2\delta^2 C(\log(d))^3 + 4C\left(\|\nabla f(\theta)\|_1^2 + \sigma^2\right)(\log(d))^2, \qquad (5.48)
\end{aligned}
$$

where we used L-smoothness of F, see Assumption A5.6, in arriving at (5.47) and the following inequality in the last step above:

$$\mathbb{E}[\|\Delta\|_\infty^k] \leq C(2\log(d))^{\frac{k}{2}}, \qquad (5.49)$$

for some universal constant C (see Lemma 3.1 in (Balasubramanian and Ghadimi, 2022a) for a proof).

The first claim follows. For the second claim, notice that

$$
\begin{aligned}
&\|\nabla f_\delta(\theta) - \nabla f(\theta)\|_2 \\
&\leq \sqrt{s}\|\nabla f_\delta(\theta) - \nabla f(\theta)\|_\infty \\
&\leq \frac{\sqrt{s}}{(2\pi)^{\frac{d}{2}}\delta}\int_{-\infty}^{\infty}|f(\theta+\delta u) - f(\theta) - \delta\langle\nabla f(\theta)),u\rangle|\ \|u\|_\infty \exp\left(-\frac{\|u\|^2}{2}\right)du \\
&\leq \frac{\sqrt{s}}{(2\pi)^{\frac{d}{2}}}\frac{\delta L}{2}\int_{-\infty}^{\infty}\|u\|_\infty^3\exp\left(-\frac{\|u\|^2}{2}\right)du \\
&\leq C\delta L\sqrt{2s}(\log d)^{\frac{3}{2}}, \qquad (5.50)
\end{aligned}
$$

where the final inequality used (5.49). □

The main result that presents a non-asymptotic bounds for the RSG algorithm with sparsity assumptions, is given below.

Theorem 5.8. Assume conditions of Proposition 5.3 hold. Suppose RSG algorithm is run with the stepsize $a(k)$ and the perturbation

constant δ_k set as follows:

$$a = \min\left\{\frac{1}{2sLC_b}, \frac{1}{\sqrt{m}}\right\}, \quad \delta = \frac{1}{\sqrt{m}},$$

where C_b is specified in Proposition 5.3. Then, choosing θ_R uniformly at random from $\{\theta_1, \ldots, \theta_m\}$, we have

$$\mathbb{E}\left[\|\nabla f(\theta_R)\|_1^2\right] \leq \frac{8s^2LC_bD_f}{m} + \frac{4sD_f}{\sqrt{m}} + 4s\left[\frac{C_c}{m^2} + \frac{C_d}{m\sqrt{m}}\right],$$

where C_a, C_b are defined in Proposition 5.3, $C_c = \frac{1}{2}\left(C^2L^2(2s)(\log(d))^3\right)$, $C_d = \frac{LC_a}{2}$, and D_f is defined in (5.9).

The bound above is $O\left(\frac{(\log(d))^3}{\sqrt{m}}\right)$. The poly-logarithmic dependence on d is an improvement over the corresponding $\sqrt{d}$ for the smooth sample performance case handled in the previous section.

Proof. Using L-smoothness of f in the ℓ_∞-norm from Assumption A5.6, we have

$$\begin{aligned} f(\theta_{k+1}) &\leq f(\theta_k) + \langle \nabla f(\theta_k), \theta_{k+1} - \theta_k \rangle + \frac{L}{2}\|\theta_{k+1} - \theta_k\|_\infty^2 \\ &\leq f(\theta_k) - a\langle \nabla f(\theta_k), \widehat{\nabla} f(\theta_k) \rangle + \frac{La^2}{2}\|\widehat{\nabla} f(\theta_k)\|_\infty^2, \end{aligned}$$

Taking conditional expectation w.r.t. the sigma field $\mathcal{F}_k$ as in earlier proofs, and using Proposition 5.3, we obtain

$$\begin{aligned} &\mathbb{E}_k[f(\theta_{k+1})] \\ &\leq f(\theta_k) - a\|\nabla f(\theta_k)\|^2 + a\langle \nabla f(\theta_k), \nabla f(\theta_k) - \nabla f_\delta(\theta_k) \rangle \\ &\quad + \frac{La^2}{2}\mathbb{E}_k[\|\widehat{\nabla} f(\theta_k)\|_\infty^2] \\ &\leq f(\theta_k) - \frac{a}{2}\|\nabla f(\theta_k)\|_2^2 + \frac{a}{2}\|\nabla f(\theta_k) - \nabla f_\delta(\theta_k)\|_2^2 \\ &\quad + \frac{La^2}{2}\mathbb{E}_k[\|\widehat{\nabla} f(\theta_k)\|_\infty^2]. \end{aligned}$$

Using (5.44), (5.45), and $\|\nabla f(\theta)\|_1 \leq \|\nabla f(\theta)\|_2\sqrt{s}$, we obtain

$$\mathbb{E}[f(\theta_{k+1})] \leq f(\theta_k) - \frac{a}{2s}\|\nabla f(\theta_k)\|_1^2 + \frac{a}{2}\left(C^2\delta^2L^2(2s)(\log(d))^3\right) + \frac{La^2}{2}(C_a + C_b\|\nabla f(\theta_k)\|_1^2).$$

Thus,

$$\left(\frac{a}{2s} - \frac{La^2}{2}C_b\right)\|\nabla f(\theta_k)\|_1^2 \leq f(\theta_k) - \mathbb{E}_k[f(\theta_{k+1})] + \frac{a}{2}\left(C^2\delta^2L^2(2s)(\log(d))^3\right) + \frac{La^2}{2}C_a.$$

Recall that $C_c = \frac{1}{2}\left(C^2L^2(2s)(\log(d))^3\right)$, and $C_d = \frac{L}{2}C_a$. Using these constants, the inequality above can be rewritten as follows:

$$\left(\frac{a}{2s} - \frac{La^2}{2}C_b\right)\|\nabla f(\theta_k)\|_1^2 \leq f(\theta_k) - \mathbb{E}[f(\theta_{k+1})] + C_c a\delta^2 + C_d a^2. \quad (5.51)$$

Taking total expectations, using $a \leq \frac{1}{2sLC_b}$ and $\mathbb{E}\left[f\left(\theta_k\right)\right] \geq f(\theta^*)$, for all $k \geq 1$, we obtain

$$\frac{a}{4s}\mathbb{E}\|\nabla f(\theta_k)\|_1^2 \leq \mathbb{E}\left(f(\theta_k) - f(\theta^*)\right) + C_c a\delta^2 + C_d a^2. \quad (5.52)$$

Since θ_R is picked uniformly at random from $\{\theta_1, \ldots, \theta_m\}$ and $a \leq 1/L$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f\left(\theta_R\right)\|_1^2\right] &= \frac{1}{m}\sum_{k=1}^{m}\mathbb{E}\,\|\nabla f\left(\theta_k\right)\|_1^2 \\
&\leq \frac{4s}{ma}\left[D_f + C_c a\delta^2 + C_d a^2\right] \\
&\leq \frac{4sD_f}{m}\max\left\{2sLC_b, \sqrt{m}\right\} + 4s\left[\frac{C_c\delta^2}{m} + \frac{aC_d}{m}\right] \\
&\leq \frac{8s^2LC_bD_f}{m} + \frac{4sD_f}{\sqrt{m}} + 4s\left[\frac{C_c}{m^2} + \frac{C_d}{m\sqrt{m}}\right].
\end{aligned}$$

The claim follows. □

5.5 Biased function measurements

In this section, we discuss a variant in which the function measurements are not unbiased and, instead, feature an estimation error component that can be controlled by increasing the batch size. We provide two motivating examples to illustrate this model variant.

Example 5.1. Consider a model in which the function measurements have an error term with a positive mean. In this model, the objective f is obtained as a solution to the following sub-problem over the optimization variable y that belongs to a convex and compact set $\mathcal{Y}$:

$$f(\theta) = \min_{y \in \mathcal{Y}} \mathbb{E}[H_\theta(y, \xi)], \forall \theta \in \mathbb{R}^d. \tag{5.53}$$

In practical applications, owing to computational considerations, a closed-form solution of the sub-problem defined above cannot be computed. A computationally efficient alternative is to perform gradient descent (GD) for a few steps, say m, and use the GD iterate as a proxy the function measurement. More precisely, let $F(\theta, m)$ denote an approximate solution of (5.53), where m denotes a batch-size parameter. Motivated by the GD approximation, we use the following form for $F(\theta, m)$:

$$F(\theta, m) = \min_{y \in \mathcal{Y}} \mathbb{E}[H_\theta(y, \xi)] + \epsilon(m), \forall \theta \in \mathbb{R}^d, \tag{5.54}$$

where ϵ is a 'positive' estimation error term. Choosing a larger batch size m implies that the subproblem in (5.53) can be solved more accurately (e.g. with more GD steps), leading to a lower estimation error $\epsilon(m)$.

The next example shows that biased function measurements appear naturally in the context of estimation of risk measures from i.i.d. samples.

Example 5.2. For a random variable X, recall that $\mathrm{VaR}_\alpha(X)$ and $\mathrm{CVaR}_\alpha(X)$, at a pre-specified level $\alpha \in (0, 1)$ are defined by

$$\mathrm{VaR}_\alpha(X) = \inf\{\xi : \mathbb{P}[X \leq \xi] \geq \alpha\}, \text{ and}$$
$$\mathrm{CVaR}_\alpha(X) = V_\alpha(X) + \frac{1}{1-\alpha}\mathbb{E}\left[X - V_\alpha(X)\right]^+,$$

where $[X]^+ = \max(0, X)$. If the distribution underlying X is continuous, then $\text{CVaR}_\alpha(X) = \mathbb{E}[X|X \geq \text{VaR}_\alpha(X)]$.

We now describe a well-known estimate of CVaR using m i.i.d. samples $\{X_i, i = 1, \ldots, m\}$. Note that CVaR estimation requires an estimate of VaR. Let $\hat{V}_{m,\alpha}$ and $\hat{C}_{m,\alpha}$ denote the estimates of VaR and CVaR. These quantities are defined as follows (see (Serfling, 2009)):

$$\hat{V}_{m,\alpha} = X_{[\lfloor m\alpha \rfloor]}, \hat{C}_{m,\alpha} = \frac{1}{m} \sum_{i=1}^{m} \frac{X_i \mathbb{I}\left\{X_i \geq \hat{V}_{m,\alpha}\right\}}{(1-\alpha)}. \tag{5.55}$$

In the above, $X_{[i]}$ denotes the ith order statistic, $\forall i$. Notice that $\mathbb{E}\left(\hat{C}_{m,\alpha}\right) \neq \text{CVaR}_\alpha(X)$, since the VaR estimate in (5.55) is not unbiased. However, a recent CVaR concentration result in (Prashanth and Bhat, 2022) shows that if the underlying r.v. X is σ-sub-Gaussian[4], then, for any $\epsilon > 0$, the following inequality holds:

$$\mathbb{P}(|\hat{C}_{m,\alpha} - \text{CVaR}_\alpha(X)| > \epsilon) \leq c_1 \exp(-c_2 m \epsilon^2 (1-\alpha)^2), \tag{5.56}$$

where constants c_1, c_2 depend on σ. Using (5.56), we have

$$\begin{aligned} &\mathbb{E}\left|\hat{C}_{m,\alpha} - \text{CVaR}_\alpha(X)\right| \\ &= \int_0^\infty \mathbb{P}(|\hat{C}_{m,\alpha}(X) - \text{VaR}_\alpha(X)| > \epsilon) d\epsilon \leq \frac{c_3}{\sqrt{m}}, \end{aligned} \tag{5.57}$$

where $c_3 > 0$ is an absolute constant.

In both the examples illustrated above, the common element is biased function measurements. Using such measurements, one could construct gradient estimates using the simultaneous perturbation (SP) method that was discussed earlier in Chapter 3. We make this construction precise below.

Let $y^+(m) = f(\theta + \delta\Delta) + \xi^+(m)$, and $y^-(m) = f(\theta - \delta\Delta) + \xi^-(m)$. Here $\xi^\pm(m)$ are the estimation errors assuming a batch size of m, δ is a perturbation constant, and $\Delta = \left(\Delta^1, \ldots, \Delta^d\right)^\top$ is a d-dimensional standard Gaussian vector. For the two examples discussed above, it

[4] A r.v. X with mean μ is said to be σ-sub-Gaussian for some $\sigma > 0$ if $\mathbb{E}[\exp(\lambda(X - \mu))] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$, for any $\lambda \in \mathbb{R}$.

is apparent that the estimation error is $\mathcal{O}(\frac{1}{\sqrt{m}})$ in expectation, if m samples are used for estimation of f at $(\theta \pm \delta\Delta)$ input parameters.

A gradient estimate is formed using two function evaluations (i.e., y^+ and y^-) as follows:

$$g(\theta, \delta, m) = \Delta \left[\frac{y^+(m) - y^-(m)}{\delta} \right], \tag{5.58}$$

where Δ is a d-dimensional Gaussian vector composed of standard normal r.v.s. Recall that the estimate defined above is referred to variously as Gaussian smoothed functional, and Gaussian smoothing. Assuming that the underlying function f is three-times continuously differentiable, we have

$$f(\theta \pm \delta\Delta) = f(\theta) \pm \delta\Delta^\intercal \nabla f(\theta) + \frac{\delta^2}{2}\Delta^\intercal \nabla^2 f(\theta)\Delta + \mathcal{O}(\delta^3).$$

Hence,

$$\begin{aligned}\mathbb{E}\left[\Delta\left(\frac{f(\theta+\delta\Delta) - f(\theta-\delta\Delta)}{2\delta}\right)\right] &= \mathbb{E}\left[\Delta\Delta^\intercal\right]\nabla f(\theta) + O(\delta^2) \\ &= \nabla f(\theta) + \mathcal{O}(\delta^2),\end{aligned}$$

where we used the fact that $\mathbb{E}\left[\Delta\Delta^\intercal\right] = I_d$, since Δ is a d-dimensional standard Gaussian vector. Combining the equality above with the fact that the estimation error is $\mathcal{O}(\frac{1}{\sqrt{m}})$, we obtain

$$\|\mathbb{E}\left[g\left(\theta, \delta, m\right)\right] - \nabla f\left(\theta\right)\| \leq c_1\delta^2 + \frac{c_2}{\sqrt{m}},$$

for some constants $c_1, c_2 > 0$. This satisfies the requirement (a) in **(O1)**.

A similar argument works for the case of a convex and smooth objective as well. In addition, a variety of distributions can be employed for the random perturbations, as discussed in Chapter 3.

Motivated by the discussion above, we define a biased gradient oracle with an estimation error component below.

(O1) Biased gradient oracle

Input: $\theta \in \mathbb{R}^d$, perturbation constant $\delta > 0$, and batch size $m > 0$.

Output: a gradient estimate $g(\theta, \delta, m) \in \mathbb{R}^d$ that satisfies

(a) $\|\mathbb{E}_\xi\left[g\left(\theta,\delta,m\right)\right]-\nabla f\left(\theta\right)\|_\infty \leq c_1\delta^2+\dfrac{c_3}{\delta\sqrt{m}}$,

(b) $\mathbb{E}_\xi\left[\,\|g\left(\theta,\delta,m\right)-\mathbb{E}_\xi\left[g\left(\theta,\xi,\delta,m\right)\right]\|^2\,\right] \leq \dfrac{c_2}{\delta^2}$,

for some constants $c_1, c_2, c_3 > 0$.

In the oracle defined above, the parameter δ is used to tradeoff bias and variance in the gradient estimates, while the parameter m is motivated by practical models where mini-batching is used for estimating the objective function. To elaborate, the function measurements are biased, however, one could choose larger values of m to increase the accuracy of the function measurements.

Using the gradient estimate from the oracle defined above, one can implement a stochastic gradient algorithm with the following update iteration:

$$\theta_{k+1} = \theta_k - a(k)\, g\left(\theta_k, \delta_k, m_k\right), \tag{5.59}$$

where δ_k is the perturbation constant and m_k the batch size at time instant k.

Following the proof technique from Section 5.1, it can be shown that the algorithm (5.59) satisfies the following bounds (see Exercise 1 below):

$$\mathbb{E}\left\|\nabla f\left(x_R\right)\right\|^2 \leq \frac{C}{m^{1/3}}, \tag{5.60}$$

for some constant C.

5.6 Minimax lower bound

In the analysis so far, we have observed that the convergence proofs rely on two properties of the gradient estimates formed using the simultaneous perturbation method, namely the bias and variance bounds in (4.3). Moreover, using such gradient estimates, we obtained a non-asymptotic bound of the order $O(1/m^{1/3})$ in the previous section. We now establish that this bound is not improvable in a minimax sense for any algorithm that is fed inputs from a biased gradient oracle, which is formalized below.

(O1) Biased gradient oracle

Input: $\theta \in \mathbb{R}^d$, perturbation constant $\delta > 0$.

Output: a gradient estimate $\widehat{\nabla} f(\theta) \in \mathbb{R}^d$ that satisfies

(a) $\|\mathbb{E}\left[\widehat{\nabla} f(\theta)\right] - \nabla f(\theta)\| \leq C_1 \delta^2$,

(b) $\mathbb{E}\left\|\widehat{\nabla} f(\theta) - \mathbb{E}\left[\widehat{\nabla} f(\theta)\right]\right\|^2 \leq \dfrac{C_2}{\delta^2}$,

for some constants $C_1, C_2 > 0$.

For the lower bound, we consider a setting where an optimization algorithm is required to select a point $\hat{\theta}_m \in \mathcal{K}$ after querying the oracle **(O1)** m times. The algorithm's performance is quantified using the *optimization error*, defined as

$$\Delta_m = \mathbb{E}\left[f(\hat{\theta}_m)\right] - \inf_{\theta \in \mathcal{K}} f(\theta), \tag{5.61}$$

where $\mathcal{K} \subset \mathbb{R}^d$ is a convex body, i.e., a nonempty closed convex set with a non-empty interior, and f is the objective function that is convex and L-smooth. We use $\mathcal{F}$ to denote the the set of convex and L-smooth functions with domain including $\mathcal{K}$.

The *worst-case error* is defined as follows:

$$\Delta^{\mathcal{A}}_{\mathcal{F},m}(C_1, C_2) = \sup_{f \in \mathcal{F}} \sup_{\gamma \in \Gamma_1(f, C_1, C_2)} \Delta^{\mathcal{A}}_m(f, \gamma)\,, \tag{5.62}$$

where $\Delta^{\mathcal{A}}_m(f, \gamma)$ is the optimization error that $\mathcal{A}$ suffers after m rounds of interaction with f through an oracle γ, and $\Gamma_1(f, C_1, C_2)$ denotes the set of **(O1)** oracles with constants C_1, C_2 satisfying the requirements **(O1)**a–**(O1)**b.

The *minimax error* is defined as

$$\Delta^*_{\mathcal{F},n}(C_1, C_2) = \inf_{\mathcal{A}} \Delta^{\mathcal{A}}_{\mathcal{F},n}(C_1, C_2),$$

where $\mathcal{A}$ ranges through all algorithms that interact with f through an oracle.

The main result that establishes a minimax lower bound is stated below[5].

[5]The reader is encouraged to read Appendix E before diving into the proof, as KL-divergence and Pinsker's inequality are essential to understanding of the derivation.

Theorem 5.9. Let $m > 0$ be an integer, $p, q > 0$, $C_1, C_2 > 0$, $\mathcal{K} \subset \mathbb{R}^d$ convex, closed, with $[+1, -1]^d \subset \mathcal{K}$. Then, for any algorithm that observes m random elements from a **(O1)** oracle, the minimax error satisfies the following bound:

$$\Delta^*_{\mathcal{F},m}(C_1, C_2) \geq K_1 \sqrt{d}\, C_1^{\frac{2}{3}} C_2^{\frac{1}{3}} m^{-\frac{1}{3}},$$

where K_1 is a universal constant.

Proof. First, we establish the lower bound for the one-dimensional case with $\mathcal{F}$ denoting the set of L smooth and convex functions with domain $\mathcal{K}$ that includes $[-1, 1]$, and $L \geq 1/2$. For brevity, let Δ^*_m denote the minimax error $\Delta^*_m(\mathcal{F}, c_1, c_2)$. Throughout the proof, a d-dimensional normal distribution with mean μ and covariance matrix Σ is denoted by $\mathsf{N}(\mu, \Sigma)$.

We begin by defining two functions $f_+, f_- \in \mathcal{F}$ with associated biased gradient oracles γ_+, γ_- such that the expected error of any deterministic algorithm can be bounded from below for the case when the environment is chosen uniformly at random from $\{(f_+, \gamma_+), (f_-, \gamma_-)\}$. By Yao's principle (Yao, 1977), the same lower bound applies to the minimax error Δ^*_m even when randomized algorithms are also allowed.

We consider the class of biased gradient oracles the construct a a random gradient estimate, when given input (θ, δ), as follows:

$$\widehat{\nabla} f(\theta, \delta) = \overline{\gamma}(\theta, \delta) + \xi \tag{5.63}$$

with some map $\overline{\gamma} : \mathcal{K} \times [0, 1) \to \mathbb{R}$, where ξ is a zero-mean normal random variable with variance $C_2 \delta^{-2}$, satisfying **(O1)**b. The map $\overline{\gamma}$ which will be chosen such that the bias requirement in **(O1)**a is satisfied.

Next, we define the two target functions and their associated oracles[6]. For $v \in \{\pm 1\}$, let

$$f_v(\theta) := \epsilon\,(\theta - v) + 2\epsilon^2 \ln\left(1 + e^{-\frac{\theta - v}{\epsilon}}\right), \quad x \in \mathcal{K}\,. \tag{5.64}$$

[6]With a slight abuse of notation, we will use interchangeably the subscripts $+$ $(-)$ and $+1$ (-1) for any quantities corresponding to these two environments, e.g., f_+ and f_{+1} (respectively, f_- and f_{-1}).

The idea underlying these functions is that they approximate $\epsilon|\theta - v|$, but with a prescribed smoothness. The first and second derivatives of f_v are

$$f_v'(\theta) = \epsilon \, \frac{1 - e^{-\frac{\theta - v}{\epsilon}}}{1 + e^{-\frac{\theta - v}{\epsilon}}}\,, \qquad \text{and} \qquad f_v''(\theta) = \frac{2e^{-\frac{\theta - v}{\epsilon}}}{\left(1 + e^{-\frac{\theta - v}{\epsilon}}\right)^2}.$$

From the above calculation, it is easy to see that $0 \le f''(\theta) \le 1/2$. Thus, f_v is $\frac{1}{2}$-smooth, and so $f_v \in \mathcal{F}$.

For $f_v, v \in \{-1, +1\}$, the gradient oracle we consider is defined as

$$\gamma_v(\theta, \delta) = \overline{\gamma}_v(\theta, \delta) + \xi_\delta,$$

with $\xi_\delta \sim \mathsf{N}(0, \frac{C_2}{\delta^2})$ selected independently for every query, where $\overline{\gamma}_v$ is a biased estimate of the gradient f_v'. We define the "bias" in $\overline{\gamma}_v$ to move the gradients closer to each other: The idea is to shift f_+' and f_-' towards each other, with the shift depending on the allowed bias $C_1\delta^2$. In particular, since $f_+' \le f_-'$, f_+' is shifted up, while f_-' is shifted down. However, the shifted up version of f_+' is clipped for positive x so that it never goes above the shifted down version of f_-'. By moving the curves towards each other, algorithms which rely on the obtained oracles will have an increasingly harder time (depending on the size of the shift) to distinguish whether the function optimized is f_+ or f_-. Since

$$0 \le f_-'(\theta) - f_+'(\theta) \le \sup_x f_-'(\theta) - \inf_x f_+'(\theta) = 2\epsilon\,,$$

we don't allow shifts larger than ϵ, leading to the following formal definitions:

$$\overline{\gamma}_+(\theta, \delta) = \begin{cases} f_+'(\theta) + \min(\epsilon, C_1\delta^2)\,, & \text{if } x < 0\,; \\ \min\left\{f_+'(\theta) + \min(\epsilon, C_1\delta^2), f_-'(\theta) - \min(\epsilon, C_1\delta^2)\right\}, & \text{else}\,, \end{cases} \tag{5.65}$$

and

$$\overline{\gamma}_-(\theta, \delta) =$$

$$\begin{cases} f'_-(\theta) - \min(\epsilon, C_1\delta^2)\,, & \text{if } x > 0\,; \\ \max\left\{f'_-(\theta) - \min(\epsilon, C_1\delta^2), f'_+(\theta) + \min(\epsilon, C_1\delta^2)\right\}, & \text{else}\,. \end{cases} \tag{5.66}$$

We claim that the oracle γ_v based on these functions satisfies the conditions imposed in **(O1)**. The variance condition **(O1)**b is trivially satisfied. To see that the bias is $C_1\delta^2$, notice that $\gamma_v(\theta,\delta) = -\gamma_{-v}(-x,\delta)$ and $f'_v(\theta) = -f'_{-v}(-x)$. Thus, $|\overline{\gamma}_+(\theta,\delta) - f'_+(\theta)| = |\overline{\gamma}_-(-x,\delta) - f'_-(-x)|$, hence it suffices to consider $v = +1$. The bias condition trivially holds for $x < 0$. For $x \geq 0$, using that $f'_+(\theta) \leq f'_-(\theta)$, we get

$$f'_+(\theta) - \min(\epsilon, C_1\delta^2) \leq \overline{\gamma}_+(\theta,\delta) \leq f'_+(\theta) + \min(\epsilon, C_1\delta^2),$$

showing $|\overline{\gamma}_+(\theta,\delta) - f'_+(\theta)| \leq C_1\delta^2$. Thus, γ_v is indeed a biased gradient oracle with the required properties.

To bound the performance of any algorithm in minimizing $f_v, v \in \{\pm 1\}$, notice that f_v is minimized at $\theta^*_v = v$, with $f_v(v) = 2\epsilon^2 \ln 2$. Next we show that if θ has the opposite sign of v, the difference $f_v(\theta) - f_v(\theta^*_v)$ is "large". This will mean that if the algorithm cannot distinguish between $v = +1$ and $v = -1$, it necessarily chooses a highly suboptimal point for either of these cases.

Since vf_v is decreasing on $\{\theta : \theta v \leq 0\}$, we have

$$M_v := \min_{x : xv \leq 0} f_v(\theta) - f_v(v) = f_v(0) - f_v(v) = \epsilon\left(-v + 2\epsilon \ln \frac{1 + e^{\frac{v}{\epsilon}}}{2}\right).$$

Let $h(v) = -v + 2\epsilon \ln \dfrac{1 + e^{\frac{v}{\epsilon}}}{2}$. Simple algebra shows that h is an even function, that is, $h(v) = h(-v)$. Indeed,

$$h(v) = -v + 2\,\epsilon\,\ln\left(e^{\frac{v}{\epsilon}}\frac{1 + e^{-\frac{v}{\epsilon}}}{2}\right) = -v + 2\,\epsilon\,\frac{v}{\epsilon} + 2\,\epsilon\,\ln\frac{1 + e^{-\frac{v}{\epsilon}}}{2} = h(-v)\,.$$

Specifically, $h(1) = h(-1)$ and thus

$$M_+ = M_- = \epsilon\left(-1 + 2\epsilon \ln \frac{1 + e^{\frac{1}{\epsilon}}}{2}\right).$$

From the foregoing, when $\theta v \leq 0$ and $\epsilon < \dfrac{1}{4\ln 2}$, we have

$$f_v(\theta) - f_v(\theta^*_v) \geq \epsilon\left(-1 + 2\epsilon \ln \frac{1 + e^{\frac{1}{\epsilon}}}{2}\right) > \frac{\epsilon}{2}.$$

Hence,

$$f_v(\theta) - f_v(\theta_v^*) \geq \frac{\epsilon}{2}\mathbb{I}\{\theta v < 0\}. \tag{5.67}$$

Given the above definitions and (5.67), by Yao's principle, the minimax error (5.62) is lower bounded by

$$\Delta_m^* \geq \inf_{\mathcal{A}} \mathbb{E}[f_V(\hat{X}_m) - \inf_{x\in X} f_V(\theta)] \geq \inf_{\mathcal{A}} \frac{\epsilon}{2}\mathbb{P}(\hat{X}_m V < 0), \tag{5.68}$$

where $V \in \{\pm 1\}$ is a random variable, $\hat{X}_m$ is the estimate of the algorithm after n queries to the oracle γ_V for f_V, the infimum is taken over all deterministic algorithms, and the expectation is taken with respect to the randomness in V and the oracle. More precisely, the distribution above is defined as follows:

Consider a fixed biased gradient oracle γ satisfying (5.63) and a deterministic algorithm $\mathcal{A}$. Let $\theta_t^{\mathcal{A}}$ (respectively, $\delta_t^{\mathcal{A}}$) denote the map from the algorithm's past observations that picks the point (respectively, accuracy parameter δ), which are sent to the oracle in round t. Define the probability space $(\Omega, \mathcal{B}, P_{\mathcal{A},\gamma})$ with $\Omega = \mathbb{R}^d \times \{-1, 1\}$, its associated Borel sigma algebra $\mathcal{B}$, where the probability measure $P_{\mathcal{A},\gamma}$ takes the form $P_{\mathcal{A},\gamma} := p_{\mathcal{A},\gamma} N(\lambda \times m)$, where λ is the Lebesgue measure on $\mathbb{R}^n$, m is the counting measure on $\{\pm 1\}$ and $p_{\mathcal{A},\gamma}$ is the density function defined by

$$\begin{aligned}
&p_{\mathcal{A},\gamma}(g_{1:n}, v) \\
&= \frac{1}{2}\Big(p_{\mathcal{A},\gamma}(g_m \mid g_{1:m-1}) \cdot \ldots \cdot p_{\mathcal{A},\gamma}(g_{m-1} \mid g_{1:m-2}) \cdot \ldots \cdot p_{\mathcal{A},\gamma}(g_1)\Big) \\
&= \frac{1}{2}\Big(p_{\mathcal{N}}(g_m - \overline{\gamma}(\theta_m^{\mathcal{A}}(g_{1:m-1}), \delta_m^{\mathcal{A}}(g_{1:m-1})), c_2(\delta_m^{\mathcal{A}}(g_{1:m-1}))) \cdot \ldots \cdot \\
&\qquad\qquad p_{\mathcal{N}}(g_1 - \overline{\gamma}(\theta_1^{\mathcal{A}}, \delta_1^{\mathcal{A}}), c_2(\delta_1^{\mathcal{A}}))\Big),
\end{aligned}$$

where $v \in \{-1, 1\}$ and $p_{\mathcal{N}}(\cdot, \sigma^2)$ is the density function of a $\mathsf{N}(0, \sigma^2)$ random variable. Then the expectation in (5.68) is defined w.r.t. the distribution $\mathbb{P} := \frac{1}{2}\left(P_{\mathcal{A},\gamma_+}\mathbb{I}\{v = +1\} + P_{\mathcal{A},\gamma_-}\mathbb{I}\{v = -1\}\right)$ and $V : \Omega \to \{\pm 1\}$ is defined by $V(g_{1:n}, v) = v$.[7] Define $\mathbb{P}_+(\cdot) := \mathbb{P}(\cdot \mid V = 1)$,

[7] Here, we are slightly abusing the notation as $\mathbb{P}$ depends on $\mathcal{A}$, but the dependence

$\mathbb{P}_-(\cdot) := \mathbb{P}(\cdot \mid V = -1)$. From (5.68), we obtain

$$\Delta_m^* \geq \inf_{\mathcal{A}} \frac{\epsilon}{4} \left(\mathbb{P}_+(\hat{X}_m < 0) + \mathbb{P}_-(\hat{X}_m > 0) \right), \tag{5.69}$$

$$\geq \inf_{\mathcal{A}} \frac{\epsilon}{4} \left(1 - \|\mathbb{P}_+ - \mathbb{P}_-\|_{\mathrm{TV}} \right), \tag{5.70}$$

$$\geq \inf_{\mathcal{A}} \frac{\epsilon}{4} \left(1 - \left(\frac{1}{2} D_{\mathrm{kl}} \left(P_+ \| P_- \right) \right)^{\frac{1}{2}} \right), \tag{5.71}$$

where (5.69) uses the definitions of $\mathbb{P}_+$ and $\mathbb{P}_-$, $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance, (5.70) follows from its definition, while (5.71) follows from Pinsker's inequality. It remains to upper bound $D_{\mathrm{kl}}\left(P_+ \| P_-\right)$.

Define G_t to be the tth observation of $\mathcal{A}$. Thus, $G_t : \Omega \to \mathbb{R}$, with $G_t(g_{1:n}, v) = g_t$. Let $P_+^t(g_1, \ldots, g_t)$ denote the joint distribution of $G_1, \ldots, G_t$ conditioned on $V = +1$. Let $P_+^t(\cdot \mid g_1, \ldots, g_{t-1})$ denote the distribution of G_t conditional on $V = +1$ and $G_1 = g_1, \ldots, G_{t-1} = g_{t-1}$. Define $P_{-j}^t(\cdot \mid g_1, \ldots, g_{t-1})$ in a similar fashion. Then, by the chain rule for KL-divergences, we have

$$D_{\mathrm{kl}}\left(P_+ \| P_-\right) = \sum_{t=1}^{m} \int_{\mathbb{R}^{t-1}} D_{\mathrm{kl}} \left(P_+^t(\cdot \mid g_{1:t-1}) \| P_-^t(\cdot \mid g_{1:t-1}) \right) dP_+^t(g_{1:t-1}). \tag{5.72}$$

By the oracle's definition on $V = +1$ we have $G_t \sim \mathsf{N}(\overline{\gamma}_+(\theta_t^{\mathcal{A}}(G_{1:t-1}), \delta_t^{\mathcal{A}}(G_{1:t-1})), c_2(\delta_t^{\mathcal{A}}(G_{1:t-1})))$, i.e., $P_+^t(\cdot \mid g_{1:t-1})$ is the normal distribution with mean $\overline{\gamma}_+(\theta_t^{\mathcal{A}}(G_{1:t-1}), \delta_t^{\mathcal{A}}(G_{1:t-1}))$ and variance $c_2(\delta_t^{\mathcal{A}}(G_{1:t-1}))$. Using the shorthands $\theta_t^{\mathcal{A}} := x_t^{\mathcal{A}}(g_{1:t-1})$, $\delta_t^{\mathcal{A}} := \delta_t^{\mathcal{A}}(g_{1:t-1})$, we have

$$D_{\mathrm{kl}} \left(P_+^t(\cdot \mid g_{1:t-1}) \| P_-^t(\cdot \mid g_{1:t-1}) \right) = \frac{(\overline{\gamma}_+(\theta_t^{\mathcal{A}}, \delta_t^{\mathcal{A}}) - \overline{\gamma}_-(\theta_t^{\mathcal{A}}, \delta_t^{\mathcal{A}}))^2}{2 c_2(\delta_t^{\mathcal{A}})},$$

as the KL-divergence between normal distributions $\mathsf{N}(\mu_1, \sigma^2)$ and $\mathsf{N}(\mu_2, \sigma^2)$ is equal to $\dfrac{(\mu_1 - \mu_2)^2}{2\sigma^2}$.

is suppressed. In what follows, we will define several other distributions derived from $\mathbb{P}$, which will all depend on $\mathcal{A}$, but for brevity this dependence will also be suppressed. The point where the dependence on $\mathcal{A}$ is eliminated will be called to the reader's attention.

It remains to upper bound the numerator. For $(\theta, \delta) \in \mathbb{R} \times (0, 1]$, first note that
$\gamma_+(\theta, \delta) \leq \gamma_-(\theta, \delta)$. Hence,

$$\begin{aligned}
|\gamma_+(\theta,\delta) - \gamma_-(\theta,\delta)| &= \gamma_-(\theta,\delta) - \gamma_+(\theta,\delta) \\
&< \sup_x \gamma_-(\theta,\delta) - \inf_x \gamma_+(\theta,\delta) \\
&= \lim_{x\to\infty} \gamma_-(\theta,\delta) - \lim_{x\to-\infty} \gamma_+(\theta,\delta) \\
&= \epsilon - \epsilon \wedge C_1\delta^2 - (-\epsilon + \epsilon \wedge C_1\delta^2) \\
&= 2\epsilon - 2\epsilon \wedge C_1\delta^2 \\
&\leq 2(\epsilon - C_1\delta^2)^+ , \qquad (5.73)
\end{aligned}$$

where $(u)^+ = \max(u, 0)$ is the positive part of u.

From the above, using the abbreviations $\theta_t^{\mathcal{A}} = x_t^{\mathcal{A}}(g_{1:t-1})$ and $\delta_t^{\mathcal{A}} = \delta_t^{\mathcal{A}}(g_{1:t-1})$ (effectively fixing $g_{1:t-1}$ for this step),

$$\begin{aligned}
D_{\mathrm{kl}}\left(P_+^t(\cdot \mid g_{1:t-1}) \| P_-^t(\cdot \mid g_{1:t-1})\right) &< \frac{2\{(\epsilon - C_1(\delta_t^{\mathcal{A}})^2)^+\}^2 (\delta_t^{\mathcal{A}})^2}{C_2} \qquad (5.74) \\
&\leq \sup_{\delta>0} \frac{2\{(\epsilon - C_1\delta^2)^+\}^2 \delta^2}{C_2} , \qquad (5.75)
\end{aligned}$$

where inequality (5.74) follows from (5.73). Notice that the right-hand side of the above inequality does not depend on the algorithm anymore.

Now, observe that $\sup_{\delta>0}\{(\epsilon - C_1\delta^2)^+\}^2\delta^2 = \sup_{(\epsilon/C_1)^{1/p} \geq \delta > 0} (\epsilon - C_1\delta^2)^2\delta^2$. From this observation, we obtain

$$\delta_* = \left(\frac{2\epsilon}{6C_1}\right)^{1/2} . \qquad (5.76)$$

Note that $C_1\delta_*^2 \leq \epsilon$, hence $\max_{\delta>0}\{(\epsilon - C_1\delta^2)^+\}^2\delta^2 = (\epsilon - C_1\delta_*^2)^2\delta_*^2$. Plugging (5.75) into (5.72) and using this last observation we obtain

$$D_{\mathrm{kl}}(P_+ \| P_-) \leq \frac{2m}{C_2} (\epsilon - C_1\delta_*^2)^2 \delta_*^2 . \qquad (5.77)$$

Note that the above bound holds uniformly over all algorithms $\mathcal{A}$. Substituting the above bound into (5.71), we obtain

$$\Delta_m^* \geq \frac{\epsilon}{4}\left(1 - \sqrt{m}\frac{(\epsilon - C_1\delta_*^2)\delta_*}{\sqrt{C_2}}\right) = \frac{\epsilon}{4}\left(1 - \sqrt{m}K_1\epsilon^{\frac{3}{2}}\right) , \qquad (5.78)$$

where $K_1 = \frac{4}{6\sqrt{C_2}} \left(\frac{2}{6C_1}\right)^{\frac{1}{2}}$.

By choosing $\epsilon = \left(\frac{2}{5\sqrt{m}K_1}\right)^{\frac{2}{3}}$, we see that

$$\Delta_m^* \geq \frac{9}{20} \left(\frac{1}{25}\right)^{1/3} C_1^{1/3} C_2^{1/3} m^{-1/3}. \tag{5.79}$$

Generalization to N dimensions: To prove the d-dimensional result, we introduce a new device which allows us to relate the minimax error of the d-dimensional problem to that of the 1-dimensional problem. The main idea is to use separable d-dimensional functions and oracles and show that if there exists an algorithm with a small loss for a rich set of separable functions and oracles, then there exists good one-dimensional algorithms for the one-dimensional components of the functions and oracles.

This device works as follows: First we define one-dimensional functions. For $1 \leq i \leq d$, let $\mathcal{K}_i \subset \mathbb{R}$ be nonempty sets, and for each $v_i \in V := \{\pm 1\}$, let $f_v^{(i)} : \mathcal{K}_i \to \mathbb{R}$. Let $\mathcal{K} = \times_{i=1}^d \mathcal{K}_i$ and for $v = (v_1, \ldots, v_d) \in V^d$, let $f_v : \mathcal{K} \to \mathbb{R}$ be defined by

$$f_v(\theta) = \sum_{i=1}^d f_{v_i}^{(i)}(\theta_i), \qquad \theta \in \mathcal{K}. \tag{5.80}$$

Without the loss of generality, we assume that $\inf_{\theta_i \in \mathcal{K}_i} f_{v_i}^{(i)}(\theta_i) = 0$, and hence $\inf_{x \in \times_{i=1}^d \mathcal{K}_i} f_v(\theta) = 0$, so that the optimization error of the algorithm producing $\hat{X}_n \in \mathcal{K}$ as the output is $f_v^{(i)}(\hat{X}_{n,i})$ and $f_v(\hat{X}_n)$, respectively. We also define a d-dimensional *separable* oracle γ_v as follows: The oracle is obtained from "composing" the d one-dimensional oracles, $(\gamma_{v_i}^{(i)})_i$. In particular, the ith component of the response of γ_v given the history of queries $(\theta_t, \delta_t, \ldots, \theta_1, \delta_1) \in (\mathcal{K} \times [0,1))^t$ is defined as the response of $\gamma_{v_i}^{(i)}$ given the history of queries $(\theta_{t,i}, \delta_t, \ldots, \theta_{1,i}, \delta_1) \in (\mathcal{K}_i \times [0,1))^t$. This definition is so far unclear about the randomization of the oracles. In fact, it turns out that the one-dimensional oracles can even use the same randomization (i.e., their output can depend on the same single uniformly distributed random variable U), but they could also

use separate randomization: our argument will not depend on this. Let $\Gamma^{(i)}(f_{v_i}^{(i)}, c_1, c_2)$ denote a non-empty set of biased gradient oracles for objective function $f_{v_i}^{(i)} : \mathcal{K}_i \to \mathbb{R}$, and let us denote by $\Gamma_{\text{sep}}(f_v, c_1, c_2)$ the set of separable oracles for the function f_v defined above. We also define $\mathcal{F}_{\text{sep}} = \{f \,:\, f(\theta) = \sum_{i=1}^{d} f_{v_i}^{(i)}(\theta_i), x \in \mathcal{K}, v_i \in V_i\}$, the set of componentwise separable functions. Note that when $\|\cdot\| = \|\cdot\|_2$ is used in the definition of type-I oracles then $\Gamma_{\text{sep}}(f_v, C_1/\sqrt{d}, C_2/d) \subset \Gamma(f_v, C_1, C_2)$.

Let an algorithm $\mathcal{A}$ interact with an oracle γ. We will denote the distribution of the output $\hat{X}_n$ of $\mathcal{A}$ at the end of n rounds by $F_{\mathcal{A},\gamma}$ (we fix n, hence the dependence of F on n is omitted). Thus, the expected optimization error of $\mathcal{A}$ on a function f with zero optimal value is

$$L^{\mathcal{A}}(f, \gamma) = \int f(\theta) F_{\mathcal{A},\gamma}(dx)\,.$$

Note that this definition applies both in the one and the d-dimensional cases. For $v \in V^d$, we introduce the abbreviation

$$L^{\mathcal{A}}(v) = L^{\mathcal{A}}(f_v, \gamma_v)\,.$$

We also define

$$\tilde{L}_i^{\mathcal{A}}(v) = \int f_{v_i}^{(i)}(\theta_i) F_{\mathcal{A},\gamma_v}(dx)$$

so that

$$L^{\mathcal{A}}(v) = \sum_{i=1}^{d} \tilde{L}_i^{\mathcal{A}}(v)\,.$$

Also, for $v_i \in V$ and a one-dimensional algorithm $\mathcal{A}$, we let

$$L_i^{\mathcal{A}}(v_i) = L^{\mathcal{A}}(f_{v_i}^{(i)}, \gamma_{v_i}^{(i)})\,.$$

Note that while the domain of $\tilde{L}_i^{A}$ is V^d, the domain of $L_i^{\mathcal{A}}$ is V, while both express an expected error measured against $f_{v_i}^{(i)}$. In fact, $\tilde{L}_i^{A}$ depends on v because the algorithm $\mathcal{A}$ uses the d-dimensional oracle γ_v, which depends on v (and not only on v_i) and thus algorithm $\mathcal{A}$ could use information returned by $\gamma_{v_j}^{(j)}$, $j \neq i$. In a way our proof shows

that using this information cannot help a d-dimensional algorithm on a separable problem, a claim that we find rather intuitive, and which we now formally state (see (Hu *et al.*, 2016) for a detailed proof).

Lemma 5.10. Let $(f_v)_{v\in V^d}, f_v \in \mathcal{F}_{\text{sep}}$, $(\gamma_v)_{v\in V^d}, \gamma_v \in \Gamma_{\text{sep}}(f_v, c_1, c_2)$ be separable for some arbitrary functions c_1, c_2, and let $\mathcal{A}$ be any d-dimensional algorithm. Then there exist d one-dimensional algorithms, $\mathcal{A}_i^*$, $1 \le i \le d$ (using only one-dimensional oracles), such that

$$\max_{v\in V} L^{\mathcal{A}}(v) \ge \max_{v_1\in V_1} L_1^{\mathcal{A}_1^*}(v_1) + \cdots + \max_{v_d\in V_d} L_d^{\mathcal{A}_d^*}(v_d)\,. \tag{5.81}$$

Now, let

$$\mathcal{F}^{(i)} = \{f_{v_i} \,:\, v_i \in V\}, \qquad i = 1, \ldots, d\,.$$

The next result follows easily from the previous lemma:

Lemma 5.11. Let $\|\cdot\| = \|\cdot\|_2$ in the definition of the type-I oracles. Then, we have that

$$\Delta^*_{\mathcal{F}_{\text{sep}},n}(c_1, c_2) \ge \sum_{i=1}^{d} \Delta^*_{\mathcal{F}^{(i)},n}(c_1/\sqrt{d}, c_2/d)\,.$$

Let $\mathcal{K} \subset \mathbb{R}^d$, such that $\times_i \mathcal{K}_i \subset \mathcal{K}$, $\{\pm 1\} \subset \mathcal{K}_i \subset \mathbb{R}$, $\mathcal{F}_d = \mathcal{F}_{L,0}(\mathcal{K})$, where recall that $L \ge 1/2$. For any $1 \le i \le d$, $\theta_i \in \mathcal{K}_i$,

$$f_{v_i}^{(i)}(\theta_i) := \epsilon\,(\theta_i - v_i) + 2\epsilon^2 \ln\left(1 + e^{-\frac{\theta_i - v_i}{\epsilon}}\right)\,. \tag{5.82}$$

i.e., $f_{v_i}^{(i)}$ is like in the one-dimensional lower bound proof (cf. equation 5.64). Note that $f_v \in \mathcal{F}_d$ since f_v is separable, so its Hessian is diagonal and from our earlier calculation we know that $0 \le \dfrac{\partial^2}{\partial\theta_i^2} f_{v_i}^{(i)}(\theta_i) \le 1/2$. Let $\Delta_m^{(d)*}$ denote the minimax error $\Delta^*_{\mathcal{F}_d,n}\left(C_1\delta^2, \dfrac{C_2}{\delta^2}\right)$ for the d-dimensional family of functions $\mathcal{F}_d$. Let $\mathcal{F}^{(i)} = \{f_{-1}^{(i)}, f_{+1}^{(i)}\}$. As it was noted above, $f_v \in \mathcal{F}_d$ for any $v \in \{\pm 1\}^d$. Hence, by Lemma 5.11,

$$\Delta_m^{(d)*} \ge \sum_{i=1}^{d} \Delta^*_{\mathcal{F}^{(i)},m}\left(\frac{C_1}{\sqrt{d}}\,\delta^2, \frac{C_2}{d}\delta^{-2}\right)\,. \tag{5.83}$$

Plugging the lower bound derived in (5.79) for the one-dimensional setting into the bound in (5.83), we obtain a $\sqrt{d}$-times bigger lower bound for the d-dimensional case. In particular, we obtain

$$\Delta_m^{(d)*} \geq \frac{9}{10}\left(\frac{C_1 C_2}{25}\right)^{1/3}\sqrt{d}m^{-1/3}.$$

□

5.7 Bandit convex optimization

The bounds presented in this chapter relate to bandit convex optimization (BCO) — a topic that is not dealt in detail directly in this book. In this section, we show the connection between minimizing a smooth convex function in a zeroth-order setting and BCO.

In the BCO setting, the environment chooses sequence $\{f_1, \ldots, f_m\}$ of convex loss functions over a common domain $\mathcal{K}$, and a bandit algorithm chooses a sequence of points $\{\theta_1, \ldots, \theta_m\}$ iteratively. The expected regret R_m incurred by the algorithm is defined as follows:

$$R_n = \mathbb{E}\left[\sum_{t=1}^{m} f_t(\theta_t)\right] - \inf_{\theta\in\mathcal{K}} \sum_{t=1}^{m} f_t(\theta).$$

A simple stochastic gradient algorithm for this setting would update as follows:

$$\theta_{t+1} = \theta_t - a(t)\widehat{\nabla} f_t(\theta_t), \tag{5.84}$$

where $\widehat{\nabla} f_t(\theta_t)$ is an estimate of the gradient $\nabla f_t(\theta_t)$. To form this gradient estimate, the bandit algorithm is given access to a function observation at a point of its choice, say $\tilde{\theta}_t$ in round t. Notice that the algorithm's query point $\tilde{\theta}_t$ can be different from the point θ_t recommended (and used in calculating regret R_n).

In (Flaxman *et al.*, 2005; Saha and Tewari, 2011), the authors employ a one point gradient estimate, along the lines described in Section 3.3.1. While (Flaxman *et al.*, 2005) established a regret bound of $O(m^{3/4})$, it was later improved to $O(m^{3/4})$ by Saha and Tewari, 2011.

If the BCO setting allows two function observations for each f_t, then using a two-point gradient estimate, it is possible to obtain a

regret bound of $O(\sqrt{m})$, see (Agarwal *et al.*, 2010). In (Shamir, 2017), the authors consider a variant of the two-point gradient estimate (see Section 3.2) and obtain a $O\left(\sqrt{\frac{d}{m}}\right)$ regret bound. This bound has the optimal dimension dependence.

A simple scheme to convert a regret-minimizing bandit algorithm to one that optimizes a smooth convex function in a zeroth-order setting is to employ averaging. More precisely, let $\bar{\theta}_m = \frac{1}{m}\sum_{i=1}^{m}\theta_t$ denote the average point, where θ_t is the point chosen by the bandit algorithm. The average point $\bar{\theta}_m$ satisfies

$$\mathbb{E}\left[f(\bar{\theta}_m)\right] - \inf_{\theta\in\mathcal{K}} f(\theta) \leq R_m,$$

where the bandit algorithm is fed the same function f in each round $t = 1, \ldots, m$.

We end this section with the remark that for a bandit algorithm with inputs from a biased gradient oracle such as the one described in Section 5.6, the best achievable regret bound is $\Omega(m^{2/3})$, and this is equivalent to the bound of $O(1/m^{1/3})$ on the optimization error that we obtained in Theorem 5.9, see also (Hu *et al.*, 2016).

5.8 Exercises

Exercise 1. Prove the bound in (5.60) while making the necessary smoothness assumptions on the objective. Specify the choice of parameters $a(k), m_k, \delta_k$.

Exercise 2. Given a dataset $D_n = \{(a_i, y_i); i = 1, .., n\}$ with $a_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, consider the linear regression problem of finding the minimizer x^* of the following objective:

$$J(x) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - x^T a_i)^2. \tag{5.85}$$

Answer the following:

(a) Find the gradient and Hessian of J at a given point x.

(b) Does the function J have a minimizer? Is it unique?

(c) Write down the update rule for a gradient descent (GD) algorithm to find the minimizer x^* of J.

