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 1–200. 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 analysis as well as finite sample analysis of the procedures used for convex and non-convex objectives.

We present algorithms that either estimate the gradient in gradient-based schemes or estimate both the gradient and Hessian in Newton-type procedures using random direction methods. As mentioned these approaches estimate the gradient and/or Hessian using sample observations and hence are zeroth order methods. We provide both asymptotic convergence guarantees in the general setup as well as asymptotic

Prashanth L. A. and Shalabh Bhatnagar (2024), "Gradient-based algorithms for zeroth-order optimization", : Vol. xx, No. xx, pp 1–200. DOI: 10.1561/XXXXXXXXX.

normality results for various 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 finally also present a couple of interesting applications of these methods in the domain of reinforcement learning. A large portion of this work is motivated from our own contributions to this domain.

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 can be of interest to readers working in allied areas such as reinforcement learning, and first-order stochastic smooth optimization. We provide a detailed coverage of zeroth-order gradient estimates, including classic ones such as SPSA, SF, and more recent ones such as RDSA and generalized SPSA. The convergence analysis includes both asymptotic guarantees via the ODE and DI approaches, as well as non-asymptotic bounds. The convergence analysis should be of interest to students as well as researchers working in the broad area of stochastic optimization and machine learning.

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 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 likely be extremely computationally intensive, more so, since we are interested in obtaining the optima of objective functions over continuously valued sets.

We shall primarily be concerned here with the problem of finding 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

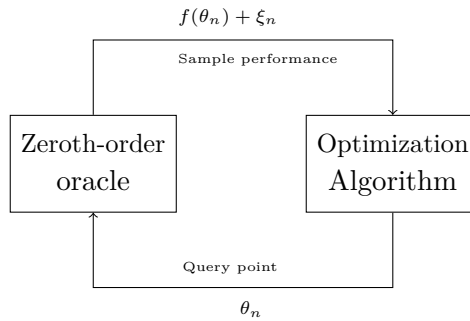


Figure 1.1: Model-free optimization framework

'noisy' data recursively 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.

One of the primary goals in model-free, i.e., purely data-driven or simulation-based optimization methods is to find the minima of a real-valued function f that is often not known analytically. Thus, the goal is 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. This is the basic setting of simulation or data driven optimization. It is important to note here that the noisy observations $f(\theta_n) + \xi_n$ above cannot be separated into the objective function $f(\theta_n)$ and the noise component ξ_n to infer the objective function value directly from the given noise corrupted data. 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 in general. Further, finding one of the minimizers is usually sufficient in most cases (as opposed to finding all of them). It must be noted however that finding a global minimum, in this setting, is far more computationally intensive in general, as compared to 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 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 computationally expensive. Thus, in addition to searching for optima, a simulation optimization algorithm has to ensure that the number of evaluations is minimum.

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).

Another area of practical interest for zeroth-order optimization algorithms is reinforcement learning (RL) (Sutton and Barto, 2018; Bertsekas and Tsitsiklis, 1996). 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 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 are a popular solution approach for such problems. The basis for such algorithms is the policy gradient theorem, which motivates the uses of likelihood ratio based gradient estimates. While such an approach of obtaining unbiased gradient estimates works in an 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.

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 it's inventors, H.Robbins and S.Monro, see (Robbins and Monro, 1951),

was designed to find the zeros of an unknown function $h : \mathcal{R}^d \to \mathcal{R}^d$. 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 $\theta_0 \in \mathcal{R}^d$. 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 result in reducing the noise-effects, thereby resulting 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 certain function as well as finding a minimum of an objective function both under noisy observations. In the former case, $h(\theta)$ in (1.2) can have the form $h(\theta) = g(\theta) - \theta$ for some function $g : \mathcal{R}^d \to \mathcal{R}^d$, while in the latter, $h(\theta)$ can be of the form $h(\theta) = -\nabla f(\theta)$ for some function $f : \mathcal{R}^d \to \mathcal{R}$. 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 estimates 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 gradient estimates obtained from noisy objective function observations.

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 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 below. This algorithm used 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), 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 algorithm

Consider the following stochastic approximation scheme:

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

where $\hat{\nabla} f(\theta_n)$ is a noisy estimate of the gradient of $f(\theta_n)$, with $f : \mathcal{R}^d \to \mathcal{R}$ 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 $i = 1, \ldots, d$,

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

where, $\hat{\nabla}_i f(\theta_n)$ 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 in the ith place and 0 elsewhere. Here $\xi_i^+(n)$ (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

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

This happens because the first and the third terms in the Taylor's expansion get cancelled as a consequence of the balanced nature of the iterates. The term comprising $o(\delta)$ contributes to the bias in the gradient estimates. In relation to (1.3), if $\delta \to 0$ as $n \to \infty$, 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 has the undesirable effect of constraining the choice of the step-size sequence $\{a(n)\}$. For a fixed δ, 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 $O(\epsilon)$.

A disadvantage with the above gradient estimator is that it requires $2d$ function measurements or simulations in order 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 for the gradient has been proposed that uses only two function measurements regardless of the value of d.

$$\hat{\nabla}_i f(\theta_n) = \frac{(f(\theta_n + \delta\Delta(n)) + \xi^+(n) - f(\theta_n - \delta\Delta(n)) - \xi^-(n))}{2\delta\Delta_i(n)}. \tag{1.7}$$

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:

$$\frac{f(\theta_n + \delta\Delta(n)) - f(\theta_n - \delta\Delta(n))}{2\delta\Delta_i(n)} = \frac{\Delta(n)^T \nabla f(\theta_n)}{\Delta_i(n)} + o(\delta)$$
$$= \nabla_i f(\theta_n) + \sum_{j \neq i} \frac{\Delta_j(n) \nabla_j f(\theta_n)}{\Delta_i(n)} + o(\delta). \tag{1.8}$$

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

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

Hence, we obtain

$$\left\| \mathbb{E}\left[\widehat{\nabla} f(\theta_n) \mid \theta_n\right] - \nabla f(\theta_n) \right\| \leq C\delta^2, \tag{1.9}$$

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

$$\hat{\nabla}_i f(\theta_n) = \frac{f(\theta_n + \delta\Delta(n)) + \xi^+(n)}{\delta\Delta_i(n)}. \tag{1.10}$$

A Taylor's expansion in the above 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 (1.8). 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 term, 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 are available in (Bhatnagar *et al.*, 2003). 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 algorithm

Recall that a SG algorithm involves the following update iteration:

$$\theta_{n+1} = \theta_n - a_n \widehat{\nabla} f(\theta_n), \tag{1.11}$$

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 (1.11) is *not* scale-invariant, i.e., changing θ to $B\theta$ for some matrix B, would imply a change in the update (1.11). 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 algorithm overcomes the shortcomings of a first-order SG algorithm mentioned above. A general gradient-search algorithm involves an update rule of the form:

$$\theta_{n+1} = \theta_n - a_n B(\theta_n)^{-1} \nabla f(\theta_n), \tag{1.12}$$

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 (1.12) reduces to the first-order gradient algorithm (1.11).

(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) Newton algorithm.

In the following, we focus on the Newton 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 (1.11) (or alternatively the case of $B(\theta) = I$ in (1.12)). 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

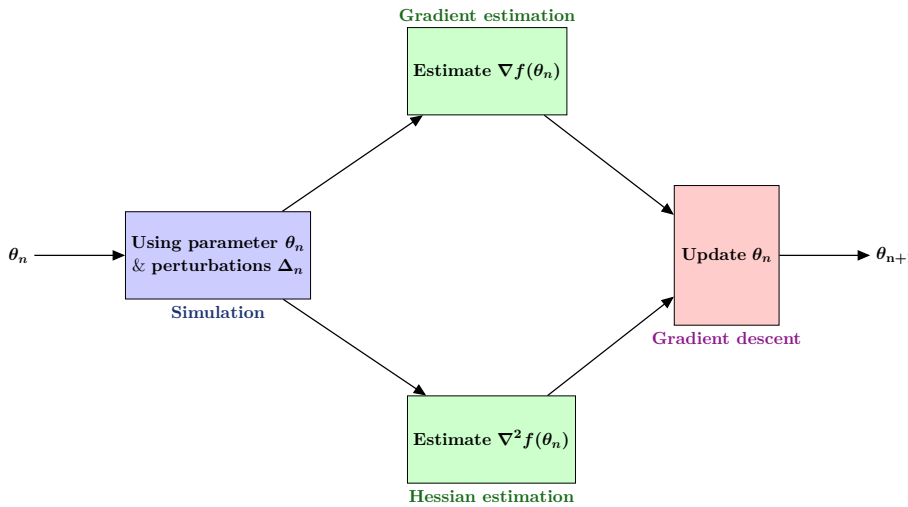


Figure 1.2: Overall flow of a second-order stochastic gradient algorithm

minimize a quadratic model of f, while SG algorithm (1.11) 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:

$$\theta_{n+1} = \theta_n - a_n \left(\overline{H}_n\right)^{-1} \widehat{\nabla} f(\theta_n). \tag{1.13}$$

The topic of gradient estimation is handled in Chapter 3, while Chapter 6 focuses on Hessian estimation, and subsequently, in Chapter 7, we shall perform a convergence analysis of (1.13), where we use zeroth-order estimates of both the gradient and the Hessian.

To understand the problem of Hessian estimation, we now discuss

a finite difference approximation, which requires $O(N^2)$ function measurements. Simultaneous perturbation trick brings this number down to a constant, irrespective of the problem dimension. 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:

$$\frac{df(\theta)}{d\theta} \approx \left(\frac{f(\theta+\delta) - f(\theta-\delta)}{2\delta}\right). \tag{1.14}$$

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

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

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

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:

$$\frac{d^2f(\theta)}{d\theta^2} \approx \frac{\left(\dfrac{f(\theta+\delta+\delta) - f(\theta+\delta-\delta)}{2\delta}\right) - \left(\dfrac{f(\theta-\delta+\delta) - f(\theta-\delta-\delta)}{2\delta}\right)}{2\delta} \tag{1.15}$$

As before, using Taylor series expansions, it can 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 $i, j \in \{1, \ldots, d\}$,

$$\nabla^2_{ij} f(\theta) \approx \frac{1}{4\delta^2}\Big(f(\theta + \delta e_i + \delta e_j) + f(\theta + \delta e_i - \delta e_j) - (f(\theta - \delta e_i + \delta e_j) - f(\theta - \delta e_i - \delta e_j))\Big). \tag{1.16}$$

Such an approach requires $4N^2$ 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 noisy case as follows: Suppose we have the following function measurements: For any $i, j \in \{1, \ldots, d\}$,

$$y_1 = f(\theta + \delta e_i + \delta e_j) + \xi_{1ij}, y_2 = f(\theta + \delta e_i - \delta e_j) + \xi_{2ij}, \tag{1.17}$$

$$y_3 = f(\theta - \delta e_i + \delta e_j) + \xi_{3ij} \text{ and } y_4 = f(\theta - \delta e_i - \delta e_j) + \xi_{4ij}. \tag{1.18}$$

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

$$\widehat{H}_{ij} = \left(\frac{y_1 - y_2 - y_3 + y_4}{4\delta^2} \right), \forall i, j \tag{1.19}$$

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

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

While the bias of the estimator is one the lower side, with explicit control via the δ parameter, the problem is in the number of function measurements. The latter number is $4N^2$, 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 constant number of function measurements, 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 data as it becomes available and are model-free procedures. In Chapter 2, we provide an 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 a general setting and 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 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 estimates. The unified estimates feature abstract random perturbations that are required to satisfy certain conditions to ensure that the bias and variance of the estimate are 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). In this chapter, we analyze the bias and variance of the aforementioned estimators in the convex as well as 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, and provide a non-asymptotic bound

that quantifies the convergence rate. For the asymptotic convergence, we use the theory of differential inclusions to establish that the stochastic gradient algorithm converges to a chain recurrent set of a differential inclusion.