(d) Let A be the $n \times d$ matrix whose i^{th} row is a_i^T. Assume A^TA is positive definite and let $\mu > 0$ denote its minimum eigenvalue. Show that the gradient descent iterate, say x_n, after n iterations, satisfies the following bounds:

$$\|x_n - x^*\|^2 \leq (x_0 - x^*)^T(I - \alpha A^TA)^{2n}(x_0 - x^*), \text{ and}$$
$$J(x_n) - J(x^*) \leq (x_0 - x^*)^T(I - \alpha A^TA)^{2n}A^TA(x_0 - x^*),$$

where α is the constant stepsize used by the GD algorithm.

(e) What is the optimal value of α that minimizes the bounds specified in the part above? Justify your choice for α.

Exercise 3. Let $f(x) = \sum_{i=1}^{m} f_i(x)$, where f is a L-smooth function, and $\|\nabla f_i(x)\|^2 \leq \sigma^2$, for $i = 1, \ldots, m$. Do note that f is *not* necesarily convex.

Answer the following:

(a) For minimizing f, write the update iteration of the SGD algorithm with stepsize denoted by $a(k)$ and iterate by x_k.

(b) Show the following bound holds for SGD algorithm from the part above:

$$\mathbb{E}\left[f(x_{k+1}) - f(x_k)\right] \leq -a(k)\|\nabla f(x_k)\|_2^2 + \frac{1}{2}a(k)^2L\sigma^2. \quad (5.86)$$

(c) Fix n, set $\alpha = \dfrac{c}{\sqrt{n}}$ for some constant c. Is there a choice for the constant c such that the following bound holds for the SGD algorithm with stepsize α:

$$\min_{0 \leq k \leq n-1} \mathbb{E}\left[\|\nabla f(x_k)\|_2^2\right] \leq \sqrt{\frac{2(f(x_0) - f(x^*))L\sigma^2}{n}},$$

where x^* is a global minimum of f. Show your work in arriving at the bound above for suitable α.

(d) Is the bound in the part above the best achievable using a stochastic gradient algorithm? Or can it be improved?

Exercise 4. Generalize the minimax lower bound in Theorem 5.9 to the following biased gradient oracle variant with a gradient estimate that satisfies the following properties:

$$\|\mathbb{E}\left[\widehat{\nabla} f(\theta)\right] - \nabla f(\theta)\| \leq C_1 \delta^p,$$
$$\mathbb{E}\left\|\widehat{\nabla} f(\theta) - \mathbb{E}\left[\widehat{\nabla} f(\theta)\right]\right\|^2 \leq \frac{C_2}{\delta^q},$$

for some constants $C_1, C_2 > 0$.

In particular, for the convex and L-smooth case, show that the minimax error satisfies

$$\Delta^*_{\mathcal{F},m}(C_1, C_2) = \Omega(m^{-\frac{p}{2p+q}}),$$

and for the strongly convex case

$$\Delta^*_{\mathcal{F},m}(C_1, C_2) = \Omega(m^{-\frac{p}{p+q/2}}).$$

5.9 Bibliographic remarks

The presentation of non-asymptotic upper as well as lower bounds is based on recent research on analysis of SG algorithms in a zeroth-order setting. In the following, we provide some references section-wise.

5.1,5.2 RSG algorithm was proposed and analyzed in (Ghadimi and Lan, 2013). We follow this reference for the unbiased gradient information, while specialize the results in Bhavsar and Prashanth, 2022 for the biased case. A special case worth considering is $f(\theta) = \mathbb{E}_\zeta(F(\theta, \zeta))$, where ζ denotes the noise element. One can obtain an improved rate of $O(1/\sqrt{m})$ when F is assumed to be L-smooth. This implies f is L-smooth, but the converse is not true. Recall that in the latter case, we could obtain $O(1/m^{1/3})$ bound. For the convex case, one could employ a geometric step-size rule to derive a $O(1/m)$ bound for the optimization error in the zeroth-order setting. The reader is referred to Section IV of (Bhavsar and Prashanth, 2022) for the details. The approach adopted in

the aforementioned reference in arriving at a last iterate bound is inspired from (Jain *et al.*, 2021).

5.3 For the strongly-convex case, we have used the analysis in the survey article (Bottou *et al.*, 2018). This applies to the unbiased gradient information case, while the biased case requires careful handling of the bias-variance trade-off parameter. For the bound on SG with biased gradient information, we rely on the proof technique from (Frikha and Menozzi, 2012), and do the necessary modifications to handle the bias in gradient estimates.

5.6 The presentation of the lower bound is based on the results in (Hu *et al.*, 2016).

6

Hessian estimation and a stochastic Newton algorithm

Recall that a stochastic Newton algorithm's update is the following:

$$\theta_{n+1} = \theta_n - a_n \left(\overline{H}_n\right)^{-1} \widehat{\nabla} f(\theta_n), \tag{6.1}$$

where $\widehat{\nabla} f(\theta_n)$ and $\overline{H}_n$ denote the gradient and Hessian estimates, respectively. The topic of gradient estimation was handled in Chapter 3, while this chapter focuses on Hessian estimation. In the next chapter, we shall perform a convergence analysis of (6.1), where we use zeroth-order estimates of both the gradient and the Hessian.

The Hessian estimate $\overline{H}_n$ is usually arrived at by explicit averaging of previously obtained estimates, i.e., $\frac{1}{n}\sum_{k=1}^{n} \widehat{H}_k$, with $\widehat{H}_k$ denoting the Hessian estimate formed using a certain number of function measurements in iteration k of (6.1). Alternatively, one can employ stochastic approximation with a more general step size to arrive at an average of $\widehat{H}_k$, $k = 1, \ldots, n$, implicitly. The focus of this chapter is to form $\widehat{H}_k$, using function measurements. For simplicity, we drop the dependence on the iteration number k. The convergence analysis of (6.1) in the next chapter would make the Hessian estimate iteration-dependent.

Figure 6.1: The interaction of a second-order stochastic gradient algorithm with an estimator that estimates the Hessian at the input point θ, with perturbation constant δ.

6.1 The estimation problem

As illustrated in Figure 6.1, the second-order algorithm would ask for Hessian estimates (in addition to gradient estimates — a topic that is already covered) in each update iteration. For simplicity, henceforth we drop the dependence on the iteration number n of (6.1) and instead, consider the problem of obtaining an estimate $\widehat{H}$ of the Hessian at a given point $\theta \in \mathbb{R}^d$, using multiple function measurements.

We first describe the classic FDSA scheme, which was proposed by Fabian, 1971. This scheme requires $O(d^2)$ function observations to estimate the Hessian. Subsequently, we introduce the simultaneous perturbation trick to Hessian estimation and describe the following well-known variants that require a constant number of function observations, irrespective of the dimension d:

SPSA: We consider two variants (both balanced) that require four and three function measurements, respectively;

SF: We present two variants that require one and two function measurements, respectively. Both methods are based on the idea of Gaussian smoothed functional, which was considered earlier in Chapter 3 in the context of gradient estimation;

RDSA: A scheme that requires three function measurements.

6.2 FDSA for Hessian estimation

Consider a scalar variable θ. A finite difference approximation of the first derivative for this simple case of a scalar parameter θ is:

$$\frac{df(\theta)}{d\theta} \approx \left(\frac{f(\theta+\delta) - f(\theta-\delta)}{2\delta} \right). \tag{6.2}$$

Assuming the objective is smooth, and employing Taylor series expansions of $f(\theta+\delta)$ and $f(\theta-\delta)$ around θ, we obtain:

$$f(\theta \pm \delta) = f(\theta) \pm \delta \frac{df(\theta)}{d\theta} + \frac{\delta^2}{2} \frac{d^2 f(\theta)}{d\theta^2} + O(\delta^3),$$

$$\text{Thus, } \frac{f(\theta+\delta) - f(\theta-\delta)}{2\delta} = \frac{df(\theta)}{d\theta} + O(\delta^2).$$

From the above, it is easy to see that the estimate (6.2) converges to the true gradient $\dfrac{df(\theta)}{d\theta}$ in the limit as $\delta \to 0$.

This idea can be extended to estimate the second derivative by applying a finite difference approximation to the derivative in (6.2) as follows:

$$\frac{d^2 f(\theta)}{d\theta^2} \approx \frac{\left(\dfrac{f(\theta+\delta+\delta) - f(\theta+\delta-\delta)}{2\delta} \right) - \left(\dfrac{f(\theta-\delta+\delta) - f(\theta-\delta-\delta)}{2\delta} \right)}{2\delta} \tag{6.3}$$

As before, using Taylor series expansions, it can be shown that the RHS above is a good approximation to the second derivative.

For the case of a vector parameter, one needs to perturb each coordinate separately, leading to the following scheme for estimating the Hessian $\nabla^2 f(\theta)$: For any $i, j \in \{1, \ldots, d\}$,

$$\nabla^2_{ij} f(\theta) \approx \frac{1}{4\delta^2} \Big(f(\theta + \delta e_i + \delta e_j) + f(\theta + \delta e_i - \delta e_j) - (f(\theta - \delta e_i + \delta e_j) - f(\theta - \delta e_i - \delta e_j)) \Big). \tag{6.4}$$

Such an approach requires $4d^2$ number of function measurements to form the Hessian estimate. In the next section, we overcome this limitation by

employing the simultaneous perturbation trick. Before that, we extend the estimate in (6.4) to the case of noisy function measurements. Suppose we have the following function measurements: For any $i, j \in \{1, \ldots, d\}$,

$$y_{ij}^{++} = f(\theta + \delta e_i + \delta e_j) + \xi_{ij}^{++},\ y_{ij}^{+-} = f(\theta + \delta e_i - \delta e_j) + \xi_{ij}^{+-}, \quad (6.5)$$

$$y_{ij}^{-+} = f(\theta - \delta e_i + \delta e_j) + \xi_{ij}^{-+} \text{ and } y_{ij}^{--} = f(\theta - \delta e_i - \delta e_j) + \xi_{ij}^{--}. \quad (6.6)$$

Using these function measurements, we form the Hessian estimate $\widehat{H}$ as follows:

$$\widehat{H}_{ij} = \left(\frac{y_{ij}^{++} - y_{ij}^{+-} - y_{ij}^{-+} + y_{ij}^{--}}{4\delta^2} \right), \forall i, j \quad (6.7)$$

We analyze the bias of the estimator defined above, under the following assumptions:

A6.1. f is four-times continuously differentiable[1] with $\left|\nabla^4_{i_1,i_2,i_3,i_4} f(\theta)\right| < \infty$, for $i_1, i_2, i_3, i_4 = 1, \ldots, d$ and for all $\theta \in \mathbb{R}^d$.

A6.2. $\mathbb{E}\left[\xi_{ij}^{++} \middle| \theta\right] = \mathbb{E}\left[\xi_{ij}^{+-} \middle| \theta\right] = \mathbb{E}\left[\xi_{ij}^{-+} \middle| \theta\right] = \mathbb{E}\left[\xi_{ij}^{--} \middle| \theta\right] = 0$ for $i, j = 1, \ldots, d$.

The four-times continuously differentiability assumption on f in A6.1 allows Taylor series expansions, while A6.2 ensures the noise factors vanish in the bias analysis. Under A6.1–A6.2, we have

$$\begin{aligned}\mathbb{E}[\widehat{H}_{ij} \mid \theta] &= \frac{1}{4\delta^2}\Big(f(\theta + \delta e_i + \delta e_j) + f(\theta + \delta e_i - \delta e_j) \\ &\qquad - (f(\theta - \delta e_i + \delta e_j) - f(\theta - \delta e_i - \delta e_j)) \Big) \\ &= \nabla^2_{ij} f(\theta) + O(\delta^2).\end{aligned}$$

The final equality can be arrived at using Taylor series expansions followed by straightforward simplifications.

[1] Here $\nabla^4 f(\theta) = \dfrac{\partial^4 f(\theta)}{\partial\theta^\intercal \partial\theta^\intercal \partial\theta^\intercal \partial\theta^\intercal}$ denotes the fourth derivative of f at θ and $\nabla^4_{i_1,i_2,i_3,i_4} f(\theta)$ denotes the (i_1, i_2, i_3, i_4)th entry of $\nabla^4 f(\theta)$, for $i_1, i_2, i_3, i_4 = 1, \ldots, d$.

6.3 SPSA for Hessian estimation

6.3.1 Four measurements Hessian estimator

In this section, we present the Hessian estimation scheme from (Spall, 2000). Let Δ be a d-vector of symmetric, ± 1-valued Bernoulli r.v.s, as in the case of first-order SPSA (see Section 3.2). Suppose $G(\theta \pm \delta\Delta)$ are approximations to the gradient of f at $\theta \pm \delta\Delta$. Then, the simultaneous perturbation trick suggests the following Hessian estimate:

$$\widehat{H} = \Delta^{-1} \frac{G(\theta + \delta\Delta) - G(\theta - \delta\Delta)}{4\delta}, \tag{6.8}$$

where $\Delta^{-1} \stackrel{\triangle}{=} (1/\Delta_1, \ldots, 1/\Delta_d)^T$.

What remains to be specified are the gradient estimates for input parameters $\theta + \delta\Delta$. For forming this estimate, we use the simultaneous perturbation trick again, i.e.,

$$G(\theta \pm \delta\Delta) = \hat{\Delta}^{-1} \frac{(y^{++} - y^{+})}{\delta},$$

where $\hat{\Delta}$ denote a second independent set of perturbations having the same distribution as Δ,

$$y^{++} = f(\theta + \delta\Delta + \delta\hat{\Delta}) + \xi^{++}, y^{-+} = f(\theta - \delta\Delta + \delta\hat{\Delta}) + \xi^{-+},$$
$$y^{+} = f(\theta + \delta\Delta) + \xi^{+}, \text{ and } y^{-} = f(\theta - \delta\Delta) + \xi^{-}.$$

We can reuse these samples to form a SPSA-based gradient estimate as follows:

$$\widehat{\nabla} f(\theta) = \Delta^{-1} \left(\frac{y^{+} - y^{-}}{2\delta} \right).$$

We require the gradient as well as the Hessian estimates to implement a Newton step, as given in (6.1). The bias of the gradient estimate given above is analyzed in Chapter 3.

For the bias bound of the Hessian estimator defined in (6.8), we require the following assumption on the noise elements.

A6.3. Let $\Delta = (\Delta_1, \ldots, \Delta_d)^{\mathsf{T}}$ and $\hat{\Delta} = (\hat{\Delta}_1, \ldots, \hat{\Delta}_d)^{\mathsf{T}}$ be two independent d-vectors of mutually independent, symmetric, ± 1-valued Bernoulli

r.v.s. Further, given θ, $\{\xi^{++}, \xi^{-+}, \xi^{+}, \xi^{-}\}$ is independent of $\Delta, \hat{\Delta}$. In addition,

$$\mathbb{E}\left[\xi^{++}\middle|\theta\right] = \mathbb{E}\left[\xi^{-+}\middle|\theta\right] = \mathbb{E}\left[\xi^{+}\middle|\theta\right] = \mathbb{E}\left[\xi^{-}\middle|\theta\right] = 0.$$

Lemma 6.1. Assume A6.1 and A6.3. Then, for any $i, j \in \{1, \ldots, d\}$, we have

$$\left|E\left[\widehat{H}_{ij}\middle|\theta\right] - \nabla^2_{i,j} f(\theta)\right| = O(\delta^2),$$

where $\widehat{H}_{ij}$ and $\nabla^2_{i,j} f(\cdot)$ denote the (i,j)th entry in the Hessian estimate $\widehat{H}$ and the true Hessian $\nabla^2 f(\cdot)$, respectively.

Proof. Using A6.2, we have

$$E\left[\widehat{H}_{ij}\middle|\theta\right] = E\Bigg[\left[\frac{f(\theta+\delta\Delta+\delta\hat{\Delta}) - f(\theta+\delta\Delta)}{2\delta\Delta_i\delta\hat{\Delta}_j}\right] - \left[\frac{f(\theta-\delta\Delta+\delta\hat{\Delta}) - f(\theta-\delta\Delta)}{2\delta\Delta_i\delta\hat{\Delta}_j}\right]\Bigg|\ \theta\Bigg]. \tag{6.9}$$

Since f satisfies A6.1, we employ Taylor series expansions to obtain

$$f(\theta \pm \delta\Delta + \delta\hat{\Delta}) = f(\theta \pm \delta\Delta) + \delta\sum_{k=1}^{d}\hat{\Delta}_k\nabla_k f(\theta \pm \delta\Delta) + \frac{1}{2}\delta^2\sum_{k=1}^{d}\sum_{l=1}^{d}\hat{\Delta}_k\nabla^2_{k,l} f(\theta \pm \delta\Delta)\hat{\Delta}_l + O(\delta^3).$$

Using (6.9) and the expansion above, we have

$$E\left[\widehat{H}_{ij}\middle|\theta\right] = E\Bigg[\frac{\nabla_i f(\theta+\delta\Delta) - \nabla_i f(\theta-\delta\Delta)}{2\delta\Delta_j} + \sum_{k\neq i}\frac{\hat{\Delta}_k}{\hat{\Delta}_i}\frac{\nabla_k f(\theta+\delta\Delta) - \nabla_k f(\theta-\delta\Delta)}{2\delta\Delta_j} + \delta\sum_{k=1}^{d}\sum_{l=1}^{d}\frac{\hat{\Delta}_k(\nabla^2_{k,l} f(\theta+\delta\Delta) - \nabla^2_{k,l} f(\theta-\delta\Delta))\hat{\Delta}_l}{4\delta\Delta_j\hat{\Delta}_i} + O(\delta^2)\ \Big|\ \theta\Bigg] \tag{6.10}$$

Expanding $\nabla_i f(\theta \pm \delta\Delta)$ around $\nabla_i f(\theta)$, we obtain

$$\frac{\nabla_i f(\theta+\delta\Delta) - \nabla_i f(\theta-\delta\Delta)}{2\delta\Delta_j} = \nabla^2_{i,j} f(\theta) + \sum_{l\neq j} \frac{\Delta_l}{\Delta_j} \nabla^2_{l,j} f(\theta) + O(\delta^3).$$

The second term on the RHS of (6.10) can be simplified in an analogous fashion.

The third term on the RHS of of (6.10) can be simplified as follows:

$$\begin{aligned}
&\delta \sum_{k=1}^{d}\sum_{l=1}^{d} \frac{\hat{\Delta}_k(\nabla^2_{k,l} f(\theta+\delta\Delta) - \nabla^2_{k,l} f(\theta-\delta\Delta))\hat{\Delta}_l}{4\delta\Delta_j\hat{\Delta}_i} \\
&= \delta \sum_{k=1}^{d}\sum_{l=1}^{d}\sum_{m=1}^{d} \frac{\hat{\Delta}_k \Delta(m) \nabla^3_{k,l,m} f(\theta)\hat{\Delta}_l}{2\hat{\Delta}_i\Delta_j} + O(\delta^2).
\end{aligned}$$

In the above, we used the following equality:

$$\frac{\nabla^2_{k,l} f(\theta+\delta\Delta) - \nabla^2_{k,l} f(\theta-\delta\Delta)}{4\delta\Delta_j} = \sum_{m=1}^{d} \frac{\Delta(m)\nabla^3_{k,l,m} f(\theta)}{2\Delta_j} + O(\delta^2)$$

Using the simplified forms for each of the terms on the RHS of (6.10), we have

$$\begin{aligned}
E\left[\widehat{H}_{ij}\middle|\,\theta\right] &= E\Bigg[\nabla^2_{i,j} f(\theta) + \sum_{l\neq i}\frac{\Delta_l}{\Delta_i}\nabla^2_{i,l} f(\theta) + \sum_{k\neq j}\frac{\hat{\Delta}_k}{\hat{\Delta}_j}\nabla^2_{k,i} f(\theta) \\
&\quad + \sum_{k\neq i}\sum_{l\neq i}\frac{\hat{\Delta}_k}{\hat{\Delta}_i}\frac{\Delta_l}{\Delta_j}\nabla^2_{k,l} f(\theta) + \delta\sum_{k,l,m=1}^{d}\frac{\hat{\Delta}_k\Delta(m)\nabla^3_{k,l,m} f(\theta)\hat{\Delta}_l}{2\hat{\Delta}_i\Delta_j} + O(\delta^2) \mid \theta\Bigg] \\
&= \nabla^2_{i,j} f(\theta) + \sum_{l\neq j} E\left[\frac{\Delta_l}{\Delta_j} \mid \theta\right]\nabla^2_{i,l} f(\theta) + \sum_{k\neq i} E\left[\frac{\hat{\Delta}_k}{\hat{\Delta}_i} \mid \theta\right]\nabla^2_{k,i} f(\theta) \\
&\quad + \sum_{k\neq i}\sum_{l\neq j} E\left[\frac{\hat{\Delta}_k}{\hat{\Delta}_i}\frac{\Delta_l}{\Delta_j}\,\middle|\,\theta\right]\nabla^2_{k,l} f(\theta) \\
&\qquad + \delta\sum_{k=1}^{d}\sum_{l=1}^{d}\sum_{m=1}^{d} E\left[\frac{\hat{\Delta}_k\hat{\Delta}_l\Delta(m)}{2\hat{\Delta}_j\Delta_i}\,\middle|\,\theta\right]\nabla^3_{k,l,m} f(\theta) + O(\delta^2).
\end{aligned}$$

Since $\Delta, \hat{\Delta}$ are independent vectors of zero mean, symmetric Bernoulli r.v.s, each term involving an expectation on the RHS above vanishes. The claim follows. □

6.3.2 Three measurements Hessian estimator

We now present a variation to 2SPSA, where the number of function measurements required for forming the Hessian estimate is brought down to three. This scheme was proposed by Bhatnagar and Prashanth, 2015, and can be motivated by using the following balanced approximation to the second derivative in the case of a scalar parameter:

$$\frac{d^2 f(\theta)}{d\theta^2} \approx \frac{\left(\frac{f(\theta+\delta)-f(\theta)}{\delta}\right) - \left(\frac{f(\theta)-f(\theta-\delta)}{\delta}\right)}{\delta} = \left(\frac{f(\theta+\delta)+f(\theta-\delta)-2f(\theta)}{\delta^2}\right). \tag{6.11}$$

The extension to a vector parameter is performed by using the following function measurements:

$$y^{++} = f(\theta+\delta\Delta+\delta\hat{\Delta})+\xi^{++}, y^{--} = f(\theta-\delta\Delta-\delta\hat{\Delta})+\xi^{--}, \text{ and } y = f(\theta)+\xi.$$

Using $y^{\pm}$ and y, together with two random perturbation vectors Δ and $\hat{\Delta}$ (as in the previous section), the Hessian estimate $\widehat{H}$ is formed as follows:

$$\widehat{H}_{ij} = \left(\frac{y^{++}+y^{--}-2y}{\delta^2 \Delta_i \hat{\Delta}_j}\right), \forall i,j. \tag{6.12}$$

Reusing the sample measurements, we form the gradient estimate as follows:

$$\widehat{\nabla} f(\theta) = \Delta^{-1} \frac{y^{++}-y^{--}}{2\delta}.$$

For the noise elements to vanish in the bias analysis of the Hessian estimator above, we make the following assumption.

A6.4. Let $\Delta = (\Delta_1, \ldots, \Delta_d)^T$ and $\hat{\Delta} = (\hat{\Delta}_1, \ldots, \hat{\Delta}_d)^{\mathrm{T}}$ be two independent d-vectors of mutually independent, symmetric, ± 1-valued Bernoulli r.v.s. Further, given θ, $\{\xi, \xi^{++}, \xi^{--}\}$ is independent of $\Delta, \hat{\Delta}$. In addition,
$\mathbb{E}\left[\xi^{++} \middle| \theta\right] = \mathbb{E}\left[\xi^{--} \middle| \theta\right] = \mathbb{E}\left[\xi \middle| \theta\right] = 0.$

Lemma 6.2. Assume A6.1 and A6.4. Then, for any $i, j \in \{1, \ldots, d\}$, we have

$$\left| E\left[\widehat{H}_{ij} \mid \theta\right] - \nabla^2_{i,j} f(\theta) \right| = O(\delta^2) \text{ a.s.}$$

Proof. We first consider the case when $i, j \in \{1, \ldots, d\}$, $i \neq j$. Let

$$\hat{f}(\theta, \Delta, \hat{\Delta}) = f(\theta + \delta\Delta + \delta\hat{\Delta}) + f(\theta - \delta\Delta - \delta\hat{\Delta}) - 2f(\theta).$$

Then, using suitable Taylor's expansions, we obtain

$$\begin{aligned}
\frac{\hat{f}(\theta, \Delta, \hat{\Delta})}{\delta^2 \Delta_i \hat{\Delta}_j} &= \frac{(\Delta + \hat{\Delta})^T \nabla^2 f(\theta)(\Delta + \hat{\Delta})}{\Delta_i \hat{\Delta}_j} + O(\delta^2) \\
&= \sum_{l=1}^{d} \sum_{m=1}^{d} \frac{\Delta_l \nabla^2_{lm} f(\theta) \Delta_m}{\Delta_i \hat{\Delta}_j} + 2 \sum_{l=1}^{d} \sum_{m=1}^{d} \frac{\Delta_l \nabla^2_{lm} f(\theta) \hat{\Delta}_m}{\Delta_i \hat{\Delta}_j} \\
&\quad + \sum_{l=1}^{d} \sum_{m=1}^{d} \frac{\hat{\Delta}_l \nabla^2_{lm} f(\theta) \hat{\Delta}_m}{\Delta_i \hat{\Delta}_j} + O(\delta^2).
\end{aligned}$$

It is now easy to see that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{l=1}^{d} \sum_{m=1}^{d} \frac{\Delta_l \nabla^2_{lm} f(\theta) \Delta_m}{\Delta_i \hat{\Delta}_j} \middle| \theta \right] \\
&= \mathbb{E}\left[\sum_{l=1}^{d} \sum_{m=1}^{d} \frac{\hat{\Delta}_l \nabla^2_{lm} f(\theta) \hat{\Delta}_m}{\Delta_i \hat{\Delta}_j} \middle| \theta \right] = 0 \text{ a.s., and} \\
&\mathbb{E}\left[\sum_{l=1}^{d} \sum_{m=1}^{d} \frac{\Delta_l \nabla^2_{lm} f(\theta) \hat{\Delta}_m}{\Delta_i \hat{\Delta}_j} \mid \theta \right] = \nabla^2_{i,j} f(\theta) \text{ a.s.}
\end{aligned}$$

Thus,

$$\mathbb{E}\left[\frac{\hat{f}(\theta, \Delta, \hat{\Delta})}{\delta^2 \Delta_i \hat{\Delta}_j} \middle| \theta \right] = 2\nabla^2_{i,j} f(\theta) + O(\delta^2).$$

The case when $i = j \in \{1, \ldots, d\}$ follows in a similar manner. The claim follows after observing that

$$\mathbb{E}\left[\widehat{H}_{ij} \middle| \theta \right] = \mathbb{E}\left[\frac{\hat{f}(\theta, \Delta, \hat{\Delta})}{\delta^2 \Delta_i \hat{\Delta}_j} \middle| \theta \right].$$

The equality above holds since the noise elements ξ^{++}, ξ^{--}, ξ satisfy A6.4. □

6.4 Gaussian smoothed functional for Hessian estimation

We now present a couple of Hessian estimation procedures from (Bhatnagar, 2007) that are based on Gaussian smoothing.

6.4.1 One-Measurement SF (1SF) Estimator

We begin with a one-measurement Hessian estimator $D^2_{\delta,1}f(\theta)$ that uses one function measurement with the same perturbed parameter as the one-measurement gradient SF procedure. We shall later provide a two-sided Hessian estimator as well that estimates both the Hessian and the gradient using two function measurements. We shall continue to assume A6.1.

As with gradient SF, we begin by taking a convolution of the objective function Hessian with a multi-variate Gaussian density functional. Through an integration-by-parts argument applied twice, the same is seen to be a convolution of the objective function with a scaled Gaussian density functional. Let

$$D^2_{\delta,1}f(\theta) = \int G_\delta(\theta - \Delta')\nabla^2_{\Delta'}f(\Delta')d\Delta', \tag{6.13}$$

denote the convolution of the Hessian $\nabla^2_{\Delta'}f(\Delta')$ with the d-dimensional multivariate normal p.d.f.

$$G_\delta(\theta - \Delta') = \frac{1}{(2\pi)^{d/2}\delta^d}\exp\left(-\frac{1}{2}\sum_{i=1}^{d}\frac{(\theta_i - \Delta'_i)^2}{\delta^2}\right),$$

where $\theta, \Delta' \in \mathcal{R}^d$.

A6.5. Let $\Delta = (\Delta_1, \ldots, \Delta_d)^T$ where $\Delta_i \sim N(0,1)$, $i = 1, \ldots, d$ and with Δ_i independent of Δ_j, $\forall i \neq j$. Further, given θ, ξ^+ is independent of Δ. Further, $\mathbb{E}\left[\xi^+ \middle| \theta\right] = 0$.

Let $y^+ = f(\theta + \delta\Delta) + \xi^+$ denote a noisy function measurement, where ξ^+ denotes the measurement noise. The one-simulation (1SF) Hessian estimator is then the following:

$$\hat{H}(\theta) = \frac{(\Delta\Delta^{\intercal} - I)}{\delta^2} y^{+}. \tag{6.14}$$

The reason for having this form for the Hessian estimator will become evident in Proposition 6.1.

As with earlier estimates, we can reuse the function measurement y^+ to form a gradient estimate as follows:

$$\widehat{\nabla} f(\theta) = \Delta \left(\frac{y^{+}}{\delta} \right).$$

Proposition 6.1 (Stein's Lemma for Hessian Estimation).

$$D^2_{\delta,1} f(\theta) = \frac{1}{\delta^2} E\left[(\Delta\Delta^T - I) f(\theta + \delta\Delta) \right],$$

where the expectation above is taken w.r.t. the d-dimensional multivariate normal p.d.f. $G(\Delta)$ corresponding to the random vector of d independent $N(0,1)$–distributed random variables.

Proof. Upon integrating by parts, one obtains

$$D^2_{\delta,1} f(\theta) = \int \nabla_\theta G_\delta(\theta - \Delta') \nabla_{\Delta'} f(\Delta') d\Delta'. \tag{6.15}$$

Now

$$\nabla_\theta G_\delta(\theta - \Delta') = -\frac{(\theta - \Delta')}{\delta^2} G_\delta(\theta - \Delta').$$

Upon substituting the above in (6.15) and performing integration-by-parts, we obtain

$$D^2_{\delta,1} f(\theta) = -\frac{1}{\delta^2} \int \nabla_\theta ((\theta - \Delta') G_\delta(\theta - \Delta')) f(\Delta') d\Delta'.$$

A change of variables then gives

$$D^2_{\delta,1} f(\theta) = -\frac{1}{\delta^2} \int \nabla_{\Delta'} (\Delta' G_\delta(\Delta')) f(\theta - \Delta') d\Delta'. \tag{6.16}$$

We now evaluate $\nabla_{\Delta'}(\Delta' G_\delta(\Delta')) = \nabla_{\Delta'}((\Delta'_1 G_\delta(\Delta'), \ldots, \Delta'_N G_\delta(\Delta'))$. Note that

$$\nabla_{\Delta'}(\Delta' G_\delta(\Delta')) =$$

$$\begin{bmatrix} \nabla_{\Delta'_1}(\Delta'_1 G_\delta(\Delta')) & \nabla_{\Delta'_2}(\Delta'_1 G_\delta(\Delta')) & \cdots & \nabla_{\Delta'_d}(\Delta'_1 G_\delta(\Delta')) \\ \nabla_{\Delta'_1}(\Delta'_2 G_\delta(\Delta')) & \nabla_{\Delta'_2}(\Delta'_2 G_\delta(\Delta')) & \cdots & \nabla_{\Delta'_d}(\Delta'_2 G_\delta(\Delta')) \\ \cdots & \cdots & \cdots & \cdots \\ \nabla_{\Delta'_1}(\Delta'_d G_\delta(\Delta')) & \nabla_{\Delta'_2}(\Delta'_d G_\delta(\Delta')) & \cdots & \nabla_{\Delta'_d}(\Delta'_d G_\delta(\Delta')) \end{bmatrix}$$

$$= \begin{bmatrix} \left(1 - \frac{\Delta'^2_1}{\delta^2}\right) & -\frac{\Delta'_1 \Delta'_2}{\delta^2} & \cdots & -\frac{\Delta'_1 \Delta'_d}{\delta^2} \\ -\frac{\Delta'_2 \Delta'_1}{\delta^2} & \left(1 - \frac{\Delta'^2_2}{\delta^2}\right) & \cdots & -\frac{\Delta'_2 \Delta'_d}{\delta^2} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{\Delta'_d \Delta'_1}{\delta^2} & -\frac{\Delta'_d \Delta'_2}{\delta^2} & \cdots & \left(1 - \frac{\Delta'^2_d}{\delta^2}\right) \end{bmatrix} G_\delta(\Delta')$$

$$= \left(I - \frac{\Delta' \Delta'^T}{\delta^2}\right) G_\delta(\Delta') \stackrel{\triangle}{=} \check{H}(\Delta') G_\delta(\Delta').$$

From (6.16), we have

$$D^2_{\delta,1} f(\theta) = -\frac{1}{\delta^2} \int \check{H}(\Delta') G_\delta(\Delta') f(\theta - \Delta') d\Delta'.$$

Let $\Delta \stackrel{\triangle}{=} \Delta'/\delta$. Then $d\Delta' = \delta^d d\Delta$. From (6.16), we then obtain

$$D^2_{\delta,1} f(\theta) = \frac{1}{\delta^2} \int \bar{I}(\Delta) \left(\frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2} \sum_{i=1}^{d} (\Delta_i)^2) \right) f(\theta - \delta\Delta) d\Delta, \tag{6.17}$$

where

$$\bar{I}(\Delta) \stackrel{\triangle}{=} (\Delta\Delta^T - I). \tag{6.18}$$

Note that Δ_i, $i = 1, \ldots, d$ are independent $N(0,1)$ distributed random variables. Now since Δ and $-\Delta$ have the same distribution, one obtains

$$D^2_{\delta,1} f(\theta) = \frac{1}{\delta^2} E\left[(\Delta\Delta^T - I) f(\theta + \delta\Delta) \right].$$

The claim follows. □

Proposition 6.2. Under Assumptions A6.1 and A6.5, we have that

$$\| E[\hat{H}(\theta)|\theta] - \nabla^2 f(\theta) \| \leq O(\delta).$$

Proof. From the definition of $\hat{H}(\theta)$,

$$E[\hat{H}(\theta)|\theta)] = \frac{1}{\delta^2} E[\bar{I}(\Delta)(f(\theta + \delta\Delta) + \xi^+)|\theta]$$

$$= D^2_{\delta,1} f(\theta) + \frac{1}{\delta^2} E[(I - \Delta\Delta^\intercal)\xi^+|\theta].$$

The second term on the RHS equals zero in the light of Assumption A6.5. Now, from Proposition 6.1, we have that

$$D^2_{\delta,1} f(\theta) = \mathbb{E}\left[\frac{1}{\delta^2}\bar{I}(\Delta) f(\theta + \delta\Delta) \mid \theta\right],$$

where $\Delta = (\Delta_1, \ldots, \Delta_d)^T$ is a vector of independent $N(0,1)$ random variates and the expectation is taken w.r.t. the density of Δ. Using a Taylor series expansion of $f(\theta + \delta\Delta)$ around θ, one obtains

$$\begin{aligned} D^2_{\delta,1} f(\theta) &= E\Big[\frac{1}{\delta^2}\bar{I}(\Delta)(f(\theta) + \delta\Delta^\intercal \nabla f(\theta) \\ &\qquad + \frac{\delta^2}{2}\Delta^\intercal \nabla^2 f(\theta)\Delta + o(\delta^2) \mid \theta\Big] \\ &= \frac{1}{\delta^2} E[\bar{I}(\Delta) f(\theta) \mid \theta] + \frac{1}{\delta} E[\bar{I}(\Delta)\Delta^\intercal \nabla f(\theta) \mid \theta] \\ &\qquad + \frac{1}{2} E[\bar{I}(\Delta)\Delta^\intercal \nabla^2 f(\theta)\Delta \mid \theta] + O(\delta). \end{aligned} \tag{6.19}$$

Now observe that $E[\bar{I}(\Delta)] = 0$ (the matrix of all zero elements) with $E[\bar{H}(\Delta)]$. Hence the first term on the RHS of (6.19) equals zero. Now consider the second term on the RHS of (6.19). Note that

$$E[\bar{I}(\Delta)\Delta^\intercal \nabla f(\theta) \mid \theta] = \mathbb{E}\left[\begin{array}{cccc} (\Delta_1^2 - 1)\Delta^\intercal \nabla f(\theta) & \Delta_1\Delta_2\Delta^\intercal \nabla f(\theta) & \cdots & \Delta_1\Delta_d\Delta^\intercal \nabla f(\theta) \\ \Delta_2\Delta_1\Delta^\intercal \nabla f(\theta) & (\Delta_2^2 - 1)\Delta^\intercal \nabla f(\theta) & \cdots & \Delta_2\Delta_d\Delta^\intercal \nabla f(\theta) \\ \cdots & \cdots & \cdots & \cdots \\ \Delta_d\Delta_1\Delta^\intercal \nabla f(\theta) & \Delta_d\Delta_2\Delta^\intercal \nabla f(\theta) & \cdots & (\Delta_d^2 - 1)\Delta^\intercal \nabla f(\theta) \end{array} \mid \theta\right]. \tag{6.20}$$

One can verify that expectation of each term (conditioned on θ) within the matrix above equals zero since $E[\Delta_i] = \mathbb{E}[\Delta_i^3] = 0$ and $\mathbb{E}[\Delta_i^2] = 1$, $\forall i = 1, \ldots, d$. Also, Δ_i is independent of Δ_j for all $i \neq j$. Hence the second term on the RHS of (6.19) equals zero as well. Consider now the third term on the RHS of (6.19). Note that

$$\frac{1}{2}\mathbb{E}[\bar{H}(\Delta)\Delta^{\mathsf{T}}\nabla^2 f(\theta)\Delta \mid \theta] =$$

$$\frac{1}{2}\mathbb{E}\left[\begin{array}{ccc} (\Delta_1^2-1)\sum_{i,j=1}^{d}\nabla_{ij}f(\theta)\Delta_i\Delta_j & \cdots & \Delta_1\Delta_d\sum_{i,j=1}^{d}\nabla_{ij}f(\theta)\Delta_i\Delta_j \\ \Delta_2\Delta_1\sum_{i,j=1}^{d}\nabla_{ij}f(\theta)\Delta_i\Delta_j & \cdots & \Delta_2\Delta_d\sum_{i,j=1}^{d}\nabla_{ij}f(\theta)\Delta_i\Delta_j \\ \cdots & \cdots & \cdots \\ \Delta_d\Delta_1\sum_{i,j=1}^{d}\nabla_{ij}f(\theta)\Delta_i\Delta_j & \cdots & (\Delta_d^2-1)\sum_{i,j=1}^{d}\nabla_{ij}f(\theta)\Delta_i\Delta_j \end{array} \mid \theta\right]. \quad (6.21)$$

Consider now the term corresponding to the first row and first column above. Note that

$$\begin{aligned} &\mathbb{E}[(\Delta_1^2-1)\sum_{i,j=1}^{d}\nabla_{ij}f(\theta)\Delta_i\Delta_j \mid \theta] \\ &= \mathbb{E}[\Delta_1^2\sum_{i,j=1}^{d}\nabla_{ij}f(\theta)\Delta_i\Delta_j \mid \theta] - \mathbb{E}[\sum_{i,j=1}^{d}\nabla_{ij}f(\theta)\Delta_i\Delta_j \mid \theta]. \end{aligned} \quad (6.22)$$

The first term on the RHS of (6.22) equals

$$\begin{aligned} &\mathbb{E}[\Delta_1^4\nabla_{11}f(\theta) \mid \theta] + \mathbb{E}[\sum_{i=j,i\neq 1}\Delta_1^2\Delta_i^2\nabla_{ij}f(\theta) \mid \theta] \\ &+\mathbb{E}[\sum_{i\neq j,i\neq 1}\Delta_1^2\Delta_i\Delta_j\nabla_{ij}f(\theta) \mid \theta] = 3\nabla_{11}f(\theta) + \sum_{i=j,i\neq 1}\nabla_{ij}f(\theta), \end{aligned}$$

since $\mathbb{E}[\Delta_1^4] = 3$. The second term on RHS of (6.22) equals $-\sum_{i=1}^{d}\nabla_{ii}f(\theta)$. Adding the above two terms, one obtains

$$\mathbb{E}[(\Delta_1^2-1)\sum_{i,j=1}^{d}\nabla_{ij}f(\theta)\Delta_i\Delta_j \mid \theta] = 2\nabla_{11}f(\theta).$$

Consider now the term in the first row and second column of the matrix in (6.21). Note that

$$\begin{aligned}
&\mathbb{E}[\Delta_1\Delta_2 \sum_{i,j=1}^{d} \nabla_{ij} f(\theta)\Delta_i\Delta_j \mid \theta] \\
&= 2\mathbb{E}[\Delta_1^2\Delta_2^2\nabla_{12} f(\theta) \mid \theta] + \mathbb{E}[\sum_{(i,j)\notin\{(1,2),(2,1)\}} \Delta_1\Delta_2\Delta_i\Delta_j\nabla_{ij} f(\theta) \mid \theta] \\
&= 2\nabla_{12} f(\theta).
\end{aligned}$$

Proceeding in a similar manner, it is easy to verify that the (i,j)th term $(i,j \in \{1,\ldots,d\})$ in the matrix in (6.21) equals $2\nabla_{ij} f(\theta)$. Substituting the above back in (6.21), one obtains

$$\frac{1}{2}\mathbb{E}[\bar{I}(\Delta)\Delta^{\mathsf{T}}\nabla^2 f(\theta)\Delta] = \nabla^2 f(\theta).$$

Thus, (6.19) now becomes

$$D^2_{\delta,1} f(\theta) = \nabla^2 f(\theta) + O(\delta).$$

The claim follows. □

6.4.2 Two-measurement SF (2SF) estimator

We now present the balanced form of the Hessian estimator from Bhatnagar, 2007 that requires only two function measurements. Let

$$D^2_{\delta,2} f(\theta) = E\left[\frac{1}{2\delta^2}\bar{I}(\Delta)(f(\theta+\delta\Delta) + f(\theta-\delta\Delta)) \mid \theta\right],$$

with $\bar{I}(\Delta)$ as in (6.18). We now present the balanced form of the Hessian estimator based on two function measurements. Let $y^+ = f(\theta+\delta\Delta)+\xi^+$ and $y^- = f(\theta+\delta\Delta)+\xi^-$, respectively, where ξ^+ and ξ^- denote the measurement noise in y^+ and y^-. The 2SF Hessian estimator is then the following:

$$\hat{H}(\theta) = \frac{(\Delta\Delta^{\mathsf{T}} - I)}{2\delta^2}(y^+ + y^-). \tag{6.23}$$

A gradient estimate re-using the function measurements $y^{\pm}$ is given by

$$\widehat{\nabla} f(\theta) = \Delta\left(\frac{y^+ - y^-}{2\delta}.\right)$$

This gradient estimate's bias was analyzed earlier in Chapter 3. We analyze the bias of the Hessian estimator in (6.23). For this analysis, we shall continue to assume A6.1. In addition, we have the following assumption on the measurement noise:

A6.6. Let $\Delta = (\Delta_1, \ldots, \Delta_d)^T$ where $\Delta_i \sim N(0,1)$, $i = 1, \ldots, d$ and with Δ_i independent of Δ_j, $\forall i \neq j$. Further, given θ, ξ^+ and ξ^- are independent of Δ and they are also independent of each other. Further, $\mathbb{E}\left[\xi^+ \middle| \theta\right] = \mathbb{E}\left[\xi^- \middle| \theta\right] = 0$.

Proposition 6.3. Under Assumptions A6.1 and A6.6, we have that

$$\left\| E[\hat{H}(\theta)|\theta] - \nabla^2 f(\theta) \right\| \leq O(\delta^2).$$

Proof. From (6.23), note that

$$E[\hat{H}(\theta)|\theta)] = \frac{1}{2\delta^2} E[\bar{I}(\Delta)((f(\theta+\delta\Delta)+\xi^+) + (f(\theta-\delta\Delta)+\xi^-)|\theta]$$

$$= D^2_{\delta,2} f(\theta) + \frac{1}{2\delta^2} E[(I - \Delta\Delta^\intercal)(\xi^+ + \xi^-)|\theta].$$

The second term on the RHS equals zero in the light of Assumption A6.6.

We now consider the first term on the RHS above. Using Taylor series expansions of $f(\theta+\delta\Delta)$ and $f(\theta-\delta\Delta)$ around θ, one obtains

$$f(\theta+\delta\Delta) = f(\theta) + \delta\Delta^\intercal \nabla f(\theta) + \frac{\delta^2}{2}\Delta^\intercal \nabla^2 f(\theta)\Delta + \frac{\delta^3}{6}\nabla^3 f(\theta)(\Delta\otimes\Delta\otimes\Delta) + O(\delta^4)$$

$$f(\theta-\delta\Delta) = f(\theta) - \delta\Delta^\intercal \nabla f(\theta) + \frac{\delta^2}{2}\Delta^\intercal \nabla^2 f(\theta)\Delta - \frac{\delta^3}{6}\nabla^3 f(\theta)(\Delta\otimes\Delta\otimes\Delta) + O(\delta^4).$$

From the foregoing, one obtains

$$D^2_{\delta,2} f(\theta) = E\left[\frac{1}{2\delta^2}\bar{I}(\Delta)\left(2f(\theta) + \delta^2\Delta^\intercal\nabla^2 f(\theta)\Delta + O(\delta^4)\right) \mid \theta\right].$$

It has been shown in the proof of Proposition 6.2 that $\mathbb{E}[\bar{I}(\Delta) f(\theta) \mid \theta] = 0$ and $\frac{1}{2}\mathbb{E}[\bar{I}(\Delta)\Delta^\intercal\nabla^2 f(\theta)\Delta \mid \theta] = \nabla^2 J(\theta)$, respectively. We thus have

$$D^2_{\delta,2} f(\theta) = \nabla^2 f(\theta) + O(\delta^2).$$

The claim follows. □

6.5 RDSA for Hessian estimation

In this section, the random perturbations are chosen using an asymmetric Bernoulli distribution. More precisely, we choose Δ_i, $i = 1, \ldots, d$, i.i.d. as follows:

$$\Delta_i = \begin{cases} -1 & \text{w.p. } \dfrac{(1+\epsilon)}{(2+\epsilon)}, \\ 1+\epsilon & \text{w.p. } \dfrac{1}{(2+\epsilon)}, \end{cases} \tag{6.24}$$

where $\epsilon > 0$ is a constant that can be chosen to be arbitrarily small. Note that, for any $i = 1, \ldots, d$, $\mathbb{E}\Delta_i = 0$, $\mathbb{E}(\Delta_i)^2 = 1 + \epsilon$ and $\mathbb{E}(\Delta_i)^4 = \dfrac{(1+\epsilon)(1+(1+\epsilon)^3)}{(2+\epsilon)}$. Henceforth, we will use τ to denote $E(\Delta_i)^4$.

Suppose we have the following function measurements:

$$y^+ = f(\theta + \delta\Delta) + \xi^+, y^- = f(\theta - \delta\Delta) + \xi^-, \text{ and } y = f(\theta) + \xi.$$

We would like to obtain a Hessian estimate $\widehat{H}$ that is not too far from the true Hessian $\nabla^2 f(\theta)$. Suppose we use the three measurements above, together with a matrix M (to be specified later) to form $\widehat{H}$ as follows:

$$\widehat{H} = M\left(\frac{y^+ + y^- - 2y}{\delta^2}\right) \tag{6.25}$$

$$= M\left[\left(\frac{f(\theta+\delta\Delta) + f(\theta-\delta\Delta) - 2f(\theta)}{\delta^2}\right) + \left(\frac{\xi^+ + \xi^- - 2\xi}{\delta^2}\right)\right]$$

$$= M\left(\Delta^\intercal \nabla^2 f(\theta)\Delta + O(\delta^2) + \left(\frac{\xi^+ + \xi^- - 2\xi}{\delta^2}\right)\right). \tag{6.26}$$

We form a gradient estimate using $y^\pm$ as follows:

$$\widehat{\nabla} f(\theta) = \frac{1}{(1+\epsilon)}\Delta\left(\frac{y^+ - y^-}{2\delta}\right).$$

The bias of this estimator is analyzed in Chapter 3.

We now deconstruct the Hessian estimate in (6.26). Taking expectations on both sides of (6.26), we observe that the last term in (6.26)

vanishes, while the first and second term remain. However, we do not have the true Hessian in the first term and it would be nice to recover $\nabla^2 f(\theta)$ from this term via a suitable matrix M and the following definition for M achieves this goal:

$$M = \begin{bmatrix} \frac{1}{\kappa}\left((\Delta_1)^2 - (1+\epsilon)\right) & \cdots & \frac{1}{2(1+\epsilon)^2}\Delta_1\Delta_d \\ \frac{1}{2(1+\epsilon)^2}\Delta_2\Delta_1 & \cdots & \frac{1}{2(1+\epsilon)^2}\Delta_2\Delta_d \\ \cdots & \cdots & \cdots \\ \frac{1}{2(1+\epsilon)^2}\Delta_d\Delta_1 & \cdots & \frac{1}{\kappa}\left((\Delta_d)^2 - (1+\epsilon)\right) \end{bmatrix}, \quad (6.27)$$

where $\kappa = \tau\left(1 - \frac{(1+\epsilon)^2}{\tau}\right)$ and $\tau = E(\Delta_i)^4 = \frac{(1+\epsilon)(1+(1+\epsilon)^3)}{(2+\epsilon)}$, for any $i = 1, \ldots, d$.

While the definition of M above looks complicated, the motivation behind such a definition can be seen through the following calculation that established that the first term, i.e., $M\left(\Delta^\intercal \nabla^2 f(\theta)\Delta\right)$ in (6.26) turns out to be the true Hessian evaluated at θ.

As before, we make the following assumption to ensure noise elements vanish in the analysis of the RDSA Hessian estimator (6.25).

A6.7. Let $\Delta = (\Delta_1, \ldots, \Delta_d)^T$ be a d-vector of mutually independent, asymmetric, Bernoulli r.v.s satisfying (6.24). Further, given θ, $\{\xi, \xi^+, \xi^-\}$ is independent of Δ. In addition, $\mathbb{E}\left[\xi^+ \middle| \theta\right] = \mathbb{E}\left[\xi^- \middle| \theta\right] = \mathbb{E}\left[\xi \middle| \theta\right] = 0$.

Lemma 6.3. (**Bias in Hessian estimate**) Assume A6.1 and A6.7. Then, $\widehat{H}$ defined according to (6.27) satisfies the following bound for any $i, j = 1, \ldots, d$,

$$\left|\mathbb{E}\left[\widehat{H}(i,j)\middle| \theta\right] - \nabla^2_{ij} f(\theta)\right| = O(\delta^2). \quad (6.28)$$

From the above lemma, it is evident that the bias in the Hessian estimate above is of the same order as that of the other balanced estimators in the previous sections.

Proof. By a Taylor's expansion, we obtain

$$f(\theta \pm \delta\Delta) = f(\theta) \pm \delta\Delta^\intercal \nabla f(\theta) + \frac{\delta^2}{2}\Delta^\intercal \nabla^2 f(\theta)\Delta$$
$$\pm \frac{\delta^3}{6}\nabla^3 f(\theta)(\Delta \otimes \Delta \otimes \Delta) + \frac{\delta^4}{24}\nabla^4 f(\tilde{\theta}^+)(\Delta \otimes \Delta \otimes \Delta \otimes \Delta).$$

Hence,

$$\begin{aligned}
&\frac{f(\theta + \delta\Delta) + f(\theta - \delta\Delta) - 2f(\theta)}{\delta^2} \\
=&\Delta^\intercal \nabla^2 f(\theta)\Delta + O(\delta^2) \\
=&\sum_{i=1}^{d}\sum_{j=1}^{d} \Delta_i \Delta_j \nabla^2_{ij} f(\theta) + O(\delta^2) \\
=&\sum_{i=1}^{d} (\Delta_i)^2 \nabla^2_{ii} f(\theta) + 2\sum_{i=1}^{d-1}\sum_{j=i+1}^{d} \Delta_i \Delta_j \nabla^2_{ij} f(\theta) + O(\delta^2).
\end{aligned}$$

Now, taking the conditional expectation of the Hessian estimate $\widehat{H}$ and observing that $\mathbb{E}[\xi^+ + \xi^- - 2\xi \mid \theta] = 0$ by A6.7, we obtain the following:

$$\begin{aligned}
\mathbb{E}[\widehat{H} \mid \theta] =&\mathbb{E}\left[M \left(\sum_{i=1}^{d-1} (\Delta_i)^2 \nabla^2_{ii} f(\theta) \right.\right. \\
&\left.\left. +2\sum_{i=1}^{d}\sum_{j=i+1}^{d} \Delta_i \Delta_j \nabla^2_{ij} f(\theta) + O(\delta^2) \right) \middle| \theta \right]. \qquad (6.29)
\end{aligned}$$

Note that the $O(\delta^2)$ term inside the conditional expectation above remains $O(\delta^2)$ even after the multiplication with M. We analyse the diagonal and off-diagonal terms in the multiplication of the matrix M with the scalar above, ignoring the $O(\delta^2)$ term.

Diagonal terms in (6.29):

Recall that τ denotes the fourth moment $E(\Delta_i)^4$, for any $i = 1, \ldots, d$. Consider the lth diagonal term inside the conditional expectation in (6.29):

$$\frac{1}{\tau(1 - \frac{(1+\epsilon)^2}{\tau})} \mathbb{E}\left(\left((\Delta_l)^2 - (1+\epsilon) \right) \left(\sum_{i=1}^{d} (\Delta_i)^2 \nabla^2_{ii} f(\theta) \right.\right.$$

$$+2\sum_{i=1}^{d-1}\sum_{j=i+1}^{d}\Delta_i\Delta_j\nabla^2_{ij}f(\theta)\Bigg)\Bigg|\,\theta\Bigg)$$

$$=\frac{1}{\tau(1-\frac{(1+\epsilon)^2}{\tau})}\mathbb{E}\left((\Delta_l)^2\sum_{i=1}^{d}(\Delta_i)^2\nabla^2_{ii}f(\theta)\,\middle|\,\theta\right)$$

$$-\frac{(1+\epsilon)}{\tau(1-\frac{(1+\epsilon)^2}{\tau})}\mathbb{E}\left(\sum_{i=1}^{d}(\Delta_i)^2\nabla^2_{ii}f(\theta)\,\middle|\,\theta\right) \tag{6.30}$$

From the distributions of Δ_i, Δ_j and the fact that Δ_i is independent of Δ_j for $i<j$, it is easy to see that $\mathbb{E}\left((\Delta^l_n)^2\sum_{i=1}^{d-1}\sum_{j=i+1}^{d}\Delta_i\Delta_j\nabla^2_{ij}f(\theta)\,\middle|\,\theta\right)=0$ and $\mathbb{E}\left(\sum_{i=1}^{d-1}\sum_{j=i+1}^{d}\Delta_i\Delta_j\nabla^2_{ij}f(\theta)\,\middle|\,\theta\right)=0$. Thus, the conditional expectations of the second and fourth terms on the RHS of (6.30) are both zero.

The first term on the RHS of (6.30) can be simplified as follows:

$$\frac{1}{\tau(1-\frac{(1+\epsilon)^2}{\tau})}\mathbb{E}\left((\Delta_l)^2\sum_{i=1}^{d}(\Delta_i)^2\nabla^2_{ii}f(\theta)\,\middle|\,\theta\right)$$

$$=\frac{1}{\tau(1-\frac{(1+\epsilon)^2}{\tau})}\mathbb{E}\Bigg((\Delta_l)^4\nabla^2_{ll}f(\theta)+\sum_{i=1,i\neq l}^{d}(\Delta_l)^2(\Delta_i)^2\nabla^2_{ii}f(\theta)\Bigg)$$

$$=\frac{1}{(1-\frac{(1+\epsilon)^2}{\tau})}\left(\nabla^2_{ll}f(\theta)+\frac{(1+\epsilon)^2}{\tau}\sum_{i=1,i\neq l}^{d}\nabla^2_{ii}f(\theta)\right). \tag{6.31}$$

For the second equality above, we have used the fact that $\mathbb{E}[(\Delta_l)^4]=\tau$ and $\mathbb{E}[(\Delta_l)^2(\Delta_i)^2]=\mathbb{E}[(\Delta_l)^2]\mathbb{E}[(\Delta_i)^2]=(1+\epsilon)^2$, $\forall l\neq i$.

The second term in (6.30) with the conditional expectation and without the negative sign can be simplified as follows:

$$\frac{(1+\epsilon)}{\tau(1-\frac{(1+\epsilon)^2}{\tau})}\mathbb{E}\left(\sum_{i=1}^{d}(\Delta_i)^2\nabla^2_{ii}f(\theta)\,\middle|\,\theta\right)$$

$$=\frac{(1+\epsilon)}{\tau(1-\frac{(1+\epsilon)^2}{\tau})}\sum_{i=1}^{d}\mathbb{E}\left[(\Delta_i)^2\right]\nabla^2_{ii}f(\theta)$$

$$= \frac{(1+\epsilon)^2}{\tau(1-\frac{(1+\epsilon)^2}{\tau})} \sum_{i=1}^{d} \nabla_{ii}^2 f(\theta). \quad (6.32)$$

Combining (6.31) and (6.32), the correctness of the Hessian estimate follows for the diagonal terms.

Off-diagonal terms in (6.29)

Consider the (k,l)th term in (6.29), with $k < l$. We obtain

$$\begin{aligned}
&\frac{1}{2(1+\epsilon)^2}\mathbb{E}\left[\Delta_k\Delta_l\left(\sum_{i=1}^{d}(\Delta_i)^2\nabla_{ii}^2 f(\theta) + 2\sum_{i=1}^{d-1}\sum_{j=i+1}^{d}\Delta_i\Delta_j\nabla_{ij}^2 f(\theta)\right)\middle|\,\theta\right] \\
=&\frac{1}{2(1+\epsilon)^2}\sum_{i=1}^{d}\mathbb{E}\left(\Delta_k\Delta_l(\Delta_i)^2\right)\nabla_{ii}^2 f(\theta) \\
&+\frac{1}{(1+\epsilon)^2}\sum_{i=1}^{d-1}\sum_{j=i+1}^{d}\mathbb{E}\left(\Delta_k\Delta_l\Delta_i\Delta_j\right)\nabla_{ij}^2 f(\theta) \quad (6.33) \\
=&\nabla_{kl}^2 f(\theta).
\end{aligned}$$

Note that the first term on the RHS of (6.33) equals zero since $k \neq l$. The claim follows. □

6.6 Summary

Tables 6.1 and 6.2 summarize the various gradient and Hessian estimates discussed in this chapter.

6.7 Asymptotic convergence of stochastic Newton algorithms

We consider the following coupled sequence of updates for the analysis of the Hessian recursion:

$$\theta_{n+1} = \Gamma\left(\theta_n - a(n)\Theta\left(\overline{H}_n\right)^{-1}\widehat{\nabla} f(\theta_n)\right), \quad (6.34)$$

$$\overline{H}_{n+1} = \overline{H}_n + b(n)(\hat{H}_n - \overline{H}_n), \quad (6.35)$$

Property → **Hessian estimate** ↓	**# measurements**	**Bias**
FDSA (6.7)	$4d^2$	$O(\delta^2)$
SPSA (6.8) and its variant (6.12)	4 3	$O(\delta^2)$
SF (6.14) and its variant (6.23)	1 2	$O(\delta)$ $O(\delta^2)$
RDSA (6.25)	3	$O(\delta^2)$

Table 6.1: A summary of the Hessian estimates presented in this chapter, along with their bias bounds.

where the quantity $\hat{H}_n$ in (6.35) can correspond to any of the simultaneous perturbation Hessian estimators described in Chapter 6. Also, $\widehat{\nabla} f(\theta_n)$ could be any of the simultaneous perturbation gradient estimators described in Chapter 3. As can be seen, it makes better sense from a computational perspective to have a similar class of estimators for both gradient and Hessian estimation. Thus, for instance, if one incorporates two-measurement SF for gradient estimation, the same two measurements can then also be used as in (6.23) for estimating the objective function Hessian. We consider here the case of a diminishing $\{\delta_n\}$, i.e., $0 < \delta_n \to 0$ as $n \to \infty$.

In (6.34), $\Gamma : \mathbb{R}^d \to C \subset \mathbb{R}^d$ is a projection operator, where C is a compact and convex set. Also, in the above, $\Theta : \mathbb{R}^{d \times d} \to \{\text{positive definite and symmetric } d \times d \text{ matrices}\}$ is a projection operator that projects any $d \times d$ matrix to the space of positive definite and symmetric matrices. Such an operator is needed to ensure that the algorithm proceeds in a descent direction. If a matrix A is already positive definite and symmetric, then Θ is chosen such that $\Theta(A) = A$ itself.

The operator Θ can be characterized using methods such as the

Algorithm	Gradient estimate	Hessian estimate
SPSA (6.8)	$\Delta^{-1}\left(\frac{y^+ - y^-}{2\delta}\right)$	$\widehat{H} = \Delta^{-1}\frac{G(\theta+\delta\Delta) - G(\theta-\delta\Delta)}{4\delta}$ $G(\theta \pm \delta\Delta) = \hat{\Delta}^{-1}\frac{(y^{++} - y^+)}{\delta}$
SPSA variant (6.12)	$\Delta^{-1}\left(\frac{y^{++} - y^{--}}{2\delta}\right)$	$\widehat{H}_{ij} = \left(\frac{y^{++} + y^{--} - 2y}{\delta^2 \Delta_i \hat{\Delta}_j}\right)$
SF (6.14)	$\Delta\left(\frac{y^+}{\delta}\right)$	$\hat{H} = \frac{(\Delta\Delta^\intercal - I)}{\delta^2} y^+$
SF variant (6.23)	$\Delta\left(\frac{y^+ - y^-}{2\delta}\right)$	$\hat{H} = \frac{(\Delta\Delta^\intercal - I)}{2\delta^2}(y^+ + y^-)$
RDSA (6.25)	$\Delta\left(\frac{y^+ - y^-}{2\delta(1+\epsilon)}\right)$	$\widehat{H} = M\left(\frac{y^+ + y^- - 2y}{\delta^2}\right)$

Table 6.2: A summary of the function measurements used and the form of gradient/Hessian estimates for the stochastic Newton algorithm (6.1). In the table, y^{++}, y^{--}, y^+, y^-, y denote the function measurements corresponding to input parameters $\theta+\delta\Delta+\delta\hat{\Delta}$, $\theta+\delta\Delta-\delta\hat{\Delta}$, $\theta+\delta\Delta$, $\theta-\delta\Delta$, and θ, respectively. The choice of random perturbation varies between rows. The matrix M used in the Hessian estimate presented in last row is defined in (6.27).

modified Choleski factorization procedure, see (Bertsekas, 1999), or the procedures in (Spall, 2000) as well as (Zhu and Spall, 2002). For a matrix A, let $(\Theta(A))^{-1}$ denote the inverse of $\Theta(A)$ which is also positive definite and symmetric.

A6.8. (i) For any two sequences of $d \times d$ matrices $\{A_n\}$ and $\{B_n\}$, $\lim_{n\to\infty} \|\Theta(A_n) - \Theta(B_n)\| = 0$ if $\lim_{n\to\infty} \|A_n - B_n\| = 0$.

(ii) We have

$$\sup_n \|\Theta(C_n)\| < \infty, \quad \sup_n \|(\Theta(C_n))^{-1}\| < \infty,$$

if $\sup_n \|C_n\| < \infty$ for a given sequence $\{C_n\}$ of $d \times d$ matrices.

The requirement in Assumption A6.8(i) can be easily imposed, see (Bertsekas, 1999; Spall, 2000; Zhu and Spall, 2002). Further, a sufficient

condition for Assumption A6.8(ii) is

$$c_1\|z\|^2 \leq z^T\Theta(C_n)z \leq c_2\|z\|^2, \tag{6.36}$$

for all $z \in \mathbb{R}^d$, $n \geq 0$. Most projection operators are seen to satisfy (6.36), see (Bhatnagar, 2005) for a detailed discussion.

A6.9. The step size schedules $\{a(n)\}$ and $\{b(n)\}$ together with the perturbation sequence $\{\delta_n\}$ of positive real numbers satisfy the following:

$$\sum_n a(n) = \sum_n b(n) = \infty, \tag{6.37}$$

$$\sum_n \left(\frac{a(n)}{\delta_n}\right)^2 < \infty; \ \sum_n \left(\frac{b(n)}{\delta_n^2}\right)^2 < \infty, \tag{6.38}$$

$$\lim_{n\to\infty}\left(\frac{a(n)}{b(n)}\right) = 0. \tag{6.39}$$

Note that (6.37) ensures that the algorithm does not exhibit premature convergence as trajectories obtained by putting the parameters in (6.34) and (6.35) along time points obtained from the sequences $\{a(n)\}$ and $\{b(n)\}$, respectively, and obtaining continuous linear interpolations of these. This helps in arguing that these continuously interpolated trajectories asymptotically track the limit points of corresponding ODEs provided the noise in the sample observations vanishes asymptotically. The latter happens from (6.38). The first condition in (6.38) is the same as the corresponding condition for gradient based schemes (see Chapter 4. The second condition is necessitated from the form of the Hessian estimators in Chapter 6 where δ_n^2 appears in the denominator of the Hessian estimator, see for instance, (6.23). As with gradient-based schemes, one can show convergence of the resulting martingale sequence obtained from the Hessian estimator under the second condition in (6.38). Finally, (6.39) results in a difference in timescales within the recursions (6.34)-(6.35). In particular, it ensures that the Hessian update (6.35) proceeds on a faster scale as compared to the θ-recursion (6.34) that makes use of the inverse of the projected Hessian update.

A6.10. The function f is four-times continuously differentiable with $\left|\nabla^4_{i_1,i_2,i_3,i_4} f(\theta)\right| < \infty$, for $i_1, i_2, i_3, i_4 = 1, \ldots, d$ and for all $\theta \in \mathbb{R}^d$.

Assumption A6.10 is the same as Assumption A6.1 (restated here for ease of reference).

A6.11. We have

(i)
$$\left\| E\left[\widehat{\nabla} f(\theta_n) \middle| \theta_n \right] - \nabla f(\theta_n) \right\| = O(\delta_n^2),$$
where $\widehat{\nabla} f(\theta_n)$ and $\nabla f(\theta_n)$ are the gradient estimate and the true gradient respectively.

(ii)
$$\left\| E\left[\widehat{H}_n \middle| \theta_n \right] - \nabla^2 f(\theta_n) \right\| = O(\delta_n^2),$$
where $\widehat{H}_n$ and $\nabla^2 f(\theta_n)$ are respectively the Hessian estimate and the true Hessian respectively.

Assumptions A6.11(i) and A6.11(ii) have been shown to hold for the various gradient and Hessian estimators based on random perturbations in Chapters 3 and 6 respectively.

Lemma 6.4. The sequence of Hessian updates $\{\overline{H}_n\}$ is uniformly bounded with probability one. In other words, $\sup_n \|\overline{H}_n\| < \infty$ a.s.

Proof. Note that (6.35) can be rewritten as

$$\overline{H}_{n+1} = \overline{H}_n + b(n)(\nabla^2 f(\theta_n) + \xi_n + M_{n+1} - \overline{H}_n), \tag{6.40}$$
$$= \overline{H}_n + b(n)(\nabla^2 f(\theta_n) + \xi_n - \overline{H}_n) + b(n) M_{n+1}, \tag{6.41}$$

where $\xi_n = E[\hat{H}_n|\theta_n] - \nabla^2 f(\theta_n)$. From Assumption A6.11, $\xi_n = O(\delta_n^2) \to 0$ as $n \to \infty$. Let us ignore for a moment, the term $b(n)M_{n+1}$ in (6.41). Then, from Assumption A6.10 and the fact that $\theta_n \in C$ (a compact set), it follows that $\sup_n \|\nabla^2 f(\theta_n)\| < \infty$. It also follows from Assumption A6.9 that $b(n) \to 0$ as $n \to \infty$. Thus, outside a set of probability zero, $\exists N_0 \geq 1$ such that $\overline{H}_{n+1}$ (upon ignoring the $b(n)M_{n+1}$ term in (6.41)) can be viewed as a convex combination of $\overline{H}_n$ and a uniformly bounded quantity. Finally, observe that $M_{n+1} = \hat{H}_n - E[\hat{H}_n|\theta_n]$ is a martingale difference term. It can now be argued as before, using the step-size

condition (6.38), and the martingale convergence theorem for square-integrable martingales, see Theorem B.7, that $\sum_m b(m)M_{m+1} < \infty$ a.s. Thus, $\{\overline{H}_n\}$ as in (6.41) is uniformly bounded almost surely. The claim follows. □

As described in Chapter 2.7, the system of ODEs corresponding to (6.34)-(6.35) when viewed from the faster timescale are the following:

$$\dot{\theta}(t) = 0, \tag{6.42}$$

$$\dot{\overline{H}}(t) = \nabla^2 f(\theta(t)) - \overline{H}(t). \tag{6.43}$$

In the light of (6.42), as also discussed in Chapter 2.7, one may let $\theta(t) \equiv \theta$, $\forall t$. Thus, (6.43) can then be rewritten as

$$\dot{\overline{H}}(t) = \nabla^2 f(\theta) - \overline{H}(t). \tag{6.44}$$

The ODE (6.44) has $\overline{H}^*(\theta) = \nabla^2 f(\theta)$ as its unique globally asymptotically stable attractor. By Assumption A6.10 and the fact that $\theta \in C$, a compact set, $\overline{H}^*(\theta)$ is Lipschitz continuous in θ.

Lemma 6.5. We have

$$\lim_{n\to\infty} \left\| \overline{H}_n - \overline{H}^*(\theta_n) \right\| = 0, \text{ a.s.}$$

Proof. An application of Theorem 2.2 on (6.35) gives us

$$\|\overline{H}_n - E[\hat{H}_n|\theta_n]\| \to 0 \text{ a.s.},$$

as $n \to \infty$. The claim follows from an application of the triangle inequality and Assumption A6.11. □

Lemma 6.6. We have

$$\left\| \Theta(\overline{H}_n)^{-1} - \Theta(\overline{H}^*(\theta_n))^{-1} \right\| = \left\| \Theta(\overline{H}_n)^{-1} - \Theta(\nabla^2 f(\theta_n))^{-1} \right\| \to 0,$$

as $n \to \infty$, a.s.

Proof. The equality in the claim follows because $\overline{H}^*(\theta_n) = \nabla^2 f(\theta_n)$. Now note that

$$\left\| \Theta(\overline{H}_n)^{-1} - \Theta(\nabla^2 f(\theta_n))^{-1} \right\|$$

$$
\begin{aligned}
&= \left\| \Theta(\nabla^2 f(\theta_n))^{-1} \left(\Theta(\nabla^2 f(\theta_n)) \Theta(\overline{H}_n)^{-1} - I \right) \right\| \\
&= \left\| \Theta(\nabla^2 f(\theta_n))^{-1} \left(\Theta(\nabla^2 f(\theta_n)) \Theta(\overline{H}_n)^{-1} - \Theta(\overline{H}_n) \Theta(\overline{H}_n)^{-1} \right) \right\| \\
&= \left\| \Theta(\nabla^2 f(\theta_n))^{-1} \left(\Theta(\nabla^2 f(\theta_n)) - \Theta(\overline{H}_n) \right) \Theta(\overline{H}_n)^{-1} \right\| \\
&\leq \left\| \Theta(\nabla^2 f(\theta_n))^{-1} \right\| \left\| \Theta(\nabla^2 f(\theta_n)) - \Theta(\overline{H}_n) \right\| \left\| \Theta(\overline{H}_n)^{-1} \right\| \\
&\leq \sup_n \left\| \Theta(\nabla^2 f(\theta_n))^{-1} \right\| \sup_n \left\| \Theta(\overline{H}_n)^{-1} \right\| \left\| \Theta(\nabla^2 f(\theta_n)) - \Theta(\overline{H}_n) \right\| \\
&\longrightarrow 0 \text{ as } n \to \infty, \text{ a.s.}
\end{aligned}
$$

The first inequality follows from the property on induced matrix norms, cf. Proposition A.12 of (Bertsekas and Tsitsiklis, 1989). Note also the following: (i) From Assumption A6.10 and the fact that $\theta \in C$ (a compact set), $\sup_n \|\nabla^2 f(\theta_n)\| \leq \bar{K} < \infty$, for some $\bar{K} > 0$ and by Assumption A6.8(ii), $\sup_n \|\Theta(\nabla^2 f(\theta_n))^{-1}\| < \infty$ a.s. (ii) By Lemma 6.4, $\sup_n \|\overline{H}_n\| < \infty$ a.s. Thus, by Assumption A6.8(ii), $\sup_n \|\Theta(\overline{H}_n)^{-1}\| < \infty$ a.s. (iii) Finally, $\|\Theta(\overline{H}_n) - \Theta(\nabla^2 f(\theta_n))\| \to 0$ as $n \to \infty$ from Assumption A6.8(i) and Lemma 6.5. □

We now shift our attention to the slower timescale recursion (6.34). Consider the following ODE associated with (6.34):

$$
\dot{\theta}(t) = -\bar{\Gamma} \left(\Theta(\nabla^2 f(\theta(t)))^{-1} \nabla f(\theta(t)) \right). \tag{6.45}
$$

Here, $\bar{\Gamma} : \mathcal{C}(C) \to \mathcal{C}(\mathcal{R}^d)$ is defined according to

$$
\bar{\Gamma}(v(x)) = \lim_{\eta \to 0} \left(\frac{\Gamma(x + \eta v(x)) - x}{\eta} \right), \tag{6.46}
$$

for any continuous $v : C \to \mathcal{R}^d$.

Let $A \subset H \stackrel{\triangle}{=} \{\theta \in \mathbb{R}^d | \nabla f(\theta) = 0\}$ be the set of globally asymptotically stable attractors for the ODE (6.45). In fact, $V(\cdot) = f(\cdot)$ serves as a Lyapunov function for this ODE since

$$
\begin{aligned}
\frac{dV(\theta)}{dt} &= \nabla V(\theta)^T \dot{\theta} \\
&= \nabla f(\theta)^T \bar{\Gamma} \left(-\Theta(\nabla^2 f(\theta(t)))^{-1} \nabla f(\theta(t)) \right) \leq 0,
\end{aligned}
$$

since $\Theta(\nabla^2 f(\theta))^{-1}$ is a positive definite matrix for each θ.

Theorem 6.7. The recursion (6.34) satisfies $\theta_n \to A$ a.s., as $n \to \infty$.

Proof. As a consequence of Lemma 6.6, we have

$$\begin{aligned}\theta_{n+1} &= \Gamma\left(\theta_n - a(n)\Theta(\overline{H}_n)^{-1}\widehat{\nabla} f(\theta_n)\right)\\ &= \Gamma\left(\theta_n - a(n)(\Theta(\overline{H}^*(\theta_n))^{-1}\nabla f(\theta_n) - \kappa_n)\right)\\ &= \Gamma\left(\theta_n - a(n)(\Theta(\nabla^2 f(\theta_n))^{-1}\nabla f(\theta_n) - \kappa_n)\right),\end{aligned}$$

where $\kappa_n = \Theta(\nabla^2 f(\theta_n))^{-1}\nabla f(\theta_n) - \Theta(\overline{H}_n)^{-1}\widehat{\nabla} f(\theta_n)$. Now note that we can rewrite κ_n as

$$\begin{aligned}\kappa_n &= \Theta(\nabla^2 f(\theta_n))^{-1}(\nabla f(\theta_n) - \widehat{\nabla} f(\theta_n))\\ &\quad + (\Theta(\nabla^2 f(\theta_n))^{-1} - \Theta(\overline{H}_n)^{-1})\widehat{\nabla} f(\theta_n)\\ &= \Theta(\nabla^2 f(\theta_n))^{-1}(\nabla f(\theta_n) - \widehat{\nabla} f(\theta_n))\\ &\quad + (\Theta(\nabla^2 f(\theta_n))^{-1} - \Theta(\overline{H}_n)^{-1})\nabla f(\theta_n)\\ &\quad + (\Theta(\nabla^2 f(\theta_n))^{-1} - \Theta(\overline{H}_n)^{-1})(\widehat{\nabla} f(\theta_n) - \nabla f(\theta_n)).\end{aligned}$$

Thus,

$$\begin{aligned}\|\kappa_n\| &\le \left\|\Theta(\nabla^2 f(\theta_n))^{-1}\right\| \left\|\nabla f(\theta_n) - \widehat{\nabla} f(\theta_n)\right\|\\ &\quad + \left\|\Theta(\nabla^2 f(\theta_n))^{-1} - \Theta(\overline{H}_n)^{-1}\right\| \|\nabla f(\theta_n)\|\\ &\quad + \left\|\Theta(\nabla^2 f(\theta_n))^{-1} - \Theta(\overline{H}_n)^{-1}\right\| \left\|\widehat{\nabla} f(\theta_n) - \nabla f(\theta_n)\right\|.\end{aligned}$$

From Assumption A6.11(i), $\left\|\widehat{\nabla} f(\theta_n) - \nabla f(\theta_n)\right\| = O(\delta_n^2) \to 0$ as $n \to \infty$ and from Lemma 6.6, $\left\|\Theta(\nabla^2 f(\theta_n))^{-1} - \Theta(\overline{H}_n)^{-1}\right\| \to 0$ as $n \to \infty$. Thus, $\left\|\Theta(\nabla^2 f(\theta_n))^{-1} - \Theta(\overline{H}_n)^{-1}\right\| \left\|\widehat{\nabla} f(\theta_n) - \nabla f(\theta_n)\right\| \to 0$ as $n \to \infty$. Further, from Assumption A6.10 and the fact that $\theta_n \in C$, $\forall n$, $\sup_n \|\nabla f(\theta_n)\| \le \check{M} < \infty$ for some $\check{M} > 0$. Likewise from Assumption A6.10 together with the fact that $\theta_n \in C$ (a compact set) and Assumption A6.8(ii), it follows that $\sup_n \left\|\Theta(\nabla^2 f(\theta_n))^{-1}\right\| \le \check{N} < \infty$, for some $\check{N} > 0$. Thus, $\|\kappa_n\| \to 0$ as $n \to \infty$ a.s.

Now observe that $\nabla f(\theta)$ is continuous in θ (by Assumption A6.10). Further, $\nabla^2 f(\theta)$ is continuous in θ by Assumption A6.10 and $\Theta(\nabla^2 f(\theta))$

is also continuous by Assumption A6.8. It can also be shown as in Lemma 6.6 that $\Theta(\nabla^2 f(\theta))^{-1}$ is continuous. Thus, the function $F(\theta) = \Theta(\nabla^2 f(\theta))^{-1}\nabla f(\theta)$ is continuous. Thus, Assumption A2.9 holds. Now, from Assumption A6.9, it follows that $\sum_n a(n) = \infty$, $a(n) \to 0$ as $n \to \infty$, thereby satisfying Assumption A2.10. Further, using the identification $\beta_n = \kappa_n$ with $\|\kappa_n\| \to 0$ a.s. as $n \to \infty$, Assumption A2.11 can be seen to be satisfied. Finally, Assumption A2.12 is trivially satisfied since $\eta_n = 0$, $\forall n$ (in our case). The claim now follows from Theorem 2.5 (the Kushner-Clark theorem for projected stochastic approximation, cf. Chapter 5 of (Kushner and Clark, 1978)). □

6.8 Bibliographic remarks

Hessian estimation

In (Fabian, 1971), the author analyzes a finite differences Hessian estimation scheme with $O(d^2)$ function measurements. In an importance advance, the author in (Spall, 2000) brings the idea of simultaneous perturbation for Hessian estimation, using random perturbations similar to those employed in SPSA. The advantage with this scheme is the drastic reduction in the number of function measurements to four, irrespective of the dimension. Subsequent advances that we presented in Sections 6.3.2, 6.5 are based on (Bhatnagar and Prashanth, 2015) and (Prashanth *et al.*, 2017), respectively.

Gaussian smoothed functional — an idea explored in Chapter 3 for estimating gradients, can be extended to estimate the Hessian as well. In Section 6.4.1 and 6.4.2, we presented two Gaussian SF schemes for Hessian estimation, and these are adapted from (Bhatnagar, 2007). Proposition 6.1 is extracted from the proof of the bias of 1SF estimation in (Bhatnagar, 2007), and this result has also been separately shown in later works, cf. (Erdogdu, 2016; Balasubramanian and Ghadimi, 2022b). These works provide the connection of the result in Proposition 6.1 to the classic Stein's identity, which includes a first as well as second-order variant, see (Stein, 1972; Stein, 1981) and also (Balasubramanian and Ghadimi, 2022b, Theorem 1.2). Proposition 6.1 is central to the analysis of SF1 as well as SF2 estimators, in particular, to provide bounds of

$O(\delta)$ and $O(\delta^2)$ on the bias of these estimators, respectively.

Zeroth-order Stochastic Newton

There is considerable work on Newton-based algorithms though not as much as for gradient-based schemes. In some early work, the Hessian is estimated using finite difference approximations that are in turn finite difference estimates of the gradients (Fabian, 1971). Such a scheme however requires $O(d^2)$ samples of the objective function at each update epoch. In (Ruppert, 1985), it is assumed that the objective function gradients are known and these are in turn used to estimate the Hessian at each update instant. Zeroth-order simultaneous perturbation Hessian estimates have been developed for the first time in (Spall, 2000) and Newton-based algorithms studied. The Hessian estimator here requires four function measurements. A procedure for projecting the Hessian to the space of positive definite and symmetric matrices is proposed. In (Zhu and Spall, 2002), another method for projecting the eigenvalues to the positive half line is proposed for the algorithm in (Spall, 2000). Certain feedback and weighting mechanisms for obtaining improved Hessian estimates have been proposed in (Spall, 2009).

Building on the work in (Spall, 2000), four Newton algorithms have been developed and studied in (Bhatnagar, 2005) that require four, three, two and one simulations, respectively, for estimating the Hessian regardless of the parameter dimension d. In (Bhatnagar, 2007), two smoothed functional algorithms based on Gaussian perturbations have been presented that require one and two function measurements, respectively. In (Bhatnagar and Prashanth, 2015), a balanced SPSA based Hessian estimator is presented that requires three function measurements. Efficient ways of obtaining the Hessian inverse – a direct method and another procedure based on the Sherman-Morrison-Woodbury identity are also proposed here. The latter technique has also been made use of to obtain an efficient procedure in (Rastogi *et al.*, 2016). In (Ghoshdastidar *et al.*, 2014a), a family of Newton algorithms based on q-Gaussian perturbations is presented. Here, one gets a wide range of smoothing functionals depending on the value of q with Gaussian, Cauchy and Uniform emerging as special cases for different values of

the q parameter.

In (Spall, 2009), the sequence $\{b(n\}$ is optimized for asymptotic variance of the Hessian estimator in terms of the perturbation sensitivity parameters $\delta_n, n \geq 0$. We do not consider here this variance optimization problem in terms of the step-sizes $b(n), n \geq 0$. In (Zhu *et al.*, 2019), an efficient method for reducing the number of floating point operations from $O(d^3)$ to $O(d^2)$ is presented that is based on the symmetric indefinite matrix factorization approach presented in (Bunch and Parlett, 1971). The method seems highly effective and stable, especially in high-dimensional problems.

7

Escaping saddle points

In Chapters 4 and 6, we provided theoretical guarantees that establish asymptotic convergence to a stationary point of the objective function f. On the other hand, in Chapter 5, we established convergence of zeroth-order stochastic gradient algorithms to an approximate stationary point in the non-asymptotic regime. However, these results are not sufficient in a non-convex optimization setting since local maxima and saddle points are also stationary points in addition to local minima. We shall refer to such undesirable stationary points collectively as saddle points, as in the recent literature (Jin *et al.*, 2017; Ge *et al.*, 2015; Jin *et al.*, 2021).

It is desirable to escape saddle points and converge to a local minimum. In several ML applications, it may be enough to avoid saddle points and converge to local minima, as such points may be as good as global minima in many applications. A few concrete applications that possess such a characteristic are as follows: low rank matrix factorization (Jin *et al.*, 2017), tensor decomposition (Ge *et al.*, 2015), matrix sensing (Bhojanapalli *et al.*, 2016), dictionary learning (Sun *et al.*, 2016), matrix completion (Ge *et al.*, 2016), robust principal component analysis (Ge *et al.*, 2017), and a sub-class of neural networks (Kawaguchi, 2016). In

each of these applications, local minima are as good as global minima, while there are innumerable first-order stationary points that are not local minima. Moreover, at each such saddle point, there is a direction of escape corresponding to a negative eigenvalue.

We discuss two schemes to escape saddle points. This goal is also referred to as avoidance of traps, cf. (Borkar, 2003; Barakat *et al.*, 2021; Gadat and Gavra, 2022). The first scheme is the vanilla ZSG algorithm presented earlier. We show that, when the noise in the function measurements is rich, then the ZSG algorithm, which employs the unified gradient estimate, converges to a local minimum asymptotically. We shall use the ODE approach for this result. The second scheme is a variant of the stochastic Newton algorithm, and incorporates a cubic-regularization term. We establish the convergence of the cubic-regularized Newton algorithm to an approximate second-order stationary point (SOSP) in the non-asymptotic regime. An SOSP would be a local minimum when the objective satisfies a strict saddle condition, made precise later.

The rest of this chapter is organized as follows: In Section 7.1, we introduce first and second-order stationary points. In Section 7.2, we present an asymptotic result for escaping saddle points for ZSG algorithm under assumptions on the measurement noise. In Section 7.3, we discuss two algorithms for escaping saddle points in a setting where exact gradient and Hessian measurements are available. The first algorithm uses curvature information, while the second one adds extraneous noise so that the iterates do not get stuck at a saddle point. In Section 7.4, we present the cubic-regularized Newton algorithm with zeroth-order gradient/Hessian estimates and provide a non-asymptotic sample complexity bound for identifying approximate SOSPs.

7.1 First and second-order stationary points

The non-asymptotic bounds in Chapter 5 were shown to converge to an approximate stationary point. Recall that, at a first-order stationary point (FOSP), say $\bar{\theta}$, the gradient vanishes, i.e., $\nabla f(\bar{\theta}) = 0$. An ϵ-approximation to FOSP is a point $\bar{\theta}$ that satisfies $\left\|\nabla f(\bar{\theta})\right\| \leq \epsilon$.

Finding an FOSP is not sufficient for a non-convex objective function f, as such a point is not necessarily a local minimum. As a simple

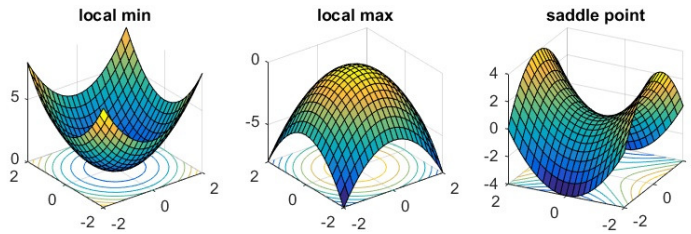


Figure 7.1: Illustration of three types of first-order stationary points. The image is sourced from offconvex.org

example, consider $f(\theta_1, \theta_2) = \theta_1^2 - \theta_2^2$. Then, $\nabla f(0,0) = 0$, implying $(0,0)$ is an FOSP. However, the origin is clearly not a local minimum because $f(0,\epsilon) < f(0,0)$ for any ϵ.

As illustrated in Figure 7.1, an FOSP could potentially be a saddle point. In order to avoid such points and find local optima of f, we need information with regard to the curvature of the underlying objective. The notion of second-order stationary point (SOSP) formalizes this idea and aids in escaping saddle points.

At a second-order stationary point (SOSP), say $\bar{\theta}$, we have $\nabla f(\bar{\theta}) = 0$ and $\lambda_{\min}\left(\nabla^2 f(\bar{\theta})\right) \geq 0$, where as before, $\lambda_{\min}(A)$ denotes the smallest eigenvalue of the $d \times d$-matrix A.

For the non-asymptotic analysis, an ϵ-version of SOSP is defined below.

Definition 7.1 (ϵ**-second-order stationary point**). Fix $\epsilon > 0$. Let θ_R be the output of a stochastic iterative algorithm for solving (1.1). Then, θ_R is said to be an ϵ-SOSP in expectation if

$$\max\left\{\sqrt{\mathbb{E}\|\nabla f(\theta_R)\|}, \frac{-1}{\sqrt{\rho}}\mathbb{E}\lambda_{\min}\left(\nabla^2 f(\theta_R)\right)\right\} \leq \sqrt{\epsilon}, \tag{7.1}$$

where ρ is a positive parameter.

From the definition above, in expectation, θ_R can be inferred to be a point where the size (or norm) of the objective gradient is small, and the Hessian at θ_R is nearly positive semi-definite. Thus, θ_R is an approximation to a point where the objective gradient vanishes and the

Hessian is positive semi-definite. We next elaborate on the subtleties behind finding such a point.

If $\nabla f(\theta) = 0$ and $\nabla^2 f(\theta) \succ 0$, then one can conclude that θ is a local minimum. On the other hand, if $\nabla f(\theta) = 0$ and $\nabla^2 f(\theta) \succeq 0$, then one has to go beyond a SOSP and look at the third derivatives to infer if θ is a local minimum or not, and so on. Such a process is not amenable for optimization using gradient-based methods, as there is no end to calculating higher-order derivatives with the hope of finding a local minimum. Staying within the realm of first and second derivatives (or the gradient and Hessian), we would like to understand conditions that guarantee that a candidate point is a local minimum and not a saddle point. At an FOSP, if the Hessian is positive definite (resp. negative definite), we have a local minimum (resp. local maximum). If the Hessian is indefinite, i.e., has both positive and negative eigenvalues, then we have arrived at a saddle point and the negative eigenvalues can be used to move away from such a point. On the other hand, if the Hessian is degenerate, i.e., either positive or negative semi-definite, then the optimization process becomes hard, in particular, to find local minima. More precisely, it is well-known that finding a local minimum is an NP-hard problem, see (Anandkumar and Ge, 2016). However, if the saddle points are strict, i.e., Hessian is not degenerate, then there exist polynomial time algorithms for finding a local minimum. The strict saddle condition is as follows:

$$\nabla f(\theta) = 0 \text{ and } \lambda_{\min}(\nabla^2 f(\theta)) < 0. \tag{7.2}$$

When the strict saddle condition (7.2) holds, an SOSP will be a local minimum since at a saddle point one can find a direction corresponding to a negative eigenvalue where the function decreases, whereas at an SOSP no such directions exist.

When the strict saddle condition (7.2) is satisfied, the Hessian $\nabla^2 f(\theta)$ has at least one negative eigenvalue and this gives a direction for a method using second-order information to escape from saddle points. The cubic-regularized Newton algorithm presented in Section 7.4 uses the Hessian estimates to escape from a saddle point, and find an ϵ-SOSP, as formalized in Definition 7.1, while using $O\left(\frac{1}{\epsilon^{3.5}}\right)$ samples.

The table below summarizes the conditions for FOSP, SOSP and their approximate variants.

Type	Condition
FOSP θ	$\nabla f(\theta) = 0$
ϵ-FOSP θ	$\|\nabla f(\theta)\| \leq \epsilon$
SOSP θ	$\nabla f(\theta) = 0$ and $\nabla^2 f(\theta) \succeq 0$
ϵ-SOSP θ	$\|\nabla f(\theta)\| \leq \epsilon$ and $\nabla^2 f(\theta) \succeq -\sqrt{\rho\epsilon}\,\mathbb{I}$

We conclude this section with a simple example, where SOSPs coincide with global minima.

Example 7.1 (Matrix factorization)**.** For a given positive semi-definite matrix M, consider the following objective function:

$$\min_{\theta \in \mathbb{R}^d} \left\{ f(\theta) = \frac{1}{2} \|\theta\theta^\top - M\|_F^2 \right\}. \tag{7.3}$$

Let $M = U\Lambda U^\top$, where Λ is a diagonal matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$, and the matrix U contains the eigenvectors of M. Assume $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. Let $u_1, u_2, \ldots, u_d$ denote the eigenvectors of M corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_d$.

The gradient and Hessian of f are given by

$$\nabla f(\theta) = \|\theta\|_2^2 \theta - M\theta, \text{ and} \tag{7.4}$$

$$\nabla^2 f(\theta) = \|\theta\|_2^2 I + 2\theta\theta^\top - M. \tag{7.5}$$

From the Hessian expression, it is apparent that the function $f(\theta)$ is non-convex. Setting $\nabla f(\theta) = 0$, we obtain $(\|\theta\|_2^2 I - M)\theta = 0$ or $M\theta = \|\theta\|_2^2 \theta$. Thus, the stationary points are zero and $\pm\sqrt{\lambda_i} u_i$, for $i = 1, \ldots, d$.

For $\theta = \sqrt{\lambda_i} u_i$, notice that $f(\theta) = -\lambda_i^2 + \|M\|_F^2$. Thus, f is minimized at $\pm\sqrt{\lambda_1} u_1$. Moreover, using the expression of the Hessian above, a straightforward calculation shows that $\nabla^2 f(\theta) \succeq 0$ at $\pm\sqrt{\lambda_1} u_1$ and not at $\pm\sqrt{\lambda_i} u_i$ for $i \neq 1$. For the remaining local minima, say $\tilde{x}$, since $\lambda_1 > \lambda_2$, we have a direction of escape using the top eigenvector u_1, since $u_1 \nabla^2 f(\tilde{x})^\top u_1 \leq \lambda_2 - \lambda_1 < 0$.

7.2 Asymptotic escaping of saddle points for ZSG algorithm

A common trick to escape saddle points is to add extraneous noise so that a stochastic gradient algorithm does not get stuck and instead, converges to a local minima. This approach is the adopted in (Jin *et al.*, 2017; Ge *et al.*, 2015; Jin *et al.*, 2021). In particular, such a scheme, referred to as perturbed gradient descent, involves the following update iteration:

$$\theta_{n+1} = \theta_n - a(n)\left(\widehat{\nabla} f(\theta_n) + \zeta_n\right), \tag{7.6}$$

where ζ_n is extraneous noise that is injected into the stochastic gradient algorithm, and is usually sampled from a zero-mean multivariate Gaussian vector with covariance matrix $\sigma^2 I$. The extraneous noise ensures that the iterate θ_n, governed by (7.6), does not converge to an unstable equilibrium of the underlying ODE $\dot{\theta}(t) = -\nabla f(\theta(t))$, implying escape from saddle points. We shall explore this idea in more detail in the next section.

In this section, we adopt a different viewpoint, which is to show convergence to local minima for the case where the noise in the gradient estimates is rich in all directions, which in turn does not let the ZSG algorithm get stuck at an undesirable saddle point. Recall ZSG uses the following update:

$$\theta_{n+1} = \theta_n - a(n)\left(\widehat{\nabla} f(\theta_n)\right), \tag{7.7}$$

We shall use the unified gradient estimator described in Chapter 3. For the sake of readability, we recall this estimator below.

$$\widehat{\nabla} f(\theta_n) = \left(\frac{y_n^+ - y_n^-}{2\delta}\right) V_n, \tag{7.8}$$

where $y_n^+ = f(\theta_n + \delta U_n) + \xi_n^+$, and $y_n^- = f(\theta_n - \delta U_n) + \xi_n^-$. Notice that, unlike the asymptotic convergence analysis from Chapter 4, we employ a constant sensitivity parameter $\delta > 0$ and not a diminishing one. Such a choice aids the main result of this section, which establishes avoidance of saddle points for the update (7.7).

Under certain conditions on the measurement noise $\{\xi_n^{\pm}\}$, one can avoid injecting noise artificially, and instead directly establish conver-

gence to local minima, owing to the noise in the gradient estimator. The additional assumption on measurement noise is made precise below.

A7.1. $\exists\, c_3, c_4 > 0$ such that $c_3 \leq \mathbb{E}_k|\xi_k^+ - \xi_k^-|$, where $\mathbb{E}_k(\cdot)$ is shorthand notation for $\mathbb{E}(\cdot \mid \mathcal{F}_k)$. In addition, $|\xi_k^+ - \xi_k^-| \leq c_4$, $\forall k$.

The assumption above ensures that the noise in function measurements is rich in all directions.

Consider the following ODE:

$$\dot{\theta}(t) = -\mathbb{E}\left[\left(\frac{f(\theta(t)+\delta U) - f(\theta(t)-\delta U)}{2\delta}\right) V \,\middle|\, \theta(t)\right], \tag{7.9}$$

where the expectation is over the joint distribution of U, V.

Theorem 7.1. Suppose the conditions of Proposition 3.1 and A7.1 hold. Further, assume $\|V_k\| \leq B_0$ a.s. for all k. Set $a(k) = \frac{c_5}{k^\alpha}$ and $\delta_k = \delta$, $\forall k$, for some constants $c_5, \delta > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right]$. Then, $\{\theta_k\}$ governed by (7.7), converges to the stable critical points of the ODE (7.9).

The result above says that the stochastic gradient algorithm (7.7) avoids unstable critical points of the ODE (7.9). However, the stable critical points of this ODE are not necessarily the local minima of the objective f. A related ODE is $\dot{\theta}(t) = -\nabla f(\theta(t))$. By the bias bounds in Chapter 3, we know that

$$\left\|\mathbb{E}\left[\left(\frac{f(\theta+\delta U) - f(\theta-\delta U)}{2\delta}\right) V \,\middle|\, \theta\right] - \nabla f(\theta)\right\| = O(\delta^2).$$

While the bias could in principle add spurious points (that are not local minima of f) to the limit set of (7.9), it is possible to find a δ_0 for any $\epsilon > 0$ such that for all $\delta \leq \delta_0$, the algorithm governed by (7.7) converges almost surely to an ϵ-neighborhood of the local minima of f, cf. Theorem 2.4 of (Bhatnagar *et al.*, 2003).

For the proof, we require a result from (Pemantle, 1990). We adapt this result to a gradient update and state it below.

Theorem 7.2 (Avoidance of traps)**.** Consider the following stochastic approximation update iteration:

$$\theta_{k+1} = \theta_k + a(k)h(\theta_k) + \psi_k. \tag{7.10}$$

Suppose the following conditions hold.

(B1) $\frac{c_5}{k^\alpha} \leq a(k) \leq \frac{c_6}{k^\alpha}$ for some constants $c_5, c_6 > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right]$;

(B2) $\mathbb{E}_k\left[(\psi_k \cdot \vartheta)^+\right] \geq c_7/k^\alpha$ for some $c_7 > 0$ and every unit vector ϑ. Here $(a \cdot b)$ denotes the dot product between a and b, and $(a)^+ = \max(a, 0)$;

(B3) $\|\psi_k\| \leq c_8/k^\alpha$ for some $c_8 > 0$.

Suppose $h \in C^2$. Then, $\{\theta_k\}$ governed by (7.10), converges to the stable critical points of the ODE $\dot{\theta}(t) = h(\theta(t))$.

We now prove Theorem 7.1.

Proof. We first rewrite the update rule (7.7) as follows:

$$\begin{aligned} \theta_{k+1} &= \theta_k - a(k)\widehat{\nabla} f(\theta_k) \\ &= \theta_k - a(k)\nabla f(\theta_k) - \psi_k, \end{aligned} \tag{7.11}$$

where $\psi_k = a(k)\left[\frac{\xi_k^+ - \xi_k^-}{\delta} V_k\right]$.

The convergence of (7.11) to a local minimum can be inferred from Theorem 7.2 provided that conditions (B1)–(B3) of (Pemantle, 1990) are satisfied.

It is easy to see that $a(k)$ defined in the theorem statement satisfies condition (B1).

We now show that condition (B2) holds. Consider the unit vector ϑ with the ith entry as 1. Letting V_k^i denote the ith entry of the vector V_k, we have

$$\mathbb{E}_k[(\psi_k \cdot \vartheta)^+] = \mathbb{E}_k\left[\frac{(a(k)(\xi_k^+ - \xi_k^-)V_k^i)^+}{\delta}\right]$$

$$\begin{aligned}
&\overset{(b)}{\geq} \mathbb{E}_k\left[\frac{a(k)(\xi_k^+ - \xi_k^-)V_k^i + a(k)|(\xi_k^+ - \xi_k^-)\,V_k^i|}{2\delta}\right] \\
&\overset{(c)}{=} \mathbb{E}_k\left[\frac{a(k)|\xi_k^+ - \xi_k^-|\,|V_k^i|}{2\delta}\right] \\
&\overset{(d)}{\geq} \frac{c_5 c_3 \min\limits_{i=1\ldots,d} \mathbb{E}|V_k^i|}{2\delta k^\alpha}.
\end{aligned}$$

In the above, we used the fact that $\max(x,y) = \dfrac{x+y+|x-y|}{2}$ to infer the equality in (b). To infer the equality in (c), we used $\mathbb{E}_k[(\xi_k^+ - \xi_k^-)V_k^i] = 0$, which holds since $\mathbb{E}_k[\xi_k^+ - \xi_k^-] = 0$ and V_k is independent of $\mathcal{F}_k$. Finally, A7.1 allows us to infer (d). Thus, condition (B2) holds.

We now turn to verifying condition (B3). Notice that

$$\|\psi_k\| \leq \frac{a(k)}{\delta}\|(\xi_k^+ - \xi_k^-)V_k\| \leq \frac{c_4 c_5 B_0}{\delta k^\alpha},$$

where we used the following facts:
(a) $\|(\xi_k^+ - \xi_k^-)\| \leq c_4$ from A7.1; (b) $\|V_k\| \leq B_0$ by assumptions in the theorem statement; and (c) $a(k) = \dfrac{c_5}{k^\alpha}$. Thus, condition (B3) holds.

The verification of conditions (B1)–(B3) above together with the fact $f \in \mathbb{C}^3$ (by assumption) imply that (7.11) avoids unstable critical points of the ODE (7.9), by an invocation of Theorem 7.2. □

7.3 Escaping saddle points with exact gradient/Hessian measurements

In this section, we operate with exact gradient and/or Hessian measurements. We use this setting to illustrate the main algorithmic ideas to find an SOSP.

We make the following smoothness assumption for the sake of algorithmic development as well as analysis.

A7.2. There exist positive scalars L_1, L_2 such that

$$\begin{aligned}
\|\nabla f(\theta_1) - \nabla f(\theta_2)\| &\leq L_1\,\|\theta_1 - \theta_2\|\,, \text{ and} \\
\left\|\nabla^2 f(\theta_1) - \nabla^2 f(\theta_2)\right\| &\leq L_2\,\|\theta_1 - \theta_2\|\,.
\end{aligned} \tag{7.12}$$

In addition, we shall assume a finite lower bound for the objective f, made precise in the assumption below.

A7.3. There exists a $\bar{f} > -\infty$ s.t. $f(\theta) \geq \bar{f}$ for all $\theta \in \mathbb{R}^d$.