In Chapter 5, we present the non-asymptotic analysis for 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. This bound 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 convex and 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 rate 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 ZSG algorithm are optimal up to 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), its later enhancements in (Bhatnagar, 2005; Bhatnagar and Prashanth, 2015a). We also describe second-order smoothed functional (Bhatnagar, 2007) and second-order RDSA (Prashanth *et al.*, 2017) schemes. We analyze the bias of these Hessian estimates, and establish that each of these aforementioned schemes result in an asymptotically unbiased Hessian estimate.

In Chapter 7, we analyze a stochastic Newton algorithm using

gradient/Hessian estimates based on the simultaneous perturbation method. The theoretical guarantees include the asymptotic 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 8, we provide applications of simultaneous perturbation methods in a 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. 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 to local optima in the 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.

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, 2007). 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 (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 two-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) Hessian using zeroth order estimators though we also consider two-timescale algorithms. We present 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 equation (ODE) or differential inclusions. Our current text also covers many recent algorithms on top of those 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 recursive 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 a constant δ leads to recursive 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 recursions will also be seen to almost surely track asymptotically the underlying 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 $\theta_n = \frac{1}{m}\sum_{k=1}^{n} X_k$ be the sample mean computed using these n samples. We now derive an iterative scheme for updating the sample mean.

$$\begin{aligned}\theta_{n+1} &= \frac{1}{n+1}\sum_{k=1}^{n+1} X_k = \frac{n}{n+1}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) + \frac{1}{n+1}X_{n+1} \\ &= \frac{n}{n+1}\theta_m + \frac{1}{n+1}X_{n+1} \\ \theta_{n+1} &= \theta_m + \frac{1}{n+1}\left(X_{n+1} - \theta_m\right). \end{aligned} \tag{2.3}$$

The update rule above is a stochastic approximation scheme with step-size $a(n) = \frac{1}{n+1}$.

Strong law of large numbers says the following:

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

Rewriting the update rule (2.3), we obtain

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

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

From an application of Kushner Clark lemma, to be presented later (see Theorem 2.3 below), it can be shown that $\theta_n \to \mu$ as $n \to \infty$ for more general step-sizes that satisfy

$$a(n) > 0, a(n) \to 0, \text{ and } \sum_n a(n) = \infty. \tag{2.4}$$

2.2.2 Stochastic gradient algorithm using unbiased gradient information

Consider the following problem:

$$\theta^* \in \arg\min_{\theta} f(\theta), \tag{2.5}$$

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

A stochastic gradient algorithm for solving (2.5) would update as follows:

$$\theta_{n+1} = \theta_n - a(n)\widehat{\nabla} f(\theta_n). \tag{2.6}$$

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 (2.4) above).

Here we shall assume unbiased gradient information is available, i.e., $\mathbb{E}(\widehat{\nabla} f(\theta_n) \mid \theta_n) = \nabla f(\theta_n)$. In this case, the algorithm in (4.4) 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 (4.4), the function of interest is ∇f. If the gradient estimates $\widehat{\nabla} f(\theta_n)$ have bounded variance, then the algorithm in (4.4) can be shown to converge to the stationary points of f. We make this claim precise later in Section 4.1.

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.

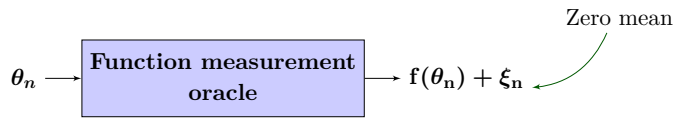


Figure 2.1: Simulation optimization

The stochastic gradient algorithm updates as follows:

$$\theta_{n+1} = \theta_n - a(n)\widehat{\nabla} f(\theta_n), \tag{2.7}$$

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 (1.7), 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 (1.7).

Under suitable assumptions, θ_n governed by (2.7) 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

$$\|f(x) - f(y)\| \leq \alpha \|x - y\|, \tag{2.8}$$

for any $x, y \in \mathbb{R}^d$. Here $\alpha \in (0, 1)$, and $\|\cdot\|$ is the ℓ_2-norm.

Assuming the underlying space is Euclidean (and hence 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 $\theta_0 \in \mathbb{R}^n$ 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:

$$\theta_{n+1} = (1 - a(n))\theta_n + a(n)(f(\theta_n) + \xi_{n+1}), \tag{2.9}$$

A simple setting is where $\{\xi_m\}$ 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; (ii) step sizes satisfy standard stochastic approximation conditions, see (2.4); and (iii) noise ξ_m is a martingale difference sequence that has bounded varaince, or satisfies a linear growth condition (see Assumption A2.8 in the next chapter).

Remark 2.1. The stochastic fixed point iteration algorithm discussed above would not necessarily converge if the modulus of contraction $\alpha = 1$ in (2.8). In this case, a fixed point is not even guaranteed to exist, e.g., consider $f(\theta) = x + 1$. Alternatively, more than one fixed points may exists (e.g., $f(\theta) = x$), or only one fixed point exists (e.g. $f(\theta) = -x$). Under an additional assumption that at least one fixed point exists, the stochastic fixed point iteration (2.9) 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.*, 2020).

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:

$$q_{n+1} = q_n + a_n \left(\mathbb{I}\{X_{n+1} \leq q_n\} - \alpha\right). \tag{2.10}$$

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 is as follows: 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 (2.10). 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 (2.10) is equivalent to

$$q_{n+1} = q_n + a_n \left(Y_{n+1} - \alpha\right). \tag{2.11}$$

Using a variant of Kushner Clark lemma it is possible to establish almost sure convergence of q_n to $q_\alpha(X)$, and we omit the details.

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 one popular risk measure defined by

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

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, let

$$v(\xi, X) := \xi + \frac{1}{1-\alpha}(X - \xi)_+ \text{ and } V(\xi) = \mathbb{E}[v(\xi, X)] \tag{2.12}$$

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 $\{X_1, \ldots, X_n\}$:

$$\text{VaR: } q_{n+1} = q_n - a_n(1 - \frac{1}{1-\alpha}\mathbb{I}\{X_{n+1} \geq q_n\}), \tag{2.13}$$

$$\text{CVaR: } \psi_{n+1} = \psi_n - \frac{1}{n+1}\left(\psi_n - v(q_n, X_{n+1})\right). \tag{2.14}$$

In the above, (2.13) can be seen as a gradient descent rule, while (2.14) can be seen as a plain averaging update.

An interesting question is whether the stochastic gradient-based estimation scheme in (2.13) converges faster than the root-finding estimation scheme in (2.10).

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:

$$\theta_{n+1} = \theta_n + a(n)(h(\theta_n) + \beta_n + \eta_n). \tag{2.15}$$

Let $L(\{\theta_n, n \geq 0\})$ denote the limit set of the sequence $\theta_n, n \geq 0$ obtained from (2.15). Consider the following ODE associated with (2.15):

$$\dot{\theta}(t) = h(\theta(t)). \tag{2.16}$$

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

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

We now state the following result given as Theorem 1.2 in (Benaïm, 1996)): For the algorithm (2.15), we make the following assumptions:

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 (2.16) with domain of attraction $\check{\Omega}$.

We now discuss these assumptions. Assumption A2.1 ensures that the ODE (2.16) is well-posed. Assumption A2.2 ensures that the bias β_n vanishes asymptotically. Assumption A2.3 contains standard stochastic approximation conditions on the step sizes $\{a(n)\}$. From the viewpoint that a linearly interpolation of the stochastic approximation iterates approximates a trajectory of the ODE, the conditions on the stepsizes can be understood as follows: $\sum_n \alpha_n = \infty$ ensures that the entire time axis is covered since we $a(n)$ can be seen as the time steps (along the x-axis) with the corresponding stochastic approximation iterate θ_n along the y-axis; The condition $a(n) \to 0$ ensures the discretization errors can be ignored. Assumption A2.4 imposes conditions on the noise M_{n+1} that ensure the effect of noise is asymptotically negligible. 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 when one does not have an attractor in the underlying system.

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_{m+1} is

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

If we choose a constant stepsize, i.e., $a(m) = a\ \forall m$, then, $Var(\theta_{m+1}) \geq a^2\sigma^2$. Thus, with a constant stepsize, $\theta_m \not\longrightarrow \theta^*$ almost surely, motivating the need for having a diminishing stepsize that vanishes asymptotically. However, such a stepsize 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 θ_m is within a certain radius of the initial point θ_0. This will be problematic if θ^* lies outside this radius. Hence, we need $\sum_\tau a(\tau) = \infty$.

The main result, which is Theorem 2.3.1 of (Kushner and Clark, 1978), that establishes convergence of (2.15) is given below.

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 (2.15) 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 also given in (Benaïm, 1996) [Theorem 1.2].

Proof. Recall the stochastic recursion (2.15):

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

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 (2.15) can be written in continuous time as

$$\theta^0(t) = \theta^0(0) + \int_0^t h(\bar{\theta}^0(\tau))d\tau + \beta^0(t) + \eta^0(t), \ t \geq 0. \tag{2.17}$$

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 $0 < T < \infty$,

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

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

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

Thus,

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

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

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

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

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

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, (2.17) can be equivalently written as follows: For $t \ge 0$,

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

$$= \theta^n(0) + \int_0^t h(\theta^n(\tau))d\tau + \epsilon^n(t) + \beta^n(t) + \eta^n(t), \tag{2.18}$$

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),

$$\|\epsilon^n(t)\| \leq L \int_0^t \|\bar{\theta}^0(t(n)+\tau)) - \theta^n(\tau))\|d\tau, \tag{2.19}$$

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,

$$\|\theta^n(t) - \theta^0(t+t(n))\| \leq \int_0^t \|h(\bar{\theta}^0(t(n)+\tau))\|d\tau + \|\beta^n(t)\| + \|\eta^n(t)\|. \tag{2.20}$$

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

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

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 (2.20), 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 (2.19), 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 (2.16) as

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

which is the integral form of the ODE (2.16).

Now note that under Assumption A2.6, $\theta^* \in \mathbb{R}^d$ is an attractor for the ODE (2.16). 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 (2.16). 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 $\{\theta_{n_{mj}}\}$ 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 (2.16). 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 $\check{\theta}(\cdot)$ that again satisfies (2.16). Then $\check{\theta}(0) \in \partial N_{\epsilon_1}(\theta^*)$ and $\check{\theta}(t) \in \bar{N}_{\epsilon_2}(\theta^*) \backslash N_{\epsilon_1}(\theta^*)$. This contradicts that θ^* is asymptotically stable. The claim follows. □

Remark 2.2. A more formal argument on the tracking of the iterate sequence to the underlying ODE (2.16) is provided in Chapter 2 of Borkar, 2022. We briefly sketch that argument here for completeness.

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 (2.16) 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, 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 (2.16) 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.

Theorem 2.3 (Generalized Kushner and Clark Theorem)**.** Under A2.1–A2.5, $L(\{\theta_n, n \geq 0\})$ is a connected internally chain recurrent set for the ODE (2.16).

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.

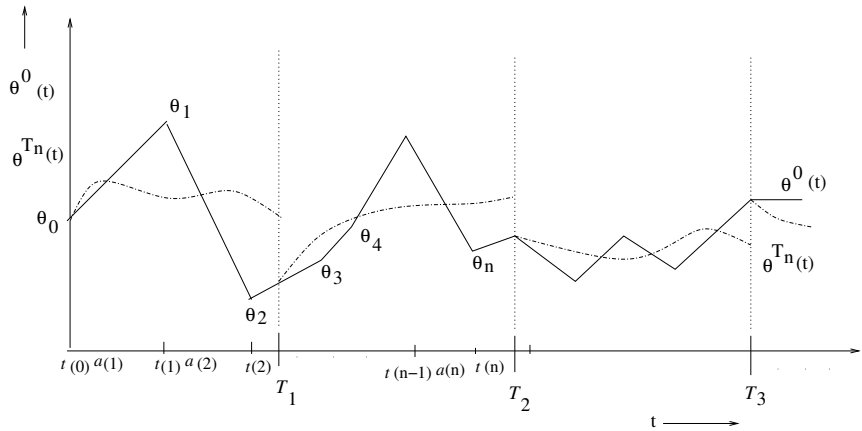


Figure 2.2: The continuously interpolated algorithm's trajectory $\theta^0(t)$ represented by the solid line asymptotically tracks the ODE's trajectory (the dashed-dotted line) $\theta^{T_n}(t)$ suitably reset to the algorithm's trajectory after every (regular) time interval say T instants long. On the 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.