7.3.1 Hessian-aided scheme

Recall that, at an ϵ-SOSP, we have $\|\nabla f(\theta)\| \leq \epsilon$ and $\lambda_{\min}(\nabla^2 f(\theta)) \geq -\sqrt{\epsilon}$. Finding a point with $\|\nabla f(\theta)\| \leq \epsilon$ is easy and a simple GD algorithm would achieve this goal. This claim is made precise below for a GD update iteration given by

$$\theta_{k+1} = \theta_k - a\nabla f(\theta_k). \tag{7.13}$$

Since f is L_1-smooth and setting $a < \dfrac{1}{L_1}$, we have

$$\begin{aligned} f(\theta_{k+1}) &\leq f(\theta_k) + \nabla f(\theta_k)^\mathsf{T}(\theta_{k+1} - \theta_k) + \frac{L_1}{2}\|\theta_{k+1} - \theta_k\|^2 \\ &= f(\theta_k) - a\|\nabla f(\theta_k)\|^2 + \frac{a^2 L_1}{2}\|\nabla f(\theta_k)\|^2 \\ &\leq f(\theta_k) - \frac{a}{2}\|\nabla f(\theta_k)\|^2. \end{aligned} \tag{7.14}$$

Thus, using the GD update (7.13), one could get to a point, say $\bar{\theta}$, satisfying $\left\|\nabla f(\bar{\theta})\right\| \leq \epsilon$. However, such a point could be a saddle, and needs to be escaped from. A natural alternative is to use second-order information to move away from a potential saddle point. From an algorithmic viewpoint, one could perform a GD step (7.13) when the gradient is large, i.e., $\|\nabla f(\theta)\| > \epsilon$, and on arriving at a point with a small gradient, inspect the Hessian to infer if an SOSP is found. More precisely, let $\lambda_k := \lambda_{\min}(\nabla^2 f(\theta_k))$. If $\lambda_k \geq -\sqrt{\epsilon}$, then an ϵ-SOSP is found, since we inspect the Hessian $\nabla^2 f(\theta_k)$ only if $\|\nabla f(\theta)\| \leq \epsilon$. Otherwise, find an eigenvector, say u_k, corresponding to λ_k, with the additional constraint that $\|u_k\| = 1$ and $u_k^\mathsf{T}\nabla f(\theta_k) \leq 0$. Using this eigenvector, perform the following update iteration:

$$\theta_{k+1} = \theta_k + a(k)u_k. \tag{7.15}$$

Using Taylor expansions and the update given above, we obtain

$$f(\theta_{k+1}) \leq f(\theta_k) + a(k)\nabla f(\theta_k)^\mathsf{T}u_k + \frac{1}{2}a(k)^2 u_k^\mathsf{T}\nabla^2 f(\theta_k)u_k + \frac{1}{6}L_2 a(k)^3\|u_k\|^3.$$

Setting $a(k) = \dfrac{2|\lambda_k|}{L_2}$, and using $u_k^\top \nabla f(\theta_k) \leq 0$, we obtain

$$f(\theta_{k+1}) \leq f(\theta_k) - \frac{1}{2}\frac{4\lambda_k^2}{L_2^2}|\lambda_k| + \frac{1}{6}L_2\frac{8|\lambda_k|^3}{L_2^3}. \tag{7.16}$$

Thus, during each iteration of an algorithm that performs either (7.13) or (7.15), the function value drops. For a GD step (7.13), the decrease in function value is given by (7.14) and for the other step involving curvature information from the Hessian, the decrease in function value is given by (7.16). Now, the algorithm on termination, returns an SOSP. The termination in a finite number of iterations can be argued by the fact that the function value decreases in each iteration, and the maximum decrease is $f(\theta_0) - f(\theta^*)$. Such a calculation would lead to an $O(1/\epsilon^2)$ number of iterations for finding an ϵ-SOSP.

Based on the discussion above, a two-step algorithm for finding SOSPs is given as a pseudocode in Algorithm 3.

Theorem 7.3. Suppose Assumptions A7.2 to A7.3 hold. Then, Algorithm 3 will find an ϵ-SOSP in $O\left(\dfrac{\max(2L_1, \frac{3}{2}L_2)(f(\theta_0) - f^\star)}{\epsilon^2}\right)$ steps.

The algorithm above has two drawbacks. First, it requires explicit Hessian computation. The cubic-regularized Newton algorithm in the next section overcomes this drawback by working with Hessian-vector products. The second drawback involves the computational overhead resulting from the update (7.15), which requires examining the eigenvalues of the Hessian. Even with Hessian vector products, the implementation is computationally expensive when compared to a GD update. We next discuss an alternative that is a variant of GD, which finds an SOSP.

7.3.2 Perturbed GD

Recall from the discussion in the section above that a GD step is appropriate when the gradient norm is large. On the other hand, when the gradient norm is small, then we have either found an SOSP or else a saddle point. To avoid the latter case, the algorithm from the section

Algorithm 3: Hessian-aided gradient descent

Input: Initial point $\theta_1 \in \mathbb{R}^d$, step sizes $\{a(k)\}$.
for $k = 0, 1, \ldots$ **do**
  if $\|\nabla f(\theta_k)\| > \epsilon$ **then**
    $\theta_{k+1} = \theta_k - a(k)\nabla f(\theta_k)$;
  end if
  else
    Let $\lambda_k := \lambda_{\min}(\nabla^2 f(\theta_k))$;
    if $\lambda_k \geq -\sqrt{\epsilon}$ **then**
      return θ_k;
      `// ϵ-SOSP found`
    end if
    else
      Find eigenvector u_k corresponding to λ_k s.t. $\|u_k\| = 1$ and $u_k^\intercal \nabla f(\theta_k) \leq 0$;
      $\theta_{k+1} = \theta_k + a(k)u_k$;
    end if
  end if
end for

above inspected the Hessian, in particular, to infer if the minimum eigenvalue satisfies the SOSP condition or not. A computationally efficient alternative is to inject noise artificially when the latter condition holds, i.e., when gradient norm is small and the iterate has not moved much for many iterations. This idea forms the basis for perturbed GD[1], with pseudocode in Algorithm 4.

Algorithm 4 performs a regular GD step when the gradient is large, and as seen before, such a step would ensure a decrease in function value. However, when the gradient norm $\|\nabla f(\theta_k)\|$ is small, the algorithm may be either at a SOSP, or at a saddle point. To avoid the latter case, Algorithm 4 adds noise from an isotropic distribution. More precisely,

[1]Here "perturbed" is not to be confused with "random perturbations" underlying a simultaneous perturbation-based gradient estimation approach. Instead, here "perturbed" refers to the fact that a GD iterate is forced out of potential saddle points by noise factors.

Algorithm 4: Perturbed gradient descent

Input: Initial point $\theta_1 \in \mathbb{R}^d$, step size a, perturbation factors ζ_k, threshold parameters t_{thres} and g_{thres}.

for $t = 0, 1, \ldots$ **do**

  if $\|\nabla f(\theta_k)\| \leq g_{thres}$ *and* $t - t_{noise} > t_{thres}$ **then**

    $\tilde{\theta}_k \leftarrow \theta_k, \quad t_{\text{noise}} \leftarrow t;$

    $\theta_{k+1} = \tilde{\theta}_k + \zeta_k;$

  end if

  if $t - t_{noise} = t_{thres}$ *and* $f(\theta_k) - f(\tilde{\theta}_{t_{noise}}) > -f_{thres}$ **then**

    return $\tilde{\theta}_{t_{\text{noise}}}$;

  end if

  $\theta_{k+1} \leftarrow \theta_k - a\nabla f(\theta_k);$

end for

as an intermediate step, when $\|\nabla f(\theta_k)\| \leq g_{\text{thres}}$ for some threshold parameter g_{thres}, and no noise has been added for a certain threshold t_{thres} number of iterations, Algorithm 4 would perturb the iterate as follows:

$$\theta_{k+1} = \theta_k + \zeta_k, \tag{7.17}$$

where ζ_k is extraneous noise that could be chosen from an isotropic distribution. In essence, Algorithm 4 performs regular GD between two instants where the parameter is perturbed. The t_{thres} parameter ensures that these instants are separated in time well enough.

For a careful choice of parameters g_{thres}, t_{thres}, f_{thres} and the distribution of ζ_k, it can be shown that perturbed GD finds an ϵ-SOSP in $O(\log^4(d)/\epsilon^2)$ iterations. The result below makes this claim precise.

Theorem 7.4 (Theorem 3 of (Jin *et al.*, 2017)). Suppose Assumptions A7.2 to A7.3 hold. Set perturbed GD algorithm's parameters as follows: $\chi = 3\max\{\log\left(\frac{d\ell\Delta_f}{c\epsilon^2\delta}\right), 4\}$, $a = \frac{c}{\ell}$, $g_{\text{thres}} =$

$\frac{\sqrt{c}}{\chi^2} \cdot \epsilon$, $f_{\text{thres}} = \frac{c}{\chi^3} \cdot \sqrt{\frac{\epsilon^3}{\rho}}$, $t_{\text{thres}} = \frac{\chi}{c^2} \cdot \frac{\ell}{\sqrt{\rho\epsilon}}$ $t_{\text{noise}} = -t_{\text{thres}} - 1$. Further, let the extraneous noise ζ_k in Algorithm 4 be sampled uniformly from the surface of sphere with radius $r = \frac{\sqrt{c}}{\chi^2} \cdot \frac{\epsilon}{\ell}$. Then, there exists an absolute constant $c_{\max}$ such that, for any $\delta > 0, \epsilon \leq \frac{\ell^2}{\rho}$, $\Delta_f \geq f(\theta_0) - f^\star$, and constant $c \leq c_{\max}$, perturbed GD will output an ϵ-SOSP, with probability $1 - \delta$, and terminate in the following number of iterations:

$$O\left(\frac{L_2(f(\theta_0) - f^\star)}{\epsilon^2} \log^4\left(\frac{dL_2\Delta_f}{\epsilon^2\delta}\right)\right).$$

Proof. We provide a brief sketch of the main proof ideas below. We refer the reader to (Jin *et al.*, 2017) for the complete proof.

Recall from the previous section that a GD step results in a function decrease given by

$$f(\theta_{k+1}) \leq f(\theta_k) - \frac{a}{2} \|\nabla f(\theta_k)\|^2 . \tag{7.18}$$

Next, if θ_k satisfies $\|\nabla f(\theta_t)\| \leq g_{\text{thres}}$ and $\lambda_{\min}(\nabla^2 f(\theta_k)) \leq -\sqrt{\rho\epsilon}$, then adding one perturbation step (7.17) followed by t_{thres} GD steps, we have

$$f(\theta_{k+t_{\text{thres}}}) - f(\theta_k) \leq -f_{t_{\text{thres}}} \text{ with high probability.}$$

Thus, if Algorithm 4 is at a saddle point, then perturb and GD steps ensure a decrease in function value.

Next, at an SOSP, Algorithm 4 would either remain there, which is the favorable case, or move away, in which case there is a function value decrease.

Thus, there is a function value decrease with GD/perturbed GD steps, and the total number of iterations can be inferred using the average function value decrease per iteration. This calculation would be along similar lines as in the previous section for Algorithm 3 in the sense that the maximum decrease is $f(\theta_0) - f^*$, and the algorithm either stops (with an SOSP), or continues to decrease the function value between iterations. □

We remark that Algorithm 4 has been extended to the case with stochastic gradients in (Jin *et al.*, 2021). In particular, the aforementioned reference considers a setting where the gradient estimates are unbiased, and the noise in these estimates satisfy a certain sub-Gaussianity requirement. Under these conditions, the authors establish convergence of a variant Algorithm 4 with noisy gradient estimates to an approximate SOSP with high probability. However, to the best of our knowledge, a similar result is not available for Algorithm 4 in the zeroth-order setting, where the gradient estimates have a bias-variance tradeoff.

7.4 Cubic-regularized stochastic Newton

The standard Newton step is given by

$$\theta_{k+1} = \theta_k - \nabla^2 f(\theta_k)^{-1} \nabla f(\theta_k).$$

This is equivalent to finding a θ that minimizes a second-order approximation, i.e., the following:

$$\theta_{k+1} = \quad \arg\min_{\theta \in \mathbb{R}^d} \left\{ \langle \nabla f(\theta_k), \theta - \theta_k \rangle + \frac{1}{2} \left\langle \nabla^2 f(\theta_k)(\theta - \theta_k), \theta - \theta_k \right\rangle \right\}.$$

For the case of a convex objective, the Newton method finds minima efficiently as compared to a gradient method, since the former uses a second-order approximation. However, for a non-convex objective, the Newton method may not necessarily escape from saddle points. A fix is to have an incremental algorithm that performs either a gradient or a Newton step adaptively, with the decision for the type of the step based on the gradient norm at the current point. In particular, gradient steps for large gradients and Newton steps otherwise. Such a two-step algorithm, analyzed for deterministic optimization in (Wright and Recht, 2022, Section 3.6), finds an ϵ-SOSP in $O\left(\frac{1}{\epsilon^3}\right)$ number of iterations, where each iteration is either a gradient or Newton step. An elegant alternative to achieve the same effect as the two-step algorithm discussed above is cubic regularization, which is described next.

The cubic regularized Newton step adds a cubic term to the auxiliary

Algorithm 5: Cubic-regularized stochastic Newton (CR-SN)

Input: Initial parameter $\theta_0 \in \mathbb{R}^d$, a non-negative sequence $\{\alpha_k\}$, positive integer sequences $\{m_k\}$ and $\{b_k\}$, and an iteration limit $N \geq 1$.

for $k = 1, \ldots, N$ **do**

`/* Gradient and Hessian estimation */`

Obtain m_k gradient estimates $\{\widehat{\nabla} f(\theta_k, l), l = 1, \ldots, m_k\}$;

Obtain b_k Hessian estimates $\{\widehat{\nabla}^2 f(\theta_k, l), l = 1, \ldots, b_k\}$;

Form estimates $\bar{g}_k, \bar{\mathcal{H}}_k$ as averages of the m_k gradient and b_k Hessian estimates, i.e.,

$$\bar{g}_k = \frac{1}{m_k} \sum_{l=1}^{m_k} \widehat{\nabla} f(\theta_k, l), \;\; \bar{\mathcal{H}}_k = \frac{1}{b_k} \sum_{l=1}^{b_k} \widehat{\nabla}^2 f(\theta_k, l).$$

`/* Cubic-regularized Newton step */`

Compute

$$\theta_k = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \tilde{f}^k(\theta) \equiv \tilde{f}(\theta, \theta_{k-1}, \bar{\mathcal{H}}_k, \bar{g}_k, \alpha_k) \right\}, \text{where}$$

$$\tilde{f}(x, y, \mathcal{H}, g, \alpha) = \langle g, x - y \rangle + \frac{1}{2} \langle \mathcal{H}(x - y), x - y \rangle + \frac{\alpha}{6} \|x - y\|^3 . \tag{7.19}$$

end for

Output: Parameter θ_N

function in the following manner:

$$\theta_{k+1} = \arg\min_{\theta \in \mathbb{R}^d} \Big\{ \langle \nabla f(\theta_k), \theta - \theta_k \rangle + \frac{1}{2} \left\langle \nabla^2 f(\theta_k)(\theta - \theta_k), \theta - \theta_k \right\rangle + \frac{\alpha}{6} \|\theta - \theta_k\|^3 \Big\},$$

where $\alpha \in \mathbb{R}^+$ is the regularization parameter. Since the gradient and Hessian of f are not directly available, we obtain $\widehat{\nabla} f(\theta_k, l), l = 1, \ldots, m_k$ estimates of the gradient and $\widehat{\nabla}^2 f(\theta_k, l), l = 1, \ldots, b_k$ estimates of the Hessian at θ_k. We use an average of these estimates, denoted by $\bar{g}_k$ and $\bar{\mathcal{H}}_k$, respectively, to solve the cubic sub-problem (7.19) in each

round of cubic-regularized stochastic Newton (CR-SN) algorithm, whose pseudocode is presented in Algorithm 5. The mean-squared error (MSE) of the gradient estimate $\bar{g}_k$ is $O\left(\frac{1}{m_k}\right)$, whereas the corresponding bound for the Hessian estimate $\bar{\mathcal{H}}_k$ is $O\left(\frac{1}{b_k}\right)$, see Lemma 7.7 below. We require the MSE to vanish asymptotically in order to ensure convergence of CR-SN to an ϵ-SOSP of the objective.

We consider CR-SN under two different settings. In the first setting, the gradient and Hessian estimates are unbiased, whereas in the second setting, these estimates are biased. In the next section, we establish convergence of CR-SN to an approximate SOSP in the first setting, and subsequently, extend the analyses to cover the second setting.

7.4.1 The case of unbiased gradient/Hessian information

In this setting, the gradient/Hessian estimates satisfy the following assumption:

A7.4. Let $\mathcal{F}_k = \sigma(\theta_i, i \leq k)$. Recall $\mathbb{E}_k$ denotes the expectation conditioned on $\mathcal{F}_k$. For any $k \geq 1$, we have

1. $\mathbb{E}_k\left[\widehat{\nabla} f(\theta_k)\right] = \nabla f\left(\theta_k\right), \mathbb{E}_k\left[\widehat{\nabla}^2 f(\theta_k)\right] = \nabla^2 f\left(\theta_k\right).$
2. $\mathbb{E}_k\left[\left\|\widehat{\nabla} f(\theta_k) - \nabla f\left(\theta_k\right)\right\|^2\right] \leq \sigma_1^2,$
   $\mathbb{E}_k\left[\left\|\widehat{\nabla}^2 f(\theta_k) - \nabla^2 f\left(\theta_k\right)\right\|^2\right] \leq \sigma_2^2,$ for some $\sigma_1, \sigma_2 \geq 0.$

The assumption above is satisfied in a risk-neutral RL setting, for instance, see (Maniyar *et al.*, 2024). On the other hand, in a risk-sensitive RL application, obtaining unbiased gradient/Hessian information is not feasible. Instead, one can use simultaneous perturbation-based gradient/Hessian estimators that are formed using function measurements. Such a setting involves biased gradient/Hessian estimates, which we shall analyze in the next section.

The result establishes convergence of Algorithm 5 to an ϵ-SOSP.

Theorem 7.5. Suppose Assumptions A7.4 and A7.2 hold. Let $\{\theta_1, \ldots, \theta_N\}$ be computed by Algorithm 5 with the following parameters:

$$\alpha_k = 3L_2, \quad N = \frac{12\sqrt{L_2}(f(\theta_0) - f^*)}{\epsilon^{\frac{3}{2}}}, \tag{7.20}$$

$$m_k = \frac{25\sigma_1^2}{4\epsilon^2}, \quad b_k = \frac{36\sqrt[3]{30(1+2\log 2d)}d^{\frac{2}{3}}\sigma_2^2}{L_2\epsilon}. \tag{7.21}$$

Let θ_R be picked uniformly at random from $\{\theta_1, \ldots, \theta_N\}$. Then,

$$5\sqrt{\epsilon} \geq \max\left\{\sqrt{\mathbb{E}\|\nabla f(\theta_R)\|}, \frac{-5}{6\sqrt{L_2}}\mathbb{E}\lambda_{\min}\left(\nabla^2 f(\theta_R)\right)\right\}, \tag{7.22}$$

where L_1 and L_2 are specified in Assumption A7.2.

As we reduce the parameter ϵ, the batch sizes m_k, b_k as well as the number of iterations N can be seen to increase. Alternatively, the batch sizes increase with N, and hence are not to be viewed as constants.

Proof of Theorem 7.5

The proof proceeds through a sequence of lemmas while following the technique from (Balasubramanian and Ghadimi, 2022a) and (Maniyar *et al.*, 2024).

Lemma 7.6. Let $\bar{\theta} = \arg\min_{x \in \mathbb{R}^d} \tilde{f}(x, \theta, \mathcal{H}, g, \alpha)$. Then, we have

$$g + \mathcal{H}(\bar{\theta} - \theta) + \frac{\alpha}{2}\left\|\bar{\theta} - \theta\right\|(\bar{\theta} - \theta) = 0, \tag{7.23}$$

$$\mathcal{H} + \frac{\alpha}{2}\left\|\bar{\theta} - \theta\right\| I_d \succeq 0. \tag{7.24}$$

where I_d is the identity matrix.

Proof. See (Nesterov and Polyak, 2006). □

The result below provides error bounds for the gradient and Hessian estimates, which are sample averages.

Lemma 7.7. Let $\bar{g}_k$ and $\bar{\mathcal{H}}_k$ be computed as in Algorithm 5, and assume $m_k \geq 1$, $b_k \geq 4(1 + 2\log 2d)$. Then,

$$\mathbb{E}\|\bar{g}_k - \nabla f(\theta_{k-1})\|^2 \leq \frac{\sigma_1^2}{m_k}, \tag{7.25}$$

$$\mathbb{E}\left\|\bar{\mathcal{H}}_k - \nabla^2 f(\theta_{k-1})\right\|^3 \leq \frac{4\sqrt{15(1+2\log 2d)}d\sigma_2^3}{b_k^{\frac{3}{2}}}. \tag{7.26}$$

Proof. Using Assumption A7.4, we have

$$\begin{aligned}&\mathbb{E}\|\bar{g}_k - \nabla f(\theta_{k-1})\|^2 \\ &= \mathbb{E}\left\|\frac{1}{m_k}\sum_{l=1}^{m_k}\left(\widehat{\nabla} f(\theta_{k-1}, l) - \nabla f(\theta_{k-1})\right)\right\|^2 \leq \frac{\sigma_1^2}{m_k}.\end{aligned}$$

This establishes the first bound in (7.25). Now we turn to proving the second bound in (7.25). By Theorem 1 in (Tropp, 2016), we have

$$\mathbb{E}\left\|\bar{\mathcal{H}}_k - \nabla^2 f(\theta_{k-1})\right\|^2 \leq \frac{2C(d)}{b_k^2}\left(\left\|\sum_{l=1}^{b_k}\mathbb{E}\Delta_{k,l}^2\right\| + C(d)\mathbb{E}\max_{l=1,\ldots,b_k}\|\Delta_{k,l}\|^2\right), \tag{7.27}$$

where $\Delta_{k,l} = \widehat{\nabla}^2 f(\theta_{k-1}, l) - \nabla^2 f(\theta_{k-1})$ and $C(d) = 4(1 + 2\log 2d)$. It is easy to see that

$$\mathbb{E}\|\Delta_{k,l}\|^2 \leq \mathbb{E}\left\|\widehat{\nabla}^2 f(\theta_{k-1}, l)\right\|^2 \leq \sigma_2^2, \quad \text{and} \tag{7.28}$$

$$\left\|\sum_{l=1}^{b_k}\mathbb{E}\Delta_{k,l}^2\right\| \leq \sum_{l=1}^{b_k}\left\|\mathbb{E}\Delta_{k,l}^2\right\| \leq \sum_{l=1}^{b_k}\mathbb{E}\|\Delta_{k,l}\|^2. \tag{7.29}$$

Using (7.28) and (7.29) in (7.27), we obtain

$$\mathbb{E}\left\|\bar{\mathcal{H}}_k - \nabla^2 f(\theta_{k-1})\right\|^2 \leq \frac{2C(d)}{b_k^2}\left(b_k\sigma_2^2 + C(d)\sigma_2^2\right) \leq \frac{4C(d)}{b_k}\sigma_2^2,$$

where in the last inequality we use the assumption that $b_k \geq C(d)$. Let $\|\cdot\|_F$ denote the Frobenius norm. Using Holder's inequality, we obtain

$$\begin{aligned}&\mathbb{E}\left\|\bar{\mathcal{H}}_k - \nabla^2 f(\theta_{k-1})\right\|^3 \\ &\leq \mathbb{E}\left\|\bar{\mathcal{H}}_k - \nabla^2 f(\theta_{k-1})\right\| \cdot \left\|\bar{\mathcal{H}}_k - \nabla^2 f(\theta_{k-1})\right\|_F^2\end{aligned}$$

$$\leq \left(\mathbb{E} \left\| \bar{\mathcal{H}}_k - \nabla^2 f(\theta_{k-1}) \right\|^2 \cdot \mathbb{E} \left\| \bar{\mathcal{H}}_k - \nabla^2 f(\theta_{k-1}) \right\|_F^4 \right)^{\frac{1}{2}}. \tag{7.30}$$

Note that $\bar{\mathcal{H}}_k - \nabla^2 f(\theta_{k-1}) = \frac{1}{b_k} \sum_{l=1}^{b_k} \Delta_{k,l}$. Hence, we have

$$\begin{aligned} \mathbb{E} \left\| \bar{\mathcal{H}}_k - \nabla^2 f(\theta_{k-1}) \right\|_F^4 &= \mathbb{E} \left\| \frac{1}{b_k} \sum_{l=1}^{b_k} \Delta_{k,l} \right\|_F^4 = \frac{1}{b_k^4} \mathbb{E} \left\| \sum_{l=1}^{b_k} \Delta_{k,l} \right\|_F^4 \\ &\leq \frac{3\mathbb{E} \|\Delta_{k,l}\|_F^4}{b_k^2}, \end{aligned}$$

where the final inequality comes from Rosenthal's inequality (cf. Lemma 16 in (Maniyar *et al.*, 2024)).

For a random matrix $Z \in \mathbb{R}^{d \times d}$, it can be shown that (see Lemma 15 in (Maniyar *et al.*, 2024))

$$\mathbb{E} \|Z - \mathbb{E}Z\|^4 \leq 5\mathbb{E} \|Z\|^4.$$

Using the inequality above in conjunction with the fact that $\|\cdot\|_F \leq \sqrt{d} \|\cdot\|$, we obtain the following for any $l \in \{1, \ldots, b_k\}$:

$$\begin{aligned} \mathbb{E} \left\| \bar{\mathcal{H}}_k - \nabla^2 f(\theta_{k-1}) \right\|_F^4 &\leq \frac{3d^2 \mathbb{E} \|\Delta_{k,l}\|^4}{b_k^2} \leq \frac{15d^2 \mathbb{E} \left\| \widehat{\nabla}^2 f(\theta_{k-1}, l) \right\|^4}{b_k^2} \\ &\leq \frac{15d^2 \sigma_2^4}{b_k^2}, \end{aligned}$$

which when combined with (7.30) leads to the second bound in (7.25). □

We next state a result that will be used in a subsequent lemma.

Lemma 7.8. If for any two matrices A and B, and a scalar c, we have[2]

$$A \preceq B + cI, \tag{7.31}$$

where I is the identity matrix of appropriate dimension, then the following holds:

$$c \geq \lambda_{max}(A) - \|B\|. \tag{7.32}$$

[2]Here, $A \succeq B$ denotes a matrix inequality in the positive semi-definite (p.s.d) sense, i.e., indicating $A - B$ is p.s.d.

Proof. See Appendix F of (Maniyar *et al.*, 2024). □

Lemma 7.9. Let $\{\theta_k\}$ be computed by Algorithm 5. Then, we have

$$\sqrt{\mathbb{E}\|\theta_k - \theta_{k-1}\|^2} \geq \max\left\{\sqrt{\frac{\mathbb{E}\|\nabla f(\theta_k)\| - \delta_k^g - \delta_k^{\mathcal{H}}}{L_2 + \alpha_K}}, \right. \\ \left. \frac{-2}{\alpha_k + 2L_2}\left[\mathbb{E}\lambda_{\min}\left(\nabla^2 f(\theta_k)\right) + \sqrt{2(\alpha_k + L_2)\delta_k^{\mathcal{H}}}\right]\right\},$$

where $\delta_k^g, \delta_k^{\mathcal{H}} > 0$ are chosen such that

$$\begin{aligned} &\mathbb{E}\|\nabla f(\theta_{k-1}) - \bar{g}_k\|^2 \leq \left(\delta_k^g\right)^2, \quad \text{and} \\ &\mathbb{E}\left\|\nabla^2 f(\theta_{k-1}) - \bar{\mathcal{H}}_k\right\|^3 \leq \left(2(L_2 + \alpha_k)\delta_k^{\mathcal{H}}\right)^{\frac{3}{2}}. \end{aligned} \tag{7.33}$$

Proof. Notice that

$$\begin{aligned} &\|\nabla f(\theta_k)\| \\ &\leq \left\|\nabla f(\theta_k) - \nabla f(\theta_{k-1}) - \nabla^2 f(\theta_{k-1})(\theta_k - \theta_{k-1})\right\| + \|\nabla f(\theta_{k-1}) - \bar{g}_k\| \\ &+ \left\|\nabla^2 f(\theta_{k-1}) - \bar{\mathcal{H}}_k\right\| \|\theta_k - \theta_{k-1}\| + \frac{\alpha_k}{2}\|\theta_k - \theta_{k-1}\|^2 \\ &\leq \frac{(L_2 + \alpha_k)}{2}\|\theta_k - \theta_{k-1}\|^2 + \|\nabla f(\theta_{k-1}) - \bar{g}_k\| \\ &\quad + \left\|\nabla^2 f(\theta_{k-1}) - \bar{\mathcal{H}}_k\right\| \|\theta_k - \theta_{k-1}\| \\ &\leq (L_2 + \alpha_k)\|\theta_k - \theta_{k-1}\|^2 + \|\nabla f(\theta_{k-1}) - \bar{g}_k\| + \frac{\left\|\nabla^2 f(\theta_{k-1}) - \bar{\mathcal{H}}_k\right\|^2}{2(L_2 + \alpha_k)}, \end{aligned}$$

where we used Young's inequality in the last step. Taking expectations and using (7.33), we have

$$\frac{(\mathbb{E}\|\nabla f(\theta_k)\| - \delta_k^g - \delta_k^{\mathcal{H}})}{L_2 + \alpha_k} \leq \mathbb{E}\|\theta_k - \theta_{k-1}\|^2. \tag{7.34}$$

By the inequality in Lemma 7.23, and the fact that f is smooth by Assumption A7.2, we have

$$\begin{aligned} \nabla^2 f(\theta_k) &\succeq \nabla^2 f(\theta_{k-1}) - L_2\|\theta_k - \theta_{k-1}\| I_d \\ &= \nabla^2 f(\theta_{k-1}) - \bar{\mathcal{H}}_k + \bar{\mathcal{H}}_k - L_2\|\theta_k - \theta_{k-1}\| I_d \end{aligned}$$

$$\succeq \nabla^2 f(\theta_{k-1}) - \bar{\mathcal{H}}_k - \frac{(\alpha_k + 2L_2)\,\|\theta_k - \theta_{k-1}\|}{2} I_d,$$

implying

$$\frac{(\alpha_k + 2L_2)\,\|\theta_k - \theta_{k-1}\|}{2} \geq \lambda_{min}(\nabla^2 f(\theta_{k-1}) - \bar{\mathcal{H}}_k) - \lambda_{\min}\left(\nabla^2 f(\theta_k)\right). \tag{7.35}$$

Taking expectations on both sides, and using the definition of $\delta_k^{\mathcal{H}}$ in (7.33), we have

$$\sqrt{\mathbb{E}\|\theta_k - \theta_{k-1}\|^2} \geq \mathbb{E}\|\theta_k - \theta_{k-1}\| \tag{7.36}$$

$$\geq \frac{-2}{\alpha_k + 2L_2}\left[\mathbb{E}\lambda_{\min}\left(\nabla^2 f(\theta_k)\right) + \sqrt{2(\alpha_k + L_2)\delta_k^{\mathcal{H}}}\right]. \tag{7.37}$$

The main claim follows by combining the above inequality with (7.34). □

Lemma 7.10. Let $\{\theta_k\}$ be computed by Algorithm 5 for a given iteration limit $N \geq 1$. Then,

$$\mathbb{E}\|\theta_R - \theta_{R-1}\|^3 \leq \frac{36}{\sum_{k=1}^N \alpha_k} \times \left[f(\theta_0) - f^* + \sum_{k=1}^N \frac{4\left(\delta_k^g\right)^{\frac{3}{2}}}{\sqrt{3\alpha_k}} + \sum_{k=1}^N \left(\frac{18\sqrt[4]{2}}{\alpha_k}\right)^2 \left((L_2 + \alpha_k)\delta_k^{\mathcal{H}}\right)^{\frac{3}{2}}\right], \tag{7.38}$$

where R is a random variable whose probability distribution $P_R(\cdot)$ is supported on $\{1, \ldots, N\}$ and given by

$$P_R(R = k) = \frac{\alpha_k}{\sum_{k=1}^N \alpha_k}, \qquad k = 1, \ldots, N, \tag{7.39}$$

and $\delta_k^g, \delta_k^{\mathcal{H}} > 0$ are defined as before in (7.33).

Proof. Using Assumption A7.2, (7.19) and the fact that $\alpha_k \geq L_2$, we have

$$f(\theta_k) \leq f(\theta_{k-1}) + \tilde{f}^k(\theta_k) + \|\nabla f(\theta_{k-1}) - \bar{g}_k\|\,\|\theta_k - \theta_{k-1}\| + \frac{1}{2}\left\|\nabla^2 f(\theta_{k-1}) - \bar{\mathcal{H}}_k\right\|\|\theta_k - \theta_{k-1}\|^2. \tag{7.40}$$

Further,

$$\begin{aligned}\tilde{f}^k(\theta_k) &= -\frac{1}{2}\left\langle \bar{\mathcal{H}}_k(\theta_k - \theta_{k-1}), (\theta_k - \theta_{k-1})\right\rangle - \frac{\alpha_k}{3}\left\|\theta_k - \theta_{k-1}\right\|^3 \\ &\leq -\frac{\alpha_k}{12}\left\|\theta_k - \theta_{k-1}\right\|^3. \end{aligned} \tag{7.41}$$

Combining (7.40) and (7.41), we obtain

$$\begin{aligned}\frac{\alpha_k}{12}\left\|\theta_{k-1} - \theta_k\right\|^3 &\leq f(\theta_{k-1}) - f(\theta_k) + \left\|\nabla f(\theta_{k-1}) - \bar{g}_k\right\| \left\|\theta_k - \theta_{k-1}\right\| \\ &\quad + \frac{1}{2}\left\|\nabla^2 f(\theta_{k-1}) - \bar{\mathcal{H}}_k\right\| \left\|\theta_k - \theta_{k-1}\right\|^2 \\ &\leq f(\theta_{k-1}) - f(\theta_k) + \frac{4}{\sqrt{3\alpha_k}}\left\|\nabla f(\theta_{k-1}) - \bar{g}_k\right\|^{\frac{3}{2}} \\ &\quad + \left(\frac{9\sqrt{2}}{\alpha_k}\right)^2 \left\|\nabla^2 f(\theta_{k-1}) - \bar{\mathcal{H}}_k\right\|^3 + \frac{\alpha_k}{18}\left\|\theta_k - \theta_{k-1}\right\|^3, \end{aligned} \tag{7.42}$$

where the last inequality follows from the fact $ab \leq \dfrac{a^p}{\lambda^p p} + \dfrac{\lambda^q b^q}{q}$ for p, q satisfying $\dfrac{1}{p} + \dfrac{1}{q} = 1$ and $\lambda > 0$.

We now take expectation on both sides of (7.42) and use (7.33) to obtain

$$\begin{aligned}&\frac{\alpha_k}{36}\mathbb{E}\left\|\theta_k - \theta_{k-1}\right\|^3 \\ &\leq f(\theta_{k-1}) - f(\theta_k) + \frac{4\left(\delta_k^g\right)^{\frac{3}{2}}}{\sqrt{3\alpha_k}} + \left(\frac{18\sqrt[4]{2}}{\alpha_k}\right)^2 \left((L_2 + \alpha_k)\delta_k^{\mathcal{H}}\right)^{\frac{3}{2}}. \end{aligned} \tag{7.43}$$

Summing over $k = 1, \ldots, N$, dividing both sides by $\displaystyle\sum_{k=1}^{N} \alpha_k$ and noting (7.39), we obtain the bound in (7.38). □

Proof of Theorem 7.5

Proof. First, note that by (7.20), Lemma 7.25, we can ensure that (7.33) is satisfied by $\delta_k^g = 2\epsilon/5$ and $\delta_k^{\mathcal{H}} = \epsilon/144$. Moreover, by Lemma 7.10, we

have

$$\mathbb{E}\|\theta_R - \theta_{R-1}\|^3 \leq \frac{12}{L_2}\left[\frac{f(\theta_0) - f^*}{N} + \frac{4\,(2/5)^{\frac{3}{2}}}{3\sqrt{L_2}}\epsilon^{\frac{3}{2}} + \frac{18^2\sqrt{2}}{9\cdot 6^3\sqrt{L_2}}\epsilon^{\frac{3}{2}}\right] \tag{7.44}$$

$$\leq \frac{1}{L_2^{\frac{3}{2}}}\left[\frac{12\sqrt{L_2}(f(\theta_0) - f^*)}{N} + 6.88\epsilon^{\frac{3}{2}}\right]$$

$$\leq \frac{8\epsilon^{\frac{3}{2}}}{L_2^{\frac{3}{2}}}. \tag{7.45}$$

The inequality in (7.45) follows by substituting the value of N specified in the theorem statement. Furthermore, from Lemma 7.9 and using Lyapunov inequality i.e.,

$$\left[\mathbb{E}\|\theta_R - \theta_{R-1}\|^2\right]^{1/2} \leq \left[\mathbb{E}\|\theta_R - \theta_{R-1}\|^3\right]^{1/3} \leq \frac{2\epsilon^{\frac{1}{2}}}{L_2^{\frac{1}{2}}}.$$

Using the bound above in conjunction with (7.34) and (7.36), we obtain

$$\sqrt{\mathbb{E}\|\nabla f(\theta_k)\|} \leq \sqrt{\left(16 + \frac{2}{5} + \frac{1}{144}\right)\epsilon} \leq 5\sqrt{\epsilon},$$

and

$$\frac{\mathbb{E}\left[-\lambda_{\min}\left(\nabla^2 f(\theta_k)\right)\right]}{\sqrt{L_2}} \leq \left(5 + \frac{1}{3\sqrt{2}}\right)\sqrt{\epsilon} \leq 6\sqrt{\epsilon}.$$

The main result in (7.22) follows from the two inequalities above.

Finally, note that the total number of required samples to obtain such a solution is bounded by

$$\sum_{k=1}^{N} m_k = O\left(\frac{1}{\epsilon^{\frac{7}{2}}}\right), \qquad \sum_{k=1}^{N} b_k = O\left(\frac{d^{\frac{2}{3}}}{\epsilon^{\frac{5}{2}}}\right).$$

□

7.4.2 The case of biased gradient/Hessian information

We now consider the case where Assumption A7.4 does not hold. Instead, an algorithm has access to zeroth-order observations.

For simplicity, we consider the setting where $f(\theta) = \mathbb{E}\left[F(\theta, \xi)\right]$, and the sample performance F is smooth, as specified in Assumption A5.4. For this setting, we employ the Gaussian smoothing approach to form the gradient and Hessian estimates in Algorithm 5. Let

$$\widehat{\nabla} f(\theta) = \Delta \left[\frac{F\left(\theta + \delta\Delta, \xi\right) - F\left(\theta, \xi\right)}{\delta} \right], \tag{7.46}$$

$$\widehat{\nabla}^2 f(\theta) = \Delta \left[\frac{F\left(\theta + \delta\Delta, \xi\right) + F\left(\theta - \delta\Delta, \xi\right) - 2F\left(\theta, \xi\right)}{2\delta^2} \left(\Delta\Delta^\intercal - I\right) \right]. \tag{7.47}$$

In the above, we have used common random noise to form the gradient estimate $\widehat{\nabla} f(\theta)$ and Hessian estimate $\widehat{\nabla}^2 f(\theta)$. In Algorithm 5, we require sample averages of these quantities. Let $\widehat{\nabla} f(\theta, l)$, $l = 1, \ldots, m$, and $\widehat{\nabla}^2 f(\theta, l)$, $l = 1, \ldots, b$, denote m and b independent samples of the quantities defined in (7.46) and (7.47), respectively. Then, as in Algorithm 5, we form the following sample average estimates, but with the difference that the individual gradient/Hessian estimates are biased:

$$\bar{g}_k = \frac{1}{m_k} \sum_{l=1}^{m_k} \widehat{\nabla} f(\theta_k, l), \;\; \bar{\mathcal{H}}_k = \frac{1}{b_k} \sum_{l=1}^{b_k} \widehat{\nabla}^2 f(\theta_k, l). \tag{7.48}$$

It can be shown that the averaged gradient estimate $\bar{g}_k$ and Hessian estimate $\bar{\mathcal{H}}_k$ satisfy the following bounds:

$$\begin{aligned} \mathbb{E}\|\bar{g}_k - \nabla f(\theta_{k-1})\|^2 &\leq \frac{2(d+5)(B^2+\sigma^2)}{m_k} + \frac{\delta^2 L^2 (d+3)^3}{2m_k}, \\ \mathbb{E}\left\|\bar{\mathcal{H}}_k - \nabla^2 f(\theta_{k-1})\right\|^2 &\leq \frac{240\sqrt{15(1+2\log 2d)}(d+16)^3 L^3}{b_k^{\frac{3}{2}}} \\ &\quad + 3L_2^2 (d+16)^5 \delta^2. \end{aligned} \tag{7.49}$$

The reader is referred to Lemmas 1 and 8 of (Balasubramanian and Ghadimi, 2022a) for the proof.

Next, by using completely parallel arguments to the proof of Theorem 7.5, with the bounds in (7.49) replacing those in Lemma 7.7, one can establish convergence to ϵ-SOSP guarantee within $O\left(\frac{1}{\epsilon^{\frac{3}{2}}}\right)$ number of iterations, which in turn translates to $O\left(\frac{1}{\epsilon^{\frac{7}{2}}}\right)$ gradient evaluations

and $O\left(\frac{d^{\frac{2}{3}}}{\epsilon^{\frac{5}{2}}}\right)$ Hessian evaluations.

7.5 Bibliographic remarks

7.1 First-order stationary points are standard in optimization literature and have been the topic of analysis in several papers involving stochastic gradient algorithms, cf. (Ghadimi and Lan, 2013; Bhavsar and Prashanth, 2022). The SOSP notion is based on (Nesterov and Polyak, 2007), and this notion has been used extensively in ML literature over the last decade, cf. (Jin *et al.*, 2017).

7.2 Avoidance of traps for a general stochastic approximation algorithm has received a lot of research attention, cf. (Pemantle, 1990; Brandiere and Duflo, 1996; Borkar, 2003; Barakat *et al.*, 2021; Gadat and Gavra, 2022). In (Borkar, 2003), an estimate for the lock-in probability, i.e., probability of convergence to an attractor given that the iterate-sequence is in its domain of attraction after a sufficiently long time is obtained and this is then used to argue an avoidance of traps result. In the case when the iterate-sequence has Markov noise in addition, (Karmakar and Bhatnagar, 2021) derive a lock-in probability lower bound while such bounds in the case of stochastic recursive inclusions (involving set-valued maps) are obtained in (Yaji and Bhatnagar, 2019). Our treatment in Section 7.2 leading to the traps avoidance claim in Theorem 7.1 for a SG algorithm with the unified gradient estimate is an adaptation of the corresponding result in (Mondal *et al.*, 2024).

In relation to the algorithm (7.6), an interesting early work is (Gelfand and Mitter, 1991) that builds on ideas from simulated annealing (Kirkpatrick *et al.*, 1983). In the context of our setting, the following recursion is considered:

$$\theta_{n+1} = \theta_n - a(n)\hat{\nabla} f(\theta_n) + b(n)\zeta_n,$$

where f is in general a C^2 and non-convex function satisfying certain additional conditions. Further, $a(n) = A/n$ and $b(n) =$

$\sqrt{B}/\sqrt{n \log\log n}, n \geq 1$, with $A, B > 0$, are two step-size schedules and $\{\zeta_n\}$ is a sequence of independent Gaussian vectors with zero mean and covariance matrix as the identity matrix. By analyzing an underlying stochastic differential equation, it is shown under some conditions in (Gelfand and Mitter, 1991), that the parameter sequence $\{\theta_n\}$ converges in probability to the set of global minima of the function f by avoiding convergence to local minima. This approach is thus a powerful technique to obtain asymptotic convergence to global minima though it can be slow in practice. Finally, in (Maryak and Chin, 2001), two-measurement SPSA estimates have also been used for $\hat{\nabla} f(\theta)$ and convergence to global minima claimed using the result in (Gelfand and Mitter, 1991). For a sub-class of non-convex objective functions, it is possible to obtain global convergence guarantees, without addition of extraneous noise. As an example, the reader is referred to (Karandikar and Vidyasagar, 2024), where the authors establish global convergence guarantees for "invex" functions, whose stationary points are global minimizers.

7.3 The two part algorithm in Section 7.3.1 is based on Section 3.6 of (Wright and Recht, 2022). The perturbed GD algorithm in Section 7.3.2 is based on (Jin *et al.*, 2017).

7.4 Cubic-regularized Newton algorithm was first proposed in (Nesterov and Polyak, 2007) in the context of deterministic optimization. Subsequently, it was analyzed in the stochastic optimization setting with unbiased gradient/Hessian information in (Tripuraneni *et al.*, 2018). Extension of stochastic cubic-regularized Newton to a zeroth-order setting was done in (Balasubramanian and Ghadimi, 2022a). A more recent RL application of the cubic-regularized Newton approach in the context of policy gradient methods is (Maniyar *et al.*, 2024).

The auxiliary problem (7.19) can be solved efficiently using gradient descent, see (Carmon *et al.*, 2016; Tripuraneni *et al.*, 2018; Maniyar *et al.*, 2024) for the details. Moreover, computationally efficient extensions to a setting where the objective is approximated

using a neural network is feasible with Hessian-vector products, see (Maniyar *et al.*, 2024).

8

Applications to reinforcement learning

Reinforcement learning (RL) refers to a class of model-free algorithms that have become widely popular for problems of decision making under uncertainty. In most real-life situations, it is hard to have precise knowledge of the system model, and this is where RL algorithms are immensely useful as these are largely data-driven algorithms.

In this chapter, we consider specifically two settings of reinforcement learning algorithms, both involving policy gradient (PG) based approaches (Sutton *et al.*, 1999) for reinforcement learning. These algorithms typically assume that the decision making policy is parameterized and have traditionally involved obtaining unbiased gradient estimators of the performance objective, many times the value function, w.r.t the aforementioned policy parameters. However, recent work suggests that zeroth order gradient estimators can result in improved performance, see for instance, (Salimans *et al.*, 2017; Mania *et al.*, 2018). In many other situations, such as in the case of actor-critic algorithms that also involve policy gradient algorithms but where the algorithm's parameters are updated as soon as a data sample becomes available, using zeroth order methods prove to be particularly effective, see for instance, (Bhatnagar and Kumar, 2004; Abdulla and Bhatnagar, 2007).

8.1 REINFORCE with an SPSA Gradient Estimate

REINFORCE is one of the popular algorithms in reinforcement learning that is based on Monte-Carlo or trajectory-based estimates of the value function. It has been first presented in (Williams, 1992), see also (Sutton and Barto, 2018, Chapter 13). We however study an application of zeroth-order gradient estimation in this setting. The work that we present in this section is based on (Bhatnagar, 2023).

8.1.1 The Basic Setting

By a Markov decision process, we mean a controlled stochastic process $\{X_n\}$ whose evolution is governed by an associated control-valued sequence $\{Z_n\}$. It is assumed that the random variables $X_n, n \geq 0$ take values in a set S called the state-space. Let $A(s)$ denote the set of all feasible actions in state $s \in S$ and $A \stackrel{\triangle}{=} \cup_{s \in S} A(s)$ denote the set of all actions. When the state (X_n) is say s and a feasible action a is chosen, the next state (X_{n+1}) seen is s' with a probability $p(s'|s,a) \stackrel{\triangle}{=} P(X_{n+1} = s' \mid X_n = s, Z_n = a)$, $\forall n$. We assume these probabilities do not depend on n. Such a process satisfies the controlled Markov property, i.e.,

$$P(X_{n+1} = s' \mid X_n, Z_n, \ldots, X_0, Z_0) = p(s' \mid X_n, Z_n) \text{ a.s.}$$

By an admissible policy or simply a policy, we mean a sequence of functions $\pi = \{\mu_0, \mu_1, \mu_2, \ldots\}$, with each $\mu_i : S \to A$, $i \geq 0$, such that $\mu_i(s) \in A(s)$, $\forall s \in S$. The policy π is a decision rule which specifies that if at instant k, the state is i, then the action chosen under π would be $\mu_k(i)$. A stationary policy π is one for which $\mu_k = \mu_l \stackrel{\triangle}{=} \mu$, $\forall k, l = 0, 1, \ldots$. In other words, under a stationary policy, the function that decides the action-choice in a given state does not depend on time instant n. Many times, instead of calling $\pi = \{\mu, \mu, \mu, \ldots\}$ a stationary policy, we simply refer to the function μ itself as the stationary policy.

Associated with any transition to a state s' from a state s under action a, is a 'single-stage' cost $g(s,a,s')$ where $g : S \times A \times S \to \mathbb{R}$ is called the cost function. The goal of the decision maker is to select actions $a_k, k \geq 0$, in response to the system states $s_k, k \geq 0$, so as to

minimize a long-term cost objective. We assume here that the number of states and actions is finite. In particular, we let $1, \ldots, p$ denote the set of non-terminal or regular states and t be the terminal or goal state. We let $S = \{1, 2, \ldots, p\}$ denote the set of all non-terminal states. Further, let $S^+ = \{1, \ldots, p, t\} = S \cup \{t\}$.

In this section, we are concerned with the stochastic shortest path problem, see (Bertsekas, 2012; Bertsekas, 2019), where we assume that under any policy, there is a positive probability of hitting the terminal state in at most p steps starting from any initial state, that would in turn signify that the problem would terminate in a finite though random amount of time. Such policies are called proper policies (see Definition 8.1 and Assumption A8.1).

Under a given policy π, define

$$V_\pi(s) = E_\pi \left[\sum_{k=0}^{T} g(X_k, \mu_k(X_k), X_{k+1}) \mid X_0 = s \right], \ s \in S,$$

where $0 < T < \infty$ is a finite random time at which the process enters the terminal state. Here $E_\pi[\cdot]$ indicates that all actions are chosen according to policy π depending on the system state. We assume that there is no action that is feasible in the terminal state t and thus once the process reaches t, it terminates.

Let Π denote the set of all admissible policies. The goal here is to find the optimal value function $V^*(i), i \in S$, defined by

$$V^*(i) = \min_{\pi \in \Pi} V_\pi(i) = V_{\pi^*}(i), \ i \in S,$$

where π^* denotes the optimal policy, i.e., the one that minimizes $V_\pi(i)$, $i \in S$, over all policies π. A related goal here would be to find the policy π^*. It turns out that in these problems, there exist stationary policies that are optimal. Thus, it is sufficient to search for an optimal policy within the class of stationary policies.

Definition 8.1. A stationary policy μ is called a proper policy if

$$\hat{p}_\mu \stackrel{\triangle}{=} \max_{s=1,\ldots,p} P(X_p \neq t \mid X_0 = s, \mu) < 1.$$

In other words, regardless of the initial state s (assumed non-terminal for obvious reasons), there is a positive probability of termination after at most p stages when using a proper policy.

Assuming that all stationary policies are proper, the optimal value function satisfies the Bellman equation

$$V^*(s) = \min_{a \in A(s)} \sum_{j \in S^+} p(j \mid s, a)(g(s, a, j) + V^*(j)), \ s \in S, \tag{8.1}$$

where by convention $V^*(t) = 0$. It can be shown, see (Bertsekas, 2012), that an optimal stationary proper policy exists.

An admissible policy (and so also a stationary policy) can be randomized as well. A randomized admissible policy or simply a randomized policy is a sequence of distributions $\psi = \{\phi_0, \phi_1, \ldots\}$ with each $\phi_i : S \to P(A)$. Even though A is the set of all actions, for any $s \in S$, a randomized policy would provide a distribution $\phi_i(s) = (\phi_i(s, a), a \in A(s))$ with $\phi_i(s, a) \geq 0, \forall a \in A(s)$, and $\sum_{a \in A(s)} \phi_i(s, a) = 1$. This also implies that $\phi_i(s, a) = 0$, $\forall a \notin A(s)$. A stationary randomized policy is one for which $\phi_j = \phi_k \stackrel{\triangle}{=} \phi$, $\forall j, k = 0, 1, \ldots$. In this case, we simply call ϕ to be a stationary randomized policy. By the foregoing, since an optimal stationary proper policy exists, an optimal stationary randomized policy that is also proper would exist as well.

8.1.2 The Reinforcement Learning Problem

We consider now the RL setting where we do not assume any knowledge of the system model, i.e., the transition probabilities $p(s' \mid s, a)$, and in their place, we assume that we have access to data (either real or simulated). The data that is available is over trajectories of states, actions, single-stage costs and next states until termination.

We assume that trajectories of states and actions are available either as real data or from a simulation device. Let G_k denote the sum of costs until termination on a trajectory starting from instant k. In other words, $G_k = \sum_{j=k}^{T-1} g_j$ where $g_j \equiv g(s_j, a_j, s_{j+1})$, where s_j, s_{j+1} are the states visited on this trajectory at time instants $j, j+1$, and a_j is the

action chosen at instant j in the trajectory. Note that if all actions are chosen according to a policy ϕ, then the value function (under ϕ) would be obtained as

$$V_\phi(s) = E_\phi[G_k \mid X_k = s],\ s \in S. \tag{8.2}$$

Further, by convention (cf. Bertsekas, 2019; Sutton and Barto, 2018), we let $V_\phi(t) = 0$, for any policy ϕ.

We consider here a class of stationary randomized policies that are parameterized by a parameter $\theta = (\theta_1, \ldots, \theta_d)^\mathsf{T} \in C \subset \mathbb{R}^d$ where C is a compact and convex subset of $\mathbb{R}^d$. We shall denote such a policy $\phi_\theta \stackrel{\triangle}{=} (\phi_\theta(s), s \in S)$, where for any $s \in S$, $\phi_\theta(s) = (\phi_\theta(s,a), a \in A(s))$ is a distribution over $A(s)$ when θ is the given parameter. We make the following assumption:

A8.1. All stationary randomized policies ϕ_θ parameterized by $\theta \in C$ are proper.

The REINFORCE algorithm of (Sutton and Barto, 2018) is a Monte-Carlo procedure based on the policy gradient method. The original algorithm uses a procedure for estimating the performance gradient that is based on an interchange of the gradient and expectation operators. We apply here a two-simulation but one-sided SPSA-based procedure for estimating the performance gradient that does not require the aforementioned interchange of operators. As mentioned, this procedure will however require two system simulations. We explain the algorithm in more detail below.

Let $\Gamma : \mathcal{R}^d \to C$ denote a projection operator that projects any $x = (x_1, \ldots, x_d)^\mathsf{T} \in \mathcal{R}^d$ to its nearest point in C. Thus, if $x \in C$, then $\Gamma(x) \in C$ as well. For ease of exposition, let's assume that C is a d-dimensional rectangle having the form $C = \prod_{i=1}^{d} [a_{i,\min}, a_{i,\max}]$, where $-\infty < a_{i,\min} < a_{i,\max} < \infty$, $\forall i = 1, \ldots, d$. A convenient way to identify $\Gamma(x)$ is as follows: Note that $\Gamma(x) = (\Gamma_1(x_1), \ldots, \Gamma_N(x_N))^\mathsf{T}$, where the ith component operator $\Gamma_i : \mathcal{R} \to [a_{i,\min}, a_{i,\max}]$ is specified by $\Gamma_i(x_i) = \min(a_{i,\max}, \max(a_{i,\min}, x_i))$, $i = 1, \ldots, d$. Also, let $\mathcal{C}(C)$ denote the space of all continuous functions from C to $\mathcal{R}^d$.

Let $\theta(n)$ and $(\theta(n) + \delta\Delta(n)), n \geq 0$ be two parameter sequences where $\theta(n) = (\theta_1(n), \ldots, \theta_d(n))^\mathsf{T} \in \mathcal{R}^d$, $\delta > 0$ is a small constant and $\Delta(n) = (\Delta_1(n), \ldots, \Delta_d(n))^\mathsf{T}, n \geq 0$, with $\Delta_i(n), i = 1, \ldots, d, n \geq 0$ being independent random variables distributed according to $\Delta_i(n) = \pm 1$ w.p. 1/2. The updates $\theta(n)$ of the parameter θ are obtained using an algorithm that will be explained below.

Algorithm (8.3) below is used to update the parameter $\theta \in C \subset \mathbb{R}^d$. For a given $n \geq 0$, let χ^n and χ^{n+} respectively denote the state-action trajectories $\chi^n = \{s_0^n, a_0^n, s_1^n, a_1^n, \ldots, s_{T-1}^n, a_{T-1}^n, s_T^n\}$ and $\chi^{n+} = \{s_0^{n+}, a_0^{n+}, s_1^{n+}, a_1^{n+}, \ldots, s_{T^+-1}^{n+}, a_{T^+-1}^{n+}, s_{T^+}^{n+}\}$, respectively, where χ^n is governed by the parameter $\theta(n)$ and χ^{n+} is governed by $\theta(n) + \delta\Delta(n)$. The instant T (resp. T^+) denotes the termination instant in the trajectory χ^n (resp. χ^{n+}). Thus, in both χ^n, χ^{n+}, $s_T^n = s_{T+}^{n+} = t$, i.e., each episode ends once the terminal or goal state is reached. Note that the various actions in the trajectory χ^n are chosen according to the policy $\phi_{\theta(n)}$ (depending on the states visited in the trajectory). Similarly, the actions in the trajectory χ^{n+} are chosen according to the policy $\phi_{\theta(n)+\delta\Delta(n)}$. The initial states in the two trajectories are kept the same, i.e., $s_0^n = s_0^{n+}$, and sampled from a given initial distribution $\nu = (\nu(i), i \in S)$ over states.

Let $G^n = \sum_{k=0}^{T-1} g_k^n$ and $G^{n+} = \sum_{k=0}^{T^+-1} g_k^{n+}$ denote the sums of costs until termination on the two trajectories that are governed with parameters $\theta(n)$ and $\theta(n) + \delta\Delta(n)$, respectively, where $g_k^n \equiv g(s_k^n, a_k^n, s_{k+1}^n)$ and $g_k^{n+} \equiv g(s_k^{n+}, a_k^{n+}, s_{k+1}^{n+})$.

The update rule that we consider here is the following:

For $n \geq 0, i = 1, \ldots, d$,

$$\theta_i(n+1) = \Gamma_i\left(\theta_i(n) - a(n)\left(\frac{G^{n+} - G^n}{\delta\Delta_i(n)}\right)\right). \tag{8.3}$$

We assume here that $\{a(n)\}$ satisfy the following assumption:

A8.2. The step-size sequence $\{a(n)\}$ satisfies $a(n) > 0$, $\forall n$. Further,

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

A8.3. The stationary randomized policy ϕ_θ, $\forall \theta \in C$ is twice continuously differentiable in θ and has a bounded third derivative.

While A8.2 is a standard requirement on the step-sizes, see the previous chapters, A8.3 is a requirement on the smoothness of the parameterized stationary randomized policies that can be seen to be satisfied by many classes of policies. For instance, the popularly used parameterized Gibbs or Boltzmann policy given by

$$\phi_\theta(s,a) = \frac{\exp(\theta^\intercal \psi_{s,a})}{\sum_{b\in A(s)} \exp(\theta^\intercal \psi_{s,b})},$$

with prescribed state-action features $\psi_{s,a} \in \mathbb{R}^d$, $s \in S$, $a \in A(s)$, can be seen to satisfy this requirement.

As soon as a parameter update is available, two trajectories – governed by the nominal and perturbed parameters, respectively, are generated with the initial state in the perturbed trajectory the same as that in the nominal trajectory and with the initial state sampled according to a given distribution ν.

8.1.3 Convergence Analysis

We begin by rewriting the algorithm (8.3) as follows:

$$\theta_i(n+1) = \Gamma_i\left(\theta_i(n) - a(n)E\left[\frac{G^{n+} - G^n}{\delta\Delta_i(n)} \mid \mathcal{F}_n\right] + M^i_{n+1}\right), \quad (8.4)$$

where

$$M^i_{n+1} = \frac{G^{n+} - G^n}{\delta\Delta_i(n)} - E\left[\frac{G^{n+} - G^n}{\delta\Delta_i(n)} \mid \mathcal{F}_n\right].$$

Here, we let $\mathcal{F}_n \stackrel{\triangle}{=} \sigma(\theta(m), m \le n, \Delta(m), \chi^m, \chi^{m+}, m < n), n \ge 1$ be a sequence of increasing sigma fields and with $\mathcal{F}_0 = \sigma(\theta(0))$. Let $M_n = (M^1_n, \ldots, M^d_n)^\intercal$, $n \ge 0$. Here we let $\|\cdot\|$ denote the Euclidean norm.

Lemma 8.1. $(M_n, \mathcal{F}_n), n \ge 0$ is a martingale difference sequence.

Proof. Notice that

$$M_n = \frac{G^{(n-1)+} - G^{(n-1)}}{\delta\Delta_i(n-1)} - E\left[\frac{G^{(n-1)+} - G^{(n-1)}}{\delta\Delta_i(n)} \mid \mathcal{F}_{n-1}\right].$$

The first term on the RHS above is clearly measurable $\mathcal{F}_n$ while the second term is measurable $\mathcal{F}_{n-1}$ and hence measurable $\mathcal{F}_n$ as well. Further, from Assumption A8.1, each M_n is integrable. Finally, it is easy to verify that

$$E[M_{n+1} \mid \mathcal{F}_n] = 0.$$

The claim follows. □

In the following, for simplicity, we denote $V_{\phi_\theta}(s)$ as $V_\theta(s)$ itself for any $\theta \in C$. If ϕ_θ is a twice continuously differentiable function of θ, it can be shown that $V_\theta(s)$ is also a twice continuously differentiable function of θ for any state s.

Proposition 8.1. We have

$$E\left[\frac{G^{n+} - G^n}{\delta\Delta_i(n)} \mid \mathcal{F}_n\right] = \sum_{s\in S} \nu(s)\nabla_i V_{\theta(n)}(s) + o(\delta) \text{ a.s.}$$

Proof. Note that

$$E\left[\frac{G^{n+} - G^n}{\delta\Delta_i(n)} \mid \mathcal{F}_n\right] = E\left[E\left[\frac{G^{n+} - G^n}{\delta\Delta_i(n)} \mid \mathcal{G}_n\right] \mid \mathcal{F}_n\right],$$

where $\mathcal{G}_n \stackrel{\triangle}{=} \sigma(\theta(m), \Delta(m), m \leq n, \chi^m, \chi^{m+}, m < n), n \geq 1$ be a sequence of increasing sigma fields with $\mathcal{G}_0 = \sigma(\theta(0), \Delta(0))$. It is clear that $\mathcal{F}_n \subset \mathcal{G}_n, \forall n \geq 0$. Now,

$$E\left[\frac{G^{n+} - G^n}{\delta\Delta_i(n)} \mid \mathcal{G}_n\right] = \frac{1}{\delta\Delta_i(n)}\left(E[G^{n+} \mid \mathcal{G}_n] - E[G^n \mid \mathcal{G}_n]\right).$$

Let $s_0^n = s_0^{n+} = s$ denote the initial state in both the trajectories χ^n and χ^{n+}, respectively. Recall that the initial state s is chosen randomly from the distribution ν. Thus,

$$\begin{aligned} E[G^n \mid \mathcal{G}_n] &= \sum_s \nu(s) E[G^n \mid s_0^n = s, \phi_{\theta(n)}] \\ &= \sum_s \nu(s) V_{\theta(n)}(s). \end{aligned}$$

Similarly,

$$E[G^{n+} \mid \mathcal{G}_n] = \sum_s \nu(s) E[G^{n+} \mid s_0^{n+} = s, \phi_{\theta(n)+\delta\Delta(n)}]$$

$$= \sum_s \nu(s) V_{\theta(n)+\delta\Delta(n)}(s).$$

Thus,

$$E\left[\frac{G^{n+} - G^n}{\delta\Delta_i(n)} \mid \mathcal{G}_n\right] = \sum_s \nu(s)\left(\frac{V_{\theta(n)+\delta\Delta(n)}(s) - V_{\theta(n)}(s)}{\delta\Delta_i(n)}\right) \text{ a.s.}$$

Thus,

$$E\left[\frac{G^{n+} - G^n}{\delta\Delta_i(n)} \mid \mathcal{F}_n\right] = \sum_s \nu(s) E\left[\frac{V_{\theta(n)+\delta\Delta(n)}(s) - V_{\theta(n)}(s)}{\delta\Delta_i(n)} \mid \mathcal{F}_n\right].$$

Using a Taylor's expansion of $V_{\theta(n)+\delta\Delta(n)}(s)$ around $\theta(n)$ gives us

$$V_{\theta(n)+\delta\Delta(n)}(s_n) = V_{\theta(n)}(s_n) + \delta\Delta(n)^{\mathsf{T}}\nabla V_{\theta(n)}(s_n)$$

$$+\frac{\delta^2}{2}\Delta(n)^{\mathsf{T}}\nabla^2 V_{\theta(n)}(s_n)\Delta(n) + o(\delta^2).$$

Thus,

$$\frac{V_{\theta(n)+\delta\Delta(n)}(s_n) - V_{\theta(n)}(s_n)}{\delta\Delta_i(n)} = \nabla_i V_{\theta(n)}(s_n) + \sum_{k\neq i}\frac{\Delta_k(n)}{\Delta_i(n)}\nabla_k V_{\theta(n)}(s_n)$$

$$+\frac{\delta}{2}\sum_{j,k=1}^{d}\frac{\Delta_j(n)\nabla^2_{j,k}V_{\theta(n)(s_n)}\Delta_k(n)}{\Delta_i(n)} + o(\delta). \tag{8.5}$$

Now,

$$E\left[\left(\frac{V_{\theta(n)+\delta\Delta(n)}(s_n) - V_{\theta(n)}(s_n)}{\delta\Delta_i(n)}\right) \mid \mathcal{F}_n\right] = \nabla_i V_{\theta(n)}(s_n) + o(\delta). \tag{8.6}$$

This follows from the following two observations:

1. The second term on the RHS of (8.5) gives us

$$E\left[\sum_{k\neq i}\frac{\Delta_k(n)}{\Delta_i(n)}\nabla_k V_{\theta(n)}(s_n) \mid \mathcal{F}_n\right] = E\left[\sum_{k\neq i}\frac{\Delta_k(n)}{\Delta_i(n)}\right]\nabla_k V_{\theta(n)}(s_n) = 0,$$

from the properties of the sequence $\Delta_l(n), l = 1, \ldots, d$.

2. The third term on the RHS of (8.5) gives us

$$\frac{\delta}{2}E\left[\sum_{j,k=1}^{d}\frac{\Delta_j(n)\nabla^2_{j,k}V_{\theta(n)}\Delta_k(n)}{\Delta_i(n)}\mid\mathcal{F}_n\right]$$

$$=\frac{\delta}{2}\sum_{j,k=1}^{d}E\left[\frac{\Delta_j(n)\Delta_k(n)}{\Delta_i(n)}\right]\nabla^2_{j,k}V_{\theta(n)}(s_n)=0.$$

This can be seen by analysing all the cases in the summation: (i) $j\neq k\neq i$, (ii) $j\neq k=i$, (iii) $j=i\neq k$, (iv) $j=k\neq i$, and (v) $j=k=i$, respectively, using again the properties of the sequence $\Delta_l(n), l=1,\ldots,d$.

The claim follows. □

In the light of (8.6), we can rewrite (8.3) as follows:

$$\theta(n+1)=\Gamma(\theta(n)-a(n)(\sum_s\nabla V_{\theta(n)}(s)+\eta(n)+\beta(n))),\tag{8.7}$$

where

$$\eta(n)=\left(\frac{G^+_n-G_n}{\delta\Delta_i(n)}\right)-E\left[\left(\frac{G^+_n-G_n}{\delta\Delta_i(n)}\right)\mid\mathcal{F}_n\right]$$

and $\beta(n)=(\beta_1(n),\ldots,\beta_d(n))$ with

$$\beta_i(n)=E\left[\left(\frac{G^+_n-G_n}{\delta\Delta_i(n)}\right)\mid\mathcal{F}_n\right]-\sum_s\nu(s)\nabla_iV_{\theta(n)}(s).$$

From Proposition 8.1, it can be seen that $\beta(n)=o(\delta)$. It is now easy to see that (8.7) has the same form as (4.2).

Lemma 8.2. The function $\nabla v_\theta(s)$ is Lipschitz continuous in θ. Further, $\exists$ a constant $K_1>0$ such that $\|\nabla v_\theta(s)\|\leq K_1(1+\|\theta\|)$.

Proof. It can be shown under A8.3 (see for instance Chapter 13 of (Sutton and Barto, 2018) that $v_\theta(s)$ is continuously differentiable in θ and satisfies

$$\nabla v_\theta(s)=\sum_{y\in S}\sum_{k=0}^{\infty}P^k_\theta(s,y)\sum_{a\in A(y)}\nabla\phi_\theta(a\mid y)q_\theta(y,a),$$

where $P_\theta^k(s,y)$ is the probability of going from state s to state y in k steps under policy ϕ_θ and $q_\theta(y,a) = E_\theta[G_n \mid X_n = y, Z_n = a]$ is the value of the state-action tuple (y,a) when actions in states subsequent to state y follow the policy ϕ_θ. It can also be shown as in Theorem 3 of (Furmston *et al.*, 2016) that $\nabla^2 v_\theta(s)$ exists and is continuous. Since θ takes values in C, a compact set, it follows that $\nabla^2 v_\theta(s)$ is bounded and thus $\nabla v_\theta(s)$ is Lipschitz continuous.

Finally, let $L_1^s > 0$ denote the Lipschitz constant for the function $\nabla v_\theta(s)$. Then, for a given $\theta_0 \in C$,

$$\begin{aligned}\|\nabla v_\theta(s)\| - \|\nabla v_{\theta_0}(s)\| &\leq \|\nabla v_\theta(s) - \nabla v_{\theta_0}(s)\| \\ &\leq L_1^s \|\theta - \theta_0\| \\ &\leq L_1^s \|\theta\| + L_1^s \|\theta_0\| .\end{aligned}$$

Thus,

$$\|\nabla v_\theta(s)\| \leq \|\nabla v_{\theta_0}(s)\| + L_1^s \|\theta_0\| + L_1^s \|\theta\| .$$

Let $K_s \stackrel{\triangle}{=} \|\nabla v_{\theta_0}(s)\| + L_1^s \|\theta_0\|$ and $K_1 \stackrel{\triangle}{=} \max(K_s, L_1^s, s \in S)$. Since $|S| < \infty$, $K_1 < \infty$. Thus, $\|\nabla v_\theta(s)\| \leq K_1(1 + \|\theta\|)$. □

Lemma 8.3. The martingale sequence $(M_n, \mathcal{F}_n)$, $n \geq 0$) satisfies

$$E[\|M_{n+1}\|^2 \mid \mathcal{F}_n] \leq \hat{L}(1 + \|\theta(n)\|^2),$$

for some constant $\hat{L} > 0$.

Proof. Note that

$$\begin{aligned}\|M_{n+1}\|^2 &= \sum_{i=1}^{d} (M_{n+1}^i)^2 \\ &= \frac{(G^{n+} - G^n)^2}{\delta^2} + \frac{1}{\delta^2}\left(E\left[\frac{G^{n+} - G^n}{\Delta_i(n)} \mid \mathcal{F}_n\right]\right)^2 \\ &\quad - 2\frac{G^{n+} - G^n}{\delta \Delta_i(n)} E\left[\frac{G^{n+} - G^n}{\delta \Delta_i(n)} \mid \mathcal{F}_n\right].\end{aligned}$$

Thus,

$$E[\|M_{n+1}\|^2 \mid \mathcal{F}_n] = E\left[\frac{(G^{n+} - G^n)^2}{\delta^2} \mid \mathcal{F}_n\right] - \left(E\left[\frac{G^{n+} - G^n}{\delta \Delta_i(n)} \mid \mathcal{F}_n\right]\right)^2 .$$

It now follows from Assumption A8.1 and the fact that all single-stage costs are bounded, that $E[\|M_{n+1}\|^2 \mid \mathcal{F}_n] \leq \check{K}$ almost surely. In fact from Proposition 8.1 and Lemma 8.2, it follows that

$$\left(E\left[\frac{G^{n+} - G^n}{\delta\Delta_i(n)} \mid \mathcal{F}_n\right]\right)^2 = \left(\sum_{s\in S} \nu(s)\nabla_i V_{\theta(n)}(s)\right)^2 + o(\delta) \leq K_\delta,$$

for some $K_\delta < \infty$. It will thus follow that

$$E[\|M_{n+1}\|^2 \mid \mathcal{F}_n] \leq \check{K}(1 + \|\theta(n)\|^2).$$

□

Define now a sequence $Z_n, n \geq 0$ according to

$$Z_n = \sum_{m=0}^{n-1} a(m)M_{m+1},$$

$n \geq 1$ with $Z_0 = 0$.

Lemma 8.4. $(Z_n, \mathcal{F}_n)$, $n \geq 0$ is an almost surely convergent martingale sequence.

Proof. It is easy to see that Z_n is $\mathcal{F}_n$-measurable $\forall n$. Further, it is integrable for each n and moreover $E[Z_{n+1} \mid \mathcal{F}_n] = Z_n$ almost surely since $(M_{n+1}, \mathcal{F}_n)$, $n \geq 0$ is a martingale difference sequence by Lemma 8.1. It is also square integrable from Lemma 8.3. The quadratic variation process of this martingale will be convergent almost surely if

$$\sum_{n=0}^{\infty} E[\|Z_{n+1} - Z_n\|^2 \mid \mathcal{F}_n] < \infty \text{ a.s.}$$

Note that

$$E[\|Z_{n+1} - Z_n\|^2 \mid \mathcal{F}_n] = a(n)^2 E[\|M_{n+1}\|^2 \mid \mathcal{F}_n].$$

Thus,

$$\begin{aligned}\sum_{n=0}^{\infty} E[\|Z_{n+1} - Z_n\|^2 \mid \mathcal{F}_n] &= \sum_{n=0}^{\infty} a(n)^2 E[\|M_{n+1}\|^2 \mid \mathcal{F}_n] \\ &\leq \check{K} \sum_{n=0}^{\infty} a(n)^2(1 + \|\theta(n)\|^2),\end{aligned}$$

by Lemma 8.3. The claim now follows from Assumption A8.2 and the fact that $\theta(n) \in C, \forall n$, a compact set. Now $(Z_n, \mathcal{F}_n)$, $n \geq 0$ can be seen to be convergent from the martingale convergence theorem for square integrable martingales (see Theorem B.7). □

Consider now the following ODE:

$$\dot{\theta}(t) = \bar{\Gamma}\left(-\sum_s \nu(s)\nabla V_\theta(s)\right), \tag{8.8}$$

where $\bar{\Gamma} : \mathcal{C}(C) \to \mathcal{C}(\mathcal{R}^d)$ is as defined in (2.32).

Let $H \stackrel{\triangle}{=} \{\theta \,|\, \bar{\Gamma}\left(-\sum_s \nu(s)\nabla V_\theta(s)\right)\}$ denote the set of asymptotically stable attractors of (8.8). Let $H^\epsilon \stackrel{\triangle}{=} N^\epsilon(H) \cap C$ denote the ϵ-neighborhood of H within the set C. Here, $N^\epsilon(H) = \{\theta \mid \|\theta - \theta_0\| < \epsilon, \theta_0 \in H\}$.

The following is the main result of this section.

Theorem 8.5. Given $\epsilon > 0$, $\exists \delta_0 > 0$ such that $\forall \delta \in [0, \delta_0)$, the stochastic sequence $\{\theta(n)\}$ obtained from (8.3) converges with probability one to H^ϵ.

Proof. We shall proceed by verifying Assumptions A2.9-A2.12. Note that Assumption A2.9 has been shown in Lemma 8.2. Assumption A2.10 is an assumption on the step-size sequence $\{a(n)\}$ that has also been made for the iterates (8.3), see A8.2. Now from Lemma 8.2, it follows that $\sum_s \nu(s)\nabla v_\theta(s)$ is uniformly bounded since $\theta \in C$, a compact set. Assumption A2.11 is now verified from Proposition 8.1. Assumption A2.12 is easy to see as a consequence of Lemma 8.4. Now note that for the ODE (8.8), $F(\theta) = \sum_s \nu(s)V_\theta(s)$ serves as an associated Lyapunov function and in fact

$$\nabla F(\theta)^\mathsf{T} \bar{\Gamma}\left(-\sum_s \nu(s)\nabla V_\theta(s)\right)$$
$$= \left(\sum_s \nu(s)\nabla_\theta V_\theta(s)\right)^\mathsf{T} \bar{\Gamma}\left(-\sum_s \nu(s)\nabla V_\theta(s)\right)$$

$$\leq 0.$$

For $\theta \in C^o$ (the interior of C), it is easy to see that $\bar{\Gamma}(\sum_s \nu(s)\nabla V_\theta(s))$ $= \sum_s \nu(s)\nabla V_\theta(s)$, and

$$\nabla F(\theta)^{\mathrm{T}}\bar{\Gamma}(-\sum_s \nu(s)\nabla V_\theta(s)) < 0 \text{ if } \theta \in H^c \cap C$$
$$= 0 \text{ otherwise.}$$

For $\theta \in \delta C$ (the boundary of C), there can be spurious attractors on the boundary of C, see (Kushner and Yin, 2003), that are also contained in H. The claim now follows from Theorem 2.5. □

8.2 Simultaneous perturbation-based risk-sensitive policy gradient

We again consider a stochastic shortest path (SSP) problem, with a special cost-free absorbing state, say 0. We restrict our attention to proper policies (see Definition 8.1), which ensure state 0 is recurrent, and the remaining states are transient in the Markov chain underlying the policy considered. We define an episode as a sample path $\{x_0, \ldots, x_\tau\}$, where $x_\tau = 0$, and τ is the first passage time to state 0.

Consider a smoothly parameterized class of policies $\{\pi_\theta \mid \theta \in \mathbb{R}^d\}$. Suppose that the policy π_θ is a continuously differentiable function of the parameter θ: a standard assumption in policy gradient literature. Let $K_\theta(x_0)$ denote the total discounted cost r.v. under policy x starting in state x_0, i.e., $K_\theta(x_0) = \sum_{t=0}^{\tau-1} \gamma^t k(x_t, a_t)$, where $0 < \gamma < 1$ is the discount factor and $k(x_t, a_t)$ is the single-stage cost incurred at time instant t in state x_t on choosing action a_t. Here actions a_t are chosen according to policy π_θ, which is parameterized by θ.

The classic objective in RL is to find a policy that minimizes, in expectation, the total discounted cost. We consider a risk-sensitive RL setting, where the goal is to find a policy that optimizes a certain risk measure, i.e., the following problem:

$$\min_{\theta \in \mathbb{R}^d} \{\rho(K_\theta(x_0))\}, \tag{8.9}$$

where $\theta \in \mathbb{R}^d$ parameterizes the policy π_θ, and ρ is a risk measure. As examples, we define three risk measures below for a random variable X with CDF F.

CVaR: Recall that the VaR and CVaR at level $\alpha \in (0,1)$ are defined as

$$\mathrm{VaR}_\alpha(X) = \inf\{\xi \mid \mathbb{P}(X \leq \xi) \geq \alpha\}, \tag{8.10}$$

$$\mathrm{CVaR}_\alpha(X) = \inf_\xi \left\{\xi + \frac{1}{(1-\alpha)}\mathbb{E}(X-\xi)_+\right\}. \tag{8.11}$$

As mentioned earlier in Section 2.2.6, VaR is not a coherent risk measure, while CVaR is. Coherency includes properties, namely monotonicity, sub-additivity, positive homogeneity and translation invariance. These properties are desirable for any risk measures, esp. in the context of finance, and VaR is not sub-additive. In financial context, sub-additivity relates to diversification, which is performed to reduce the risk, e.g., a financial portfolio.

Spectral risk measure (SRM): This risk measure, proposed in (Acerbi, 2002), is defined as

$$M_\phi(X) = \int_0^1 \phi(\beta) F^{-1}(\beta) \mathrm{d}\beta, \tag{8.12}$$

where $\phi : [0,1] \to [0,\infty)$ is the risk spectrum. SRM is a coherent risk measure, when the risk spectrum is positive, increasing and integrates to one. Moreover, SRM generalizes CVaR after observing the following equivalent form, known as Acerbi's formula:

$$\mathrm{CVaR}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_\beta(X)\, d\beta. \tag{8.13}$$

In particular, CVaR is an SRM with $\varphi(\beta) = \dfrac{1}{1-\alpha}\mathbb{I}\{\beta > \alpha\}$, $\alpha \in (0,1)$. As an example, one could consider $\phi(\beta) = \dfrac{\kappa\, e^{-\kappa(1-\beta)}}{1-e^{-\kappa}}$, $\beta \in [0,1]$ for some $\kappa > 0$. From an attitude towards risk viewpoint, assuming X models the loss associated with a financial position, SRM with an exponential risk spectrum defined above, is preferable over CVaR as it assigns a higher weight to larger losses,

whereas CVaR the same weight for all losses beyond a certain quantile.

Utility-based shortfall risk (UBSR): UBSR, proposed in (Föllmer and Schied, 2002), for a given loss function l and threshold parameter λ, is defined as

$$S_\alpha(X) = \inf \{\xi \in \mathbb{R} \mid \mathbb{E}\,(l(-X - \xi)) \leq \alpha\}\,. \tag{8.14}$$

UBSR belongs to the class of convex risk measures, which subsumes coherent risk measures, since sub-additivity and positive homogeneity imply convexity. As examples of loss functions in the definition of UBSR above, one could consider the following: (i) $l(x) = \exp(\beta x)$; and (ii) $l(x) = x^2$. For the first candidate loss, UBSR turns out to be the entropic risk measure. For the second candidate loss, i.e., the square loss, UBSR can be related to CVaR, see (Giesecke *et al.*, 2008).

Notice that the optimization problem in (8.9) is non-convex in nature. For solving the problem defined above using gradient-based methods, one requires (i) an estimate of the risk measure for any given policy π_θ; and (ii) an estimate of the gradient of the risk measure w.r.t. the policy parameter θ. We elaborate on these two parts below.

We simulate m episodes simulated using the policy π_θ, and collect samples of the total cost $K_\theta(x_0)$. Let $\{K_1, \ldots, K_m\}$ denote the i.i.d. samples from the distribution of $K_\theta(x_0)$. We define the empirical distribution function (EDF) F_m of $K_\theta(x_0)$ as follows:

$$F_m(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\{K_i \leq x\}\,, \forall x \in \mathbb{R}.$$

Using the EDF, we form the estimate ρ_m of $\rho(K_\theta(x_0))$ as follows:

$$\rho_m = \rho(F_m). \tag{8.15}$$

Such an estimation scheme for an abstract risk measure has been considered earlier in (Prashanth and Bhat, 2022).

Next, we present a bound in expectation for the estimation error associated with (8.15).

Proposition 8.2. Suppose the risk measure ρ satisfies the following continuity requirement for any two distributions F, G for some $L > 0$:

$$|\rho(F) - \rho(G)| \leq LW_1(F, G), \tag{8.16}$$

where $W_1(F, G)$ is the Wasserstein distance[1] between distributions F and G. Suppose the r.v. $K_\theta(x_0)$ satisfies $\mathbb{E}[K_\theta(x_0)^2] \leq B < \infty$, for any $x \in \mathbb{R}^d$. Then,

$$\mathbb{E}\,|\rho_m - \rho(K_\theta(x_0))| \leq \frac{c}{\sqrt{m}},$$

for some constant c that depends on B.

Proof. Let F denote the distribution of $K_\theta(x_0)$. Then, we have

$$\mathbb{E}\,|\rho_m - \rho(K_\theta(x_0))| = \mathbb{E}\,|\rho(F_m) - \rho(F)| \leq LW_1(F_m, F) \leq \frac{c_1 LB}{\sqrt{m}},$$

where the final inequality follows by using Theorem 3.1 of Lei, 2020. □

The continuity requirement in (8.16) is satisfied by the three popular risk measures CVaR, UBSR and SRM, and the reader is referred to (Prashanth and Bhat, 2022) for details.

To construct an estimate of the gradient of the risk measure $\rho(K_\theta(x_0))$, one could employ the simultaneous perturbation method, e.g., the estimate in (3.15). Using this gradient estimate, and the template of RSG-BGO algorithm, we arrive at the following update iteration for the risk-sensitive policy gradient (risk-PG) algorithm:

$$\theta_{k+1} = \theta_k - a(k)\widehat{\nabla}\rho(\theta_k), \tag{8.17}$$

where γ_k is the stepsize, and $\widehat{\nabla}\rho(\theta_k)$ is an estimate of the gradient of the risk measure $\rho(K_{\theta_k}(x_0))$. To elaborate on the gradient estimation aspect of the algorithm above, we first simulate m_k trajectories of the underlying MDP with policy parameters $\theta_k + \delta_k\Delta_k$ and $\theta_k - \delta_k\Delta_k$, respectively. Here δ_k is the perturbation constant, and Δ_k is a standard Gaussian

[1] Given two cumulative distribution functions (CDFs) F and G on $\mathbb{R}$, let $\Gamma(F, G)$ denote the set of all joint distributions on $\mathbb{R}^2$ having F and G as marginals. Then the Wasserstein distance between F and G is defined by $W_1(F, G) = \left[\inf_{C \in \Gamma(F,G)} \int_{\mathbb{R}^2} |x - y| dC(x, y)\right]$.

vector. Other random perturbations are feasible, see Chapter 3. Using (8.15), we estimate the risk measures $\rho(K_{\theta_k \pm \delta_k \Delta_k}(x_0))$ corresponding to the aforementioned policy parameters, and then, use (3.15) to form $\widehat{\nabla}\rho(\theta_k)$.

For a non-asymptotic analysis of (8.17), we require smoothness of ρ as a funciton of θ. We verify this assumption for the special case of CVaR below. Let $\mathrm{C}_\alpha(K_\theta(x_0))$ denote the CVaR associated with a policy π that is parameterized by θ. Then, using the likelihood ratio method, we arrive at the following variant of the policy gradient theorem under the CVaR objective (cf. Tamar *et al.*, 2015):

$$\nabla_\theta C_\alpha(K_\theta(x_0)) = \mathbb{E}\Big[[K_\theta(x_0) - \mathrm{VaR}_\alpha(K_\theta(x_0))] \times \underbrace{\sum_{m=0}^{\tau-1} \nabla \log \pi_\theta(a_m|x_m)}_{(I)} \Big| K_\theta(x_0) \geq \mathrm{VaR}_\alpha(K_\theta(x_0))\Big].$$

In the above, $K_\theta(x_0)$ and term (I) on the RHS are Lipschitz functions due to the policy gradient assumption mentioned above. In addition, if we assume that the distribution, say F_θ, of $K_\theta(x_0)$ is a Lipschitz function in θ, then we can infer that $\nabla_\theta C_\alpha(K_\theta(x_0))$ is sum of product of Lipschitz functions, implying smoothness of CVaR as a function of θ. One could generalize this argument to the case when ρ is a coherent risk measure, and the reader is referred to (Tamar *et al.*, 2015) for details.

Using the non-asymptotic bounds for a RSG algorithm, see Section 5.5, we can infer that the iteration complexity for the risk-sensitive policy gradient algorithm (8.17) is $\mathcal{O}\left(\frac{1}{\epsilon^3}\right)$.

8.3 Bibliographic remarks

Reinforcement learning has been found to be extremely useful in problems of dynamic decision making under uncertainly when the decision maker has access to data (either from a real system or alternatively a simulation) but not the transition model. Textbook treatments of RL are available in (Sutton and Barto, 2018; Bertsekas and Tsitsiklis, 1996; Bertsekas, 2019; Meyn, 2022). We considered two different RL settings

in this chapter: (i) finding optimal policies for stochastic shortest path problems and (ii) finding optimal policies for risk sensitive control, again in the stochastic shortest path case. Both the algorithms that we presented fall under the broad category of policy gradient algorithms. For (i), we specifically considered the REINFORCE algorithm developed originally in (Williams, 1992), see also (Sutton and Barto, 2018, Chapter 13), (Sutton *et al.*, 1999) for a detailed treatment. Whereas the original algorithm required a single simulation and involved unbiased function gradients, we presented in this chapter a zeroth-order algorithm based on two-simulation SPSA gradient estimates. This algorithm has been presented and analysed in (Bhatnagar, 2023). An alternative to using trajectory-based Reinforce type policy gradient algorithms for finding the optimal policy is to use two-timescale actor-critic algorithms, where on the faster timescale, the critic recursion estimates the value function corresponding to the most recent policy parameter update most often using a temporal difference (TD) learning procedure, while along the slower recursion, the policy parameter is updated using policy gradients.

Zeroth-order stochastic gradient estimation based procedures have been found to be efficient and are not new to the literature, see for instance, (Bhatnagar and Kumar, 2004; Abdulla and Bhatnagar, 2007) in the context of actor-critic algorithms. In the context of trajectory-based policy gradient algorithms (like REINFORCE), (Salimans *et al.*, 2017; Mania *et al.*, 2018) have proposed generating multiple trajectories of data for a given parameter update and then selecting the best candidate directions over which an additional layer of sample averaging is performed which then is used as an increment for the parameter update. Even though the latter can be computationally tedious, it is found to improve the performance of the algorithm. In (Choromanski *et al.*, 2018), zeroth order algorithms based on evolutionary strategies for policy optimization again for reinforcement learning are presented. They make use of random orthogonal matrices and study specifically two constructions for the random perturbations - one based on Gaussian perturbations and another on random Hadamard Rademacher matrices.

Zeroth-order gradient-based methods have been employed for solving risk-sensitive RL problems through the policy gradient solution approach. A variety of risk measures have been tackled using zeroth-

order policy gradients, and we list a few representative works in the following: (i) SPSA for mean-variance optimization in a discounted MDP in (Prashanth and Ghavamzadeh, 2016); (ii) SF for distortion risk measures in (Vijayan and Prashanth, 2023); (iii) A simultaneous perturbation-based oracle for optimizing smooth risk measures in (Bhavsar and Prashanth, 2022). Our presentation in Section 8.2 is based the contributions from the aforementioned reference.

Appendices

A

ODEs and differential inclusions

In this appendix, we review some of the key concepts from ordinary differential equations and differential inclusions. This background material is useful for the asymptotic analysis of stochastic approximation algorithms and stochastic recursive inclusions.

A.1 Ordinary differential equations

We consider the following ODE, that is the same as (2.2). We shall describe limit sets and notions of stability for such an ODE.

$$\dot{\theta}(t) = h(\theta(t)). \tag{A.1}$$

We recall first the Gronwall inequality, a fundamental result useful for showing stability properties of ODEs, see Lemma B.1 of (Borkar, 2022), for a proof.

Lemma A.1 (Gronwall Inequality)**.** Suppose that for continuous

$u, v : [0, T] \to [0, \infty)$, for $T > 0$ and scalars $C, K \geq 0$:

$$u(t) \leq C + K \int_0^t u(s)v(s)ds, \ \forall t \in [0, T].$$

Then it follows that for all $t \in [0, T]$,

$$u(t) \leq C \exp\left(K \int_0^T v(s)ds\right).$$

A.1.1 Limit sets of ODE

We present first some basic definitions on the limit sets of ODEs. Consider the ODE (A.1) with the function $h : \mathbb{R}^d \to \mathbb{R}^d$ being Lipschitz continuous. In other words, $\exists L > 0$ (a constant) such that

$$\|h(\eta) - h(\beta)\| \leq L\|\eta - \beta\|, \ \forall \eta, \beta \in \mathbb{R}^d.$$

Definition A.1. We say that the ODE (A.1) is well-posed if for any initial condition $\theta_0 \in \mathbb{R}^d$, there is a unique solution $\theta(\cdot) \in C([0, \infty); \mathbb{R}^d)$ that is also continuous as a function of θ_0.

In the above, $C([0, \infty); \mathbb{R}^d)$ denotes the space of all continuous functions from $[0, \infty)$ to $\mathbb{R}^d$. The integral solution to the ODE (A.1) is obtained as

$$\theta(t) = \theta_0 + \int_0^t h(\theta(s))ds, \ t \geq 0. \tag{A.2}$$

If an ODE is well-posed, it has unique integral curves. The following theorem says that a sufficient condition for well-posedness of (A.1) is that the function h be Lipschitz continuous, see Theorem B.1 of (Borkar, 2022) for a proof based on the Gronwall inequality (Lemma A.1).

Theorem A.2. Suppose the function $h : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous. Then the ODE (A.1) is well-posed.

For the ODE (A.1), let $\Phi : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ be defined as the map $\Phi(t, x) \stackrel{\triangle}{=} \Phi_t(x)$ that takes $\theta(0)$ to $\theta(t)$ via the ODE (A.1). Thus,

$$\theta(t) = \Phi_t(\theta(0)) = \theta(0) + \int_{\tau=0}^t h(\Phi_\tau(\theta(0)))d\tau.$$

Assuming h is Lipschitz continuous, it follows from Theorem A.2 that the map Φ is continuous. It is easy to verify that $\{\Phi_t, t \in \mathbb{R}\}$ forms a group since $\Phi_t \circ \Phi_s = \Phi_{t+s}$, $\forall t, s \in \mathbb{R}$ and $\Phi_0 = I$ (the identity map). Thus, $\{\Phi_t, t \in \mathbb{R}\}$ is a flow of h, see (Benaïm, 1996), for a more general discussion.

Definition A.2 (Invariant sets and Periodic Points). 1. We say that $A \subset \mathbb{R}^d$ is invariant for the ODE (A.1) if $\Phi_t(A) \subset A$ for all $t \in \mathbb{R}$.

2. We say that $A \subset \mathbb{R}^d$ is positively (resp. negatively) invariant for the ODE (A.1) if $\Phi_t(A) \subset A$ for all $t \geq 0$ (resp. $t \leq 0$).

3. A point θ is a periodic point for the ODE (A.1) if $\exists T > 0$ such that $\Phi_T(\theta) = \theta$.

Note that since the flow Φ is induced by the vector field h, equilibria of (A.1) coincide with the zeros of the function $h(\cdot)$. Further, both periodic points and equilibria can be viewed as *recurrent* points.

Definition A.3 (Limit Sets of an ODE). 1. Given a trajectory $\theta(\cdot)$ of (A.1) with $\theta(0) = \theta_0 \in \mathbb{R}^d$, the orbit through θ_0 is the set

$$\mathcal{O}(\theta_0) = \{\theta(t) \in \mathbb{R}^d | t \in \mathbb{R}\}.$$

2. Given a trajectory $\theta(\cdot)$ of (A.1), the set $\mathcal{L} \stackrel{\triangle}{=} \cap_{t \geq 0} \overline{\theta([t, \infty))}$ that comprises of the set of limit points of (A.1) is called the ω-limit set of (A.1).

3. Given $\delta, T > 0$, a (δ, T)-pseudo-orbit from $\lambda \in \mathbb{R}^d$ to $\eta \in \mathbb{R}^d$ is defined as a set of k trajectories of (A.1) (for some $k < \infty$): $\{\Phi_t(\eta_i) : t \in [0, t_i],\ t_i \geq T\}$, $i = 0, 1, \ldots, k-1$, where $\eta_0, \eta_1, \ldots, \eta_k \in \mathbb{R}^d$ and such that (i) $\|\eta_0 - \lambda\| < \delta$, (ii) $\|\Phi_{t_j}(\eta_j) - \eta_{j+1}\| < \delta$, $\forall j = 0, 1, \ldots, k-1$, and (iii) $\eta_k = \eta$.

4. If a (δ, T)-pseudo-orbit exists between any $\lambda, \eta \in \mathbb{R}^d$, for every $\delta, T > 0$, we say that the flow Φ of (A.1) is chain transitive.

5. The flow Φ as above restricted to $\eta = \lambda$, for all $\lambda \in \mathbb{R}^d$ is called chain recurrent.

6. A compact invariant set $A \subset \mathcal{R}^d$ on which the flow Φ of the ODE (A.1) is chain recurrent (resp. chain transitive) is called an internally chain recurrent (resp. internally chain transitive) set for (A.1).

We now recall the following result, see (Benaïm, 1999, Proposition 5.3):

Lemma A.3. Let $A \subset \mathcal{R}^d$ be a compact invariant set for the ODE (A.1). The following are equivalent:

1. A is internally chain transitive.
2. A is connected and internally chain recurrent.

Definition A.4 (Equilibria and Attractors of an ODE).

1. A point $\theta \in \mathbb{R}^d$ is an equilibrium of the ODE (A.1) if $\Phi_t(\theta) = \theta$, $\forall t$. In other words, $h(\theta) = 0$.
2. An equilibrium $\theta \in \mathbb{R}^d$ of (A.1) is said to be isolated, if there exists an open set $U \subset \mathbb{R}^d$ such that $\theta \in U$ and there does not exist any other equilibrium $\check{\theta} \in U$.
3. A compact invariant set $A \subset \mathbb{R}^d$ is said to be Lyapunov stable or simply stable for the ODE (A.1) if given any $\epsilon > 0$, $\exists \delta > 0$ such that $d(\theta_0, A) < \delta$ implies that $d(\Phi_t(\theta(0)), A) < \epsilon$ for all $t > 0$. Here for any given $x \in \mathbb{R}^d$, $d(x, A) = \min_{\eta \in A} \|x - \eta\|$ is the distance between x and the set A.
4. A set $A \subset \mathbb{R}^d$ is an attractor for (A.1) if A is nonempty, compact and invariant. Further, A has a positively invariant open neighborhood $M \subset \mathbb{R}^d$ such that $d(\Phi_t(\theta), A) \to 0$ as $t \to \infty$ uniformly in $\theta \in M$.
5. The largest open neighborhood M for an attractor A above is called the domain of attraction of A.
6. A compact invariant $A \subset \mathbb{R}^d$ is asymptotically stable for the ODE (A.1) if it is both Lyapunov stable and an attractor.

We now mention the following important results in Lemmas A.4–A.6, see for instance, (Nandakumaran *et al.*, 2017) for a detailed treatment.

Lemma A.4. Suppose $\theta(\cdot)$ is a solution of (A.1). Then $\theta_l(t) = \theta(t+l)$ is also a solution to (A.1) for any fixed l and for all t.

Lemma A.5. Let $\theta(\cdot)$ be a solution to the ODE (A.1) and $\lim_{t\to\infty} \theta(t) = \bar{\theta}$ for some $\bar{\theta} \in \mathbb{R}^d$. Then $\bar{\theta}$ is an equilibrium of (A.1).

Proof. From Lemma A.4, for any $l > 0$, $\theta(t+l)$, $t \geq 0$, is also a solution to (A.1) and $\lim_{t\to\infty} \theta(t+l) = \bar{\theta}$. By the mean value theorem,

$$\theta(t+l) - \theta(t) = l\dot{\theta}(\bar{t}) = lh(\theta(\bar{t})),$$

for some $\bar{t} \in [t, t+l]$. Thus, as $t \to \infty$, $\bar{t} \to \infty$ as well, and $\theta(t+l) - \theta(t) \to 0$ as $t \to \infty$. By continuity, this implies that $lh(\bar{\theta}) = 0$, hence $h(\bar{\theta}) = 0$. □

Linearized System

Consider the ODE (A.1) and assume that the function $h : \mathbb{R}^d \to \mathbb{R}^d$ is twice continuously differentiable. Let $\bar{\theta} \in \mathbb{R}^d$ be an equilibrium of (A.1). Then by a Taylor's expansion, we get

$$h(\bar{\theta} + \theta) = h(\bar{\theta}) + Dh(\bar{\theta})\theta + O(\|\theta\|^2),$$

where

$$Dh(\bar{\theta}) = \begin{bmatrix} \nabla_1 h_1(\bar{\theta}) & \nabla_1 h_2(\bar{\theta}) & \cdots & \nabla_1 h_d(\bar{\theta}) \\ \nabla_2 h_1(\bar{\theta}) & \nabla_2 h_2(\bar{\theta}) & \cdots & \nabla_2 h_d(\bar{\theta}) \\ \cdots & \cdots & \cdots & \cdots \\ \nabla_d h_1(\bar{\theta}) & \nabla_d h_2(\bar{\theta}) & \cdots & \nabla_d h_d(\bar{\theta}) \end{bmatrix}$$

is the Jacobian of the function $h = (h_1, h_2, \ldots, h_d)$ evaluated at $\bar{\theta}$. Now note that $h(\bar{\theta}) = 0$. If we ignore the higher order terms $O(\|\theta\|^2)$, we get the linearized ODE:

$$\dot{\theta}(t) = Dh(\bar{\theta})\theta(t).$$

Lemma A.6. If all the eigenvalues of $Dh(\bar{\theta})$ have negative real parts, then $\bar{\theta}$ is asymptotically stable for the ODE (A.1).

Sufficient Condition for Asymptotic Stability

Before proceeding further, we give a sufficient condition for verifying asymptotic stability of an attractor $A \subset \mathbb{R}^d$ of the ODE (A.1). Let $V : M \subset \mathbb{R}^d \to \mathbb{R}$ be a non-negative, continuously differentiable function. Suppose it satisfies the following condition:

$$\langle \nabla V(\theta), h(\theta) \rangle \begin{cases} < 0 \text{ if } \theta \in M \cap A^c \\ = 0 \text{ if } \theta \in A. \end{cases}$$

The function $h(\cdot)$ above is the driving vector field of the ODE (A.1). The asymptotic stability of A follows since $\frac{d}{dt}V(\theta(t) \leq 0$ with equality only for $\theta(t) \in A$.

We now recall the Lasalle Invariance Principle, see Theorem 2 of (J. P. Lasalle and S. Lefschetz, 1961).

Theorem A.7 (Lasalle Invariance Principle)**.** Let $V(\cdot)$ as above be a Lyapunov function for the ODE (A.1). Then any trajectory $\theta(\cdot)$ of (A.1) must converge to the largest invariant set contained in $\{\theta \mid \langle \nabla V(\theta), h(\theta) \rangle = 0\}$.

Gradient Systems

Suppose the underlying system is a gradient scheme with the corresponding ODE being

$$\dot{\theta}(t) = -\nabla f(\theta(t)), \ \theta(0) = \theta_0. \tag{A.3}$$

Thus, here $h(\theta) = -\nabla f(\theta)$. Note that

$$\begin{aligned} \frac{df(\theta(t))}{dt} &= -\nabla f(\theta(t))^T \nabla f(\theta(t)) \\ &= -\|\nabla f(\theta(t))\|^2 \\ &< 0 \quad \text{if } \nabla f(\theta) \neq 0 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Assuming $f \geq 0$, the function f itself serves as a Lyapunov function with the set $H = \{\theta | \nabla f(\theta) = 0\}$ as the set of equilibrium points of

(A.3). If f is not non-negative but bounded below, i.e., $\exists C < 0$ such that $\min_x f(x) = C$. Then, one may let $V(x) = f(x) + |C|$, which will ensure that $V \geq 0$ and the above continues to hold.

We recall now Lemma 11.1 of (Borkar, 2022).

Lemma A.8. The only invariant sets that can occur as w-limit sets for the ODE (A.3) are the subsets of $H \stackrel{\triangle}{=} \{\theta \in \mathbb{R}^d | \nabla f(\theta) = 0\}$.

Lasalle Invariance Principle, see Theorem A.7, in the case of gradient systems, would say something similar as below.

Lemma A.9. Any trajectory $\theta(\cdot)$ of the ODE (A.3) with f as above must converge to the largest invariant set contained in $H \stackrel{\triangle}{=} \{\theta \mid \nabla f(\theta) = 0\}$.

A.2 Set-valued maps and differential inclusions

In many real life situations, one often encounters problems that are ill-posed, the solution is not unique, or there are uncertainties and imprecise modeling errors. Such problems arise often in stochastic control and optimization, reinforcement learning, viability theory and stochastic games. In such scenarios, one may not encounter single-valued maps at all and more general analytical techniques are needed. In this section, we present a brief background on set-valued maps and differential inclusions for which we refer primarily to the books (Aubin and Frankowska, 1990) and (Aubin and Cellina, 1984).