A2.7. $\sum_n a(n) = \infty$ 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,

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

Assumption A2.7 is stronger than Assumption A2.3. However, Assumptions A2.7 and A2.8 turn out to be sufficient conditions for the verification of Assumption A2.4 (in addition to Assumption A2.5). 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,

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

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

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

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 Chapter 3 of (Borkar, 1995), $\{\chi_n\}$ itself is almost surely convergent. Assumption A2.4 will thus follow.

We now discuss the specific case of stochastic gradient algorithms of the form (2.15) with $h(\theta) = -\nabla f(\theta)$ for which we shall use Assumptions A2.1, A2.2, A2.5 and A2.8.

Theorem 2.4 (Kushner and Clark Theorem for Gradient Search Algorithms)**.** For the recursion (2.15), under Assumptions A2.1, A2.2, A2.5, A2.7, A2.8, $L(\{\theta_n, n \geq 0\})$ is a connected internally chain recurrent set for the ODE (A.1). Further, $L(\{\theta_n, n \geq 0\}) \subset H \stackrel{\triangle}{=} \{\theta \mid \nabla f(\theta) = 0\}$.

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.

There are however many practical situations where it is difficult to verify such sufficient conditions for stability for 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 they 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 $x = (x_1, \ldots, x_d)^T \in \mathcal{R}^d$ to its nearest point in C. Also, let $\mathcal{C}(C)$ denote the space of all continuous functions from C to $\mathcal{R}^d$.

Consider the following d-dimensional stochastic recursion:

$$X_{n+1} = \Gamma(X_n + a(n)(h(X_n) + \xi_n + \beta_n)), \tag{2.21}$$

under the assumptions listed below. Also, consider the following ODE associated with (2.21):

$$\dot{X}(t) = \bar{\Gamma}(h(X(t))). \tag{2.22}$$

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

$$\bar{\Gamma}(v(x)) = \lim_{\eta \to 0} \left(\frac{\Gamma(x + \eta v(x)) - x}{\eta} \right), \tag{2.23}$$

for any continuous $v : C \to \mathcal{R}^d$. The limit in (2.23) exists and is unique since C is a convex set. In case C is not convex, the limit $\bar{\Gamma}(v(x))$ in (2.23) will not be unique in general for all x and so $\bar{\Gamma}(h(X(t)))$ will be a set of points for any $X(t)$, that is not necessarily a singleton, and so instead of the ODE (2.22), one may consider the following differential inclusion:

$$\dot{X}(t) \in \bar{\Gamma}(h(X(t))). \tag{2.24}$$

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 (2.23), note that $\bar{\Gamma}(v(x)) = v(x)$ if $x \in C^o$ (the interior of C). This is because for such an x, one can

find $\eta > 0$ sufficiently small so that $x + \eta v(x) \in C^o$ as well and hence $\Gamma(x + \eta v(x)) = x + \eta v(x)$. On the other hand, if $x \in \partial C$ (the boundary of C) is such that $x + \eta v(x) \notin C$, for any small $\eta > 0$, then $\bar{\Gamma}(v(x))$ is the projection of $v(x)$ to the tangent space of ∂C at x.

Consider now the assumptions listed below.

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

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

$$a(n) > 0 \ \forall n, \ \sum_n a(n) = \infty, \ a(n) \to 0 \text{ as } n \to \infty.$$

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

Let $t(n), n \geq 0$ be a sequence of positive real numbers defined according to $t(0) = 0$ and for $n \geq 1$, $t(n) = \sum_{j=0}^{n-1} a(j)$. By Assumption A2.10, $t(n) \to \infty$ as $n \to \infty$. Let $m(t) = \max\{n \mid t(n) \leq t\}$. Thus, $m(t) \to \infty$ as $t \to \infty$.

A2.12. There exists $T > 0$ such that $\forall \epsilon > 0$,

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

Let $K \subset \mathcal{R}^d$ denote the set of asymptotically stable attractors of (2.22). 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 $n \to \infty$.

2.4 Convergence analysis using stochastic recursive inclusions

We present here some results on stochastic recursive inclusions (SRI) from (Benaïm *et al.*, 2005). Let $J : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}^d\}$ be a Peano map. The DI (A.3) then admits typically nonunique solutions through any initial point $x_0 \in \mathbb{R}^d$.

Let $(\Omega, \mathcal{F}, P)$ be the underlying probability space with $\{\mathcal{F}_n\}$ as the filtration. A stochastic recursive inclusion (SRI) is a recursion of the following form:

$$x_{n+1} - x_n - a(n)M_{n+1} \in a(n)J(x_n), \tag{2.25}$$

where $(M_n, \mathcal{F}_n), n \geq 0$, is a martingale difference sequence.

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:

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

From the above, $W(t(n)) = x_n, \forall n$.

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) $\sup_n \|x_n\| < \infty$ almost surely.

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

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 (2.25).

A2.13. The iterates (2.25) satisfy $\sup_n \|x_n\| < \infty$ 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 (A.3). We thus have the following main result.

Theorem 2.6. The limit set of $W(\cdot)$, the continuous time affine interpolated process obtained from the stochastic recursion (2.25) with $W(0) = z$, given by $L(z) = \bigcap_{t \geq 0} \overline{\{W[t, +\infty)\}}$, is internally chain transitive for the DI (A.3).

2.4.1 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:

$$\theta_{n+1} = \theta_n + a(n) \left(h(\theta_n, X(n)) + M_{n+1} \right), \tag{2.26}$$

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 follows: $t(0) = 0$, $t(n) = \sum_{k=0}^{n-1} a(k)$, $n \geq 1$. 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 $\tilde{h}(\theta, \nu) = \int h(\theta, x)\nu(dx)$. Also, $\tilde{h}_c(\theta, \nu) = \dfrac{\tilde{h}(c\theta, \nu(c\theta))}{c}$.

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

(ii) There exists an attracting set $\mathcal{A}$ associated with the differential inclusion (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 $\bar{B}_1(0) \stackrel{\triangle}{=} \{x \mid ||x|| \leq 1\}$ is a fundamental neighborhood of $\mathcal{A}$.

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

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

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 first a simple example as an application to Theorem 2.7. The temporal difference (TD) learning algorithm in reinforcement learning 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 an 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:

$$\theta_{n+1} = \theta_n + a(n)(A(X(n))\theta_n + b(X(n))), \tag{2.27}$$

where $A(X(n))$ for any $n \geq 0$ is a $d \times d$ matrix and $b(X(n)) \in \mathbb{R}^d$ is a 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 (2.27), 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. 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,

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

Further, note 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 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 which then can be viewed 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.

2.5 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 one such scheme

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 computations in the dual-loop, particularly in the model-free setting, one often resorts to stochastic approximation with two timescales. In these algorithms, the 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, see (Sutton and Barto, 2018), in reinforcement learning (that mimics policy iteration) or the simulation optimization algorithm for optimizing long-run average cost objectives under Markov noise, see (Bhatnagar *et al.*, 2013), are instances of such algorithms.

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

$$\theta_{n+1} = \theta_n + \alpha_n(f(\theta_n, \gamma_n) + N^1_{n+1}), \tag{2.28}$$

$$\gamma_{n+1} = \gamma_n + \beta_n(g(\theta_n, \gamma_n) + N^2_{n+1}), \tag{2.29}$$

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,

$$\sum_n \alpha_n = \sum_n \beta_n = \infty, \ \sum_n \left(\alpha_n^2 + \beta_n^2\right) < \infty, \tag{2.30}$$

$$\lim_{n\to\infty} \frac{\beta_n}{\alpha_n} = 0. \tag{2.31}$$

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 σ-fields $\bar{\mathcal{F}}_n = \sigma(\theta_m, \gamma_m, N^1_m, N^2_m, m \leq n)$, $n \geq 0$, and in addition 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, (2.31) is an important requirement which results in the separation of timescales. As a consequence of (2.31), $\beta_n \to 0$ faster than $\{\alpha_n\}$. Consider now the ODEs

$$\dot{\theta} = f(\theta(t), \gamma(t)), \tag{2.32}$$

$$\dot{\gamma}(t) = 0. \tag{2.33}$$

As a consequence of (2.33), one can alternatively consider the ODE

$$\dot{\theta} = f(\theta(t), \gamma) \tag{2.34}$$

in place of (2.32), where because of (2.33), $\gamma(t) \equiv \gamma$, a constant.

A2.23. The ODE (2.34) 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

$$\dot{\gamma} = g(\mu(\gamma(t)), \gamma(t)). \tag{2.35}$$

A2.24. The ODE (2.35) has a unique globally asymptotically stable equilibrium $\gamma^\star$.

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 (2.32)-(2.33) with $\theta^s(s) = \bar{\theta}(s)$ and $\gamma^s(s) = \bar{\gamma}(s)$. Note that because of (2.33),

$\gamma^s(r) = \bar{\gamma}(s)\ \forall r \geq s$. Now (2.28)-(2.29) can be viewed as 'noisy' Euler discretizations of the ODEs (2.32)-(2.33) when the time discretization corresponds to $\{r_n\}$. This is because (2.29) 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 (2.31) implies that the term multiplying α_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 (2.35), a similar conclusion with regards to iteration (2.29) (and ODE (2.35)) can be drawn following a continuous time trajectory that is obtained with the iterates in (2.29) 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. With probability one, $(\theta_n, \gamma_n) \to (\mu(\gamma^\star), \gamma^\star)$ as $n \to \infty$.

2.6 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 is based on Theorem 8.1 of (Borkar, 2022). Finally, the reader is referred to (Benaïm *et al.*,

2005) for a detailed introduction to differential inclusions and their convergence analysis.

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, stepsizes $\{a_k\}_{k\geq 1}$.
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
Return 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 C.

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

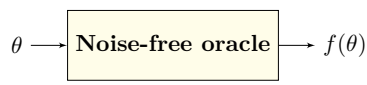
(i) f **is** L**-smooth** if for some constant $L > 0$,

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

(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:

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

Using $2N$ 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:

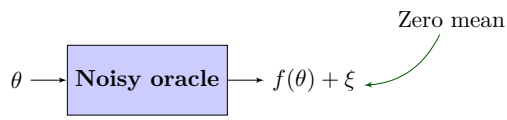
$$\left\|\widehat{\nabla} f(\theta) - \nabla f(\theta)\right\|_\infty = O(\delta^2).$$

Thus, using a two-sided estimate reduced the error to $O(\delta^2)$, while the number of sample measurements went up to $2N$ 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:

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

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

above using Taylor-series expansions as follows:

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

With $2N$ 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 $2N$ 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 $\Delta = (\Delta^1, \ldots, \Delta^d)^\intercal$ is a d-vector of symmetric, ± 1-valued Bernoulli r.v.s, i.e., $\Delta^i = +1$ w.p. $1/2$ and -1 w.p. $1/2$, for $i = 1, \ldots, d$. Consider the following estimate:

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

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

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

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 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 a 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.v.s for random perturbations. However, the trick is more generally valid, and not restricted to this choice for random perturbations. Furthermore, this trick can be used for estimating 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 $U \sim \mathcal{N}(0, I_N)$, where $\mathcal{N}(0, I_N)$ denotes the d-dimensional standard Gaussian vector, and $V = U$, we obtain the smoothed functional scheme proposed by (Katkovnik and Kulchitsky, 1972). 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).
- $U \sim \text{Unif}(\mathbb{S}_N)$, i.e., U is chosen uniformly at random on the surface of an d-dimensional unit sphere, and with $V = dU$, we obtain the random direction stochastic approximation (RDSA) scheme proposed by (Kushner and Clark, 1978), and refined by (Prashanth *et al.*, 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).
- 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 $\mathbb{E}\left[VU^\top\right] = I$, $\mathbb{E}[V] = 0$, and $\mathbb{E}\left[\|V\| \, \|U\|^3\right] < \infty$.