A.2.1 Set-valued maps

A set-valued map $x \mapsto h(x)$ is one where for any $x \in \mathbb{R}^n$, $h : \mathbb{R}^n \to \{\text{subsets of } \mathbb{R}^m\}$ and is specified via it's graph, i.e., $\text{Graph}(J) = \{(x, y) \mid y \in h(x)\}$. The domain ($\text{Dom}(h)$) and image ($\text{Im}(h)$) are respectively given by $\text{Dom}(h) = \{x \in \mathbb{R}^n \mid h(x) \neq \phi\}$ and $\text{Im}(h) = \cup_{x \in \mathbb{R}^n} h(x)$, respectively. The inverse h^{-1} of the set-valued map h (above) is also a set-valued map such that $x \in h^{-1}(y)$ if and only if $y \in h(x)$, viz., $(x, y) \in \text{Graph}(h)$.

The open ball of radius ϵ around the origin is denoted $B_\epsilon(0)$, while the closed ball is denoted $\overline{B}_\epsilon(0)$. Thus, $B_\epsilon(0) = \{x \in \mathbb{R}^n \mid \|x\| < \epsilon\}$ and

$\overline{B}_\epsilon(0) = \{x \in \mathbb{R}^n \mid \|x\| \leq \epsilon\}$. For any set $A \subset \mathbb{R}^n$, for any $\delta > 0$, we call $N_\delta(A) = \{x \in \mathbb{R}^n \mid \|x - y\| < \delta, y \in A\}$ the δ-open neighborhood or simply the neighborhood of the set A. The δ-closed neighborhood of A is likewise the set $\overline{N^\delta}(A) = \{x \mid \|x - y\| \leq \delta, y \in A\}$.

We now have the following definitions pertaining to set-valued maps. Let $h : \mathbb{R}^n \to \{\text{subsets of } \mathbb{R}^m\}$ be a set-valued map.

Definition A.5 (Continuity of Set-Valued Maps). (i) h is said to be upper semi-continuous at a point $x \in \text{Dom}(J)$ if given sequences $\{x_k\}_{k\geq 1}$ (in $\mathbb{R}^n$) and $\{y_k\}_{k\geq 1}$ (in $\mathbb{R}^m$) with $x_k \to x$, $y_k \to y$ and $y_k \in h(x_k)$, $\forall k \geq 1$, we have $y \in h(x)$. We say that h is upper semi-continuous if it is upper semi-continuous at every $x \in \text{Dom}(h)$. In other words, Graph(h) is closed.

(ii) h is said to be lower semi-continuous at a point $x \in \text{Dom}(h)$ if for any $y \in h(x)$, and any sequence of points $x_k \in \text{Dom}(h)$ converging to x, there exists a sequence of elements $y_k \in h(x_k) \to y \in h(x)$. We say that h is lower semi-continuous if it is lower semi-continuous at every $x \in \text{Dom}(h)$.

(iii) h is continuous at $x \in \text{Dom}(h)$ if it is both upper semi-continuous and lower semi-continuous at x. It is said to be continuous if and only if it is continuous at every $x \in \text{Dom}(h)$.

(iv) h is Lipschitz at $z \in \mathbb{R}^n$ if there exists $L > 0$ and $\epsilon > 0$ such that for all $x, y \in N_\epsilon(\{z\})$, we have that $h(x) \subset h(y) + L\|x - y\|B_1(0)$ where $B_1(0) = \{w \in \mathbb{R}^m \mid \|w\| < 1\}$ is a unit ball around the origin in $\mathbb{R}^m$ or more compactly $h(x) \subset N_{L\|x-y\|}(h(y))$.

It is important to note here that there exist set-valued maps that are upper semi-continuous but not lower semi-continuous and vice versa.

Definition A.6 (Peano or Marchaud Map). A set-valued map $h : \mathbb{R}^n \to \{\text{subsets of } \mathbb{R}^m\}$ is called a Peano or Marchaud map if it satisfies the following properties:

(i) For every $x \in \mathbb{R}^n$, $h(x)$ is convex and compact.

(ii) h is pointwise bounded for every $x \in \mathbb{R}^n$, i.e., for some $K > 0$ we have, $\sup_{w \in h(x)} \|w\| \leq K(1 + \|x\|)$.

(iii) h is upper semi-continuous, see Definition A.5(i).

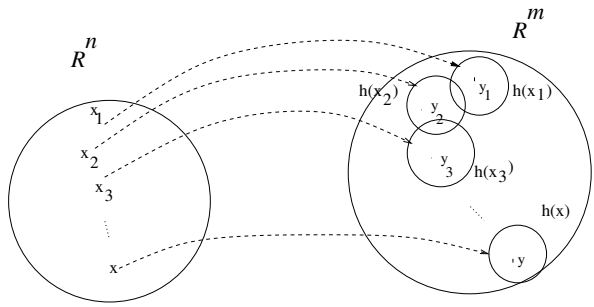


Figure A.1: A set-valued map $h : \mathbb{R}^n \to \{\text{subsets of } \mathbb{R}^m\}$ is called a Peano or marchaud map if (i) $h(x)$ is compact and convex for every x, (ii) $h(x)$ is pointwise bounded and (iii) h is upper semi-continuous.

The distance of a point $x \in \mathbb{R}^d$ to a set $A \subset \mathbb{R}^d$ (for any $d \geq 1$) is defined as $d(x, A) = \inf\{\|x - y\| \mid y \in A\}$. Notice that a point $x_0 \in \mathbb{R}^d$ is a boundary point of A if and only if $d(x, A) = d(x, A^c) = 0$.

Definition A.7 (Limsup and Liminf of Set-Valued Maps)**.** Given a set-valued map $h : \mathbb{R}^n \to \{\text{subsets of } \mathbb{R}^m\}$, we define the upper limit (Limsup) and lower limit (Liminf) of the sequence of sets $h(x_k)$ as follows:

(i) $\text{Limsup}_{x_k \to x} h(x_k) = \{y \in \mathbb{R}^m \mid \liminf_{x_k \to x} d(y, h(x_k)) = 0\}$.

(ii) $\text{Liminf}_{x_k \to x} h(x_k) = \{y \in \mathbb{R}^m \mid \lim_{x_k \to x} d(y, h(x_k)) = 0\}$.

Note that both Liminf and Limsup are closed sets. Liminf collects the limit points of $\{h(x_k)\}$ while Limsup collects its accumulation points. Further, $\text{Liminf}_{x_k \to x} h(x_k) \subset \overline{h(x)} \subset \text{Limsup}_{x_k \to x} h(x_k)$.

A.2.2 Differential inclusions

A differential inclusion (DI) can be viewed as a generalization of an ODE in the sense that it involves set-valued maps as opposed to the usual point-to-point maps and in general has the form

$$\dot{x}(t) \in h(t, x(t)), \tag{A.4}$$

where $h : \mathbb{R} \times \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}^d\}$. We shall mainly be interested with the case where $h(t,x) \stackrel{\triangle}{=} h(x)$, i.e., there is no explicit time dependence of the set-valued map h. In such a case, $h(x) \subset \mathbb{R}^d$, for any $x \in \mathbb{R}^d$. Thus, the DI in this case takes the form

$$\dot{x}(t) \in h(x(t)), \tag{A.5}$$

with $h : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}^d\}$. Any solution to (A.5) is viewed in the Caratheodory sense, i.e., as an absolutely continuous function satisfying (A.5) almost everywhere.

Definition A.8. (i) Let $K \subset \text{Dom}(h)$. A function $x : [0,T] \to \mathbb{R}^d$ is said to be viable in K if $x(t) \in K$, $\forall t \in [0,T]$.

(ii) A solution $x(\cdot)$ to (A.5) is said to be viable if for some closed subset K of $\text{Dom}(h)$, we have that $x(t) \in K$, $\forall t$.

(iii) For $K \subset \mathbb{R}^d$, given $x \in \bar{K}$ (the closure of K), the contingent cone is defined by

$$C(x,K) \stackrel{\triangle}{=} \left\{ y \in \mathbb{R}^d \mid \liminf_{k \to 0^+} \frac{d(x+ky,K)}{k} = 0 \right\}.$$

(iv) We say that a set $K \subset \text{Dom}(h)$ is a viability domain of the set-valued map h if and only if for all $x \in K$, $h(x) \cap C(x,K) \neq \phi$.

Consider the case where $K = \{x\}$. Then the contingent cone to $\{x\}$ is given by $C(x,\{x\}) = \left\{ y \mid \liminf_{k \to 0^+} \frac{d(x+ky,\{x\})}{k} = 0 \right\} = \{0\}$. Then, from Definition A.8(iv), it follows that $K = \{x\}$ is a viability domain of h if and only if $h(x) \cap \{0\} \neq \phi$ or x is a stationary solution to the inclusion $0 \in h(x)$ implying that x is an equilibrium of h. Thus, the minimal viability domains are equilibria of set-valued maps. We now recall the following results from (Aubin and Frankowska, 1990) (see Theorems 10.1.12-10.1.13 there).

Theorem A.10. Consider a Peano or Marchaud map $h : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}^d\}$. Then the limit sets of the solutions to the DI (A.5) are closed viability domains. Further, the limit of a solution $x(t)$ to the DI (A.5) (if it exists), as $t \to \infty$, is an equilibrium of h.

Theorem A.11. Let $h : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}^d\}$ be a Peano or Marchaud map. If $K \subset \mathrm{Dom}(h)$ is a compact viability domain and if $h(K)$ is convex, then there exists an equilibrium of h in K.

A.2.3 Limit Sets of Differential Inclusions

Recall that a solution to the DI (A.5) is an absolutely continuous mapping $\mathbf{x} : \mathbb{R} \to \mathbb{R}^d$ such that $\mathbf{x}(0) = x$ and $\dot{\mathbf{x}}(t) \in h(\mathbf{x}(t))$ for almost every $t \in \mathbb{R}$. The ω-limit set of a given solution $\mathbf{x}$ of the DI (A.5) with $\mathbf{x}(0) = x$ is given by $L(x) = \bigcap_{t \geq 0} \overline{\mathbf{x}([t, +\infty))}$.

Consider $\{\Phi_t\}_{t \in \mathbb{R}}$ defined by $\Phi_t(x) = \{\mathbf{x}(t) \mid \mathbf{x} \text{ is a solution to the DI (A.5) with } \mathbf{x}(0) = x\}$. Then $\{\Phi_t\}$ is the set-valued semi-flow associated with the DI (A.5). For $B \times M \subset \mathbb{R} \times \mathbb{R}^d$, we let $\Phi_B(M) = \bigcup_{t \in B, x \in M} \Phi_t(x)$. For $M \subset \mathbb{R}^d$, the ω-limit set for the DI (A.5) is specified by (cf. Benaïm *et al.*, 2005) $\omega_\Phi(M) = \bigcap_{t \geq 0} \overline{\Phi_{[t,+\infty)}(M)}$.

Definition A.9 (Invariance of Sets). Let $M \subset \mathbb{R}^d$. We say that

(i) M is strongly invariant if $M = \Phi_t(M)$ for every $t \in \mathbb{R}$.

(ii) M is quasi-invariant if $M \subset \Phi_t(M)$, $\forall t \in \mathbb{R}$.

(iii) M is semi-invariant if $\Phi_t(M) \subset M$, $\forall t \in \mathbb{R}$.

(iv) M is strongly positively invariant if $\Phi_t(M) \subset M$, $\forall t > 0$.

(v) M is invariant (for the set-valued map h) if $\forall x \in M$, $\exists$ a solution $\mathbf{x}$ to the DI (A.5) with $\mathbf{x}(0) = x_0$ and with $\mathbf{x}(\mathbb{R}) \subset M$.

Definition A.10 ((ϵ, T)-Chain). Given a set $M \subset \mathbb{R}^d$, and $x, y \in M$, by an (ϵ, T)-chain from x to y, we mean a sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, for some integer $n \geq 1$, of solutions to the DI (A.5) together with real numbers $t_1, \ldots, t_n > T$, such that

(i) $\mathbf{x}_i(s) \in M$, $\forall 0 \leq s \leq t_i$ and $i = 1, \ldots, n$,

(ii) $\|\mathbf{x}_i(t_i) - \mathbf{x}_{i+1}(0)\| \leq \epsilon$, for all $i = 1, \ldots, n-1$,

(iii) $\|\mathbf{x}_1(0) - x\| \leq \epsilon$ and $\|\mathbf{x}_n(t_n) - y\| \leq \epsilon$.

Definition A.11 (Internally Chain Transitive and Chain Recurrent Sets)**.** We define these sets as follows:

(i) The set $M \subset \mathbb{R}^d$ is said to be internally chain transitive for the DI (A.5) if M is compact and for any $x, y \in M$, there exists an (ϵ, T)-chain for any $\epsilon, T > 0$.

(ii) If the property in part (i) above holds only for all $x = y \in M$, then the set M is said to be chain recurrent.

Definition A.12 (Perturbed Solution to a DI)**.** A function $z : [0, \infty) \to \mathbb{R}^d$ is said to be a perturbed solution to (A.5) if the following hold:

(i) z is absolutely continuous.

(ii) There exists a locally integrable function $U : [0, \infty) \to \mathbb{R}^d$ such that

(a) $\lim_{t\to\infty} \sup_{0\leq v\leq T} \| \int_t^{t+v} U(s)ds\| = 0$ for all $T > 0$.

(b) $\frac{dy(t)}{dt} - U(t) \in h^{\delta(t)}(y(t))$ for almost every $t > 0$, for some $\delta : [0, \infty) \to \mathbb{R}$ such that $\delta(t) \to 0$ as $t \to \infty$. Here $h^{\delta}(y) = \{x \in \mathbb{R}^d \mid \exists z \text{ s.t. } \|z - x\| < \delta, d(x, h(z)) < \delta\}$.

We now state a couple of important results, see (Benaïm *et al.*, 2005, Lemma 3.5 and Theorem 3.6).

Lemma A.12. Any internally chain transitive set for the DI (A.5) is invariant.

Theorem A.13. Let $\mathbf{z}$ be a bounded perturbed solution to the DI (A.5) with $\mathbf{z}(0) = z$. Then the limit set of $\mathbf{z}$ given by $L(z) = \bigcap_{t\geq 0} \overline{\{\mathbf{z}[t, +\infty)\}}$ is internally chain transitive for (A.5).

Definition A.13 (Attracting/Attractor and Lyapunov Stable Sets for a DI)**.** In relation to the DI (A.5), we have the following definitions:

(i) $A \subseteq \mathbb{R}^d$ is said to be an attracting set if it is compact and there exists a neighborhood U such that for any $\epsilon > 0$, $\exists\ T(\epsilon) \geq 0$ with $\Phi_{[T(\epsilon),+\infty)}(U) \subset N^{\epsilon}(A)$. In other words, any DI trajectory initiated in U reaches the ϵ-neighborhood of A, $T(\epsilon)$ instants later and stays there forever subsequently.

(ii) The set U above is called the fundamental neighborhood of A.

(iii) An attracting set A that is also invariant is called an attractor set.

(iv) The basin of attraction of A is the set $B(A) = \{x \in \mathbb{R}^d \mid w_{\Phi}(x) \subset A\}$. In other words, this is the largest subset of $\mathbb{R}^d$ such that the DI initiated anywhere within this set has its ω-limit set contained in A.

(v) The set A is said to be Lyapunov stable if for all $\delta > 0$, $\exists\ \epsilon > 0$ such that $\Phi_{[0,+\infty)}(N^{\epsilon}(A)) \subseteq N^{\delta}(A)$.

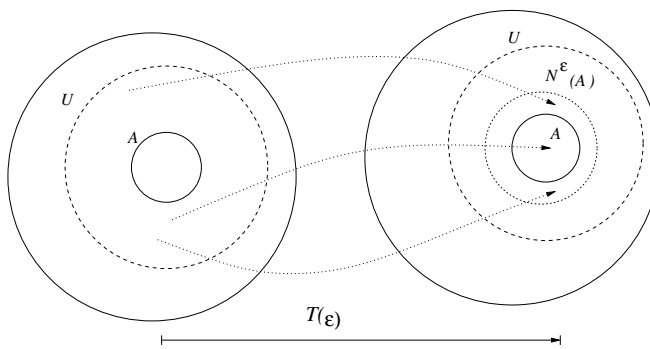


Figure A.2: The set A is attracting if (a) it is compact and (b) there is a neighborhood U of A such that given any $\epsilon > 0$, there exists $T(\epsilon) > 0$ so that any trajectory of the DI (A.5) starting in U arrives and stays within an ϵ-neighborhood of A beyond an amount of time $T(\epsilon)$ and subsequently stays in that neighborhood. Thus, $\Phi_t(U) \in N^{\epsilon}(A)$, $\forall t \in [T(\epsilon), \infty)$. The set A is an attractor if in addition it is also invariant.

A.3 Bibliographic Remarks

Differential equations have been well studied over many centuries with starting work primarily in the areas of physical and mechanical systems. Both Isaac Newton and Gottfried Leibniz are credited to have done early work in differential equations in the late 17th century. Early textbook treatments of ODEs include (Ince, 1956; Coddington *et al.*, 1956). Excellent, more recent, texts include (Arnold, 1992; Hirsch *et al.*, 2013; Nandakumaran *et al.*, 2017). An excellent treatment on differential equations with discontinuous right hand sides is given in (Filippov, 2013). Applications of such equations in many engineering domains have been well studied, for instance, see (Andronov *et al.*, 2013) for a recent English translation of a Russian text of the 1950's by these authors. Differential inclusions is a more general framework for dynamical systems that have differential equations with non-unique solutions resulting from lack of Lipschitz continuity and possibly even discontinuity of the right hand sides. Excellent texts on differential inclusions include (Aubin and Frankowska, 1990; Aubin and Cellina, 1984). We finally remark that a set-valued map h as in Definition A.6 has been referred to as Peano map in (Aubin and Frankowska, 1990) and as Marchaud map in (Benaïm *et al.*, 2005).

A.4 Exercises

Exercise 1. A function $V(\theta)$ satisfying $V(0) = 0$ and $V(\theta) > 0$ for $\theta \neq 0$ is said to be positive-definite. For $V(\theta) = a\theta_1^2 + 2\theta_1\theta_3 + a\theta_2^2 + 4\theta_2\theta_3 + a\theta_3^2$, identify the range of a that ensures positive-definiteness of V.

Exercise 2. Consider the ODE $\dot{\theta}(t) = A\theta(t)$, where $A = \begin{bmatrix} 0 & -1 \\ 1 & -1 \end{bmatrix}$. Answer the following:

(a) Is this system asymptotically stable?

(b) Suppose $V(\theta) = \theta^\mathsf{T} P\theta$ for some matrix symmetric, positive-definite P that ensures $PA + A^\mathsf{T} P$ is negative-definite. Show that V is a Lyapunov function for the linear system given above.

(c) Exhibit a Lyapunov function for the system given above through an appropriate choice of P.

Exercise 3. Consider the ODE (A.1) with a Lipschitz continuous function $h : \mathbb{R}^d \to \mathbb{R}^d$ with Lipschitz constant $L > 0$. Assume the ODE evolves over the time interval $[0, T]$ for some $T > 0$.

(i) Writing the integral form (A.2) of the ODE with two different initial conditions $\theta_1, \theta_2 \in \mathbb{R}^d$, obtain trajectories $\theta_1(t)$ and $\theta_2(t), t \in [0, T]$, respectively.

(ii) Write down an inequality (upper bound) for $\|\theta_1(t) - \theta_2(t)\|$, $t \in [0, T]$, using Lipschitz continuity of the function h.

(iii) Apply Gronwall's inequality (cf. Lemma A.1) on the inequality above using the functions $u, v : [0, T] \to [0, \infty)$, where $u(t) = \|\theta_1(t) - \theta_2(t)\|$ and $v(t) = 1$, $\forall t \in [0, T]$, to show that $u(t)$ is Lipschitz continuous as a function of the initial condition $u(0)$.

Exercise 4. Find the equilibria of the ODE system

$$\dot{\theta_1} = \theta_2, \ \dot{\theta_2} = -K \sin(\theta_1),$$

for some $K > 0$. Are these equilibria isolated?

Exercise 5. Consider the following ODE:

$$\dot{\theta}(t) = A\theta(t) + b,$$

where A is a $d \times d$ negative definite matrix, $b \in \mathbb{R}^d$ is a given vector and $\theta(t) \in \mathbb{R}^d, \forall t \geq 0$.

(i) Identify equilibria for this ODE?

(ii) Does this ODE have any attractors? If so, identify them?

(iii) Show that the following serves as a Lyapunov function for the above ODE:

$$W(\theta) = \frac{1}{2}(A\theta + b)^T(A\theta + b).$$

Exercise 6. Consider the following ODE in $\mathcal{R}^d$:

$$\dot{\theta}(t) = -P(\nabla^2 J(\theta(t)))^{-1}\nabla J(\theta(t)), \quad \theta(0) \in \mathcal{R}^d.$$

Here, $J : \mathcal{R}^d \to R$ is a twice continuously differentiable function. Further, P is an operator that uniquely maps all symmetric matrices to the space of positive definite and symmetric matrices. Let $\theta(t)$, $t \geq 0$ denote a trajectory of the above ODE. Giving precise arguments, describe the behaviour of $\theta(t)$ as $t \to \infty$? List down any assumptions that you may make.

Exercise 7. Consider the following set-valued map:

$$F(x) = \begin{cases} x & \text{if } x < 0 \\ 1 & \text{if } x > 0. \end{cases}$$

Further, for $x = 0$, $F(x) = [0, 1]$. Giving precise arguments, show whether or not

(i) F is a Peano or Marchaud map?

(ii) F is lower semicontinuous?

(iii) $\{0\}$ is strongly positively invariant for F?

(iv) $\{0\}$ is invariant for F?

Exercise 8. Let F be the set-valued map on $\mathbb{R}$ given by $F(x) = -sgn(x)$ for $x \neq 0$ and $F(0) = [-1, 1]$. Here $sgn(x) = +1$ if $x > 0$ and equals -1 otherwise.

(i) Show that $\{0\}$ is strongly positively invariant as well as invariant for the differential inclusion $\dot{x}(t) \in F(x(t))$?

(ii) Identify $\Phi_t(0)$, show whether or not $\{0\}$ is an attractor for the inclusion $\dot{x}(t) \in F(x(t))$ and if so, also identify the attracting set?

(iii) Show that $F(\cdot)$ is a Peano or Marchaud map?

B

Conditional expectations and martingales

In this appendix, we provide an introduction to conditional expectation, various notions of convergence of random variables, and martingales.

B.1 Conditional expectation

We first handle the case of a discrete r.v.

Definition B.1. The conditional probability mass function of a r.v. Y given $X = x$ is $p_{Y|X}(y \mid x) = \mathbb{P}[Y = y \mid X = x]$. Given $\{X = x\}$, the distribution of Y has a probability mass function $p_{Y|X}(y \mid x)$ and the expected value of this distribution, denoted as $\mathbb{E}[Y \mid X = x]$, is given by

$$\mathbb{E}[Y \mid X = x] = \sum_y y p_{Y|X}(y \mid x).$$

We shall use the notation $\mathbb{E}[Y \mid X]$ to denote the conditional expectation of Y given X.

Next, we extend the notion of conditional expectation to a continuous r.v.

Definition B.2. Suppose X, Y are continuous r.v.s with joint density f. Then, the conditional probability density function, denoted by $f_{Y|X}(y \mid$

x), is defined as follows:

$$f_{Y|X}(y \mid x) = \frac{f(x,y)}{f_X(x)}, \text{ for } x \text{ s.t. } f_X(x) > 0.$$

In the above, f_X denotes the marginal density of X.

The conditional expectation of Y given $\{X = x\}$, denoted as $\mathbb{E}[Y \mid X = x]$, is given by

$$\mathbb{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) dy.$$

As before, $\mathbb{E}[Y \mid X]$ denotes the conditional expectation of Y given X.

In the above definitions, we have followed a simpler definition that covers discrete and continuous r.v.s, along the lines of (Grimmett and Stirzaker, 2020). A more general definition of the conditional expectation of Y given a sigma field $\mathcal{F}$, denoted by $\mathbb{E}[Y|\mathcal{F}]$, is any random variable Z that is $\mathcal{F}$-measurable and satisfies

$$\mathbb{E}[Y\mathbb{I}_A] = \mathbb{E}[Z\mathbb{I}_A], \forall A \in \mathcal{F},$$

where $\mathbb{I}_A$ is the indicator function that takes value 1 on the set A and 0 otherwise.

Such a Z is unique almost surely, and this definition is more general in the sense that it does not rely on the existence of a conditional distribution, see (Durrett, 2019) for a detailed exposition. In the rest of this appendix, if not explicitly mentioned, equality between random variables should be interpreted in the almost sure sense.

We list some useful properties of conditional expectation.

Proposition B.1. The conditional expectation $\mathbb{E}[Y \mid X]$ satisfies the following properties:

1. $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}Y$.
2. $\mathbb{E}[\mathbb{E}[Y \mid X] g(X)] = \mathbb{E}[Y g(X)]$ for any g such that both expectations exist.
3. $\mathbb{E}[aY + bZ \mid X] = a\mathbb{E}[Y \mid X] + b\mathbb{E}[Z \mid X]$.
4. $\mathbb{E}[Y \mid X] = \mathbb{E}Y$ if X and Y are independent.

5. $\mathbb{E}[Yg(X) \mid X] = g(X)\mathbb{E}[Y \mid X]$ for g such that the expectations exist.

6. $\mathbb{E}[\mathbb{E}[Y \mid X, Z] \mid X] = \mathbb{E}[\mathbb{E}[Y \mid X] \mid X, Z] = \mathbb{E}[Y \mid X]$.

B.2 Notions of convergence of random variables

Definition B.3 (Almost sure or with probability 1 convergence)**.** Let $X_m, m \geq 0$ and X be random variables defined on a common probability space $(\Omega, \mathcal{F}, P)$. Then, $X_m \to X$ almost surely or $X_m \to X$ with probability 1 as $m \to \infty$ if $\mathbb{P}\left[w | \lim_{m\to\infty} X_m(w) = X(w)\right] = 1$.

A well-known example of almost sure convergence is the strong law of large numbers, which states that the sample mean converges almost surely to the true mean, under a bounded moment assumption.

Definition B.4 (Convergence in probability)**.** Let $X_m, m \geq 0$ and X be random variables defined on a common probability space $(\Omega, \mathcal{F}, P)$. Then, $X_m \xrightarrow{p} X$ if $\lim_{m\to\infty} \mathbb{P}[w|X_m(w) - X(w)| > \epsilon] = 0 \ \forall \ \epsilon > 0$, where $\mathbb{P}[w|X_m(w) - X(w)| > \epsilon]$ is usually written as $\mathbb{P}[|X_m - X| > \epsilon]$.

The weak law of large numbers is an example of convergence in probability for the sample mean of i.i.d. r.v.s.

Definition B.5 (L^2 or mean-squared convergence)**.** Let $X_m, m \geq 0$ and X be random variables defined on a common probability space $(\Omega, \mathcal{F}, P)$. Then $X_m \xrightarrow{L^2} X$ if $\mathbb{E}[|X_m(w) - X(w)|]^2 \to 0$ as $m \to \infty$, where $\mathbb{E}[|X_m(w) - X(w)|]^2$ is the mean squared error.

Definition B.6 (Convergence in distribution)**.** Let $X_m, m \geq 0$ and X be random variables (not necessarily defined on a common probability space). We say that $X_m \xrightarrow{d} X$ if $F_{X_m}(x) \longrightarrow F_X(x)$ at all points of continuity of F_X. Here $F_Y(\cdot)$ denotes the cumulative distribution function (CDF) of the random variable Y.

The reader is referred to (Borkar, 1995; Billingsley, 2013) for equivalent definitions of convergence in distribution.

It can be shown that

1. Almost sure convergence $\implies$ convergence in probability $\implies$ convergence in distribution.

2. Mean-squared convergence $\implies$ convergence in probability $\implies$ convergence in distribution.

For counterexamples that show that the converses of the above implications do not hold, the reader is referred to (Billingsley, 2017).

In this book, we provide almost sure convergence guarantees for the well-known gradient-based zeroth-order optimization algorithms.

B.3 Martingales

A filtration $\mathcal{F}_n$ is an increasing sequence of sigma fields. A sequence of random variables Y_n is said to be *adapted* to $\mathcal{F}_n$ if Y_n is $\mathcal{F}_n$-measurable, for all n.

A martingale is a stochastic process that is defined below.

Definition B.7. A sequence $\{Y_n, n \geq 1\}$ is a *martingale* with respect to the sequence $\{X_n, n \geq 1\}$ if, for all $n \geq 1$,

- $\mathbb{E}[|Y_n|] < \infty$;
- Y_n is adapted to $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$; and
- $\mathbb{E}[Y_{n+1}|X_1, \ldots, X_n] = Y_n$.

In particular, if $\mathbb{E}[Y_{n+1}|X_1, \ldots, X_n] = 0$, then $\{Y_n, n \geq 1\}$ is a *martingale difference* sequence.

The sequence $\{X_n\}$ can be the same as the $\{Y_n\}$ sequence for the conditions listed above to be valid.

Notice that

$$\begin{aligned}\mathbb{E}[\,Y_{n+2}|\,Y_1, Y_2, \ldots, Y_n] &= \mathbb{E}[\mathbb{E}[\,Y_{n+2}|\,Y_1, Y_2, \ldots, Y_{n+1}]|\,Y_1, Y_2, \ldots, Y_n] \\ &= \mathbb{E}[\,Y_{n+1}|\,Y_1, Y_2, \ldots, Y_n] = Y_n.\end{aligned}$$

Extending the argument, we have $\mathbb{E}[Y_{n+m}|Y_1, Y_2, \ldots, Y_n] = Y_n$, for any $m > 0$.

A few examples of martingales are given below.

Example B.1. Let $\{X_i\}$ be a sequence of random variables satisfying $\mathbb{E}[X_{i+1}|\, X_1, X_2, \ldots, X_i] = 0$, $\forall i$. Define $S_n = \sum_{i=1}^{n} X_i$. Then,

$$\begin{aligned}
&\mathbb{E}[S_{n+1}|\, S_1, S_2, \ldots, S_n] \\
&= \mathbb{E}[X_{n+1}|\, S_1, S_2, \ldots, S_n] + \mathbb{E}[S_n|\, S_1, S_2, \ldots, S_n] \\
&= \mathbb{E}[X_{n+1}|\, X_1, X_2, \ldots, X_n] + S_n = S_n.
\end{aligned}$$

Thus, $\{S_n\}$ is a martingale sequence.

Example B.2. Let $\{X_i\}$ be a sequence of i.i.d. random variables with mean one. Let $S_n = \prod_{i=1}^{n} X_i$. Then, $\{S_n\}$ is a martingale since

$$\mathbb{E}[S_{n+1}|\, S_1, S_2, \ldots, S_n] = \mathbb{E}[X_{n+1}S_n|\, S_1, S_2, \ldots, S_n] = \mathbb{E}[X_{n+1}]S_n = S_n.$$

Definition B.8. Let $\mathcal{F}_n$ be a filtration. A sequence $\{Y_n\}$ is a *martingale* w.r.t. the filtration $\mathcal{F}_n$ if, for all $n \geq 1$,

1. $\mathbb{E}[|Y_n|] < \infty$;
2. Y_n is adapted to $\mathcal{F}_n$; and
3. $\mathbb{E}[Y_{n+1}|\mathcal{F}_n] = Y_n$.

If the equality in the last condition above is replaced by a $\leq$ (resp. $\geq$), then the resulting sequence is a super (resp. sub) martingale.

Definition B.7 is retrieved by choosing $\mathcal{F}_n$ to be $\sigma(X_0, X_1, \ldots, X_n)$, which is the smallest σ-field with respect to which $X_1, \ldots, X_n$ are measurable. If Y is a martingale with respect to $\mathcal{F}$, then it is also a martingale with respect to $\mathcal{G}$ where $\mathcal{G}_n = \sigma(Y_1, \ldots, Y_n)$. This is because $\mathcal{G}_n$ is the smallest sigma algebra w.r.t which Y_n is measurable for every n, and thus, $\mathcal{G}_n \subset \mathcal{F}_n$, $\forall n$. Hence,

$$E[Y_{n+1}|\mathcal{G}_n] = E[E[Y_{n+1}|\mathcal{F}_n]|\mathcal{G}_n] = E[Y_n|\mathcal{G}_n] = Y_n \text{ a.s.}$$

Example B.3. Let $\{X_i\}$ be i.i.d. and let $S_n = \sum_{i=1}^{n} X_i$. Then,

(a) $\{S_n\}$ is a submartingale if $\mathbb{E}[X_i] \geq 0$; and

(b) $\{S_n\}$ is a supermartingale if $\mathbb{E}[X_i] \leq 0$.

Example B.4. Let $\{Z_n\}$ be a martingale sequence, and $S_n = Z_n - Z_{n-1}$. Then,

$$\begin{aligned}\mathbb{E}[S_n|S_1,\ldots,S_{n-1}] &= \mathbb{E}[Z_n|S_1,\ldots,S_{n-1}] - \mathbb{E}[Z_{n-1}|S_1,\ldots,S_{n-1}] \\ &= Z_{n-1} - Z_{n-1} = 0.\end{aligned}$$

Further, $\mathbb{E}[|S_n|] \leq \mathbb{E}[|Z_n|] + \mathbb{E}[|Z_{n-1}|] < \infty$. Thus, the sequence $\{S_n\}$ is a martingale difference.

B.3.1 Applications

Mean Estimation

Consider a r.v. Y with mean μ and variance σ^2. Suppose we are given i.i.d samples $Y_1, Y_2 \ldots Y_n$ from the distribution of Y. Let x_n denote the sample mean, i.e.,

$$x_n = \frac{1}{n}\sum_{k=1}^{n} Y_k.$$

We have

$$x_{n+1} = \frac{1}{n+1}\sum_{k=1}^{n+1} Y_{k+1} = \frac{n}{n+1}\left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right) + \frac{1}{n+1}Y_{n+1}.$$

Hence, sample mean can be iteratively computed as follows:

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

Instead of $\frac{1}{n+1}$, one can employ a more general step-size α_n satisfying standard stochastic approximation conditions, to arrive at the following update rule:

$$x_{n+1} = x_n + \alpha_n(Y_{n+1} - x_n). \tag{B.1}$$

Rewriting the equation above, we obtain

$$x_{n+1} = x_n + \alpha_n(Y_{n+1} - x_n) \tag{B.2}$$
$$= x_n + \alpha_n[(\mu - x_n) + (Y_{n+1} - \mu)] \tag{B.3}$$
$$= x_n + \alpha_n[(\mu - x_n) + w_{n+1}], \tag{B.4}$$

where $w_{n+1} = Y_{n+1} - \mu$ is the noise term. Notice that

$$\begin{aligned}\mathbb{E}[w_{n+1}|x_1, \ldots x_n] &= \mathbb{E}[w_{n+1}|Y_1, \ldots Y_n] \\ &= \mathbb{E}[Y_{n+1}|Y_1, \ldots Y_n] - \mu \\ &= \mathbb{E}[Y_{n+1}] - \mu = 0.\end{aligned}$$

Hence, $\{w_n\}$ is a martingale difference sequence.

Urn model

Suppose we have an empty urn to which we randomly add a red or a blue ball (one at a time) iteratively. Let us define

$$Y_{n+1} = \begin{cases} 1, & \text{if } (n+1)\text{th ball is red} \\ 0, & \text{otherwise} \end{cases}$$

$S_n = \sum_{k=1}^{n} Y_k$ denotes the total number of red balls. Then $x_n = \dfrac{S_n}{n}$ denotes the fraction of red balls. We have

$$\begin{aligned}x_{n+1} &= \frac{1}{n+1}\sum_{k=1}^{n+1} Y_k \\ &= (1 - \frac{1}{n+1})x_n + \frac{1}{n+1}Y_{n+1} \\ &= x_n + \alpha_n(Y_{n+1} - x_n),\end{aligned}$$

where $\alpha_n = \dfrac{1}{n+1}$, $n \geq 0$. Suppose the conditional probability that the next ball added at iterate $(n+1)$ is red, given the past, depends only on x_n, i.e.,

$$\mathbb{P}\left[Y_{n+1} = 1 | x_1 \ldots x_n\right] = p(x_n).$$

Then,

$$x_{n+1} = x_n + \alpha_n(p(x_n) - x_n) + w_{n+1},$$

where $w_{n+1} = Y_{n+1} - p(x_n)$. Notice that

$$\begin{aligned}\mathbb{E}(w_{n+1}|x_1, \dots x_n) &= \mathbb{E}(Y_{n+1} - p(x_n)|x_1, \dots x_n) \\ &= \mathbb{P}\left[Y_{n+1} = 1|x_1, \dots x_n\right] - p(x_n) \\ &= p(x_n) - p(x_n) = 0.\end{aligned}$$

Therefore, $\{w_n\}$ is a martingale difference sequence.

In the next section, we state and prove the well-known maximal inequality for martingales. This inequality will be used subsequently in the proof of the martingale convergence theorem. The latter claim helps in establishing asymptotic convergence of stochastic approximation algorithms with noise factors that are martingale differences.

B.3.2 Maximal inequality

We state and prove the *Doob-Kolmogorov Inequality* below.

Theorem B.1. If $\{S_n\}$ is a martingale with respect to $\{X_n\}$ then

$$\mathbb{P}\left(\max_{1\leq i\leq n} |S_i| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\mathbb{E}[S_n^2] \text{ for any } \epsilon > 0.$$

Proof. For the given $\epsilon > 0$, we define a partition of Ω as follows:

$$A_k \cup \left(\bigcup_{i=1}^{k} B_i\right) = \Omega,$$

where $A_0 = \Omega, A_k = \{|S_i| < \epsilon, \forall i \leq k\}$, and $B_k = A_{k-1} \cap \{|S_k| \geq \epsilon\}$. Here B_k denotes the even when $|S_i| \geq \epsilon$ for the first time with $i = k$.

The sets $B_1, \dots, B_k, A_k$ form a partition of Ω, which implies

$$\mathbb{E}[S_n^2] = \sum_{i=1}^{n} \mathbb{E}[S_n^2 I_{B_i}] + \mathbb{E}[S_n^2 I_{A_n}] \geq \sum_{i=1}^{n} \mathbb{E}[S_n^2 I_{B_i}].$$

Notice that

$$\begin{aligned}\mathbb{E}[S_n^2 I_{B_i}] &= \mathbb{E}[(S_n - S_i + S_i)^2 I_{B_i}] \\ &= \underbrace{\mathbb{E}[(S_n - S_i)^2 I_{B_i}]}_{(I)} + \underbrace{2\mathbb{E}[(S_n - S_i)S_i I_{B_i}]}_{(II)} + \underbrace{\mathbb{E}[S_i^2 I_{B_i}]}_{(III)}.\end{aligned}$$

Note that $(I) \geq 0$ and $(III) \geq \epsilon^2 P(B_i)$, because $|S_i| \geq \epsilon$ if B_i occurs. To deal with term (II), note that

$$\begin{aligned}\mathbb{E}[(S_n - S_i)S_i I_{B_i}] &= \mathbb{E}\left[S_i I_{B_i} \mathbb{E}[(S_n - S_i)|X_1, \ldots, X_i]\right] \\ &= 0,\end{aligned}$$

since B_i concerns $X_1, \ldots, X_i$ only, the inequality presented above becomes

$$\mathbb{E}[S_n^2] \geq \sum_{i=1}^{n} \epsilon^2 P(B_i) = \epsilon^2 \mathbb{P}\left[\max_{1 \leq i \leq n} |S_i| \geq \epsilon\right].$$

□

B.3.3 Martingale convergence theorem

Theorem B.2. Suppose $\{S_n\}$ is a martingale sequence satisfying $\mathbb{E}[S_n^2] < M < \infty$ for some M and $\forall\, n$. Then, there exists a r.v. S such that

1. $S_n \xrightarrow{a.s} S$ as $n \to \infty$;
2. $S_n \xrightarrow{L^2} S$ as $n \to \infty$ (mean-squared sense).

Proof. We begin with the proof of the first claim, i.e., almost sure convergence. Notice that S_m and $S_{m+n} - S_m$ are uncorrelated $\forall m, n \geq 1$ since $\mathbb{E}[S_m(S_{m+n} - S_m)] = 0$. Further,

$$\mathbb{E}[S_{m+n}^2] = \mathbb{E}[S_m^2] + \mathbb{E}[(S_{m+n} - S_m)^2] \geq \mathbb{E}[S_m^2].$$

Thus, $\{\mathbb{E}[S_n^2]\}$ is a non-decreasing sequence that is bounded above (by assumption). Choose M such that $\mathbb{E}[S_n^2] \uparrow M$ as $n \to \infty$. Now, it is enough to show that $\{S_n(\omega)\}_{n=1}$ is Cauchy convergent as it would imply almost sure convergence.
Let $C = \{\omega \mid S_n(\omega) \text{ is Cauchy convergent}\}$, i.e.,

$$C = \{\omega \mid \forall \epsilon > 0, \exists m \text{ such that } |S_{m+i}(\omega) - S_{m+j}(\omega)| < \epsilon \;\forall\, i, j \geq 1\}.$$

If $|S_{m+i} - S_m| < \epsilon$ and $|S_{m+j} - S_m| < \epsilon$ then $|S_{m+i} - S_{m+j}| < 2\epsilon$ by triangle inequality. So,

$$C = \{\omega | \forall \text{ (rational) } \epsilon > 0, \exists\, m \text{ s.t. } |S_{m+i}(\omega) - S_m(\omega)| < \epsilon,\; \forall\, i \geq 1\}$$

$$= \bigcap_{\epsilon>0} \bigcup_{m\geq 1} \{|S_{m+i} - S_m| < \epsilon, \forall i \geq 1\}$$

$$C^{\mathsf{c}} = \bigcup_{\epsilon>0} \bigcap_{m\geq 1} \{|S_{m+i} - S_m| \geq \epsilon, \text{ for some } i \geq 1\}.$$

Let $A_m(\epsilon) = \{|S_{m+i} - S_m| \geq \epsilon \text{ for some } i \geq 1\}$ then, $C^c = \bigcup_{\epsilon>0} \bigcap_{m\geq 1} A_m(\epsilon)$.
If $\epsilon \geq \epsilon'$, $A_m(\epsilon) \subseteq A_m(\epsilon')$.

We want $\mathbb{P}(C^c) = 0$. Notice that

$$0 \leq \lim_{\epsilon \downarrow 0} \mathbb{P}\left(\bigcap_m A_m(\epsilon)\right) \leq \lim_{\epsilon\downarrow 0} \lim_{m\to\infty} \mathbb{P}(A_m(\epsilon)).$$

If $\lim_{m\to\infty} \mathbb{P}(A_m(\epsilon)) = 0$ for any $\epsilon > 0$, then $\mathbb{P}(C^c) = 0$.
Let $Y_n = S_{m+n} - S_m$, for a fixed m. Then, $\{Y_n\}$ is a martingale since $\mathbb{E}[Y_{n+1}|Y_1, \ldots, Y_n] = Y_n$.
Applying the Doob-Kolmogorov inequality for Y_i, we obtain

$$\mathbb{P}(|Y_i| \geq \epsilon \text{ for some } 1 \leq i \leq n) \leq \frac{1}{\epsilon^2}\mathbb{E}\left[Y_n^2\right],$$

$$\mathbb{P}(|S_{m+i} - S_m| \geq \epsilon, \text{ for some } 1 \leq i \leq n) \leq \frac{\mathbb{E}\left[(S_{m+n} - S_m)^2\right]}{\epsilon^2},$$

$$0 \leq \mathbb{P}(A_m(\epsilon)) \leq \frac{\mathbb{E}(S_{m+n} - S_m)^2}{\epsilon^2} = \frac{\mathbb{E}[S_{m+n}^2] + \mathbb{E}[S_m^2] - 2\mathbb{E}[S_{m+n}S_m]}{\epsilon^2}.$$

Notice that

$$\begin{aligned}\mathbb{E}[S_{m+n}S_m] &= \mathbb{E}[\mathbb{E}[S_{m+n}S_m|S_1, \ldots, S_m]] \\ &= \mathbb{E}[S_m E[S_{m+n}|S_1, \ldots, S_m]] = E[S_m^2].\end{aligned}$$

Thus,

$$\begin{aligned}0 \leq \mathbb{P}\left[A_m(\epsilon)\right] &\leq \frac{\mathbb{E}[S_{m+n}^2] - E[S_m^2]}{\epsilon^2} \\ &\leq \lim_{n\to\infty} \frac{\mathbb{E}[S_{m+n}^2] - E[S_m^2]}{\epsilon^2} = \frac{M - \mathbb{E}[S_m^2]}{\epsilon^2} \\ \mathbb{P}[A_m(\epsilon)] &\leq \frac{M - \mathbb{E}[S_m^2]}{\epsilon^2}.\end{aligned}$$

As $m \to \infty$, $\mathbb{E}[S_m^2] \uparrow M$. Hence, $\lim_{m\to\infty} \mathbb{P}[A_m(\epsilon)] = 0$, implying $\mathbb{P}(C^{\mathsf{c}}) = 0$ (or) $\mathbb{P}(C) = 1$, i.e., the sequence $\{S_n\}$ is Cauchy convergent. Thus,

$\exists\, S$ such that $S_n \overset{a.s}{\to} S$ as $n \to \infty$.

We now turn to proving convergence in mean-squared sense. For this claim, we need *Fatou's Lemma*, which is stated as follows:
If $\{X_n\}$ is such that $X_n \geq 0, \forall n$, then

$$\mathbb{E}[\liminf_{n\to\infty} X_n] \leq \liminf_{n\to\infty} \mathbb{E}[X_n]$$

Notice that

$$\begin{aligned}
\mathbb{E}[(S_n - S)^2] &= \mathbb{E}[\liminf_{m\to\infty} (S_n - S_m)^2] && \text{(B.5)}\\
&\leq \liminf_{m\to\infty} \mathbb{E}[(S_n - S_m)^2] && \text{(Fatou's Lemma)}\\
&= M - \mathbb{E}[S_n^2] \overset{n\to\infty}{\longrightarrow} 0. && \text{(B.6)}
\end{aligned}$$

$\implies \mathbb{E}[(S_n - S)^2] \overset{n\to\infty}{\longrightarrow} 0$ or $S_n \overset{L^2}{\to} S$.

To arrive at the equality in (B.5), we used the following fact for a fixed n:

$$\begin{aligned}
\mathbb{E}\left[\lim_{m\to\infty} (S_n^2 + S_m^2 - 2S_m S_n)\right] &= \mathbb{E}[{S_n}^2 + S^2 - 2S_n S]\\
&= \mathbb{E}[(S_n - S)^2].
\end{aligned}$$

Further, (B.6) is justified as follows:

$$\begin{aligned}
\lim_{m\to\infty} \mathbb{E}[(S_n - S_m)^2] &= \lim_{m\to\infty} (\mathbb{E}[S_n^2] + \mathbb{E}[S_m^2] - 2\mathbb{E}[S_n S_m])\\
&= \lim_{m\to\infty} (\mathbb{E}[S_m^2] - \mathbb{E}[S_n^2])\\
&= M - \mathbb{E}[S_n^2].
\end{aligned}$$

Hence proved. □

B.3.4 More general martingale convergence results

We state here a few general martingale convergence theorems that are popular in the literature, see for instance, (Borkar, 1995, Chapter 3). As before, for a random variable X, let $X^+ \overset{\triangle}{=} \max(X, 0)$.

Theorem B.3. Let $(S_n, \mathcal{F}_n)$, $n \geq 0$, be a submartingale satisfying $\sup_n E[S_n^+] < \infty$. Then $S_n \to S$ a.s.

Theorem B.4. Let $(S_n, \mathcal{F}_n), n \geq 0$, be a martingale or a non-negative submartingale satisfying $\sup_n E[|S_n|^p] < \infty$, for some $p \in (1, \infty)$. Then there exists a random variable S such that

(i) $S_n \to S$ a.s.,

(ii) $S_n \overset{L^p}{\to} S$.

Definition B.9. 1. A sequence of random variables $\{S_n\}$ is said to be uniformly integrable (U.I.) if it is integrable and

$$\lim_{a \to \infty} \sup_n E[|S_n| I\{|S_n| \geq a\}] = 0.$$

2. A martingale $(S_n, \mathcal{F}_n), n \geq 0$, is said to be regular if there exists a random variable Y with $E[|Y] < \infty$, such that $S_n = E[Y|\mathcal{F}_n]$, $\forall n$.

Lemma B.5. $\{S_n\}$ is U.I. if and only if $\sup_n E[|S_n|] < \infty$ and

$$\lim_{P(A) \to 0} \sup_n \int_A |S_n| dP = 0.$$

Theorem B.6. Let $(S_n, \mathcal{F}_n)$, $n \geq 0$ be a martingale. Then the following are equivalent:

(i) $(S_n, \mathcal{F}_n)$, $n \geq 0$, is regular.

(ii) $\{S_n\}$ is U.I.

(iii) There exists a random variable S with $E[|S|] < \infty$ and $S_n \overset{L^1}{\to} S$.

(iv) $\sup_n E[|S_n|] < \infty$ and $S := \lim_{n \to \infty} S_n$ satisfies $S_n = E[S|\mathcal{F}_n]$, $\forall n$.

Note that the existence of the limiting random variable S in Theorem B.6(iv) follows from Theorem B.4.

We shall now consider the class of square integrable martingales $(S_n, \mathcal{F}_n), n \geq 0$, i.e., those for which $E[S_n^2] < \infty$, $\forall n$. Note that if $(S_n, \mathcal{F}_n), n \geq 0$, is a martingale, then

$$E[S_{n+1}^2 | \mathcal{F}_n] \geq (E[S_{n+1}|\mathcal{F}_n])^2$$

$$= S_n^2 \text{ a.s.}$$

The inequality above follows from the conditional Jensen's inequality, while the equality results from the martingale property. Thus, $(S_n^2, \mathcal{F}_n)$, $n \geq 0$, is a submartingale and by the Doob decomposition theorem,

$$S_n^2 = X_n + Z_n, \ n \geq 0,$$

where $(X_n, \mathcal{F}_n)$, $n \geq 0$, is a zero-mean martingale and $\{Z_n\}$ is the quadratic variation process where Z_n is obtained as

$$Z_n = \sum_{m=1}^{n} (E[S_m^2 | \mathcal{F}_{m-1}] - S_{m-1}^2) + E[S_0^2],$$

and is seen to be measurable w.r.t. $\mathcal{F}_{n-1}$, $\forall n \geq 0$, where $\mathcal{F}_{-1} = \{\Omega, \phi\}$. Note also that because $(S_n^2, \mathcal{F}_n)$, $n \geq 0$, is a submartingale, $Z_{n+1} \geq Z_n$ a.s., $\forall n$.

Theorem B.7. Let $(S_n, \mathcal{F}_n), n \geq 0$, be a square integrable martingale and let $\{Z_n\}$ be its associated quadratic variation process such that each Z_n is measurable w.r.t. $\mathcal{F}_{n-1}$. Let $Z_\infty = \lim_{n \to \infty} Z_n$ a.s. Then $\{S_n\}$ converges almost surely on the set $\{Z_\infty < \infty\}$. Also, $S_n = o(f(Z_n))$ on $\{Z_\infty = \infty\}$ for every increasing $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ satisfying

$$\int_0^\infty (1 + f(t))^{-2} dt < \infty.$$

Remark B.1. Even though the above theorems are given for scalar valued martingales, they continue to hold even with vector-valued martingales. In most of the asymptotic convergence analyses that we cover for our algorithms, Theorem B.7 is seen to be useful. Consider, for instance, the following stochastic approximation algorithm as in (2.1):

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

under the assumptions (A1)-(A4) of (Borkar, 2022, Chapter 2). We list these below for ease of reference.

(A1) The function $h : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous.

(A2) The step-size sequence $\{a(n)\}$ is a sequence of positive real numbers satisfying

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

(A3) The sequence $\{M_n\}$ forms a martingale difference sequence w.r.t. the filtration $\mathcal{F}_n \equiv \sigma(x_m, M_m, m \leq n)$, $n \geq 0$. In addition,

$$E[\|M_{n+1}\|^2|\mathcal{F}_n] \leq \check{U}(1 + \|x_n\|^2).$$

(A4) $\sup_n \|x(n)\| < \infty$ a.s.