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

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

Proposition 3.1. Assume A3.1–A3.3, 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:

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

where $C_1 = \dfrac{\tilde{B}\mathbb{E}\left[\|V\| \, \|U\|^3\right]}{6}$, and $C_2 = 4\mathbb{E}\left[\|V\|^2\right]\left(\sigma^2 + B^2\right)$.

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:

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

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,

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

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

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

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

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

where $C_2 = 4\mathbb{E}\left[\|V\|^2\right]\left(\sigma^2 + B^2\right)$. 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, and also that the function f is convex and L-smooth, as specified in 3.1. Then the gradient estimate defined in (3.6) satisfies the following bounds for any given θ:

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

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 $\left\|\widehat{\phi}(\theta, \delta)\right\| \leq C_1 \triangleq \frac{L}{2}\mathbb{E}[\|V\| \|U\|^2]$. 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 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 bound on the bias above is comparable to that obtained by the two-point estimate (3.6) was $O(\delta^2)$. However, a closer inspection in the proofs 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 for 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

$$\begin{aligned}\mathbb{E}\left[V\frac{f(\theta+\delta U)}{\delta}\right] &= \mathbb{E}\left[V\frac{f(\theta)}{\delta}\right] + \mathbb{E}\left[VU^{\mathsf{T}}\nabla f(\theta)\right] + \mathbb{E}\left[\frac{\delta}{2}VU^{\mathsf{T}}\nabla^2 f(\theta)U\right] \\ &\quad + \mathbb{E}\left[\frac{\delta^2}{12}V\nabla^3 f(\tilde{\theta}^+)(U\otimes U\otimes U)\right] \\ &= \nabla f(\theta) + \mathbb{E}\left[\frac{\delta^2}{12}V\nabla^3 f(\tilde{\theta}^+)(U\otimes U\otimes U)\right]. \quad (3.13)\end{aligned}$$

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:

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

The random perturbations U, V satisfying $\mathbb{E}VU^{\mathsf{T}} = \mathbb{I}_N$ resulted in a nearly unbiased estimator (see Proposition 3.1). Now, if U, V are chosen in a deterministic fashion, such that$V_m U_m^{\mathsf{T}}$ sums to identity over a loop, 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 $H^\mathsf{T} H = m I_m$, 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 canceling the bias terms. 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., $\Delta \sim d(0, I_d)$, 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., $\Delta \sim d(0, I_d)$.

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 $x \in \mathbb{R}^d$, define

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

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

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

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

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

Notice that

$$\frac{1}{(2\pi)^{\frac{d}{2}}} \int_{-\infty}^{\infty} \langle \nabla f(x)), u \rangle \exp\left(-\frac{\|u\|^2}{2}\right) u \, du = \nabla f(x),$$

since $\frac{1}{(2\pi)^{\frac{d}{2}}} \int_{-\infty}^{\infty} u \, u^\intercal \exp\left(-\frac{\|u\|^2}{2}\right) u \, du = \mathbb{I}_N$. Here $\mathbb{I}_N$ denotes the d-dimensional identity matrix. Using the above fact, we obtain

$\|\nabla f_\delta(x) - \nabla f(x)\|$

$$\leq \frac{1}{(2\pi)^{\frac{d}{2}}\delta}\int_{-\infty}^{\infty} |f(x+\delta u) - f(x) - \delta\langle \nabla f(x)), u\rangle|\ \|u\| \exp\left(-\frac{\|u\|^2}{2}\right) du$$

$$\leq \frac{1}{(2\pi)^{\frac{d}{2}}}\frac{\delta L}{2}\int_{-\infty}^{\infty} \|u\|^3 \exp\left(-\frac{\|u\|^2}{2}\right) du$$

$$\leq \frac{\delta L(d+3)^{\frac{3}{2}}}{2}, \tag{3.18}$$

where the penultimate inequality follows by using $|f(y) - f(x) - \langle \nabla f(x)), y - x\rangle| \leq \frac{1}{2}L\|x-y\|^2$, while the last inequality is a straightforward moment calculation for a multivariate Gaussian.

The claim in (3.16) concerning the bias of the Gaussian smoothing estimator now follows by combining (3.18) with A3.2.

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

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,

$$f(\theta) = \int F(\theta, \psi) P_\psi(d\psi)\,, \tag{3.19}$$

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 seed 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:

$$\widehat{\nabla} f(\theta) = \left(\frac{y^+ - y^-}{2\delta} \right) V. \tag{3.20}$$

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 and L-smooth, as specified in 3.1. Then the gradient estimate defined in (3.20) satisfies the following bounds for any given θ:

$$\left\| \mathbb{E}\left[\widehat{\nabla} f(\theta) \right] - \nabla f(\theta) \right\| \leq C_1 \delta, \text{ and} \tag{3.21}$$

$$\mathbb{E}\left[\left\| \widehat{\nabla} f(\theta) - \mathbb{E}\left[\widehat{\nabla} f(\theta) \right] \right\|^2 \right] \leq C_2 + C_3 \delta^2. \tag{3.22}$$

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

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

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 (3.20), 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

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

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:

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

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 $B_1 = \sup_\theta \|\nabla f(\theta)\|$. □

3.4 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), $f \in \mathcal{C}^1$	$C_1\delta$	$\frac{C_2}{\delta^2}$

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. 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.

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 the latter estimation relied on the idea of simultaneous perturbation. 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^*, when the underlying gradient estimates are biased. We shall also address the associated convergence rate question.

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 $\eta_n = \widehat{\nabla} f(\theta_n) - \mathbb{E}(\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n)$ and $\beta_n = \mathbb{E}(\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n) - \nabla f(\theta_n)$ is the error in the gradient estimate. Recall that the latter is of the

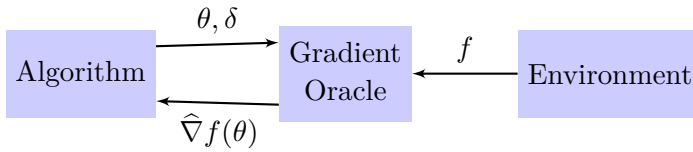


Figure 4.1: The interaction of the algorithms with a stochastic zeroth-order oracle that provides a gradient at the input point θ, with perturbation constant δ.

order $= O(\delta^2)$. Since the sensitivity parameter $\delta > 0$ is held fixed in the algorithms, 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 could 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.

We consider two broad cases for analysis. First, in each iteration of the stochastic gradient algorithm (4.1), the gradient estimates are obtained at input parameter θ_n and perturbation constant δ_n. The sequence $\{\delta_n\}$ is assumed to vanish asymptotically. This setting allows analysis using the ODE approach for stochastic approximation, and is the content of Section 4.1. Second, in each iteration of (4.1), we use the input parameter θ_n in conjunction with a time-invariant perturbation constant δ. The analysis in this setting is more sophisticated as compared

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 4.1	Analysis via ODE limit
Asymptotic convergence	constant	Theorem 4.9	Analysis via DI limit
Non-asymptotic bound	constant	Theorem 5.3	Bound on iterate sequence

to the vanishing δ_n case, and involves the theory of differential inclusions (DIs). Section 4.3 provides the DI analysis.

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

4.1 Asymptotic convergence: An ODE approach

4.1.1 Stochastic gradient algorithm using unbiased gradient information

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

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

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 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. In the next section, we present the simultaneous perturbation trick for estimating gradients from zeroth-order information. Such estimates are not unbiased, but feature a parameter that can reduce the bias at the cost of variance. Before getting to 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 (4.4) 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 (4.4), the function of interest is ∇f.

The algorithm in (4.4) 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., $\mathbb{E}(\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n) = \nabla f(\theta_n)$, where $\mathcal{F}_n = \sigma(\theta_m, m \leq n)$ denotes the underlying sigma-field. Further, there exists $\sigma > 0$ such that

$$\mathbb{E}\left[\left\|\widehat{\nabla} f(\theta_n) - \mathbb{E}\left[\widehat{\nabla} f(\theta_n)\right]\right\|^2\right] \leq \sigma^2 < \infty. \tag{4.5}$$

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 4.1. Assume A4.1–A4.4. Let $\bar{H}$ denote the largest invariant set contained in $\{\theta \mid \nabla f(\theta) = 0\}$. Then, the iterates θ_n, updated according to (4.4), satisfies

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

We now discuss the assumptions made to arrive at the result above. First, the continuity requirement in A4.1 is standard in the analysis of gradient-based algorithms. 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 information through measurements of the objective function at any query point. In the following section, we shall discuss the simultaneous perturbation trick, leading to 'nearly-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, while hard to ensure directly, is common to the analysis of stochastic approximation algorithms. A typical workaround is to employ a projection operator that keeps the iterate bounded, i.e., use the following update rule in place of (4.4):

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

where Π is a projection operator that keeps the iterate bounded within a compact and convex set, say $\Theta \subset \mathbb{R}^d$. For instance, a computationally inexpensive projection onto $\Theta = \prod_{i=1}^{d}[\theta^i_{\min}, \theta^i_{\max}]$ can be realized by setting $\Pi_i(\theta) = \min(\max(\theta^i_{\min}, \theta^i), \theta^i_{\max}),\ i \in \{1 \dots d\}$. If the projected region Θ contains all points where the gradient of the objective f vanishes, then θ_n updated according to (4.6) would converge to such a point. In the complementary case, the iterate θ_n might get stuck at the boundary of Θ. To avoid the latter situation, one could 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) provided the mean field has a globally attracting set. Such a scheme is shown to remain stable in (Yaji and Bhatnagar, 2019) and is convergent 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 — a topic that is not dealt in detail in this book. 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, 2016) and (Ramaswamy and Bhatnagar, 2021) for such conditions in the context of set-valued stochastic approximation.

Convergence analysis

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 4.1 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 again to establish asymptotic convergence.

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

$$\theta_{n+1} = \theta_n + a(n)(-\nabla f(\theta_n) + \beta_n + \eta_n). \tag{4.7}$$

The ODE associated with this scheme would be

$$\dot{\theta} = h(\theta) = -\nabla f(\theta). \tag{4.8}$$

For this ODE, $V(\theta) = f(\theta)$ serves as a Lyapunov function. Further, $\nabla V(\theta)^T h(\theta) \leq 0, \forall \theta$. In this setting, we recall the following result, see Theorem 2 – Corollary 2 of (Lasalle, 1966).

Lemma 4.2. Any trajectory $\theta(\cdot)$ of (4.8) must converge to the largest invariant set that is a subset of $M = \{\theta \mid \nabla f(\theta)^T h(\theta) = 0\}$.

Let $H \stackrel{\triangle}{=} \{\theta \mid \nabla f(\theta) = 0\}$ denote the set of equilibrium points of (4.8). The following immediately holds:

Corollary 4.3. The set M in Lemma 4.2 is the same as H.

Proof. Notice that $M = \{\theta \mid \nabla f(\theta)^T h(\theta) = 0\}$ with $h(\theta) = -\nabla f(\theta)$. Thus, $M = \{\theta \mid - \parallel \nabla f(\theta) \parallel^2 = 0\}$ where $\parallel \cdot \parallel$ denotes the Euclidean norm. The claim follows. □

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 4.2.

Theorem 4.4. Under A2.1–A2.5, $\{\theta_n\}$ given by (4.7) satisfies $\theta_n \rightarrow$

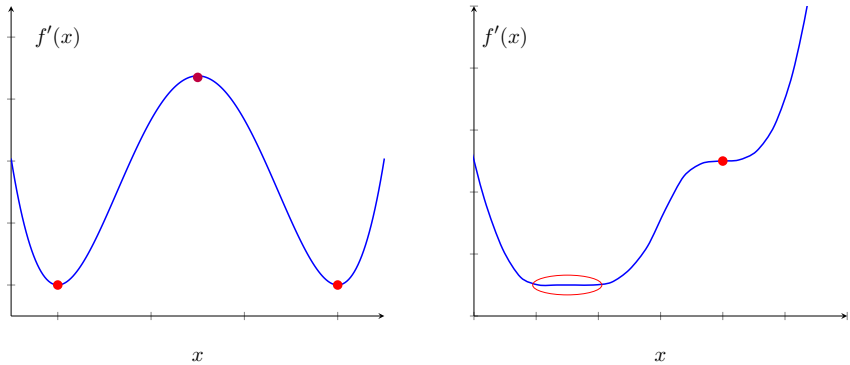


Figure 4.2: Two graphs illustrating the types of convergence for a stochastic gradient (SG) algorithm. In the left graph, an SG algorithm for minimization would converge to one of the two local minima or the local maximum indicated by the filled (red) circles, where which one it reaches depends on the 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.

$\bar{H}$, 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 4.5. Let the set H above comprise of isolated equilibrium points. Then, under conditions of Theorem 4.4, $\{\theta_n\}$ given by (4.7) satisfies $\theta_n \to \theta^*$ for some $\theta^* \in \bar{H}$.

Corollary 4.5 is useful in most practical situations where the equilibrium points of the ODE (4.8) are isolated.

Theorem 4.4 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 4.4 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 in the right graph of Figure 4.2, which contains a flat local minimal region and a saddle point. "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 now.

Since the ODE tracked by the iteration (4.7) 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 (4.7) will a.s. converge to a sample path-dependent stationary point. Under additional assumptions, one can ensure convergence to a local minimum, i.e., avoid local maxima and 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 always add extraneous i.i.d.

zero mean noise, i.e., an SA update iteration of the form

$$\theta_{n+1} = \theta_n - a_n(\widehat{\nabla} f(\theta_n) + \varphi_n), \tag{4.9}$$

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.

Proof of Theorem 4.1

For proving Theorem 4.1, we shall invoke Theorem 4.4.

Proof of Theorem 4.1. The update in (4.4) is equivalent to

$$\theta_{n+1} = \theta_n - a(n)\Big(\nabla f(\theta_n) + \eta_n\Big), \tag{4.10}$$

where $\eta_n = \widehat{\nabla} f(\theta_n) - \mathbb{E}(\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n)$ is a martingale difference term. The equivalent update rule above used the fact that $\mathbb{E}(\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n) = \nabla f(\theta_n)$, which holds by assumption A4.2.

The mean ODE underlying (4.1) is

$$\dot{\theta} = -\nabla f(\theta), \tag{4.11}$$

with limit set $H = \{\theta : \nabla f(\theta)) = 0\}$.

To apply Theorem 4.4, 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)):

$$\mathbb{P}\left(\sup_{m\geq 0} \|W_m\| \geq \epsilon\right) \leq \frac{1}{\epsilon^2} \lim_{m\to\infty} \mathbb{E}\,\|W_m\|^2 . \tag{4.12}$$

Applying the inequality above to $W_n \triangleq \sum_{i=0}^{n-1} a(i)\eta_i$, $n \geq 1$, we obtain

$$P\left(\sup_{m\geq n}\left\|\sum_{i=n}^{m} a(i)\eta_i\right\| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\mathbb{E}\left\|\sum_{i=n}^{\infty} a(i)\eta_i\right\|^2 = \frac{1}{\epsilon^2}\sum_{i=n}^{\infty} a(i)^2\mathbb{E}\left\|\eta_i\right\|^2. \tag{4.13}$$

The last equality above follows by observing that, for $m < n$, $\mathbb{E}(\eta_m\eta_n) = \mathbb{E}(\eta_m\mathbb{E}(\eta_n \mid \mathcal{F}_n)) = 0$.

Now, using the square-summability of the stepsize in A4.3 and (4.5) in A4.2, we have

$$P\left(\sup_{m\geq n}\left\|\sum_{i=n}^{m} a(i)\eta_i\right\| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\sum_{i=n}^{\infty} a(i)^2\mathbb{E}\left\|\eta_i\right\|^2 \leq \frac{\sigma^2}{\epsilon^2}\lim_{n\to\infty}\sum_{i=n}^{\infty} a(i)^2 \to 0 \text{ as } n \to \infty.$$

Thus, θ_n converges a.s. to the set $\bar{H}$ by an application of Theorem 4.4. □

4.1.2 Stochastic gradient algorithm using biased gradient information

Let $\mathcal{F}_n = \sigma(\theta_1, \ldots, \theta_n)$ denote the sigma field underlying the following stochastic gradient algorithm:

$$\theta_{n+1} = \theta_n - a(n)\widehat{\nabla} f(\theta_n), \tag{4.14}$$

where $\widehat{\nabla} f(\theta_n)$ is formed using the unified estimate from the previous chapter, which is recalled below.

$$\widehat{\nabla} f(\theta_n) = \left(\frac{y_n^+ - y_n^-}{2\delta_n}\right) V, \tag{4.15}$$

where $y_n^+ = f(\theta_n + \delta_n U) + \xi_n^+$, and $y_n^- = f(\theta_n - \delta_n U) + \xi_n^-$. The reader is referred to Chapter 3 for a variety of choices for the random perturbations U, V.

For the analysis of this algorithm, we require the following assumptions in addition to A4.4 listed earlier:

A4.5.

$$\forall n \geq 1, \left\| \mathbb{E}\left[\widehat{\nabla} f(\theta_n)\right] - \nabla f(\theta_n) \right\| \leq C_1 \delta_n^2, \text{ and}$$
$$\mathbb{E}\left[\left\| \widehat{\nabla} f(\theta_n) - \mathbb{E}\left[\widehat{\nabla} f(\theta_n)\right] \right\|^2 \right] \leq \frac{C_2}{\delta_n^2},$$

for some constants C_1 and C_2.

A4.6. The noise factors $\xi^{\pm}$ in (4.22) satisfy

$$\mathbb{E}[\xi_n^+ - \xi_n^- | \mathcal{F}_n] = 0, \quad \text{and} \quad \mathbb{E}[(\xi_n^+ - \xi^-)^2 | \mathcal{F}_n] \leq \sigma^2 < \infty, \forall n \geq 1. \tag{4.16}$$

A4.7. The objective function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$\sup_{\theta \in \mathbb{R}^d} \mathbb{E}[f(\theta \pm \delta_n \Delta(n))^2] \leq B < \infty. \tag{4.17}$$

A4.8. The step-sizes $a(n)$ and perturbation constants δ_n are positive, for all n and satisfy

$$a(n), \delta_n \to 0 \text{ as } n \to \infty, \sum_n a(n) = \infty \text{ and } \sum_n \left(\frac{a(n)}{\delta_n}\right)^2 < \infty.$$

Theorem 4.6. Assume A4.5–A4.8, A4.4, and that $f \in \mathcal{C}^3$, the set of all three times continuously differentiable functions. Let $\bar{H}$ denote the largest invariant set contained in $\{\theta \mid \nabla f(\theta) = 0\}$. Then, the iterates θ_n, updated according to (4.14), satisfy

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

Proof. We first rewrite the update rule (4.14) as follows:

$$\theta_{n+1} = \theta_n - a(n)(\nabla f(\theta_n) + \eta_n + \beta_n), \tag{4.18}$$

where $\eta_n = \widehat{\nabla} f(\theta_n) - \mathbb{E}(\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n)$ is a martingale difference term, and $\beta_n = \mathbb{E}(\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n) - \nabla f(\theta_n)$ is the bias in the gradient estimate.

Convergence of (4.14) can be inferred from Theorem 4.4, provided we verify the necessary assumptions, and we do this verification below.

- $f \in \mathcal{C}^3$ implies A2.1.

- From A4.5, we have $\beta_n = O(\delta_n^2)$. In conjunction with A4.8, we have $\beta_n \to 0$, verifying A2.2.

- Applying Doob's martingale inequality $W_n := \sum_{i=0}^{n-1} a(i)\eta_i$, $n \geq 1$, we obtain

$$\mathbb{P}\left(\sup_{m\geq n}\left\|\sum_{i=n}^{m} a(i)\eta_i\right\| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\mathbb{E}\left\|\sum_{i=n}^{\infty} a(i)\eta_i\right\|^2 = \frac{1}{\epsilon^2}\sum_{i=n}^{\infty} a(i)^2\mathbb{E}\left\|\eta_i\right\|^2. \tag{4.19}$$

  The last equality above follows by observing that, for $m < n$, $\mathbb{E}(\eta_m\eta_n) = \mathbb{E}(\eta_m\mathbb{E}(\eta_n \mid \mathcal{F}_n)) = 0$. This verifies A2.4.

  Using A4.7, it can be shown that

$$\mathbb{E}\left\|\eta_n\right\|^2 \leq \frac{C}{\delta_n^2}, \text{ for some } C < \infty. \tag{4.20}$$

  We shall establish this inequality in the next chapter for a more general gradient estimator that includes the SPSA scheme in (1.7). Now, substituting the bound obtained in (**??**) into (4.19), we obtain

$$\lim_{n\to\infty} P\left(\sup_{m\geq n}\left\|\sum_{i=n}^{m} a(i)\eta_i\right\| \geq \epsilon\right) \leq \frac{C}{\epsilon^2}\lim_{n\to\infty}\sum_{i=n}^{\infty}\frac{a(i)^2}{\delta_i^2} = 0.$$

  The equality above follows from A4.8, 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 4.4. □

4.2 Escaping saddle points

So far, we have provided theoretical guarantees that establish convergence to a stationary point of the objective function f. However, this result is 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 such saddle points and converge to a local minimum. A usual trick to achieve this objective is to add extraneous noise so that the algorithm does not converge to an unstable equilibrium, as shown 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:

$$\theta_{n+1} = \theta_n - a(n)\left(\widehat{\nabla} f(\theta_n) + \zeta_n\right), \tag{4.21}$$

where ζ_n is extraneous noise that is injected into the SG algorithm, and is usually sampled from a zero-mean multivariate Gaussian vector with covariance matrix $\sigma^2 I$. Further, we shall use the unified gradient estimate described in Chapter 3. For the sake of readabilitiy, we recall this estimator below.

$$\widehat{\nabla} f(\theta_n) = \left(\frac{y_n^+ - y_n^-}{2\delta_n}\right) V_n, \tag{4.22}$$

where $y^+ = f(\theta_n + \delta_n U_n) + \xi_n^+$, and $y_n^- = f(\theta_n - \delta_n U_n) + \xi_n^-$. For U_n, V_n satisfying certain conditions, it was shown in Chapter 3 that the bias in $\widehat{\nabla} f(\theta)$ is bounded above by $c_1\delta^2$ for some constant c_1.

Under certain conditions on the measurement noise $\{\xi_n^{\pm}\}$, one can avoid injecting noise artificially, and instead directly establish convergence to local minima, owing to the noise in the gradient estimator. The additional assumption on measurement noise is made precise below.

A4.9. $\exists\, c_3, c_4 > 0$ such that $c_3 \leq \mathbb{E}_k|\xi_k^+ - \xi_k^-|$ and $|\xi_k^+ - \xi_k^-| \leq c_4$, $\forall k$.

The assumption above ensures that the noise in function measurements is rich in all directions.