Define now

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

Then $(S_n, \mathcal{F}_n)$, $n \geq 1$ forms a martingale sequence. Using the Euclidean norm, we obtain

$$E[\|S_n\|^2] = E[S_n^T S_n] = \sum_{m=0}^{n-1} a(m)^2 \|M_{m+1}\|^2] < \infty,$$

as the cross terms of the form $a(i)a(j)E[M_{i+1}^T M_{j+1}] = 0$, for all $i \neq j$. Further, the quadratic variation process $\{Z_n\}$ associated with $(S_n, \mathcal{F}_n), n \geq 1$ is the following:

$$\begin{aligned} Z_n &= \sum_{m=1}^{n} (E[\|S_m\|^2|\mathcal{F}_{m-1}] - \|S_{m-1}\|^2) + E[\|S_0\|^2] \\ &= \sum_{m=1}^{n} E[\|(S_m - S_{m-1})\|^2|\mathcal{F}_{m-1}] + E[\|S_0\|^2] \\ &= \sum_{m=1}^{n} a(m-1)^2 E[\|M_m\|^2|\mathcal{F}_{m-1}] + E[\|S_0\|^2] \\ &\leq \check{U} \sum_{m=1}^{n} a(m-1)^2(1 + \|x_{m-1}\|^2) + E[\|S_0\|^2], \end{aligned}$$

from (A3). Thus,

$$Z_n \leq \check{U} \sum_{m=1}^{n} a(m-1)^2(1 + \sup_m \|x_{m-1}\|^2) + E[\|S_0\|^2].$$

It now follows from (A2) (the square summability of step-size condition) and (A4) (stability of the iterates) that

$$Z_n \to Z_\infty \leq \check{U} \sum_{m=1}^{\infty} a(m-1)^2(1 + \sup_m \|x_{m-1}\|^2) + E[\|S_0\|^2] < \infty \text{ a.s.},$$

from (A2) and (A4). From Theorem B.7, it will then follow that the martingale sequence $\{S_m\}$ converges almost surely.

B.4 Bibliographic remarks

The background material presented in this appendix is based on (Grimmett and Stirzaker, 2020; Borkar, 1995; Billingsley, 2017).

B.5 Exercises

Exercise 1. Suppose $X_n \xrightarrow{\mathrm{d}} X$ and $Y_n \xrightarrow{\mathrm{p}} c$ for some constant c. Show that $X_n Y_n \xrightarrow{\mathrm{d}} cX$.

Exercise 2. Suppose $X_n \xrightarrow{\mathrm{p}} X$ and $Y_n \xrightarrow{\mathrm{p}} Y$. Show that $X_n Y_n \xrightarrow{\mathrm{p}} XY$.

Exercise 3. Suppose $X_n \xrightarrow{\mathcal{L}_1} X$ and $Y_n \xrightarrow{\mathcal{L}_1} Y$. Disprove the following claim: $X_n Y_n \xrightarrow{\mathcal{L}_1} XY$.

Exercise 4. Suppose $X_n \xrightarrow{\mathcal{L}_2} X$ as $n \to \infty$. Show that

$$\text{Variance}(X_n) \to \text{Variance}(X) \text{ as } n \to \infty.$$

Exercise 5. Answer the following questions to understand the relation between convergence in probability and in distribution.

(a) Prove that convergence in probability implies convergence in distribution, and give a counterexample to show that the converse need not hold.

(b) Show that convergence in distribution to a constant random variable implies convergence in probability to that constant.

Exercise 6. Let $\{X_n\}$ and $\{Y_n\}$ be martingale sequences on a common probability space, i.e., for all n,

$$\mathbb{E}\left[|X_n| + |Y_n|\right] < \infty, \mathbb{E}\left[X_{n+1} \mid Z_1, \ldots, Z_n\right] = X_n, \text{ and}$$
$$\mathbb{E}\left[Y_{n+1} \mid Z_1, \ldots, Z_n\right] = Y_n.$$

Show that, for $m \leq n$,

$$\mathbb{E}\left[X_n Y_m \mid Z_1, \ldots, Z_m\right] = X_m Y_m.$$

Exercise 7. Let $\{X_n\}$ be a martingale sequence.

Consider the following two statements:

I: For all $n \geq 1$, $\mathbb{E}[X_n] = E[X_1]$.
II: For all $n \geq 1$, Variance(X_n) = Variance(X_1).
III: For all $n \geq 1$, Variance$(X_n) \geq$ Variance(X_1).
IV: For all $n \geq 1$, Variance$(X_n) \leq$ Variance(X_1).

Which of the statements above are true?

Exercise 8. Let $\{X_i\}$ be a i.i.d. sequence of random variables with mean zero and variance σ^2. Define $S_n = X_1 + \ldots + X_n$, and $Y_n = S_n^2 - n\sigma^2$. Show that $\{Y_n\}$ is a martingale sequence.

Exercise 9. Let $X_i, i = 1, 2, \ldots$ be a sequence of independent random variable with common mean μ and variance $\mathbb{E}\left[(X_i - \mu)^2\right] \leq k^{3/2}$. Let $\overline{X}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ be the sample mean. Does $\overline{X}_n$ converge in the mean-squared sense to μ?

Exercise 10. Let $\{X_i\}$ be a i.i.d. sequence of positive random variables with mean one. Define $Y_n = \displaystyle\prod_{i=1}^{n} X_i$.

Consider the following two statements:

I: $\{Y_n\}$ is a martingale sequence.
II: $\{\sqrt{Y_n}\}$ is a supermartingale sequence.
II: $\{\sqrt{Y_n}\}$ is a submartingale sequence.

Which of the statements above are true?

Exercise 11. Let $\{X_n\}$ be a martingale sequence, with $X_n \in [0, 1]$, $\forall n$. Does X_n converges almost surely?

Exercise 12. Let $\{X_n\}_{n\geq 1}$ be a sequence of independent random variables (r.v.s). Let $f : \mathbb{R} \to \mathbb{R}$ be a function such that $\mathbb{E}[|f(X_n)|] < \infty$, $\forall n$. Let $a_n = \mathbb{E}(f(X_n))] \neq 0$, $\forall n$. Define

$$S_n = \frac{\prod_{m=1}^{n} f(X_m)}{\prod_{m=1}^{n} a_m}, \forall n \geq 1.$$

Answer the following:

(a) Is $\mathbb{E}[|S_n|] < \infty,\ \forall n$?

(b) Is $\{S_n\}$ a martingale sequence?

Exercise 13. Suppose X_n, $n \geq 0$ is a sequence of real-valued random variables adapted to a filtration $\{\mathcal{F}_n\}$. Let $h : \mathcal{R} \to \mathcal{R}$ be a given function. Define a sequence $\{R(n)\}$ of random variables according to

$$R(n) = \sum_{m=1}^{n} \gamma(m)(h(X_m) - E[h(X_m) \mid \mathcal{F}_{m-1}]),$$

$n \geq 1$, where $\gamma(n)$, $n \geq 1$ are some positive scalars. Give suitable conditions on $\gamma(n)$, $n \geq 1$ and $h(\cdot)$ under which $(R(n), \mathcal{F}_n)$, $n \geq 1$ is (i) a martingale, (ii) a square integrable martingale, and (iii) an almost surely convergent martingale sequence?

Exercise 14. Let $(X_n, \mathcal{F}_n)$, $n \geq 0$ be a supermartingale, and define

$$\begin{aligned} Y_0 &= X_0, \\ Y_n &= Y_{n-1} + (X_n - E[X_n \mid \mathcal{F}_{n-1}]),\ n \geq 1. \end{aligned}$$

Also define

$$\begin{aligned} A_0 &= 0, \\ A_n &= A_{n-1} + (X_{n-1} - E[X_n \mid \mathcal{F}_{n-1}]),\ n \geq 1. \end{aligned}$$

(i) Express X_n in terms of Y_n and A_n for general $n \geq 0$.

(ii) What kind of a process is $(Y_n, \mathcal{F}_n)$, $n \geq 0$?

(iii) Is $\{A_n\}$ an increasing or a decreasing sequence? Prove your claim.

C

Markov chains

In this appendix, we provide an introduction to discrete time Markov chains (DTMCs), covering the main results concerning their transient and limiting behavior. This background is essential for understanding stochastic approximation algorithms, where the observation noise originates from a Markov chain. This setting was covered earlier in Section 2.6.

C.1 Introduction

Definition C.1. A stochastic process $\{X_n, n \geq 0\}$ with a countable state space $\mathcal{X}$ is a DTMC if $X_n \in \mathcal{X}, \forall n \geq 0$, and $\forall n \geq 0, i, j \in \mathcal{X}$,

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, ..X_0 = i_0) = P(X_{n+1} = j | X_n = i).$$

The condition above is the well-known Markov property, which in simple terms means the future is independent of the past, given the present.

A DTMC with countable state space S is time-homogeneous if

$$P(X_{n+1} = j | X_n = i) \stackrel{\triangle}{=} P_{i,j}(n) = P_{i,j}, \forall n \geq 0, \forall i, j \in S.$$

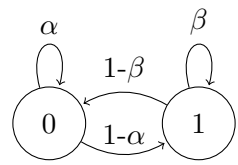
In other words, the transition probabilities are time invariant. This is the case we consider in this book. Note that

$$P_{i,j} \geq 0, \ \forall i, j \in \mathcal{X}; \ \sum_{j \in \mathcal{X}} P_{i,j} = 1.$$

When the state space is finite, we can write a transition probability matrix $M = [[P_{i,j}]]_{i,j=1,\ldots,|\mathcal{X}|}$.

We shall use the following DTMC as a running example in this appendix.

Example C.1. Consider the following two state DTMC for some $0 \leq \alpha, \beta \leq 1$: The transition probability matrix $P = \begin{bmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{bmatrix}$.



A relevant question is if the transition probability matrix is enough to derive the finite-dimensional distributions, i.e,

$$P(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n), \ i_0, i_1, \ldots, i_n \in \mathcal{X}.$$

The answer is no. The additional information required is the initial distribution, i.e., a $|\mathcal{X}|$-vector a with entries $a_i = P(X_0 = i), \forall i \in \mathcal{X}$. In such a case, it is easy to see from the Markov property that

$$P(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = a_{i_0} P_{i_0,i_1} P_{i_1,i_2} \cdots P_{i_{n-1},i_n}.$$

C.2 Transient behavior

Let $\{X_n, n \geq 0\}$ be a DTMC with state space $\mathcal{X} = \{0, 1, 2...\}$, initial distribution a and transition probability matrix P. We now derive the marginal distribution of X_n, i.e., $a_j^{(n)} \triangleq P(X_n = j), j \in \mathcal{X}$. Notice that

$$P(X_n = j) = \sum_{i \in \mathcal{X}} P(X_n = j | X_0 = i) P(X_0 = i)$$

$$= \sum_{i \in \mathcal{X}} P(X_n = j | X_0 = i) a_i$$
$$= \sum_{i \in \mathcal{X}} a_i P_{i,j}^{(n)},$$

where $P_{i,j}^{(n)} = P(X_n = j | X_0 = i),\ \forall i, j \in \mathcal{X}, n \geq 0$.

The n-step transition probabilities $P_{i,j}^{(n)}$ satisfy the following relation, known as Chapman-Kolmogorov equations.

$$P_{i,j}^{(n)} = \sum_{r \in X} P_{i,r}^{(k)} P_{r,j}^{(n-k)}, \forall i, j \in S, 0 \leq k \leq n. \tag{C.1}$$

Letting $P_{(r)}$ be the r-step transition probability matrix with entries $P_{i,j}^{(r)}$, for any $r \geq 0$, the relation in (C.1) can be compactly re-written as follows:

$$P_{(n)} = P_{(k)} P_{(n-k)}, 0 \leq k \leq n.$$

Example C.2. Consider a random walk with the following transition probabilities:

$$P_{i,i+1} = p, P_{i,i-1} = q = 1 - p, \forall i,$$

where $0 < p < 1$. To find $P_{0,0}^{(n)} = P(X_n = 0 | X_0 = 0)$, note that for an odd n, $P_{0,0}^{(n)} = 0$ as an even number of steps is necessary to return to the starting position. On the other hand, if n is even, say $n = 2k$, then of the $2k$ steps, k steps move forward and the remaining move backward, so that at the end of $2k$ steps, the DTMC is in state 0. Therefore,

$$P_{0,0}^{(2k)} = ((2k)!/k!k!) p^k q^k.$$

Example C.3. Consider a DTMC with state space $\mathcal{X} = \{1, 2, 3, 4\}$ and initial distribution $a = \{0.25, 0.25, 0.25, 0.25\}$. The transition probability matrix $P = \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 \\ 0.25 & 0.25 & 0.5 & 0 \\ 0.5 & 0 & 0.1 & 0.4 \\ 0 & 0 & 0.4 & 0.6 \end{bmatrix}$.

The marginal distribution of X_4, i.e.,

$$a^{(4)} = \begin{bmatrix} P(X_4 = 1), & P(X_4 = 2), & P(X_4 = 3), & P(X_4 = 4) \end{bmatrix},$$

can be found using the following relation:

$$a^{(4)} = aP^4.$$

Definition C.2 (First Passage times)**.** Let $\{X_n, n \geq 0\}$ be a DTMC on $\mathcal{X} = \{0, 1, 2, ..\}$. The first passage time T to a state k is defined as

$$T = min\{n \geq 0 | X_n = k\}.$$

Two interesting quantities that are related to T are: (i) The complementary CDF $P(T > n), n \geq 0$; and (ii) Probability of eventually hitting state k: $P(T < \infty)$. The example below illustrates the aforementioned quantities.

Example C.4. For the two state DTMC in Example C.1, let $T = min\{n \geq 0 | X_n = 1\}$. Then, $V_2(n) = P(T > n | X_0 = 2) = \beta^n$, and $P(T = n | X_0 = 2) = V_2(n-1) - V_2(n) = \beta^{n-1}(1-\beta)$.

Definition C.3 (Occupancy times)**.** Let $\{X_n, n \geq 0\}$ be a DTMC on $\mathcal{X} = \{0, 1, 2, ..\}$. Let $V_j^{(n)}$ denote the number of visits to state j up to time n (including 0). Then, the occupancy time of j up to n starting in state i is defined as

$$M_{i,j}^{(n)} = E(V_j^{(n)} | X_0 = i), i, j \in S, n \geq 0.$$

Let the occupancy matrix be $M^{(n)} = [M_{i,j}^{(n)}]$. Then, $M^{(n)}$ can be calculated as follows:

$$M^{(n)} = \sum_{r=0}^{n} P^r, n \geq 0.$$

Example C.5. Consider a DTMC with state space $\mathcal{X} = \{A, B, C\}$, and transition probability matrix $P = \begin{bmatrix} 0.1 & 0.2 & 0.7 \\ 0.2 & 0.4 & 0.4 \\ 0.1 & 0.3 & 0.6 \end{bmatrix}$.

Then, the occupance matrix with $n = 9$ is given by

$$M^{(9)} = \sum_{r=0}^{9} P^r = \begin{bmatrix} 2.14 & 2.74 & 5.12 \\ 1.26 & 3.95 & 4.78 \\ 1.15 & 2.85 & 6 \end{bmatrix}.$$

So far, we have seen that marginal distributions and passage times are useful in characterizing the transient behaviour of DTMCs. In the next section, we turn our attention to the limiting behavior of DTMCs.

C.3 Limiting behavior

To understand the limiting behavior of a DTMC, we pose the following two questions:

(Q1) With $P_{i,j}^{(n)}$ denoting the probability of going from state i to state j in n steps, does $P^{(n)}$ converge as $n \to \infty$?

(Q2) Recall the occupancy Matrix $M^{(n)} = \sum_{r=0}^{n} P^r$, with entries $M_{i,j}^{(n)}$ denoting the number of visits to state j starting from state i up to time n. Does $\dfrac{M_{i,j}^{(n)}}{n+1}$ converge as $n \to \infty$?

Example C.6. Consider the two state DTMC from Example C.1 with $\alpha + \beta < 2$. By an induction argument, it can be shown that

$$P^n = \frac{1}{2-\alpha-\beta}\begin{bmatrix} 1-\beta & 1-\alpha \\ 1-\beta & 1-\alpha \end{bmatrix} + \frac{(\alpha+\beta-1)^n}{2-\alpha-\beta}\begin{bmatrix} 1-\alpha & \alpha-1 \\ \beta-1 & 1-\beta \end{bmatrix}.$$

Taking limits, it is apparent that

$$\lim_{n\to\infty} P^n = \frac{1}{2-\alpha-\beta}\begin{bmatrix} 1-\beta & 1-\alpha \\ 1-\beta & 1-\alpha \end{bmatrix}.$$

Similarly, by induction, one can obtain the following result:

$$M^n = \frac{n+1}{2-\alpha-\beta}\begin{bmatrix} 1-\beta & 1-\alpha \\ 1-\beta & 1-\alpha \end{bmatrix} + \frac{1-(\alpha+\beta-1)^{n+1}}{(2-\alpha-\beta)^2}\begin{bmatrix} 1-\alpha & \alpha-1 \\ \beta-1 & 1-\beta \end{bmatrix}.$$

It is easy to see that

$$\lim_{n\to\infty} \frac{M^n}{n+1} = \frac{1}{2-\alpha-\beta}\begin{bmatrix} 1-\beta & 1-\alpha \\ 1-\beta & 1-\alpha \end{bmatrix}.$$

In this example, P^n and $\dfrac{M^n}{n+1}$ converge to the same limit. This is not in general true for all DTMCs, as the next example demonstrates.

Example C.7. Consider a three-state DTMC with transition probability matrix $P = \begin{bmatrix} 0 & 1 & 0 \\ q & 0 & p \\ 0 & 1 & 0 \end{bmatrix}$, for some $0 < p < 1$ and $q = 1 - p$. It is easy

to see that $P^{2n} = \begin{bmatrix} q & 0 & p \\ 0 & 1 & 0 \\ q & 0 & p \end{bmatrix}$, and $P^{2n+1} = P$. Thus, P^n does not converge.

On the other hand, it can be shown that

$$M^{2n} = \begin{bmatrix} 1+nq & 1+n & np \\ nq & 1+n & np \\ nq & n & 1+np \end{bmatrix} \text{ and } M^{2n+1} = \begin{bmatrix} 1+nq & 1+n & np \\ (n+1)q & 1+n & (n+1)p \\ nq & n+1 & 1+np \end{bmatrix}.$$

Thus, both $\frac{M^{2n}}{2n+1}$ and $\frac{M^{2n+1}}{2n+2}$ converge to $\begin{bmatrix} \frac{q}{2} & \frac{1}{2} & \frac{p}{2} \\ \frac{q}{2} & \frac{1}{2} & \frac{p}{2} \\ \frac{q}{2} & \frac{1}{2} & \frac{p}{2} \end{bmatrix}$ as $n \to \infty$.

To understand the limiting behavior of DTMCs, we require the notions of communicating classes, recurrence and transience. We introduce these concepts next.

Definition C.4 (Accessibility and communication)**.** A state j is said to be *accessible* from state i if $\exists n \geq 0$ such that $P_{i,j}^{(n)} > 0$. If state j is accessible from state i, we write $i \to j$. Notice that $i \to j$ implies that there exists a directed path from state i to state j in the transition diagram.
States i and j communicate if $i \to j$ and $j \to i$. We shall use $i \leftrightarrow j$ to denote that state i communicates with state j.

Using the definition above, it can be inferred that communication is an equivalence relation, i.e.,

1. $i \leftrightarrow i$ (reflexive);
2. if $i \leftrightarrow j$ then $j \leftrightarrow i$ (symmetric);
3. if $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$(transitive).

Definition C.5 (Communicating class)**.** A set C is a communicating class if the following properties hold:

1. $i \in C, j \in C \implies i \leftrightarrow j$;
2. $i \in C, i \leftrightarrow j \implies j \in C$ - this makes C maximal

In addition, if any $i \in C$ and $j \notin C$ do not communicate, then the class C is said to be closed.

Notice that if $X_n \in C$ for some n, and C is a closed communicating class, then $X_m \in C$, for all $m \geq n$.

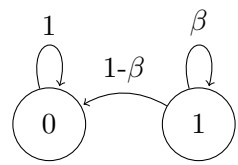
The state space of a DTMC can be partitioned as follows:

$$\mathcal{X} = C_1 \cup C_2 \cup ... \cup C_k \cup T, \tag{C.2}$$

for some $k \geq 1$, where $C_1, C_2..., C_k$ are closed communicating classes and the remaining states form T. The latter set of states are transient — a notion that we define below in Definition C.7.

Definition C.6 (Irreducibility)**.** If the state space $\mathcal{X}$ is a single closed communicating class then the DTMC is said to be irreducible.

Example C.8. For the two state DTMC from Example C.1, if $0 <= \alpha, \beta < 1$, then $\{0, 1\}$ is a closed communicating class and the DTMC is irreducible. On the other hand, if $\alpha = 1$, then we have the following transition diagram: In this case, $\{1\}$ is not a closed communicating



class, whereas $\{0\}$ is closed. The state partition, see (C.2), would be the union of closed class $\{0\}$ and $T = \{1\}$.

Recurrence and transience

Let $\tilde{T}_i = \min\{n > 0 | X_n = i\}, i \in \mathcal{X}$ denote the first time (after time instant 0) when the chain hits state i, $\tilde{u}_i = P(\tilde{T}_i < \infty | X_0 = i)$ denote the probability of returning back to state i, and $\tilde{m}_i = E(\tilde{T}_i | X_0 = i)$ denote the expected number of steps to return to state i. Using these quantities, we define the notion of recurrence/transience of a state below.

Definition C.7. A state $i \in \mathcal{X}$ is said to be recurrent if $\tilde{u}_i = 1$ and transient if $\tilde{u}_i < 1$. Further, a recurrent state is said to be positive recurrent if $\tilde{m}_i < \infty$ and null recurrent if $\tilde{m}_i = \infty$.

Note that $\tilde{m}_i = \infty$ for a transient state. The recurrence and transience properties carry over to all states within any communicating class, i.e.,

(a) i is transient, $i \leftrightarrow j \implies j$ is transient; and

(b) i is recurrent , $i \leftrightarrow j \implies j$ is recurrent.

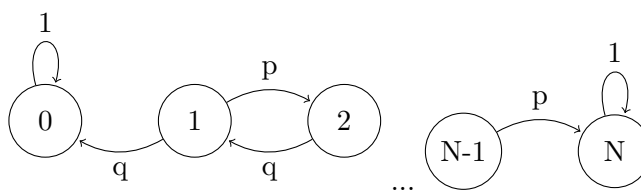
Similarly, positive and null recurrence are also class properties.

A communicating class is called

- transient if all its states are transient;
- positive recurrent if all its states are positive recurrent;
- null recurrent if all its states are null recurrent.

An irreducible DTMC is positive/null recurrent if all its states are positive/null recurrent.

Example C.9. Consider the following random walk: $P_{0,0} = P_{N,N} = 1$, $P_{i,i+1} = p$ and $P_{i,i-1} = q$, $0 < p, q < 1$, $p + q = 1$. The transition diagram is given below.



It is easy to see that 0 and N are recurrent states and the remaining states are transient.

Stationary distribution

Definition C.8. For a DTMC with transition probability matrix P, the vector $\pi = (\pi_i, i \in \mathcal{X})$ is called a stationary distribution if

1. $\pi_i \geq 0$, $\forall i$ and $\sum_i \pi_i = 1$.

2. $\pi = \pi P$.

The first condition above implies π is a distribution, while the second condition relates to stationarity. In particular, if the initial distribution of the DTMC is $X_0 \sim \pi$, then the distribution of the state X_n (at instant n) is:

$$\pi P^n = \pi P P^{n-1} = \pi P^{n-1} = \ldots = \pi.$$

The main result concerning the existence of stationary distribution is given below.

Theorem C.1. Consider an irreducible DTMC $\{X_n\}$. Then,

1. There exists a stationary distribution if and only if some state in the DTMC is positive recurrent.

2. If there exists a stationary distribution π, then every state is positive recurrent, and

$$\pi_i = \frac{1}{m_i}, \text{ where } m_i = \mathbb{E}[T_i \mid X_0 = i],$$

and $T_i = \min\{n \geq 1 \mid X_n = i\}$.

3. π is unique.

Example C.10. For the two state DTMC in Example C.1 with $\alpha+\beta < 2$, the set of equations for finding the stationary distribution are as follows:

$$\begin{aligned}
\pi_0 &= \alpha\pi_0 + (1-\beta)\pi_1, \\
\pi_1 &= (1-\alpha)\pi_0 + \beta\pi_1, \\
\pi_0 + \pi_1 &= 1.
\end{aligned}$$

Solving, we obtain $\pi_0 = \dfrac{1-\beta}{1-\alpha-\beta}$, and $\pi_1 = \dfrac{1-\alpha}{1-\alpha-\beta}$. This coincides with the limit of P^n as well as $M^n/(n+1)$, as discussed in Example C.6. We discuss convergence to stationary distribution next.

Periodicity

We require the notion of period associated with a recurrent state before we understand convergence to stationary distribution. We define this notion below.

Definition C.9. Let $\tilde{T}_i = \min\{n > 0 : X_n = i\}, i \in \mathcal{X}$. Let i be a recurrent state and d the largest positive integer such that

$$\sum_{k=1}^{\infty} P(\tilde{T}_i = kd) = 1.$$

If $d = 1$, then state i is aperiodic. On the other hand, if $d > 1$, then state i is said to be periodic with period d.

Equivalently, if i is a recurrent state with period d, then $P_{i,i}^{(n)} = 0$ for all n that are not positive integer multiples of d.

Remark C.1. Periodicity is a class property, i.e., if $i \leftrightarrow j$, then i, j have the same period.

As an example, consider a symmetric random walk, i.e., $P_{i,i+1} = P_{i,i-1} = 1/2$ for all i. In this case, it is easy to see that the period of state 0 is two and by using the fact that the chain is irreducible, all states have period 2.

Convergence to stationary distribution

In Example C.6, we observed that P^n and $\dfrac{M^n}{n+1}$ both converged, whereas in Example C.7, P^n did not converge. The distinguishing feature between these two examples is that one of the is aperiodic and the other not. The result below formalizes convergence to stationary distribution for aperiodic chains.

Theorem C.2. Let $\{X_n\}$ be an irreducible, recurrent, aperiodic DTMC. Then, for any $i, j \in \mathcal{X}$, we have

$$P_{i,j}^{(n)} \to \pi_j \text{ as } n \to \infty,$$

where π is the (unique) stationary distribution.

A easy counterexample that emphasizes the need for aperiodicity to ensure $P^{(n)}$ converges is a two state DTMC with $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

For periodic DTMCs, one can claim convergence to stationary distribution in the so-called "Cesaro sense". We formalize this statement next.

Let $\tilde{V}_n(j) = \sum_{m=1}^{n} \mathbb{I}\{X_m = j\}$ denote the occupancy measure for state j. We are interested in knowing if the time-averaged occupancy, i.e., $\frac{1}{n}\tilde{V}_n(j)$, converges to the stationary distribution as $n \to \infty$ for an irreducible recurrent DTMC. Such a law of large numbers type result is stated next.

Theorem C.3. Let $\{X_n\}$ be an irreducible, recurrent DTMC. Then, for any $j \in \mathcal{X}$, we have the following for any start state:

$$\frac{1}{n}\tilde{V}_n(j) \to \frac{\mathbb{I}\{T_j < \infty\}}{m_j} \text{ a.s. as } n \to \infty. \tag{C.3}$$

A few remarks are in order.

Remark C.2. From the result above, we have $\frac{1}{n}\tilde{V}_n(j)$ converges to $\pi_j = \frac{1}{m_j}$ if j is positive recurrent, and to 0 otherwise. The latter case includes null recurrent states, and a similar claim can be shown for transient states as well.

Remark C.3. Notice that

$$\mathbb{E}\left[\frac{1}{n}\tilde{V}_n(j) \mid X_0 = i\right] = \frac{1}{n}\sum_{m=1}^{n} \mathbb{E}\left[\mathbb{I}\{X_m = j\} \mid X_0 = i\right] = \frac{1}{n}\sum_{m=1}^{n} P^{(m)}(i,j),$$

and

$$\mathbb{E}\left[\frac{\mathbb{I}\{T_j < \infty\}}{m_j} \mid X_0 = i\right] = \frac{\mathbb{P}\left(T_j < \infty \mid X_0 = i\right)}{m_j}.$$

Thus,

$$\frac{1}{n}\sum_{m=1}^{n} P^{(m)}(i,j) \to \frac{\mathbb{P}\left(T_j < \infty \mid X_0 = i\right)}{m_j} \text{ a.s. as } n \to \infty.$$

The limit on the RHS above is zero for null recurrent states j and likewise positive for positive recurrent j.

Remark C.4. For a transient state j, $\sum_{m=1}^{\infty} P^{(m)}(i,j) < \infty$. Hence,

$$\mathbb{E}\left[\frac{1}{n}\tilde{V}_n(j) \mid X_0 = i\right] = \frac{1}{n}\sum_{m=1}^{n} P^{(m)}(i,j) \to 0 \text{ a.s. as } n \to \infty.$$

C.4 Bibliographic remarks

There are several excellent textbooks for Markov chains, for instance, (Levin and Peres, 2017; Meyn and Tweedie, 2012; Grimmett and Stirzaker, 2020; Norris, 1998; Gallager, 2013; Kulkarni, 2016). Our treatment is based on a combination of (Kulkarni, 2016) and (Grimmett and Stirzaker, 2020).

C.5 Exercises

Exercise 1. Let $\{X_n, n \geq 0\}$ be a DTMC. Then,

$$\mathbb{P}(X_0 = i, X_2 = k \mid X_1 = j) = \mathbb{P}(X_0 = i \mid X_1 = j)\,\mathbb{P}(X_2 = k \mid X_1 = j).$$

Exercise 2. Suppose $\{X_n, n \geq 0\}$ and $\{Y_n, n \geq 0\}$ are two independent DTMCs with state-space $S = \{0, 1, 2, \ldots\}$. Prove or give a counterexample to the following statements:

(a) $\{X_n + Y_n, n \geq 0\}$ is a DTMC.

(b) $\{(X_n, Y_n), n \geq 0\}$ is a DTMC.

Exercise 3. Consider a DTMC on state space $\{1, 2, 3, 4, 5\}$, with the following transition probability matrix:

$$P = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 1 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0.5 \end{bmatrix}.$$

Answer the following:

(a) Is the DTMC irreducible? Aperiodic?

(b) Let $T = \min\{n \geq 0 \mid X_n = 4\}$. Compute $\mathbb{P}(T < \infty \mid X_0 = 1)$.

Exercise 4. Consider a random walk with $p_{0,0} = p_{N,N} = 1$, and $p_{i,i+1} = p = 1 - p_{i,i-1}$ for $1 \leq i \leq N-1$. Let T be the first passage time to either 0 or N, i.e., $T = \min\{n \geq 0 \mid X_n = 0 \text{ or } N\}$.

Answer the following:

(a) In the case where $p \neq \frac{1}{2}$, show that

$$\mathbb{E}[T \mid X_0 = i] = \frac{i}{q-p} - \left(\frac{N}{q-p}\right)\left(\frac{1-\left(\frac{q}{p}\right)^i}{1-\left(\frac{q}{p}\right)^N}\right).$$

(b) Compute $\mathbb{E}[T \mid X_0 = i]$ when $p = \frac{1}{2}$.

Exercise 5. For each of the following statements, either provide a proof or disprove by exhibiting a counterexample.

(a) A finite DTMC has at least one closed communicating class.

(b) If a DTMC is periodic, then it is not positive recurrent.

(c) Every finite DTMC possesses a stationary distribution.

(d) Consider a finite DTMC with state space $\{0, 1, \ldots, K\}$. Let $T = \sup\{n \geq 0 \mid X_n = 0\}$. Then, T is a stopping time.

(e) In a DTMC, a state i is transient only if there exists a state j such that j is accessible from i, but i is not accessible from j.

(f) A finite irreducible DTMC is aperiodic if and only if $\exists n > 0$ such that $p_{i,j}^{(n)} > 0$, $\forall i, j$.

Exercise 6. Consider a DTMC space with state space $\{1, 2, \ldots, N\}$ with the following $N \times N$ transition probability matrix

$$\begin{bmatrix}
q & p & 0 & 0 & 0 & . & . & . & . & . & 0 & 0 & 0 \\
q & 0 & p & 0 & 0 & . & . & . & . & . & 0 & 0 & 0 \\
0 & q & 0 & p & 0 & . & . & . & . & . & 0 & 0 & 0 \\
. & & & & & & & & & & & & \\
. & & & & & & & & & & & & \\
. & & & & & & & & & & & & \\
0 & . & . & . & 0 & 0 & q & 0 & p & 0 & . & . & 0. \\
. & & & & & & & & & & & & \\
. & & & & & & & & & & & & \\
0 & 0 & 0 & 0 & . & . & . & . & . & 0 & q & 0 & p \\
0 & 0 & 0 & 0 & . & . & . & . & . & 0 & 0 & q & p
\end{bmatrix}$$

Answer the following:

(a) Is the DTMC irreducible ?

(b) Is the DTMC periodic ?

(c) Does the DTMC have stationary distribution. If yes, provide the same.

Exercise 7. Consider a DTMC on $\{0, 1, 2, \ldots\}$ with $p_{0,0} = 1$, and $p_{i,i-1} = q = 1 - p_{i,i}$ for $i \geq 1$.

Answer the following:

(a) Find $\mathbb{P}\left(X_n = 0, X_m \neq 0, \text{for } 0 < m < n \mid X_0 = i\right)$ for $i \geq 1$.

(b) What is the expected value of the distribution from the part above?

Exercise 8. Consider a random walk with $p_{i,i+1} = p$, and $p_{i,i-1} = q$, for $-\infty < i < \infty$. Here $0 < p < 1$ and $p + q = 1$.

Answer the following, assuming that the random walk starts at the origin:

(a) Find the probability that the random walk hits state i before hitting state $-j$, where $(i, j > 0)$.

(b) Show that the expected number of visits to the state i before hitting state 0 is $\left(\frac{p}{q}\right)^i$, when $p < q$.

(c) What would the expected value from the part above be when $p = q$?

Exercise 9. For the DTMCs with transition probability matrices listed below, identify the communicating classes, and determine their transience/recurrence. Further, for each i, j in the state space, find $\lim_{n\to\infty} p_{i,j}^{(n)}$.

(a) $P = \begin{bmatrix} 0 & 0.4 & 0.6 \\ 0.1 & 0 & 0.9 \\ 0.3 & 0.7 & 0 \end{bmatrix}$.

(b) $P = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0.3 & 0.7 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$.

Exercise 10.

Suppose there are two urns, say 1 and 2. Each urn has r balls. Among the $2r$ balls, $b \leq r$ balls are black, and the remaining $2r - b$ are white. At each trial, one ball is picked uniformly at random from each urn, and they are interchanged.

Answer the following:

(a) Model this problem as a DTMC with the state as the number of white balls in urn 1. Specify the state space and transition probabilities. Is the DTMC irreducible? Aperiodic?

(b) Compute the stationary distribution.

D

Smoothness and Convexity

In this appendix, we discuss foundations of algorithms for non-linear smooth optimization problem. The topics covered include Taylor's theorem and its applications, convex sets and convex/strongly-convex functions.

A general problem of interest here is to find a θ^* such that

$$\theta^* \in \arg\min_{x \in \mathcal{D}} f(\theta), \tag{D.1}$$

where $\mathcal{D} \subset \mathbb{R}^d$. The problem (D.1) includes the case where $\mathcal{D}$ is the whole of $\mathbb{R}^d$ as well.

The following definitions are relevant in the context of the optimization problem (D.1).

Definition D.1 (local minima)**.** A point $\theta^* \in \mathbb{R}^d$ is called a local minimum of f if there exists a neighborhood $\mathcal{N}(\theta^*, \epsilon)$ of θ^* such that $f(\theta) \geq f(\theta^*)$ for all $\theta \in \mathcal{N}(\theta^*, \epsilon) \cap \mathcal{D}$.

Definition D.2 (global minima)**.** A point $\theta^* \in \mathcal{D}$ is called a global minimum of f if $f(\theta) \geq f(\theta^*)$ for all $\theta \in \mathcal{D}$.

Definition D.3 (strict local minima)**.** A point $\theta^* \in \mathcal{D}$ is called a strict local minimum of f if there exists a neighbourhood $\mathcal{N}(\theta^*, \epsilon)$ of θ^* such that $f(\theta) > f(\theta^*)$ for all $\theta \in \mathcal{N}(\theta^*, \epsilon) \cap \mathcal{D}$ with $\theta \neq \theta^*$.

D.1 Necessary conditions for local minima

Given a point $\theta^* \in \mathcal{D}$, how does one determine whether it is a local minimum or not? The following results, which are standard in optimization literature, provide an answer to this question.

Theorem D.1 (First and second-order necessary conditions)**.** Let θ^* be a local minimum of $f : \mathcal{D} \to \mathbb{R}$ and f be continuously differentiable. Then $\nabla f(\theta^*) = 0$.
Further if f is twice continuously differentiable, then $\nabla^2 f(\theta^*)$ is a positive semi-definite matrix.

Proof. Fix $s \in \mathbb{R}^d$. Recall that θ^* is a local minimum. Then we have,

$$s^\mathsf{T} \nabla f(\theta^*) = \lim_{\delta \to 0} \frac{f(\theta^* + \delta s) - f(\theta^*)}{\delta} \geq 0.$$

Similarly, we have,

$$-s^\mathsf{T} \nabla f(\theta^*) \geq 0.$$

Combining the two equations above, we have that $\nabla f(\theta^*) = 0$.

Further, if f is twice continuously differentiable, then by Taylor series expansion, we have

$$f(\theta^* + \delta s) - f(\theta^*) = \delta s^\mathsf{T} \nabla f(\theta^*) + \frac{\delta^2}{2} s^\mathsf{T} \nabla^2 f(\theta^*) s + o(\delta^3).$$

Since $\nabla f(\theta^*) = 0$, we have

$$0 \leq \frac{f(\theta^* + \delta s) - f(\theta^*)}{\delta^2} = \frac{1}{2} s^\mathsf{T} \nabla^2 f(\theta^*) s + o(\delta).$$

Thus, as $\delta \to 0$, for all $s \in \mathbb{R}^d$, we have $s^\mathsf{T} \nabla^2 f(\theta^*) s \geq 0$, implying $\nabla^2 f(\theta^*)$ is positive semi-definite Hence proved. □

Example D.1. Consider $f(\theta) = \frac{1}{2} \theta^\mathsf{T} A \theta - b^\mathsf{T} \theta$. From the first-order necessary condition, we have that $\nabla f(\theta^*) = 0$ and $\nabla^2 f(\theta^*)$ is positive semi-definite, which is equivalent to $A\theta^* - b = 0$ and A is positive semi-definite.

We have the following cases in general:

- If A is not positive semi-definite, then f has no local minima.
- If A is positive semi-definite, then f is convex and any θ^* solving $A\theta^* - b = 0$ is a global minimum.
- if A is positive definite, then f has a unique global minimum given by $\theta^* = A^{-1}b$.
- The reader is encouraged to think about the case where A is positive semi-definite and singular. In this case, it is relevant to check if b is in the column space of A or not and reason accordingly.

D.2 Taylor's theorem

Taylor's theorem shows how a smooth function f can be approximated locally by polynomials that depend on low-order derivatives of f.

Theorem D.2. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function. Given $\theta, p \in \mathbb{R}^d$, we have

$$f(\theta + p) = f(\theta) + \int_0^1 \nabla f(\theta + \alpha p)^\intercal p d\alpha, \text{ and} \tag{D.2}$$

$$f(\theta + p) = f(\theta) + \nabla f(\theta + \alpha p)^\intercal p \text{ for some } \alpha \in (0, 1). \tag{D.3}$$

If f is twice continuously differentiable, we have

$$\nabla f(\theta + p) = \nabla f(\theta) + \int_0^1 \nabla^2 f(\theta + \alpha p) p d\alpha, \text{ and}$$

$$f(\theta + p) = f(\theta) + \nabla f(\theta)^\intercal p + \frac{1}{2} p^\intercal \nabla^2 f(\theta + \alpha p) p, \tag{D.4}$$

for some $\alpha \in (0, 1)$.

A consequence of (D.2) is that for a continuously differentiable f at θ, we have

$$f(\theta + p) = f(\theta) + \nabla f(\theta)^\intercal p + o(||p||).$$

Definition D.4 (smooth function)**.** A function $f : \mathcal{D}(\subset \mathbb{R}^d) \to \mathbb{R}$ is said to be L-smooth if for all $x, y \in \mathcal{D}$, the following condition

holds:

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||. \tag{D.5}$$

The three results below provide useful characterizations of L-smooth functions.

Lemma D.3. Let $f : \mathcal{D}(\subset \mathbb{R}^d) \to \mathbb{R}$ be a L-smooth function. Then for any $x, y \in \mathcal{D}$, we have the following:

$$f(y) \leq f(x) + \nabla f(x)^\intercal (y - x) + \frac{L}{2}||y - x||^2. \tag{D.6}$$

Lemma D.4. Suppose $f : \mathcal{D}(\subset \mathbb{R}^d) \to \mathbb{R}$ is twice continuously differentiable function. Then, $\forall \theta \in \mathcal{D}$,
(I) f is L-smooth implies $\nabla^2 f(\theta) \preceq L\mathbb{I}$
(II) conversely, if $-L\mathbb{I} \preceq \nabla^2 f(\theta) \preceq L\mathbb{I}$, then f is L-smooth.

Lemma D.5. Suppose f is twice continuously differentiable on $\mathbb{R}^d$. Then if f is L-smooth, we have $\nabla^2 f(\theta) \preceq L\mathbb{I}$ for all $\theta \in \mathcal{D}$.
Conversely, if $-L\mathbb{I} \preceq \nabla^2 f(\theta) \preceq L\mathbb{I}$, $\forall \theta \in \mathcal{D}$, then f is L-smooth.

D.3 Sufficient conditions for local minima

Theorem D.6 (Sufficient Conditions for Smooth Unconstrained Optimization)**.** Suppose that f is twice continuously differentiable and that, for some $\theta^* \in \mathbb{R}^d$, we have $\nabla f(\theta^*) = 0$, and $\nabla^2 f(\theta^*)$ is positive definite. Then θ^* is a strict local minimizer of $\min_{\theta \in \mathbb{R}^d} f(\theta)$.

Proof. See (Bertsekas, 1999). □

D.4 Convex Sets and Functions

Definition D.5. A set $\mathbb{C} \subset \mathbb{R}^d$ is a convex set if $\forall x, y \in \mathbb{C}$ and for

all $\lambda \in [0,1]$, it satisfies:

$$\lambda x + (1-\lambda)y \in \mathbb{C}. \tag{D.7}$$

Example D.2. A set $\{x \mid v^\mathsf{T} x = b\}$ with $v \in \mathbb{R}^d$ and $b \in \mathbb{R}$, is a convex set. Such a set is known as a hyperplane. With the same notation, a set $\{x \mid v^\mathsf{T} x \leq b\}$, which is known as a halfspace, is also a convex set.

Example D.3. Euclidean Balls: $B = \{\theta \mid \|x\| \leq 1\}$ is a convex set.

Definition D.6. A function $f : \Omega \to \mathbb{R}$ is a convex function if its domain Ω is convex and it satisfies the following condition for all $x, y \in \Omega$ and $\lambda \in [0,1]$:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y). \tag{D.8}$$

Further, the function f is strictly convex if the inequality is strict for $x \neq y$ and $0 < \lambda < 1$.

Note that a function f is concave/strictly concave if $-f$ is convex/strictly convex.

Lemma D.7. Suppose f is convex. Then,

1. Any local minimum is a global minimum.
2. The set of all global minima is convex.

Theorem D.8 (Necessary condition for optima)**.** Suppose that f is continuously differentiable and convex. Then if $\nabla f(\theta^*) = 0$, then θ^* is a global minimizer.

Proof. By applying Taylor's theorem,

$$f(x + \alpha(y-x)) = f(x) + \alpha \nabla f(x)^\mathsf{T}(y-x) + o(\alpha) \leq (1-\alpha) f(x) + \alpha f(y).$$

$$f(y) \geq f(x) + \nabla f(x)^\mathsf{T}(y-x) + o(1).$$

when $\alpha \downarrow 0$, $o(1)$ term vanishes, and we obtain

$$f(y) \geq f(x) + \nabla f(x)^{\mathsf{T}}(y - x).$$

Setting $x = \theta^*$ leads to

$$f(y) \geq f(\theta^*), \ \forall y.$$

Hence proved. □

We now provide useful characterizations of convex functions through the result below.

Theorem D.9. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable over an open domain. Then the following are equivalent

i) f is convex;

ii) $f(y) \geq f(x) + \nabla f(x)^{\mathsf{T}}(y - x), \forall x, y \in \mathcal{D}$;

iii) $\nabla^2 f(x) \succeq 0$, for all $x \in \mathcal{D}$.

Proof. We prove $(i) \Leftrightarrow (ii)$ then $(ii) \Leftrightarrow (iii)$.

$(i) \Rightarrow (ii)$ If f is convex, by definition

$$f(\lambda y + (1 - \lambda)x) \leq \lambda f(y) + (1 - \lambda) f(x), \forall \lambda \in [0, 1], x, y \in \text{dom}(f)$$

After rewriting, we have

$$\begin{aligned} & f(x + \lambda(y - x)) \leq f(x) + \lambda(f(y) - f(x)) \\ \Rightarrow & f(y) - f(x) \geq \frac{f(x + \lambda(y - x)) - f(x)}{\lambda}, \forall \lambda \in (0, 1] \end{aligned}$$

As $\lambda \downarrow 0$, we get

$$f(y) - f(x) \geq \nabla f^T(x)(y - x) \tag{D.9}$$

$(ii) \Rightarrow (i)$ Suppose (D.9) holds $\forall x, y \in \text{dom}(f)$. Take any $x, y \in \text{dom}(f)$ and let

$$z = \lambda x + (1 - \lambda) y$$

We have

$$f(x) \geq f(z) + \nabla f^T(z)(x - z) \tag{D.10}$$
$$f(y) \geq f(z) + \nabla f^T(z)(y - z) \tag{D.11}$$

Multiplying (D.10) by λ, (D.11) by $(1 - \lambda)$ and adding, we get

$$\begin{aligned} \lambda f(x) + (1 - \lambda) f(y) &\geq f(z) + \nabla f^T(z)(\lambda x + (1 - \lambda)y - z) \\ &= f(z) \\ &= f(\lambda x + (1 - \lambda)y). \end{aligned}$$

$(ii) \Leftrightarrow (iii)$ We prove both of these claims first in dimension 1 and then generalize.

$(ii) \Rightarrow (iii)$*(uni-variate case)* Let $x, y \in \text{dom}(f), y > x$. We have

$$f(y) \geq f(x) + f'(x)(y - x) \tag{D.12}$$
$$\text{and } f(x) \geq f(y) + f'(y)(x - y) \tag{D.13}$$
$$\Rightarrow f'(x)(y - x) \leq f(y) - f(x) \leq f'(y)(y - x)$$

using (D.12) then (D.13). Dividing LHS and RHS by $(y - x)^2$ gives

$$\frac{f'(y) - f'(x)}{y - x} \geq 0, \forall x, y, x \neq y$$

As we let $y \to x$, we get

$$f''(x) \geq 0, \forall x \in \text{dom}(f)$$

$(iii) \Rightarrow (ii)$*(uni-variate case)* Suppose $f''(x) \geq 0, \forall x \in \text{dom}(f)$. By the mean value version of Taylor's theorem we have

$$f(y) = f(x) + f'(x)(y - x) + \frac{1}{2} f''(z)(y - x)^2, \text{ for some } z \in [x, y].$$
$$\Rightarrow f(y) \geq f(x) + f'(x)(y - x).$$

Now to establish $(ii) \Leftrightarrow (iii)$ in general dimension, we recall that convexity is equivalent to convexity along all lines; i.e., $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $g(\alpha) = f(x_0 + \alpha v)$ is convex, $\forall x_0 \in \text{dom}(f)$ and $\forall v \in \mathbb{R}^n$. We just proved this happens if and only if

$$g''(\alpha) = v^T \nabla^2 f(x_0 + \alpha v) v \geq 0$$

$\forall x_0 \in \text{dom}(f), \forall v \in \mathbb{R}^n$ and $\forall \alpha$ s.t. $x_0 + \alpha v \in \text{dom}(f)$. Hence, f is convex if and only if $\nabla^2 f(x) \succeq 0$ for all $x \in \text{dom}(f)$. □

D.5 Strongly Convex Functions

Definition D.7. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be μ-strongly convex ($\mu > 0$) if for all $x, y \in \mathbb{R}^d$, then

$$f((1-\lambda)x+\lambda y) \leq (1-\lambda)f(x)+\lambda f(y)+\frac{m}{2}(1-\lambda)\left\|y-x\right\|_2^2 \quad \text{(D.14)}$$

Theorem D.10. Suppose f is continuously differentiable and μ-strongly convex, then for any $x, y \in \mathbb{R}^d$

$$f(y) \geq f(x) + \nabla f(x)^\intercal (y-x) + \frac{\mu}{2}\left\|y-x\right\|_2^2$$

Lemma D.11. Suppose that f is twice-continuously differentiable on $\mathbb{R}^d$. Then f has modulus of convexity μ if and only if $\nabla^2 f(x) \succeq \mu I$ for all x.

Proof. For any $x, u \in \mathbb{R}^d$ and $\alpha > 0$, we have from Taylor's theorem that

$$f(x+\alpha u) = f(x) + \alpha \nabla f(x)^\intercal u + \frac{1}{2}\alpha^2 u^\intercal \nabla^2 f(x+\gamma\alpha u)u,$$

for some $\gamma \in (0,1)$.

From the strong convexity property, we have

$$f(x+\alpha u) \geq f(x) + \alpha \nabla f(x)^\intercal u + \frac{\mu}{2}\alpha^2 \left\|u\right\|^2$$

By comparing the two equations above, we obtain

$$u^\intercal \nabla^2 f(x+\gamma\alpha u)u \geq \mu \left\|u\right\|^2$$

By taking $\alpha \downarrow 0$, we obtain

$$u^\intercal \nabla^2 f(x)u \geq \mu \left\|u\right\|^2.$$

Since the above is true for all $u \in \mathbb{R}^d$, we have

$$\nabla^2 f(\theta) \succeq \mu I. \quad \text{(D.15)}$$

□

D.6 Bibliographic remarks

For an introduction to convex optimization, the reader is referred to either classic textbooks such as (Boyd and Vandenberghe, 2004; Nocedal and Wright, 1999; Bertsekas, 1999), or the more recent machine learning-oriented optimization book (Wright and Recht, 2022). The material presented in this appendix is based on (Wright and Recht, 2022) and (Bertsekas, 1999).

D.7 Exercises

Exercise 1. The convex hull of a set C, denoted $\text{Conv}(C)$ is defined as

$$\text{Conv}(C) = \{\alpha_1 x_1 + \ldots + \alpha_k x_k \mid | x_i \in C, \alpha_i \geq 0, \forall i, \alpha_1 + \ldots + \alpha_k = 1\}.$$

For each of the following sets in $\mathbb{R}^2$, provide a visual depiction of their convex hulls by sketching:

1. $C = \{(0,1), (0,4), (-2,-1), (0,0), (3,-2), (-1,2)\}$.
2. Union of two unit circles, centered at $(1,1)$ and $(-1,-1)$, respectively.

Exercise 2. Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. For any three points x_1, x_2, x_3 such that $x_1 < x_2 < x_3$, show that

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_3) - f(x_2)}{x_3 - x_2}.$$

Exercise 3. Consider the following two statements:

I: If $\log f$ is convex, then f is convex.
II: If f is convex, then $\log f$ is convex.
Which of the statements above are true?

Exercise 4. Give an example of a convex function $f : \mathbb{R} \to \mathbb{R}$ that is bounded above.

Exercise 5. Let $x \in \mathbb{R}^d$, with x_i denoting the ith coordinate. Are the functions defined below convex? Justify your answer.

(a) $f(x) = \log\left(\exp(x_1) + \ldots + \exp(x_d)\right)$.

(b) $f(x) = \exp(x^\mathsf{T} A x)$, where A is a positive semi-definite matrix.

Exercise 6. Answer the following questions concerning necessary conditions for local minima:

(a) Let $f : \mathbb{R} \to \mathbb{R}$. Recall that $f'(x^*) = 0$ and $f''(x^*) \geq 0$ are the first and second-order necessary conditions for a local minimizer. In a similar spirit, derive a third-order necessary condition, assuming f is three-times continuously differentiable.

(b) Show an example function f and a point x^* that satisfies the first, second and third-order necessary conditions, but x^* is not a local minimizer of the function f.

Exercise 7. Suppose we want to minimize the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x_1, x_2) = (x_1 - x_2)^4 + x_1^2 - x_2^2 - 2x_1 + 2x_2 + 1.$$

Find points where the first-order necessary condition for a minimum is satisfied. For each of these points, characterize whether the second-order necessary condition is satisfied.

Exercise 8. Let $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ and let α_1, α_2 be two positive scalars.

(a) Prove or disprove: If f_1, f_2 are convex, then $\max(\alpha_1 f_1, \alpha_2 f_2)$ is convex.

(b) Prove or disprove: If f_1, f_2 are concave, then $\max(\alpha_1 f_1, \alpha_2 f_2)$ is concave.

Exercise 9. Suppose a function $f : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz and differentiable. Show that $\sup_x |f'(x)| \leq L$.

Exercise 10. Exhibit a real-valued function f that is L-smooth but not L-Lipschitz.

Exercise 11. Exhibit a real-valued function f that is Lipschitz and differentiable, but not smooth.

Exercise 12. Suppose $f_1 : \mathbb{R}^d \to \mathbb{R}$ is L_1-smooth and $f_2 : \mathbb{R}^d \to \mathbb{R}$ is L_2-smooth. Show that $f_1 + f_2$ is $(L_1 + L_2)$-smooth.

Exercise 13. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x_1, x_2) = ax_1^2 + 2bx_1x_2 + cx_2^2.$$

Answer the following:

(a) Prove or disprove: f is strongly convex if $a > 0$ and $c > 0$.

(b) Derive a necessary and sufficient condition for strong convexity of f. This condition should be in terms of a, b, c.

(c) Under the condition from the part above, characterize the minimizer, say x^* of f.

Exercise 14. Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is a m-strongly convex function with a L-Lipschitz gradient. Let x^* be the minimizer with corresponding function value $f^* = f(x^*)$.

(a) Let $g(x) = f(x) - \dfrac{m}{2} \|x\|^2$. Show that $g(x)$ is convex with $(L - m)$ Lipschitz continuous gradients.

(b) Using the fact that g, defined in the part above, is convex, prove the following property: For any $x, y \in \mathbb{R}^d$,

$$\begin{aligned}&(\nabla f(x) - \nabla f(y))^\intercal (x - y) \\ &\qquad \geq \frac{mL}{m + L} \|x - y\|^2 + \frac{1}{m + L} \|\nabla f(x) - \nabla f(y)\|^2 .\end{aligned}$$

E

Information theory

In this appendix, we briefly cover the necessary information theory concepts that are useful in understanding the derivation of the minimax lower bound in Section 5.6.

In the following, we assume that the underlying random variables are discrete and leave it to the reader to fill in the necessary details for the continuous extension.

E.1 Entropy

Definition E.1. Consider a discrete r.v. X taking values in the set $\mathcal{X}$ with p.m.f. p. Then, the entropy $H(X)$ is defined as

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x),$$

where the log is to base 2.

It is easy to see that $H(X) \geq 0$ for any X, since $\log p(x) \leq 0$ for $p(x) \in [0, 1]$. As an example, the entropy of a Bernoulli r.v. X with parameter p is $H(X) = -p \log p - (1-p) \log(1-p)$. Plotting $H(X)$ as a function of p, it is easy to infer that $H(X)$ is maximized at $p = 1/2$, $H(X) = 0$ at $p = 0$ and $p = 1$.

The notion of entropy has roots in information theory, as it gives the expected number of bits necessary to encode a random signal (= a random variable). We illustrate this interpretation through the following r.v.:

$$X = \begin{cases} a & \text{w.p. } 1/2, \\ b & \text{w.p. } 1/4, \\ c & \text{w.p. } 1/8, \\ d & \text{w.p. } 1/8. \end{cases}$$

If one were to design a sequence of binary questions to infer the value of the r.v. X and ask the minimum number of questions in expectation, then it would serve him/her to start with "Is $X = a$?" rather than start with "Is $X = d$?". Now using the pmf of X given above, the expected number of questions asked is $1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 3 \times \frac{1}{8} = \frac{7}{4}$. It is not a coincidence that $H(X)$ turns out to be $\frac{7}{4}$ for this r.v.

An equivalent interpretation is the following: Suppose that the value a is represented by the code "1", b by "01", c by "001" and d by "000". Assuming that the values a, b, c, d occur with probabilities given above, the average code length turns out to be the same as $H(X)$.

Definition E.2. The joint entropy $H(X, Y)$ of r.v. pair (X, Y) with joint pmf $p(x, y)$ is defined as

$$H(X, Y) = -\sum_x \sum_y p(x, y) \log p(x, y).$$

Definition E.3. The conditional entropy $H(Y \mid X)$, assuming the r.v. pair (X, Y) has joint pmf $p(x, y)$, is defined as

$$\begin{aligned} H(Y \mid X) &= \sum_x p(x) H(Y \mid X = x) \\ &= -\sum_x p(x) \sum_y p(y \mid x) \log p(y \mid x) \\ &= -\sum_x \sum_y p(x, y) \log p(y \mid x). \end{aligned}$$

Theorem E.1. $H(X,Y) = H(X) + H(Y \mid X)$.

Proof. Follows by using the definition of $H(X,Y)$ followed by a separation of terms using $p(x,y) = p(x)p(y \mid x)$ to obtain $H(X)$ and $H(Y \mid X)$. □

We are now ready to define the concept of KL-divergence, also known as relative entropy, between two probability distributions.

E.2 KL-divergence

Definition E.4. The KL-divergence $D_{\mathrm{kl}}(p\|q)$ between two pmfs p and q is defined as

$$D_{\mathrm{kl}}(p\|q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right),$$

where, $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$.

Definition E.5. The total variation distance $\|P - Q\|_{\mathrm{TV}}$ between two distributions P and Q on a common sigma field $\mathcal{X}$ is defined as

$$\|P - Q\|_{\mathrm{TV}} = \sup_{A \subset \mathcal{X}} |P(A) - Q(A)|.$$

We shall prove later that $\|P - Q\|_{\mathrm{TV}}^2 \leq \frac{1}{2} D_{\mathrm{kl}}(P\|Q)$ – a fact well-known as Pinsker's inequality.

Example E.1. Let p and q be the probability mass functions (PMFs) of Bernoulli r.v.s with parameters α and β, respectively. Then,

$$D_{\mathrm{kl}}(p\|q) = \alpha \log \frac{\alpha}{\beta} + (1-\alpha) \log \frac{1-\alpha}{1-\beta}.$$

Plugging in values $1/4$ and $1/2$ for α and β, it is easy to see that $D_{\mathrm{kl}}(p\|q)$ is not equal to $D_{\mathrm{kl}}(q\|p)$.

KL-divergence is not a metric because it is not symmetric, as shown in the above example. Moreover, KL-divergence does not satisfy the triangle inequality. However, KL-divergence is non-negative and zero if and only if the probability distributions are the same – a claim made precise below.

Lemma E.2. The KL-divergence $D_{\mathrm{kl}}(p\|q)$ between two PMFs p and q is non-negative and equals zero if and only if $p(x) = q(x), \forall x$.

Proof. Let $A = \{x \mid p(x) > 0\}$ be the support of p. Then, using Jensen's inequality for the concave log function, we have

$$\begin{aligned}
-D_{\mathrm{kl}}(p\|q) &= -\sum_{x\in A} p(x) \log\left(\frac{p(x)}{q(x)}\right) \\
&= \sum_{x\in A} p(x) \log\left(\frac{q(x)}{p(x)}\right) \\
&\leq \log\left(\sum_{x\in A} p(x)\frac{q(x)}{p(x)}\right) \\
&= \log\left(\sum_{x\in A} q(x)\right) \leq \log\left(\sum_{x} q(x)\right) \\
&= \log 1 = 0,
\end{aligned}$$

which proves the first part of the claim. For the second part, observe that log is stricly concave and hence, equality holds in Jensen's if and only if $\dfrac{p(x)}{q(x)} = 1, \forall x$. □

Definition E.6. The conditional KL-divergence between two PMFs p and q is defined as

$$D_{\mathrm{kl}}(p(y \mid x)\|q(y \mid x))) = \sum_x p(x) \sum_y p(y \mid x) \log \frac{p(y \mid x)}{q(y \mid x)}.$$

Lemma E.3. (Chain rule)

$$D_{\mathrm{kl}}(p(x,y)\|q(x,y)) = D_{\mathrm{kl}}(p(x)\|q(x)) + D_{\mathrm{kl}}(p(y \mid x)\|q(y \mid x))).$$

In addition, if x and y are independent, then

$$D_{\mathrm{kl}}(p(x,y)\|q(x,y)) = D_{\mathrm{kl}}(p(x)\|q(x)) + D_{\mathrm{kl}}(p(y)\|q(y)).$$

Proof. Notice that

$$D_{\mathrm{kl}}(p(x,y)\|q(x,y))$$

$$
\begin{aligned}
&= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y) \log \frac{p(y \mid x)}{q(y \mid x)} \\
&= D_{\mathrm{kl}}\left(p(x) \| q(x)\right) + D_{\mathrm{kl}}\left(p(y \mid x) \| q(y \mid x)\right).
\end{aligned}
$$

This proves the first claim in the lemma statement. The second claim can be easily inferred from the first. □

E.3 Pinsker's inequality

Lemma E.4. (Pinsker's inequality) Given two PMFs p and q, for any event A, we have

$$2(p(A) - q(A))^2 \leq D_{\mathrm{kl}}\left(p \| q\right).$$

Proof. Fix an event A. Then, we have

$$\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \geq p(A) \log \frac{p(A)}{q(A)}. \tag{E.1}$$

The proof of the claim above is as follows: Letting $p_A(x) = \dfrac{p(x)}{p(A)}$ and $q_A(x) = \dfrac{q(x)}{q(A)}$, we have

$$
\begin{aligned}
\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} &= p(A) \sum_{x \in A} p_A(x) \log \frac{p(A) p_A(x)}{q(A) q_A(x)} \\
&= p(A) \log \frac{p(A)}{q(A)} \sum_{x \in A} p_A(x) + p(A) \sum_{x \in A} p_A(x) \log \frac{p_A(x)}{q_A(x)} \\
&\geq p(A) \log \frac{p(A)}{q(A)},
\end{aligned}
$$

where the last inequality follows from the fact that $\sum_x p_A(x) \log \dfrac{p_A(x)}{q_A(x)} = D_{\mathrm{kl}}\left(p_A \| q_A\right) \geq 0$ and $\sum_x p_A(x) = 1$.

Letting $\alpha = p(A)$ and $\beta = q(A)$ and using (E.1), we have

$$D_{\mathrm{kl}}\left(p \| q\right) \geq \alpha \log \frac{\alpha}{\beta} + (1 - \alpha) \log \frac{1 - \alpha}{1 - \beta}$$

$$
\begin{aligned}
&= \int_{\alpha}^{\beta} \left(\frac{-\alpha}{x} + \frac{1-\alpha}{1-x} \right) dx \\
&= \int_{\alpha}^{\beta} \left(\frac{x-\alpha}{x(1-x)} \right) dx \geq \int_{\alpha}^{\beta} \frac{x-\alpha}{1/4} dx \ \ (\text{since } x(1-x) \leq 1/4) \\
&= 2(\alpha-\beta)^2.
\end{aligned}
$$

Hence proved. □

The following result is now immediate from the bound in the lemma above.

Corollary E.5. Given two PMFs p and q, we have

$$\|p-q\|_{\mathrm{TV}}^2 \leq \frac{1}{2} D_{\mathrm{kl}}\left(p \| q\right).$$

Lemma E.6. (Pinsker's inequality: a variant)
Given two PMFs p and q, for any event A, we have

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp(-D_{\mathrm{kl}}\left(p \| q\right)),$$

where $P(A)$ (resp. $Q(A^c)$) is shorthand for $\sum_{x \in A} p(x)$ (resp. $\sum_{x \in A^c} q(x)$).

Proof. Notice that

$$
\begin{aligned}
\sum_x \min(p(x), q(x)) &= \sum_{x \in A} \min(p(x), q(x)) + \sum_{x \in A^c} \min(p(x), q(x)) \\
&\leq \sum_{x \in A} p(x) + \sum_{x \in A^c} q(x) = P(A) + Q(A^c).
\end{aligned}
$$

So, it is enough to prove a lower bound on $\sum_{x \in A} \min(p(x), q(x))$. We claim that

$$\sum_x \min(p(x), q(x)) \geq \frac{1}{2} \left(\sum_x \sqrt{p(x) q(x)} \right)^2.$$

The inequality above holds because

$$\left(\sum_x \sqrt{p(x) q(x)} \right)^2 = \left(\sum_x \sqrt{\min(p(x), q(x)) \max(p(x), q(x))} \right)^2$$

$$\leq \left(\sum_x \min(p(x), q(x))\right)\left(\sum_x \max(p(x), q(x))\right)$$
$$\leq 2\sum_x \min(p(x), q(x)),$$

where the last inequality holds because

$$\sum_x \max(p(x), q(x)) = \sum_x (p(x) + q(x) - \min(p(x), q(x)))$$
$$\leq 2 - \sum_x \min(p(x), q(x)) \leq 2.$$

Now, we have

$$\left(\sum_x \sqrt{p(x)q(x)}\right)^2 = \exp\left(2\log\left(\sum_x \sqrt{p(x)q(x)}\right)\right)$$
$$= \exp\left(2\log\left(\sum_x p(x)\sqrt{\frac{q(x)}{p(x)}}\right)\right)$$
$$\geq \exp\left(2\left(\sum_x p(x)\log\sqrt{\frac{q(x)}{p(x)}}\right)\right)$$
(Jensen's inequality)
$$= \exp\left(\sum_x p(x)\log\frac{q(x)}{p(x)}\right)$$
$$= \exp\left(-D_{\mathrm{kl}}\left(p\|q\right)\right).$$

□

E.4 Bibliographic remarks

The information theory background covered here is based on the classic text book by Cover and Thomas, 2012.

E.5 Exercises

Exercise 1. For some $0 < \Delta < 1/2$, let p, q and r correspond to the PMFs of Bernoulli r.v.s with parameters $\frac{1}{2}$, $\frac{1+\Delta}{2}$ and $\frac{1-\Delta}{2}$, respectively. Then,

$$D_{\mathrm{kl}}\left(p\|q\right) \leq \Delta^2, D_{\mathrm{kl}}\left(q\|p\right) \leq 2\Delta^2, D_{\mathrm{kl}}\left(p\|r\right) \leq \Delta^2 \text{ and } D_{\mathrm{kl}}\left(r\|q\right) \leq 4\Delta^2.$$

Exercise 2. For distributions P and Q of a continuous random variable, the KL-divergence is defined to be the integral:

$$D_{\mathrm{kl}}\left(P\|Q\right) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx,$$

where p and q denote the densities of P and Q, respectively.

Answer the following:

(a) Prove Pinsker's inequality, i.e., given distributions P, Q of continuous r.v.s,

$$\|P - Q\|_{\mathrm{TV}}^2 \leq \frac{1}{2} D_{\mathrm{kl}}\left(P\|Q\right).$$

(b) Suppose that P and Q correspond to univariate Gaussian distributions with means μ_1, μ_2, and variances σ_1^2, σ_2^2, respectively. Show that

$$D_{\mathrm{kl}}\left(P\|Q\right) = \frac{1}{2}\left(\log\frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2} - 1\right) + \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2}.$$

(c) Suppose that P and Q correspond to bivariate Gaussian distributions with zero mean and covariance matrices $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 & \rho^2 \\ \rho^2 & 1 \end{bmatrix}$, where $\rho \in (0, 1)$. Calculate $D_{\mathrm{kl}}\left(P\|Q\right)$, upper bound it using the simplest possible function of ρ.

Exercise 3.

Suppose there are two coins. The first is a fair coin, while the second one is biased (i.e., it falls heads with probability $\frac{3}{4}$). Suppose n sample outcomes $X_1, \ldots, X_n$ are generated using one of the two coins and an algorithm, say $\mathcal{A}$, uses these samples to identify the source coin. Let $\hat{I}_n$ denote the index that the algorithm $\mathcal{A}$ returns as its estimate of the source coin. Let P_v (resp. $P_{v'}$) denote the law of the observed samples $(X_1, \ldots, X_n)$, when the underlying source is the fair (resp. biased) coin.

If $n < 4 \log 2$, then show that no algorithm can ensure

$$\max(P_v(\hat{I}_n = 2), P_{v'}(\hat{I}_n = 1)) \leq 0.22.$$

Hint: Use Pinsker's inequality.

References

A.Dvoretzky. (1956). "On stochastic approximation". *Proc. Third Berkeley Symp. Math. Stat. and Prob.* 1: 39–55.

Abdulla, M. S. and S. Bhatnagar. (2007). "Reinforcement learning based algorithms for average cost Markov decision processes". *Discrete Event Dynamic Systems*. 17(1): 23–52.

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

Acerbi, C. (2002). "Spectral measures of risk: A coherent representation of subjective risk aversion". *Journal of Banking & Finance*. 26(7): 1505–1518.

Agarwal, A., O. Dekel, and L. Xiao. (2010). "Optimal Algorithms for Online Convex Optimization with Multi-Point Bandit Feedback". In: *COLT*. 28–40.

Alzantot, M., Y. Sharma, S. Chakraborty, H. Zhang, C.-J. Hsieh, and M. B. Srivastava. (2019). "GenAttack: practical black-box attacks with gradient-free optimization". In: *Proceedings of the Genetic and Evolutionary Computation Conference. GECCO '19*. Prague, Czech Republic: Association for Computing Machinery. 1111–1119. ISBN: 9781450361118. DOI: 10.1145/3321707.3321749. URL: https://doi.org/10.1145/3321707.3321749.

Anandkumar, A. and R. Ge. (2016). "Efficient approaches for escaping higher order saddle points in non-convex optimization". In: *Conference on learning theory*. PMLR. 81–102.

Andronov, A. A., A. A. Vitt, and S. E. Khaikin. (2013). *Theory of Oscillators: Adiwes International Series in Physics*. Vol. 4. Elsevier.

Arnold, V. I. (1992). *Ordinary differential equations*. Springer Science & Business Media.

Artzner, P., F. Delbaen, J. Eber, and D. Heath. (1999). "Coherent measures of risk". *Mathematical Finance*. 9(3): 203–228.

Asmussen, S. and P. Glynn. (2007a). *Stochastic Simulation: Algorithms and Analysis*. Springer.

Asmussen, S. and P. W. Glynn. (2007b). *Stochastic simulation: algorithms and analysis*. Vol. 57. Springer.

Aubin, J. and A. Cellina. (1984). *Differential Inclusions: Set-Valued Maps and Viability Theory*. Springer.

Aubin, J. and H. Frankowska. (1990). *Set-Valued Analysis*. Birkhauser.

Balasubramanian, K. and S. Ghadimi. (2022a). "Zeroth-Order Nonconvex Stochastic Optimization: Handling Constraints, High Dimensionality, and Saddle Points". *Foundations of Computational Mathematics*. 22(1): 35–76. ISSN: 1615-3383.

Balasubramanian, K. and S. Ghadimi. (2022b). "Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points". *Foundations of Computational Mathematics*. 22(1): 35–76.

Barakat, A., P. Bianchi, W. Hachem, and S. Schechtman. (2021). "Stochastic optimization with momentum: Convergence, fluctuations, and traps avoidance". *Electronic Journal of Statistics*. 15(2): 3892–3947. DOI: 10.1214/21-EJS1880. URL: https://doi.org/10.1214/21-EJS1880.

Bardou, O., N. Frikha, and G. Pages. (2009). "Computing VaR and CVaR using stochastic approximation and adaptive unconstrained importance sampling". *Monte Carlo Methods and Applications*. 15(3): 173–210.

Benaïm, M. (1996). "A Dynamical System Approach to Stochastic Approximations". *SIAM J. Control Optim.* 34(2): 437–472.

Benaïm, M. (1999). "Dynamics of stochastic approximation algorithms". *Seminaire De Probabilities (Strasbourg)*. 1709: 1–68.

Benaïm, M. and M. W. Hirsch. (1996). "Asymptotic pseudotrajectories and chain recurrent flows, with applications". *J. Dynam. Differential Equations*. 8: 141–176.

Benaïm, M., J. Hofbauer, and S. Sorin. (2005). "Stochastic approximations and differential inclusions". *SIAM Journal on Control and Optimization*: 328–348.

Benaïm, M., J. Hofbauer, and S. Sorin. (2012). "Perturbations of set-valued dynamical systems, with applications to game theory". *Dynamic Games and Applications*. 2(2): 195–205.

Berahas, A. S., L. Cao, K. Choromanski, and K. Scheinberg. (2022). "A theoretical and empirical comparison of gradient approximations in derivative-free optimization". *Foundations of Computational Mathematics*. 22(2): 507–560.

Bertsekas, D. P. (2012). *Dynamic Programming and Optimal Control, Vol.II*. Athena Scientific.

Bertsekas, D. P. and J. N. Tsitsiklis. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

Bertsekas, D. (2019). *Reinforcement learning and optimal control*. Vol. 1. Athena Scientific.

Bertsekas, D. P. (1999). *Nonlinear Programming*. 2nd. Belmont, MA: Athena Scientific.

Bertsekas, D. and J. Tsitsiklis. (1989). *Parallel and distributed computation*. Prentice Hall Inc.

Bhagoji, A. N., W. He, B. Li, and D. Song. (2018). "Practical Black-Box Attacks on Deep Neural Networks Using Efficient Query Mechanisms". In: *Computer Vision – ECCV 2018*. Ed. by V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss. Cham: Springer International Publishing. 158–174. ISBN: 978-3-030-01258-8.

Bhatnagar, S., H. L. Prasad, and L. A. Prashanth. (2013). *Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods (Lecture Notes in Control and Information Sciences)*. Vol. 434. Springer.

Bhatnagar, S. (2005). "Adaptive multivariate three-timescale stochastic approximation algorithms for simulation based optimization". *ACM Transactions on Modeling and Computer Simulation*. 15(1): 74–107.

Bhatnagar, S. (2007). "Adaptive Newton-based smoothed functional algorithms for simulation optimization". *ACM Transactions on Modeling and Computer Simulation*. 18(1): 2:1–2:35.

Bhatnagar, S. and L. Prashanth. (2015). "Simultaneous perturbation Newton algorithms for simulation optimization". *Journal of Optimization Theory and Applications*. 164(2): 621–643.

Bhatnagar, S., R. Sutton, M. Ghavamzadeh, and M. Lee. (2009). "Natural Actor-Critic Algorithms". *Automatica*. 45(11): 2471–2482.

Bhatnagar, S. (2010). "An actor–critic algorithm with function approximation for discounted cost constrained Markov decision processes". *Systems & Control Letters*. 59(12): 760–766.

Bhatnagar, S. (2023). "The Reinforce Policy Gradient Algorithm Revisited". *arXiv preprint arXiv:2310.05000*.

Bhatnagar, S. and M. S. Abdulla. (2008). "Simulation-based optimization algorithms for finite-horizon Markov decision processes". *Simulation*. 84(12): 577–600.

Bhatnagar, S. and K. M. Babu. (2008). "New algorithms of the Q-learning type". *Automatica*. 44(4): 1111–1119.

Bhatnagar, S. and V. S. Borkar. (1998). "A two timescale stochastic approximation scheme for simulation-based parametric optimization". *Probability in the Engineering and Informational Sciences*. 12(4): 519–531.

Bhatnagar, S. and V. S. Borkar. (2003). "Multiscale chaotic SPSA and smoothed functional algorithms for simulation optimization". *Simulation*. 79(10): 568–580.

Bhatnagar, S., V. S. Borkar, M. Akarapu, and S. Mannor. (2006). "A Simulation-Based Algorithm for Ergodic Control of Markov Chains Conditioned on Rare Events." *Journal of Machine Learning Research*. 7(10).

Bhatnagar, S., M. C. Fu, S. I. Marcus, I. Wang, *et al.* (2003). "Two-timescale simultaneous perturbation stochastic approximation using deterministic perturbation sequences". *ACM Transactions on Modeling and Computer Simulation*. 13(2): 180–209.

Bhatnagar, S., N. Hemachandra, and V. K. Mishra. (2011a). "Stochastic approximation algorithms for constrained optimization via simulation". *ACM Transactions on Modeling and Computer Simulation (TOMACS)*. 21(3): 1–22.

Bhatnagar, S. and S. Kumar. (2004). "A simultaneous perturbation stochastic approximation-based actor-critic algorithm for Markov decision processes". *IEEE Transactions on Automatic Control*. 49(4): 592–598.

Bhatnagar, S. and K. Lakshmanan. (2012). "An online actor–critic algorithm with function approximation for constrained markov decision processes". *Journal of Optimization Theory and Applications*. 153: 688–708.

Bhatnagar, S. and K. Lakshmanan. (2016). "Multiscale Q-learning with linear function approximation". *Discrete Event Dynamic Systems*. 26: 477–509.

Bhatnagar, S., V. K. Mishra, and N. Hemachandra. (2011b). "Stochastic algorithms for discrete parameter simulation optimization". *IEEE Transactions on Automation Science and Engineering*. 8(4): 780–793.

Bhatnagar, S. and L. A. Prashanth. (2023). "Generalized simultaneous perturbation stochastic approximation with reduced estimator bias". In: *2023 57th Annual Conference on Information Sciences and Systems (CISS)*. IEEE. 1–6.

Bhavsar, N. and L. Prashanth. (2022). "Non-asymptotic bounds for stochastic optimization with biased noisy gradient oracles". *IEEE Transactions on Automatic Control*: 1–1. DOI: 10.1109/TAC.2022.3159748.

Bhojanapalli, S., B. Neyshabur, and N. Srebro. (2016). "Global optimality of local search for low rank matrix recovery". *Advances in Neural Information Processing Systems*. 29.

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.

Billingsley, P. (2017). *Probability and measure*. John Wiley & Sons.

Borkar, V. S. (1995). *Probability Theory: An Advanced Course*. New York: Springer.

Borkar, V. S. (2022). *Stochastic Approximation: A Dynamical Systems Viewpoint, 2'nd Edition*. Cambridge University Press.

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

Borkar, V. S. and S. Meyn. (1999). "The O.D.E. Method for Convergence of Stochastic Approximation and Reinforcement Learning". *SIAM J. Control Optim.* 38: 447–469.

Borkar, V. S. (2003). "Avoidance of traps in stochastic approximation". *Systems & control letters*. 50(1): 1–9.

Bottou, L., F. E. Curtis, and J. Nocedal. (2018). "Optimization methods for large-scale machine learning". *SIAM Review*. 60(2): 223–311.

Boyd, S. and L. Vandenberghe. (2004). *Convex optimization*. Cambridge university press.

Brandiere, O. and M. Duflo. (1996). "Les algorithmes stochastiques contournent-ils les pièges?" In: *Annales de l'IHP Probabilités et statistiques*. Vol. 32. No. 3. 395–427.

Bunch, J. R. and B. N. Parlett. (1971). "Direct methods for solving symmetric indefinite systems of linear equations". *SIAM Journal on Numerical Analysis*. 8(4): 639–655.

Cai, H., D. McKenzie, W. Yin, and Z. Zhang. (2022). "Zeroth-order regularized optimization (zoro): Approximately sparse gradients and adaptive sampling". *SIAM Journal on Optimization*. 32(2): 687–714.

Carmon, Y., J. Duchi, O. Hinder, and A. Sidford. (2016). "Accelerated methods for non-convex optimization." *arXiv preprint*. arXiv:1611.00756.

Cassandras, C. G. and S. Lafortune. (2008). *Introduction to discrete event systems*. Springer.

Chen, H. F., L. Guo, and A. J. Gao. (1987). "Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds". *Stochastic Processes and their Applications*. 27: 217–231.

Chen, H.-F., T. E. Duncan, and B. Pasik-Duncan. (1999). "A Kiefer-Wolfowitz algorithm with randomized differences". *IEEE Transactions on Automatic Control*. 44(3): 442–453.

Chen, J., M. I. Jordan, and M. J. Wainwright. (2020a). "HopSkipJumpAttack: A Query-Efficient Decision-Based Attack". In: *IEEE Symposium on Security and Privacy (SP)*. IEEE. 1277–1294.

Chen, P.-Y., H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh. (2017). "ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models". In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. AISec '17*. Dallas, Texas, USA: Association for Computing Machinery. 15–26. ISBN: 9781450352024. DOI: 10.1145/3128572.3140448. URL: https://doi.org/10.1145/3128572.3140448.

Chen, S., A. Devraj, A. Busic, and S. Meyn. (2020b). "Explicit mean-square error bounds for monte-carlo and linear stochastic approximation". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 4173–4183.

Chin, D. C. (1997). "Comparative study of stochastic algorithms for system optimization based on gradient approximations". *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*. 27(2): 244–249.

Choromanski, K., M. Rowland, V. Sindhwani, R. Turner, and A. Weller. (2018). "Structured evolution with compact architectures for scalable policy optimization". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. *Proceedings of Machine Learning Research*. PMLR. 970–978.

Coddington, E. A., N. Levinson, and T. Teichmann. (1956). "Theory of ordinary differential equations".

Cover, T. M. and J. A. Thomas. (2012). *Elements of information theory*. John Wiley & Sons.

Dalal, G., B. Szorenyi, G. Thoppe, and S. Mannor. (2018). "Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning". In: *Conference on Learning Theory*. 1–35.

Dippon, J. (2003). "Accelerated Randomized Stochastic Optimization". *The Annals of Statistics*. 31(4): 1260–1281.

Dong, Y., H. Su, J. Z. Wu, Z. Zhang, and J. Liu. (2020). "Improving Black-box Adversarial Attacks with a Transfer-based Prior". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 17631–17641.

Duchi, J. C., P. L. Bartlett, and M. J. Wainwright. (2012). "Randomized smoothing for stochastic optimization". *SIAM Journal on Optimization*. 22(2): 674–701.

Dunkel, J. and S. Weber. (2010). "Stochastic root finding and efficient estimation of convex risk measures". *Operations Research*. 58(5): 1505–1521.

Durrett, R. (2019). *Probability: theory and examples*. Vol. 49. Cambridge university press.

Erdogdu, M. A. (2016). "Newton-Stein method: An optimization method for glms via stein's lemma". *Journal of Machine Learning Research*. 17(215): 1–52.

Fabian, V. (1968). "On asymptotic normality in stochastic approximation". *The Annals of Mathematical Statistics*: 1327–1332.

Fabian, V. (1971). "Stochastic approximation". In: *Optimizing Methods in Statistics (ed. J.J.Rustagi)*. New York: Academic Press. 439–470.

Filippov, A. F. (2013). *Differential equations with discontinuous right-hand sides: control systems*. Vol. 18. Springer Science & Business Media.

Flaxman, A. D., A. T. Kalai, and H. B. McMahan. (2005). "Online convex optimization in the bandit setting: gradient descent without a gradient". In: *SODA*. 385–394.

Föllmer, H. and A. Schied. (2002). "Convex measures of risk and trading constraints". *Finance and stochastics*. 6(4): 429–447.

Frikha, N. and S. Menozzi. (2012). "Concentration Bounds for Stochastic Approximations". *Electronic Communications in Probability*. 17: no. 47, 1–15.

Fu, M. C., ed. (2015). *Handbook of Simulation Optimization*. Springer. 387.

Furmston, T., G. Lever, and D. Barber. (2016). "Approximate Newton Methods for Approximate Policy Search in Markov Decision Processes". *Journal of Machine Learning Research*. 17: 1–51.

Gadat, S. and I. Gavra. (2022). "Asymptotic study of stochastic adaptive algorithms in non-convex landscape". *Journal of Machine Learning Research*. 23(228): 1–54.

Gallager, R. G. (2013). *Stochastic processes: theory for applications*. Cambridge University Press.

Gasnikov, A., A. Novitskii, V. Novitskii, F. Abdukhakimov, D. Kamzolov, A. Beznosikov, M. Takac, P. Dvurechensky, and B. Gu. (2022). "The power of first-order smooth optimization for black-box nonsmooth problems". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato. Vol. 162. *Proceedings of Machine Learning Research*. PMLR. 7241–7265.

Ge, R., F. Huang, C. Jin, and Y. Yuan. (2015). "Escaping From Saddle Points – Online Stochastic Gradient for Tensor Decomposition". *Conference of Learning Theory*.

Ge, R., C. Jin, and Y. Zheng. (2017). "No spurious local minima in nonconvex low rank problems: A unified geometric analysis". In: *International Conference on Machine Learning*. PMLR. 1233–1242.

Ge, R., J. D. Lee, and T. Ma. (2016). "Matrix completion has no spurious local minimum". *Advances in neural information processing systems*. 29.

Gelfand, S. B. and S. K. Mitter. (1991). "Recursive stochastic algorithms for global optimization in R^d". *SIAM Journal on Control and Optimization*. 29(5): 999–1018.

Gerencser, L., S. D. Hill, and Z. Vago. (1999). "Optimization over discrete sets via SPSA". In: *Proceedings of the 31st conference on Winter simulation: Simulation—a bridge to the future-Volume 1*. 466–470.

Ghadimi, S. and G. Lan. (2013). "Stochastic first-and zeroth-order methods for nonconvex stochastic programming". *SIAM Journal on Optimization*. 23(4): 2341–2368.

Ghoshdastidar, D., A. Dukkipati, and S. Bhatnagar. (2014a). "Newton-based stochastic optimization using q-Gaussian smoothed functional algorithms". *Automatica*. 50(10): 2606–2614.

Ghoshdastidar, D., A. Dukkipati, and S. Bhatnagar. (2014b). "Smoothed Functional Algorithms for Stochastic Optimization Using q-Gaussian Distributions". *ACM Transactions on Modeling and Computer Simulation*. 26(3): 17:1–17:26.

Giesecke, K., T. Schmidt, and S. Weber. (2008). "Measuring the risk of large losses". *Journal of Investment Management, Fourth Quarter*.

Grimmett, G. and D. Stirzaker. (2020). *Probability and random processes*. Oxford university press.

Hegde, V., A. S. Menon, L. A. Prashanth, and K. Jagannathan. (2021). "Online Estimation and Optimization of Utility-Based Shortfall Risk". *Papers* No. 2111.08805. arXiv.org.

Hirsch, M. W., S. Smale, and R. L. Devaney. (2013). *Differential equations, dynamical systems, and an introduction to chaos*. Academic press.

Hu, X., L. A. Prashanth, A. György, and C. Szepesvári. (2016). "(Bandit) Convex Optimization with Biased Noisy Gradient Oracles". In: *Artificial Intelligence and Statistics*. 819–828.

Huang, F., L. Tao, and S. Chen. (2020). "Accelerated stochastic gradient-free and projection-free methods". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. *Proceedings of Machine Learning Research*. PMLR. 4519–4530.

Hurley, M. (1995). "Chain recurrence, semiflows, and gradients". *Journal of Dynamics and Differential Equations*: 437–456.

Ilyas, A., L. Engstrom, A. Athalye, and J. Lin. (2018). "Black-box adversarial attacks with limited queries and information". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. *Proceedings of Machine Learning Research*. PMLR. 2137–2146.

Ilyas, A., L. Engstrom, and A. Madry. (2019). "Prior Convictions: Black-Box Adversarial Attacks with Bandits and Priors". In: *7th International Conference on Learning Representations (ICLR)*.

Ince, E. L. (1956). *Ordinary differential equations*. Courier Corporation.

J. P. Lasalle and S. Lefschetz. (1961). *Stability by Liapunov's Direct Method with Applications*. New York: Academic Press.

J.R.Blum. (1954). "Approximation methods which converge with probability one". *Annals of Mathematical Statistics*: 382–386.

Jain, P., D. Nagaraj, and P. Netrapalli. (2021). "Making the Last Iterate of SGD Information Theoretically Optimal". *SIAM Journal on Optimization*. 31(2): 1108–1130.

Ji, K., Z. Wang, Y. Zhou, and Y. Liang. (2019). "Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. *Proceedings of Machine Learning Research*. PMLR. 3100–3109.

Jin, C., R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. (2017). "How to Escape Saddle Points Efficiently." *ICML*: 1724–1732.

Jin, C., P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. (2021). "On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points". *Journal of the ACM (JACM)*. 68(2): 1–29.

Karandikar, R. L. and M. Vidyasagar. (2024). "Convergence rates for stochastic approximation: Biased noise with unbounded variance, and applications". *Journal of Optimization Theory and Applications*: 1–39.

Karmakar, P. and S. Bhatnagar. (2018). "Two time-scale stochastic approximation with controlled Markov noise and off-policy temporal-difference learning". *Mathematics of Operations Research*. 43(1): 130–151.

Karmakar, P. and S. Bhatnagar. (2021). "Stochastic Approximation With Iterate-Dependent Markov Noise Under Verifiable Conditions in Compact State Space With the Stability of Iterates Not Ensured". *IEEE Transactions on Automatic Control*. 66(12): 5941–5954.

Katkovnik, V. Y. and Y. Kulchitsky. (1972). "Convergence of a class of random search algorithms". *Automation Remote Control*. 8: 1321–1326.

Kawaguchi, K. (2016). "Deep learning without poor local minima". *Advances in neural information processing systems*. 29.

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

Kirkpatrick, S., C. D. Gelatt Jr, and M. P. Vecchi. (1983). "Optimization by simulated annealing". *science*. 220(4598): 671–680.

Konda, V. R. and J. N. Tsitsiklis. (2003). "On actor-critic algorithms". *SIAM journal on Control and Optimization*. 42(4): 1143–1166.

Kornowski, G. and O. Shamir. (2024). "An algorithm with optimal dimension dependence for zero-order nonsmooth nonconvex stochastic optimization". *Journal of Machine Learning Research*. 25(122): 1–14.

Kozak, D., C. Molinari, L. Rosasco, L. Tenorio, and S. Villa. (2023). "Zeroth-order optimization with orthogonal random directions". *Mathematical Programming*. 199(1): 1179–1219.

Kulkarni, V. G. (2016). *Modeling and analysis of stochastic systems*. Chapman and Hall/CRC.

Kushner, H. J. and D. S. Clark. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. New York: Springer Verlag.

Kushner, H. J. and G. G. Yin. (2003). *Stochastic Approximation Algorithms and Applications, 2'nd Ed.* New York: Springer Verlag.

Lei, J. (2020). "Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces". *Bernoulli*. 26(1): 767–798.

Levin, D. A. and Y. Peres. (2017). *Markov chains and mixing times*. Vol. 107. American Mathematical Soc.

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

Malladi, S., T. Gao, E. Nichani, A. Damian, J. D. Lee, D. Chen, and S. Arora. (2023). "Fine-tuning language models with just forward passes". *Advances in Neural Information Processing Systems*. 36: 53038–53075.

Mania, H., A. Guy, and B. Recht. (2018). "Simple random search of static linear policies is competitive for reinforcement learning". *Advances in neural information processing systems*. 31.

Maniyar, M. P., L. Prashanth, A. Mondal, and S. Bhatnagar. (2024). "A Cubic-regularized Policy Newton Algorithm for Reinforcement Learning". In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. Ed. by S. Dasgupta, S. Mandt, and Y. Li. Vol. 238. *Proceedings of Machine Learning Research*. PMLR. 4708–4716.

Maryak, J. L. and D. C. Chin. (2001). "Global random optimization by simultaneous perturbation stochastic approximation". In: *Proceedings of the 2001 American control conference.(Cat. No. 01CH37148)*. Vol. 2. IEEE. 756–762.

Meyn, S. (2022). *Control systems and reinforcement learning*. Cambridge University Press.

Meyn, S. P. and R. L. Tweedie. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Mondal, A., L. Prashanth, and S. Bhatnagar. (2024). "Truncated Cauchy random perturbations for smoothed functional-based stochastic optimization". *Automatica*. 162: 111528. ISSN: 0005-1098. DOI: https://doi.org/10.1016/j.automatica.2024.111528.

Mou, W., C. J. Li, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. (2020). "On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration". In: *Conference on Learning Theory*. PMLR. 2947–2997.

Mukhoty, B., V. Bojkovic, W. de Vazelhes, X. Zhao, G. D. Masi, H. Xiong, and B. Gu. (2023). "Direct Training of SNN using Local Zeroth Order Method". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Nandakumaran, A., P. Datti, and R. George. (2017). *Ordinary differential equations: Principles and applications*. Cambridge University Press.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro. (2009). "Robust stochastic approximation approach to stochastic programming". *SIAM Journal on Optimization*. 19(4): 1574–1609.

Nesterov, Y. and B. Polyak. (2007). "Cubic regularization of Newton method and its global performance." *Mathematical Programming*. 112: 159–181.

Nesterov, Y. and V. Spokoiny. (2017). "Random gradient-free minimization of convex functions". *Foundations of Computational Mathematics*. 17(2): 527–566.

Nesterov, Y. and B. Polyak. (2006). "Cubic regularization of Newton method and its global performance". *Math. Program.* 108(Aug.): 177–205.

Nocedal, J. and S. J. Wright. (1999). *Numerical optimization*. Springer.

Norris, J. R. (1998). *Markov chains*. No. 2. Cambridge university press.

Pachal, S., S. Bhatnagar, and L. A. Prashanth. (2023). "Generalized Simultaneous Perturbation-based Gradient Search with Reduced Estimator Bias". *arXiv preprint arXiv:2212.10477*.

Pemantle, R. (1990). "Non-convergence to unstable points in urn models and stochastic approximations". *The Annals of Probability*. 18(2): 698–712.

Polyak, B. and A. Tsybakov. (1990). "Optimal orders of accuracy for search algorithms of stochastic optimization". *Problems in Information Transmission*: 126–133.

Polyak, B. T. and A. B. Juditsky. (1992). "Acceleration of stochastic approximation by averaging". *SIAM Journal on Control and Optimization*. 30(4): 838–855.

Powell, W. B. (2021). "Reinforcement learning and stochastic optimization".

Prashanth, L. A. and S. P. Bhat. (2022). "A Wasserstein Distance Approach for Concentration of Empirical Risk Estimates". *Journal of Machine Learning Research*. 23(238): 1–61.

Prashanth, L. A. and S. Bhatnagar. (2012). "Threshold Tuning Using Stochastic Optimization for Graded Signal Control". *IEEE Transactions on Vehicular Technology*. 61(9): 3865–3880.

Prashanth, L. A., S. Bhatnagar, N. Bhavsar, M. Fu, and S. I. Marcus. (2020). "Random Directions Stochastic Approximation With Deterministic Perturbations". *IEEE Transactions on Automatic Control*. 65(6): 2450–2465.

Prashanth, L. A., S. Bhatnagar, M. C. Fu, and S. I. Marcus. (2017). "Adaptive system optimization using random directions stochastic approximation". *IEEE Transactions on Automatic Control*. 62(5): 2223–2238.

Prashanth, L. A., A. Chatterjee, and S. Bhatnagar. (2014). "Two timescale convergent Q-learning for sleep-scheduling in wireless sensor networks". *Wireless networks*. 20: 2589–2604.

Prashanth, L. A. and M. Ghavamzadeh. (2016). "Variance-constrained actor-critic algorithms for discounted and average reward MDPs". *Machine Learning*. 105: 367–417.

Prashanth, L. A., N. Korda, and R. Munos. (2021). "Concentration bounds for temporal difference learning with linear function approximation: the case of batch data and uniform sampling". *Mach. Learn.* 110(3): 559–618.

Ramaswamy, A. and S. Bhatnagar. (2016a). "A generalization of the Borkar-Meyn theorem for stochastic recursive inclusions". *Mathematics of Operations Research.* 42(3): 648–661.

Ramaswamy, A. and S. Bhatnagar. (2018). "Analysis of gradient descent methods with nondiminishing bounded errors". *IEEE Transactions on Automatic Control.* 63(5): 1465–1471.

Ramaswamy, A. and S. Bhatnagar. (2019). "Stability of stochastic approximations with controlled Markov noise and temporal difference learning". *IEEE Transactions on Automatic Control.* 64(6): 2614–2620.

Ramaswamy, A. and S. Bhatnagar. (2021). "Analyzing approximate value iteration algorithms". *Mathematics of Operations Research.* DOI: 10.1287/moor.2021.1202.

Ramaswamy, A. and S. Bhatnagar. (2016b). "Stochastic recursive inclusion in two timescales with an application to the Lagrangian dual problem". *Stochastics.* 88(8): 1173–1187.

Rando, M., C. Molinari, L. Rosasco, and S. Villa. (2023). "An optimal structured zeroth-order algorithm for non-smooth optimization". In: *Advances in Neural Information Processing Systems.* Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc. 36738–36767.

Rando, M., C. Molinari, S. Villa, and L. Rosasco. (2024). "Stochastic zeroth order descent with structured directions". *Computational Optimization and Applications.* Oct.

Rastogi, P., J. Zhu, and J. C. Spall. (2016). "Efficient implementation of enhanced adaptive simultaneous perturbation algorithms". In: *2016 Annual Conference on Information Science and Systems (CISS).* IEEE. 298–303.

Robbins, H. and S. Monro. (1951). "A stochastic approximation method". *Ann. Math. Statist.* 22: 400–407.

Rockafellar, R. T. and S. Uryasev. (2000). "Optimization of conditional value-at-risk". *Journal of Risk.* 2: 21–42.

Royden, H. and P. Fitzpatrick. (2010). *Real Analysis*. 4th. Boston: Pearson. ISBN: 9780131437470.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method*. New York: Wiley.

Ruppert, D. (1985). "A Newton-Raphson version of the multivariate Robbins-Monro procedure". *Annals of Statistics*. 13: 236–245.

Saha, A. and A. Tewari. (2011). "Improved regret guarantees for online smooth convex optimization with bandit feedback". In: *AISTATS*. 636–642.

Salimans, T., J. Ho, X. Chen, S. Sidor, and I. Sutskever. (2017). "Evolution strategies as a scalable alternative to reinforcement learning". *arXiv preprint arXiv:1703.03864*.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*. Vol. 162. John Wiley & Sons.

Shamir, O. (2017). "An optimal algorithm for bandit and zero-order convex optimization with two-point feedback". *Journal of Machine Learning Research*. 18(52): 1–11.

Spall, J. C. (2000). "Adaptive stochastic approximation by the simultaneous perturbation method". *IEEE Trans. Autom. Contr.* 45: 1839–1853.

Spall, J. C. (2005). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Vol. 65. John Wiley & Sons.

Spall, J. C. (1992). "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation". *IEEE Transactions on Automatic Control*. 37(3): 332–341.

Spall, J. C. (1997). "A one-measurement form of simultaneous perturbation stochastic approximation". *Automatica*. 33(1): 109–112.

Spall, J. C. (2009). "Feedback and weighting mechanisms for improving Jacobian estimates in the adaptive simultaneous perturbation algorithm". *IEEE Transactions on Automatic Control*. 54(6): 1216–1229.

Stein, C. (1972). "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables". In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. Vol. 6. University of California Press. 583–603.

Stein, C. (1981). "Estimation of the mean of a multivariate normal distribution". *The annals of Statistics*: 1135–1151.

Styblinski, M. A. and T.-S. Tang. (1990). "Experiments in nonconvex optimization: stochastic approximation with function smoothing and simulated annealing". *Neural Networks*. 3: 467–483.

Sun, J., Q. Qu, and J. Wright. (2016). "Complete dictionary recovery over the sphere I: Overview and the geometric picture". *IEEE Transactions on Information Theory*. 63(2): 853–884.

Sutton, R. S. and A. W. Barto. (2018). *Reinforcement Learning, 2'nd Edition*. MIT Press.

Sutton, R. S., H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora. (2009). "Fast gradient-descent methods for temporal-difference learning with linear function approximation". In: *Proceedings of the 26th annual international conference on machine learning*. 993–1000.

Sutton, R. S., D. McAllester, S. Singh, and Y. Mansour. (1999). "Policy gradient methods for reinforcement learning with function approximation". *Advances in neural information processing systems*. 12.

Swain, J. (2017). "Simulation Software Survey-Simulation: new and improved reality show". *OR/MS Today*. 44(5): 38–49.

Tamar, A., Y. Chow, M. Ghavamzadeh, and S. Mannor. (2015). "Policy Gradient for Coherent Risk Measures". In: *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc.

Tripuraneni, N., M. Stern, C. Jin, J. Regier, and M. I. Jordan. (2018). "Stochastic Cubic Regularization for Fast Nonconvex Optimization". In: *NeurIPS*. Vol. 31. Curran Associates, Inc.

Tropp, J. A. (2016). "The Expected Norm of a Sum of Independent Random Matrices: An Elementary Approach". In: *High Dimensional Probability VII: The Cargèse Volume*. 173–202.

Tsitsiklis, J. N. and B. Van Roy. (1997). "An analysis of temporal-difference learning with function approximation". *IEEE Transactions on Automatic Control*. 42(5): 674–690.

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

Vijayan, N. and L. A. Prashanth. (2021). "Smoothed functional-based gradient algorithms for off-policy reinforcement learning: A non-asymptotic viewpoint". *Systems & Control Letters*. 155: 104988.

Vijayan, N. and L. A. Prashanth. (2023). "A policy gradient approach for optimization of smooth risk measures". In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. Ed. by R. J. Evans and I. Shpitser. Vol. 216. *Proceedings of Machine Learning Research*. PMLR. 2168–2178.

Wang, T. and Y. Feng. (2024). "Convergence rates of zeroth order gradient descent for Lojasiewicz functions". *INFORMS Journal on Computing*.

Williams, R. J. (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning". *Machine learning*. 8: 229–256.

Wright, S. J. and B. Recht. (2022). *Optimization for data analysis*. Cambridge University Press.

Yaji, V. and S. Bhatnagar. (2019). "Analysis of stochastic approximation schemes with set-valued maps in the absence of a stability guarantee and their stabilization". *IEEE Transactions on Automatic Control*. 65(3): 1100–1115.

Yaji, V. G. and S. Bhatnagar. (2018). "Stochastic recursive inclusions with non-additive iterate-dependent Markov noise". *Stochastics*. 90(3): 330–363.

Yaji, V. G. and S. Bhatnagar. (2020). "Stochastic recursive inclusions in two timescales with nonadditive iterate-dependent markov noise". *Mathematics of Operations Research*. 45(4): 1405–1444.

Yao, A. C. C. (1977). "Probabilistic computations: Toward a unified measure of complexity". In: *FOCS*. 222–227.

Zhang, Y., Y. Yao, J. Jia, J. Yi, M. Hong, S. Chang, and S. Liu. (2022). "How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=W9G_ImpHlQd.

Zhu, J., L. Wang, and J. C. Spall. (2019). "Efficient implementation of second-order stochastic approximation algorithms in high-dimensional problems". *IEEE transactions on neural networks and learning systems*. 31(8): 3087–3099.

Zhu, X. and J. C. Spall. (2002). "A modified second-order SPSA optimization algorithm for finite samples". *Int. J. Adapt. Control Signal Process.* 16: 397–409.

Index