Theorem 4.7 (*Avoidance of traps*)**.** Suppose the conditions of Proposition 3.1 and A4.9 hold. Further, assume $\|V_k\| \leq B_0$ a.s. for all k and $|f(\theta)| \leq B_1, \|\nabla f(\theta)\| \leq B_2$ for all θ. Set $a(k) = \dfrac{c_5}{k^\alpha}$ and $\delta_k = \dfrac{c_6}{k^\phi}$, for some constants $c_5, c_6 > 0$ and $\alpha \in \left(\dfrac{1}{2}, 1\right]$, $\alpha > \phi$. Then, $\{\theta_k\}$ governed by (4.1), converges to the set of local minima a.s.

Proof. We first rewrite the update rule (4.1) as follows:

$$\begin{aligned}\theta_{k+1} &= \theta_k - a(k)\widehat{\nabla} f(\theta_k) \\ &= \theta_k - a(k)\nabla f(\theta_k) - \psi_k, \end{aligned} \tag{4.23}$$

where $\psi_k = a(k)\left[\beta_k + \frac{\xi_k^+ - \xi_k}{\delta_k} V_k\right]$ and $\beta_k = \left(\frac{f(\theta_k + \delta_k U_k) - f(\theta_n - \delta_k U_k)}{2\delta_k}\right) V_k - \nabla f(\theta_k)$. From the analysis in Chapters 3 and 4, we know that $\beta_k = O(\delta_k^2)$.

The convergence of (4.23) to a local minimum can be inferred from Theorem 1 of (Pemantle, 1990) provided that conditions 5–7 of (Pemantle, 1990) are satisfied. These conditions, when applied to (4.23) are as follows:

(i) $\frac{c_5}{k^\alpha} \leq a(k) \leq \frac{c_6}{k^\alpha}$ for some constants $c_5, c_6 > 0$ and $\alpha \in (\frac{1}{2}, 1]$;

(ii) $\mathbb{E}_k\left[(\psi_k \cdot \theta)^+\right] \geq c_7/k^\alpha$ for some $c_7 > 0$ and every unit vector θ. Here $(a \cdot b)$ denotes the dot product between a and b, and $(a)^+ = \max(a, 0)$;

(iii) $\|\psi_k\| \leq c_8/k^\alpha$ for some $c_8 > 0$.

We will now show that (i)-(iii) hold here. It is easy to see that $a(k)$ defined in the theorem statement satisfies condition (i).

We now show that condition (ii) 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

$$\begin{aligned}\mathbb{E}_k[(\psi_k \cdot \vartheta)^+] &= \mathbb{E}_k[(a(k)(\xi_k^+ - \xi_k^-)V_k^i)^+ + (a(k)\beta_k)^+] \\ &\overset{(b)}{\geq} \mathbb{E}_k\left[\frac{a(k)(\xi_k^+ - \xi_k^-)V_k^i + a(k)|(\xi_k^+ - \xi_k)\, V_k^i|}{2}\right] \\ &\overset{(c)}{=} \mathbb{E}_k\left[\frac{a(k)|\xi_k^+ - \xi_k|\,|V_k^i|}{2}\right] \\ &\overset{(d)}{\geq} \frac{c_5 c_3 \min\limits_{i=1\ldots,N} \mathbb{E}|V_k^i|}{2k^\alpha}. \end{aligned}$$

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, A4.9 allows us to infer (d). Thus, condition (ii) holds.

We now turn to verifying condition (iii). Notice that

$$\|\psi_k\| \leq \frac{a(k)}{\delta_k}\|(\xi_k^+ - \xi_k)V_k\| + a(k)|\beta_k| \leq \frac{c_5 B_0}{c_6 k^{\alpha-\phi}} + \frac{2c_5 B_1 B_0}{c_6 k^{\alpha-\phi}} + \frac{c_5 B_2}{c_6 k^{\alpha}},$$

where we used the following facts:
(a) $\|(\xi_k^+ - \xi_k)\| \leq c_4$ from A4.9; (b) $\|V_k\| \leq B_0$ and $|f(\theta)| \leq B_1, \|\nabla f(\theta)\| \leq B_2$ by assumptions in the theorem statement; and (c) $a(k) = \dfrac{c_5}{k^\alpha}$ and $\delta_k = \dfrac{c_6}{k^\phi}$, with $\alpha - \phi > 0$. Thus, condition (iii) holds.

Hence, from Theorem 1 of (Pemantle, 1990), we conclude that (4.23) converges to the set of local minima a.s. □

4.3 Asymptotic convergence: A differential inclusions approach

4.3.1 Assumptions

We make the following assumptions:

A4.10. $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable. Furthermore, $\| \nabla f(\theta) \| \leq \tilde{K}(1 + \| \theta \|)$ for all $\theta \in \mathbb{R}^d$, for some $\tilde{K} > 0$.

A4.11. $\{\eta_n\}$ 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,

$$\mathbb{E}[\| \eta_n \|^2 | \mathcal{F}_n] \leq K_1(1 + \| \theta_n \|^2),$$

for some constant $K_1 > 0$.

A4.12. $a(n) > 0, \forall n$. Further, $\sum_n a(n) = \infty$ and $\sum_n a(n)^2 < \infty$.

A4.13. $\sup_n \|\theta_n\| < \infty$ w.p. 1.

Assumption A4.10 requires that the function f is continuously differentiable and that $\nabla f(\theta)$ satisfies an upper bound that grows linearly with θ. A sufficient condition for the latter to hold is that ∇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 A4.11 is on the noise sequence $\{\eta_n\}$. From the manner in which it is defined, viz., $\eta_n = \widehat{\nabla} f(\theta_n) - \mathbb{E}(\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n)$ 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 A4.12 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 A4.13 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, 2016; 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 project 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.

4.3.2 Proof of Convergence

Let $G(\theta) = \nabla f(\theta) + \overline{B}_\epsilon(0)$.

Lemma 4.8. G is a Marchaud map.

Proof. Note that, 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 A4.10. 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,

$$\begin{aligned} \| y - \nabla f(\theta) \| \leq & \| y - y_n \| + \| y_n - \nabla f(\theta_n) \| \\ & + \| \nabla f(\theta_n) - \nabla f(\theta) \| \end{aligned}$$

$$\leq \quad \epsilon + \delta.$$

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):

$$\dot{\theta}(t) \in -G(\theta(t)). \tag{4.24}$$

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 4.9. The iterates (4.2) converge to a closed connected internally chain transitive and invariant set of the DI (4.24).

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 in the case when $\epsilon = 0$:

$$\dot{\theta}_t = -\nabla f(\theta_t). \tag{4.25}$$

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 $\delta_n \to 0$. In the latter case, the square summability requirement of the step-size sequence $\{a(n)\}$ will need to be considerably tightened. More specifically, the condition $\sum_n a(n)^2 < \infty$ in A4.12 will need to be replaced by the condition $\sum_n \left(\frac{a(n)}{\delta_n}\right)^2 < \infty$ in A4.8. 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 4.10. Given $\delta > 0$, $\exists \epsilon(\delta) > 0$ such that the chain recurrent set of (4.24) is within the δ-open neighborhood of the chain recurrent set of (4.25) for all $\epsilon \leq \epsilon(\delta)$.

It follows as a consequence of Theorem 4.9 and Theorem 4.10 that (4.2) with $\epsilon < \epsilon(\delta)$ (cf. Theorem 4.10) converges almost surely to $N^{\delta}(\mathcal{M})$.

4.3.3 A Set of Stability Conditions

We now present a set of conditions from (Ramaswamy and Bhatnagar, 2016; 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 4.8 that G is a Marchaud map. For each integer $c \geq 1$, let $G_c(\theta) := \{y/c \mid y \in G(c\theta)\}$. Let $G_\infty(\theta) := \overline{co}\,\langle Limsup_{c\to\infty} G_c(\theta)\rangle$, where $Limsup_{x_n \to x} J(x_n) = \{y \in \mathbb{R}^d \mid \liminf_{x_n \to x} d(y, J(x_n)) = 0\}$, see Definition A.7. Given $A \subseteq \mathbb{R}^d$, the convex closure of A, denoted by $\overline{co}\langle A\rangle$, 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:

A4.14. $\dot{\theta}(t) \in -G_\infty(x(t))$ 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 $\sup_{\theta\in\mathcal{A}} \|\theta\| < a$.

A4.15. 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 A4.14 holds. Further, -A4.15 is satisfied when ∇f is Lipschitz continuous.

Theorem 4.11. Under A4.10–A4.12 and A4.14-A4.15, the stochastic update sequence given by (4.2) remains stable, i.e., satisfies $\sup_n \| \theta_n \| < \infty$.

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 A4.14 and A4.15 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.

4.4 Bibliographic remarks

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 4.2 leading to the traps avoidance claim in Theorem 4.7 for a SG algorithm with the unified gradient estimate is an adaptation of the corresponding result in (Mondal *et al.*, 2024).

5

Non-asymptotic analysis of stochastic gradient algorithms

We consider a SG algorithm for solving (1.1), with an update iteration is 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 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). The analysis in the former case servers 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 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, while 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 the RSG 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 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. 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, 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	$\|\|f(\theta) - f(\theta^*)\|\| \leq \epsilon$	$n \geq \frac{L}{m}\log\left(\frac{f(\theta_0) - f(\theta^*)}{\epsilon}\right)$

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, m 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 uniformly at random from $\{\theta_1, \ldots, \theta_m\}$. 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 the all the iterates $\{\theta_i,\ i = 1, \ldots, m\}$, while RSG achieves same effect, except that the averaging happens in expectation. Algorithm 2 presents the pseudocode of RSG algorithm.

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 following the simultaneous perturbation method described in Chapter 3.

Algorithm 2 RSG algorithm

Input: Initial point $\theta_1 \in \mathbb{R}^d$, iteration limit m, stepsizes $\{a_k\}_{k\geq 1}$ and probability mass function $P_R(\cdot)$ supported on $\{1, \ldots, m\}$.
Let R be a random variable with probability mass function P_R.
for $k = 1, \ldots, R$ **do**
Perform the following stochastic gradient descent update:
$\theta_{k+1} = \theta_k - a_k \widehat{\nabla} f(\theta_k)$.
end for
Return θ_R.

5.1 The non-convex case

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 provides 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., $f(\theta_m) - f(\theta^*)$ 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

$$a_k = min\left\{\frac{1}{L}, \frac{c}{\sqrt{m}}\right\}, \qquad \forall k \geq 1, \tag{5.2}$$

for some constant $c > 0$. Then, for any $m \geq 1$, we have

$$\mathbb{E}\left[\|\nabla f(\theta_R)\|^2\right] \leq \frac{2LD_f}{m} + \frac{1}{\sqrt{m}}\left[\frac{2D_f}{c} + L\sigma^2 c\right],$$

where R is uniformly distributed over $\{1, \ldots, m\}$, θ^* is an optimal solution to (1.1). and

$$D_f = f(\theta_1) - f(\theta^*). \tag{5.3}$$

In the proposition below, we prove a general result that holds for any choice of the non-increasing stepsize sequence. Subsequently, we specialize to the case of constant stepsize, to prove Theorem 5.1.

Proposition 5.1. Assume that the objective function f is L-smooth (as defined in 3.1) and assumption A5.1 holds. Suppose that the RSG algorithm is run with a non-increasing stepsize sequence satisfying $0 < a_k \leq 1/L, \forall k \geq 1$ and with the probability mass function

$$P_R(k) := Prob\{R = k\} = \frac{a_k}{\sum_{k=1}^m a_k}, \quad k = 1, \ldots, m, \tag{5.4}$$

then, for any $m \geq 1$, we have

$$\mathbb{E}\left[\|\nabla f(\theta_R)\|^2\right] \leq \frac{1}{\sum_{k=1}^m a_k}\left[\frac{2D_f}{(2 - La_1)} + L\sigma^2 \sum_{k=1}^m \frac{a_k^2}{(2 - La_k)}\right], \tag{5.5}$$

where D_f is as defined in (5.14).

Proof. Since f is L-smooth, we have

$$f(\theta_{k+1}) \leq f(\theta_k) + \langle \nabla f(\theta_k), \theta_{k+1} - \theta_k \rangle + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2$$

$$= f(\theta_k) - a_k \left\langle \nabla f(\theta_k), \widehat{\nabla} f(\theta_k) \right\rangle + \frac{L}{2} a_k^2 \left\| \widehat{\nabla} f(\theta_k) \right\|^2$$

Using $\mathbb{E}_k \left[\widehat{\nabla} f(\theta_k) \right] = \nabla f(\theta_k)$, 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

$$\begin{aligned} \mathbb{E}_k[f(\theta_{k+1})] &\leq f(\theta_k) - a_k \|\nabla f(\theta_k)\|^2 + \frac{L}{2} a_k^2 \left[\|\nabla f(\theta_k)\|^2 + \sigma^2 \right] \\ &= f(\theta_k) - \left(a_k - \frac{L}{2} a_k^2 \right) \|\nabla f(\theta_k)\|^2 + \frac{L}{2} a_k^2 \sigma^2 \end{aligned} \tag{5.6}$$

Re-arranging the terms, we obtain

$$\left(a_k - \frac{L}{2} a_k^2 \right) \|\nabla f(\theta_k)\|^2 \leq f(\theta_k) - \mathbb{E}_k[f(\theta_{k+1})] + \frac{L}{2} a_k^2 \sigma^2$$

$$a_k \|\nabla f(\theta_k)\|^2 \leq \frac{2 \left[f(\theta_k) - \mathbb{E}_k[f(\theta_{k+1})] \right]}{(2 - L a_k)} + \frac{L a_k^2 \sigma^2}{(2 - L a_k)}$$

Now summing up the above inequality from $k = 1$ to m, we obtain

$$\sum_{k=1}^{m} a_k \|\nabla f(\theta_k)\|^2 \leq 2 \sum_{k=1}^{m} \frac{\left[f(\theta_k) - \mathbb{E}_k[f(\theta_{k+1})] \right]}{(2 - L a_k)} + L\sigma^2 \sum_{k=1}^{m} \frac{a_k^2}{(2 - L a_k)}$$

Taking total expectations on both sides of above equation, we obtain

$$\begin{aligned} &\sum_{k=1}^{m} a_k \mathbb{E} \|\nabla f(\theta_k)\|^2 \\ &\leq 2 \sum_{k=1}^{m} \frac{\left[\mathbb{E}[f(\theta_k)] - \mathbb{E}[f(\theta_{k+1})] \right]}{(2 - L a_k)} + L\sigma^2 \sum_{k=1}^{m} \frac{a_k^2}{(2 - L a_k)} \\ &= 2 \left[\frac{f(\theta_1)}{(2 - L a_1)} - \sum_{k=2}^{m} \left(\frac{1}{(2 - L a_{k-1})} - \frac{1}{(2 - L a_k)} \right) \mathbb{E}[f(\theta_k)] - \frac{\mathbb{E}[f(\theta_{m+1})]}{(2 - L a_m)} \right] \\ &\quad + L\sigma^2 \sum_{k=1}^{m} \frac{a_k^2}{(2 - L a_k)} \end{aligned}$$

[1] When $\|\cdot\|$ is defined from an inner product, we have $\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2 \right] = \mathbb{E}\left[\|X\|^2 \right] - \|\mathbb{E}[X]\|^2$.

Notice that since stepsizes $\{a_k\}_{k\geq 1}$ are non-increasing, we have $\left(\frac{1}{(2-La_{k-1})}-\frac{1}{(2-La_k)}\right)\geq 0$ and using the fact that $\mathbb{E}\left[f\left(\theta_k\right)\right]\geq f(\theta^*)$, we obtain

$$\begin{aligned}
&\sum_{k=1}^{m} a_k \mathbb{E}\left\|\nabla f\left(\theta_k\right)\right\|^2 \\
&\leq 2\left[\frac{f(\theta_1)}{(2-La_1)}-f(\theta^*)\sum_{k=2}^{m}\left(\frac{1}{(2-La_{k-1})}-\frac{1}{(2-La_k)}\right)-\frac{f(\theta^*)}{(2-La_m)}\right] \\
&\quad + L\sigma^2\sum_{k=1}^{m}\frac{a_k^2}{(2-La_k)} \\
&= \frac{2\left(f(\theta_1)-f(\theta^*)\right)}{(2-La_1)}+L\sigma^2\sum_{k=1}^{m}\frac{a_k^2}{(2-La_k)}.
\end{aligned}$$

It follows from the definition of P_R in (5.4) that,

$$\mathbb{E}\left[\left\|\nabla f\left(\theta_R\right)\right\|^2\right]=\frac{\sum_{k=1}^{m} a_k \mathbb{E}\left\|\nabla f\left(\theta_k\right)\right\|^2}{\sum_{k=1}^{m} a_k}.$$

Thus, we conclude

$$\mathbb{E}\left[\left\|\nabla f\left(\theta_R\right)\right\|^2\right]\leq\frac{1}{\sum_{k=1}^{m} a_k}\left[\frac{2\left(f(\theta_1)-f(\theta^*)\right)}{(2-La_1)}+L\sigma^2\sum_{k=1}^{m}\frac{a_k^2}{(2-La_k)}\right].$$

□

Proof of Theorem 5.1

Proof. Notice that for constant stepsizes i.e., $a_k=a,\forall k\geq 1$, we have

$$\mathbb{E}\left[\left\|\nabla f\left(\theta_R\right)\right\|^2\right]=\frac{1}{m}\mathbb{E}\left[\left\|\nabla f\left(\theta_R\right)\right\|^2\right]. \tag{5.7}$$

Combining the above fact with (5.5), we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla f\left(\theta_R\right)\right\|^2\right] &\leq \frac{1}{ma}\left[\frac{2D_f}{(2-La)}+L\sigma^2 m\frac{a^2}{(2-La)}\right] \\
&\leq \frac{1}{ma}\left[2D_f+L\sigma^2 ma^2\right] \\
&= \frac{2D_f}{ma}+L\sigma^2 a
\end{aligned}$$

$$\leq \frac{2D_f}{m} max\left\{L, \frac{\sqrt{m}}{c}\right\} + L\sigma^2 \frac{c}{\sqrt{m}}$$
$$\leq \frac{2LD_f}{m} + \frac{2D_f}{c\sqrt{m}} + L\sigma^2 \frac{c}{\sqrt{m}}$$
$$= \frac{2LD_f}{m} + \frac{1}{\sqrt{m}}\left[\frac{2D_f}{c} + L\sigma^2 c\right].$$

The claim follows. □

Now, we analyze the convergence of the RSG algorithm under the condition that the step-size in (**??**) is diminishing. Specifically, we assume that the stepsizes $\{a_k\}_{k\geq 1}$ satisfy the following standard stochastic approximation conditions:

$$\sum_{k=1}^{\infty} a_k = \infty, \quad \sum_{k=1}^{\infty} a_k^2 < \infty. \tag{5.8}$$

Theorem 5.2. (**Unbiased gradients, Diminishing Stepsizes**) Suppose f is L-smooth and A5.1 holds. Suppose that the RSG algorithm is run with the probability mass function as defined in (5.4), and stepsize sequence satisfying (5.8), then, we have

$$\mathbb{E}\left[\|\nabla f(\theta_R)\|^2\right] \xrightarrow{m\to\infty} 0. \tag{5.9}$$

Proof. Recall that from (5.5), we have

$$\mathbb{E}\left[\|\nabla f(\theta_R)\|^2\right] \leq \frac{1}{\sum_{k=1}^{m} a_k}\left[\frac{2D_f}{(2-La_1)} + L\sigma^2 \sum_{k=1}^{m} \frac{a_k^2}{(2-La_k)}\right].$$

The second condition in (5.8) implies that the terms in square bracket on the RHS of above equation converges to a finite limit when m increases. Then, (5.9) follows since the first condition in (5.8) ensures that $\sum_{k=1}^{m} a_k \to \infty$ as $m \to \infty$.

□

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 $\|\nabla f(x)\|_1 \leq B, \forall x \in \mathbb{R}^N$.

A5.3. The gradient estimate $\widehat{\nabla} f(\theta_k)$ satisfies the following inequalities for all $k \geq 1$:

$$\left\| \mathbb{E}_k \left[\widehat{\nabla} f(\theta_k) \right] - \nabla f\left(\theta_k\right) \right\|_\infty \leq c_1 \delta^2, \tag{5.10}$$

and

$$\mathbb{E}_k \left[\left\| \widehat{\nabla} f(\theta_k) \right\|^2 \right] \leq \left\| \mathbb{E}_k \left[\widehat{\nabla} f(\theta_k) \right] \right\|^2 + \frac{c_2}{\delta^2}. \tag{5.11}$$

In the above, $\mathbb{E}_k$ is shorthand for $\mathbb{E}(\cdot \mid \mathcal{F}_k)$, with $\mathcal{F}_k$ denoting the sigma-field $\sigma\left(\theta_i, i < n\right)$.

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 5.3.
Suppose the objective function f is L-smooth (as defined in 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

$$a = min\left\{ \frac{1}{L}, \frac{1}{m^{2/3}} \right\}, \ \delta = \frac{1}{m^{1/6}}, \ \forall k \geq 1. \tag{5.12}$$

Then, choosing θ_R uniformly at random from $\{\theta_1, \ldots, \theta_m\}$, we have

$$\mathbb{E} \left\| \nabla f\left(\theta_R\right) \right\|^2 \leq \frac{2L(f(\theta_1) - f(\theta^*))}{m} + \frac{\mathcal{K}_1}{m^{1/3}}, \tag{5.13}$$

where $\mathcal{K}_1 = 2D_f d^{4/3} + \dfrac{4Bc_1}{d^{5/3}} + \dfrac{Lc_1^2}{d^{11/3} m} + Lc_2 d^{1/3}$, constants c_1, c_2 are defined in A5.3, B is as defined in A5.2,

$$D_f = f(\theta_1) - f(\theta^*), \tag{5.14}$$

and θ^* is a global optima of f.

From the bound in the result above, it is easy to see that an order $\mathcal{O}\left(\dfrac{1}{\epsilon^3}\right)$ iterations of the RSG algorithm are enough to find a point θ_R that satisfies $\mathbb{E} \left\| \nabla f\left(\theta_R\right) \right\|^2 \leq \epsilon$.

Proof. Since f is L-smooth, we have

$$f(\theta_{k+1}) \leq f(\theta_k) + \langle \nabla f(\theta_k), \theta_{k+1} - \theta_k \rangle + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2$$
$$\leq f(\theta_k) - a \left\langle \nabla f(\theta_k), \widehat{\nabla} f(\theta_k) \right\rangle + \frac{L}{2} a^2 \left\| \widehat{\nabla} f(\theta_k) \right\|^2. \quad (5.15)$$

Taking expectations with respect to the sigma field $\mathcal{F}_k$ on both sides of (5.15), and using (5.10) and (5.11) from A5.3, we obtain

$$\mathbb{E}_k [f(\theta_{k+1})]$$
$$\leq \mathbb{E}_k [f(\theta_k)] - a \left\langle \nabla f(\theta_k), \nabla f(\theta_k) + c_1 \delta^2 \mathbf{1}_{d\times 1} \right\rangle$$
$$+ \frac{L}{2} a^2 \left[\left\| \mathbb{E}_k \left[\widehat{\nabla} f(\theta_k) \right] \right\|^2 + \frac{c_2}{\delta^2} \right]$$
$$\leq f(\theta_k) - a \|\nabla f(\theta_k)\|^2 + c_1 \delta^2 a \mathbb{E}_k \|\nabla f(\theta_k)\|_1$$
$$+ \frac{L}{2} a^2 \left[\|\nabla f(\theta_k)\|^2 + 2 c_1 \delta^2 \mathbb{E}_k \|\nabla f(\theta_k)\|_1 + d c_1^2 \delta^4 + \frac{c_2}{\delta^2} \right] \quad (5.16)$$
$$\leq f(\theta_k) - \left(a - \frac{L}{2} a^2 \right) \|\nabla f(\theta_k)\|^2 + c_1 \delta^2 B \left(a + L a^2 \right) \quad (5.17)$$
$$+ \frac{L}{2} a^2 \left[d c_1^2 \delta^4 + \frac{c_2}{\delta^2} \right],$$

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 (5.16). 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

$$a \|\nabla f(\theta_k)\|^2 \leq \frac{2}{(2 - La)} \Big[f(\theta_k) - \mathbb{E}_k f(\theta_{k+1})$$
$$+ c_1 \delta^2 \left(a + L a^2 \right) B \Big] + \frac{L a^2}{(2 - La)} \left[d c_1^2 \delta^4 + \frac{c_2}{\delta^2} \right].$$

Now, summing up the inequality above for $k = 1$ to m, and taking expectations, we obtain

$$\sum_{k=1}^{m} a \mathbb{E}_m \|\nabla f(\theta_k)\|^2$$
$$\leq 2 \sum_{k=1}^{m} \frac{(\mathbb{E}_m f(\theta_k) - \mathbb{E}_m f(\theta_{k+1}))}{(2 - La)} + 2 \sum_{k=1}^{m} c_1 \delta^2 B \left(\frac{a + L a^2}{2 - La} \right)$$

$$
\begin{aligned}
&+ L \sum_{k=1}^{m} \frac{a^2}{(2 - La)} \left[dc_1^2 \delta^4 + \frac{c_2}{\delta^2} \right] \\
&= 2 \left[\frac{f(\theta_1)}{(2 - La_1)} - \frac{\mathbb{E}_m \left[f(\theta_{m+1}) \right]}{(2 - La_m)} \right] \\
&\quad + 2 \sum_{k=1}^{m} c_1 \delta^2 B \left(\frac{a + La^2}{2 - La} \right) + L \sum_{k=1}^{m} \frac{a^2}{(2 - La)} \left[dc_1^2 \delta^4 + \frac{c_2}{\delta^2} \right].
\end{aligned}
$$

Using $\mathbb{E}_m \left[f(\theta_k) \right] \geq f(\theta^*)$, we obtain

$$
\begin{aligned}
\sum_{k=1}^{m} a \mathbb{E}_m \left\| \nabla f(\theta_k) \right\|^2 \leq \frac{2 \left(f(\theta_1) - f(\theta^*) \right)}{(2 - La_1)} + 2 \sum_{k=1}^{m} c_1 \delta^2 B \left(\frac{a + La^2}{2 - La} \right) \\
+ L \sum_{k=1}^{m} \frac{a^2}{(2 - La)} \left[dc_1^2 \delta^4 + \frac{c_2}{\delta^2} \right].
\end{aligned}
$$

Using the fact that θ_R is picked uniformly at random from $\{\theta_1, \ldots, \theta_m\}$, we obtain

$$
\begin{aligned}
\mathbb{E} \left[\left\| \nabla f(\theta_R) \right\|^2 \right] \leq \frac{1}{ma} \left[\frac{2D_f}{(2 - La_1)} + 2B \sum_{k=1}^{m} c_1 \delta^2 \left(\frac{a + La^2}{2 - La} \right) \right. \\
\left. + L \sum_{k=1}^{m} \frac{a^2}{(2 - La)} \left[dc_1^2 \delta^4 + \frac{c_2}{\delta^2} \right] \right]. \qquad (5.18)
\end{aligned}
$$

Next, we simplify the bound obtained above by substituting the step-size and perturbation constant δ values specified in (5.12) as follows:

$$
\begin{aligned}
&\mathbb{E} \left[\left\| \nabla f(\theta_R) \right\|^2 \right] \\
&\leq \frac{1}{ma} \left[2D_f + 4maBc_1 \delta^2 + Lma^2 \left[dc_1^2 \delta^4 + \frac{c_2}{\delta^2} \right] \right] \qquad (5.19) \\
&\leq \frac{2D_f}{m} max \left\{ L, m^{2/3} \right\} + 4B \left(\frac{c_1}{m^{1/3}} \right) + L \left[\frac{dc_1^2}{m^{2/3}} + \frac{d^{5/3} c_2}{m^{-1/3}} \right] \frac{1}{(d^2 m)^{2/3}}. \qquad (5.20)
\end{aligned}
$$

In the above, the inequality (5.19) follows by using the fact that $a \leq 1/L$, while the inequality (5.20) uses the choice of δ in (5.12). The main claim follows follows by rearranging terms in (5.20). □

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 5.4.
Suppose the objective function f is L-smooth (as defined in 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 (5.12). Let θ_R be chosen uniformly at random from $\{\theta_1, \ldots, \theta_m\}$. Then, for any $m \geq 1$, we have

$$\mathbb{E}\left[f\left(\theta_R\right)\right] - f(\theta^*) \leq \frac{LD^2}{m} + \frac{\mathcal{K}_1}{m^{1/3}},$$

where $\mathcal{K}_2 = D^2 + 4\sqrt{d}Dc_1\delta^2 + \dfrac{dc_1^2\delta^4}{m} + c_2$, constants c_1 and c_2 are specified in A5.3, and

$$D = \|\theta_1 - \theta^*\|, \tag{5.21}$$

with θ^* denoting a global optima of f.

Remark 5.1. From the result above, it is apparent that an $\mathcal{O}\left(\dfrac{1}{\epsilon^3}\right)$ number of iterations are 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 5.2. For the special case of noise originating from a common random numbers sequence, it is possible to obtain an improved bound of the order $\mathcal{O}\left(\dfrac{1}{\sqrt{m}}\right)$. This improved is due to 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 5.4. We omit the details.

Proof. 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

$$\begin{aligned}\omega_{k+1}^2 &= \|\theta_{k+1} - \theta^*\|^2 \\ &= \|\theta_k - a\widehat{\nabla} f(\theta_k) - \theta^*\|^2 \\ &= \omega_k^2 - 2a \left\langle \widehat{\nabla} f(\theta_k), \theta_k - \theta^* \right\rangle + a^2 \left\| \widehat{\nabla} f(\theta_k) \right\|^2 . \end{aligned} \tag{5.22}$$

Taking expectations with respect to the sigma field $\mathcal{F}_k$ on both sides of (5.22), and using (5.10), (5.11), we obtain

$$\begin{aligned}\mathbb{E}[\omega_{k+1}^2] &\leq \mathbb{E}[\omega_k^2] - 2a \left\langle \nabla f(\theta_k), \theta_k - \theta^* \right\rangle - 2a\mathbb{E}\left[\langle \Delta_k, \theta_k - \theta^* \rangle\right] \\ &\quad + a^2 \left[\left\| \mathbb{E}_k \left[\widehat{\nabla} f(\theta_k) \right] \right\|^2 + \frac{c_2}{\delta^2} \right] \\ &\leq \mathbb{E}[\omega_k^2] - 2a \left\langle \nabla f\left(\theta_k\right), \theta_k - \theta^* \right\rangle + 2ac_1\delta^2 \|\theta_k - \theta^*\|_1 \\ &\quad + a^2 \left[\|\nabla f\left(\theta_k\right)\|^2 + 2\sqrt{d}c_1\delta^2 \|\nabla f\left(\theta_k\right)\| + dc_1^2\delta^4 + \frac{c_2}{\delta^2} \right], \end{aligned} \tag{5.23}$$

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 (5.23), we obtain

$$\begin{aligned}\mathbb{E}[\omega_{k+1}^2] &\leq \mathbb{E}[\omega_k^2] - 2a \left\langle \nabla f\left(\theta_k\right), \theta_k - \theta^* \right\rangle + 2ac_1\delta^2 \|\theta_k - \theta^*\|_1 \\ &\quad + a^2 \Bigg[L \left\langle \nabla f\left(\theta_k\right), \theta_k - \theta^* \right\rangle + 2\sqrt{d}c_1\delta^2 L \|\theta_k - \theta^*\| \\ &\qquad + dc_1^2\delta^4 + \frac{c_2}{\delta^2} \Bigg] \\ &\leq \mathbb{E}[\omega_k^2] - (2a_k - La^2) \left[f\left(\theta_k\right) - f(\theta^*) \right] \\ &\quad + 2\sqrt{d}\omega_k c_1\delta^2 a + La^2) + a^2 \left[dc_1^2\delta^4 + \frac{c_2}{\delta^2} \right], \end{aligned}$$

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 (5.21), 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

$$\begin{aligned}
&\mathbb{E}\left[f\left(\theta_R\right)\right]-f(\theta^*) \\
&\leq \frac{1}{ma}\left[\frac{D^2}{(2-La)}+2\sqrt{d}D\sum_{k=1}^{m}c_1\delta^2\frac{(a+La^2)}{(2-La)}\right. \\
&\qquad\left.+\sum_{k=1}^{m}\frac{a^2}{(2-La)}\left(dc_1^2\delta^4+\frac{c_2}{\delta^2}\right)\right], \qquad (5.24)
\end{aligned}$$

Using (5.12) in (5.24), we obtain

$$\begin{aligned}
&\mathbb{E}\left[f\left(\theta_R\right)\right] - f(\theta^*) \\
&\leq \frac{1}{ma}\left[\frac{D^2}{(2-La)} + 2\sqrt{d}D\sum_{k=1}^{m} c_1\delta^2\frac{a+La^2)}{(2-La)}\right. \\
&\qquad \left. + \sum_{k=1}^{m}\frac{a^2}{(2-La)}\left(dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right)\right] \\
&\leq \frac{1}{ma}\left[D^2 + 4\sqrt{d}Dmac_1\delta^2 + ma^2\left(dc_1^2\delta^4 + \frac{c_2}{\delta^2}\right)\right], \qquad (5.25)
\end{aligned}$$

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 (5.12) 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 C.

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:

$$\theta_{k+1} = \theta_k - a(k)\widehat{\nabla} f(\theta_k). \qquad (5.26)$$

We first state and prove a result for the case of a constant step size.

Theorem 5.5. Let f be a μ-strongly convex function. Assume A5.1. Then, the SG algorithm governed by (5.26) and with $a(k) = a$ s.t. $0 < aL < 1$, satisfies

$$\mathbb{E}[f(\theta_n) - f(\theta^*)] \leq \frac{aL\sigma^2}{2\mu} + (1 - a\mu)^{n-1}\left(f(\theta_1) - f(\theta^*) - \frac{aL\sigma^2}{2\mu}\right). \tag{5.27}$$

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

$$\mathbb{E}_k[f(\theta_{k+1})] - f(\theta_k) \leq -\frac{1}{2}a\|\nabla f(\theta_k)\|_2^2 + \frac{1}{2}a^2 L\sigma^2. \tag{5.28}$$

Since f is μ-strongly convex, the PL-condition holds, i.e.,

$$f(\theta) - f(\theta^*) \leq \frac{1}{2\mu}\|\nabla f(\theta)\|_2^2, \ \forall \theta.$$

Using PL-condition in (5.28), we obtain

$$\mathbb{E}_k[f(\theta_{k+1})] - f(\theta_k) \leq -\mu a(f(\theta_k) - f(\theta^*)) + \frac{1}{2}a^2 L\sigma^2. \tag{5.29}$$

Subtracting $f(\theta^*)$ on both sides and re-arranging, we obtain

$$\mathbb{E}_k[f(\theta_{k+1}) - f(\theta^*)] \leq (1 - a\mu)[f(\theta_k) - f(\theta^*)] + \frac{1}{2}a^2 L\sigma^2. \tag{5.30}$$

Taking expectations followed by straightforward simplifications, we obtain

$$\begin{aligned}
&\mathbb{E}[f(\theta_{k+1}) - f(\theta^*)] - \frac{aL\sigma^2}{2\mu} \\
&\leq (1 - a\mu)\mathbb{E}[f(\theta_k) - f(\theta^*)] + \frac{a^2 L\sigma^2}{2} - \frac{aL\sigma^2}{2\mu} \\
&= (1 - a\mu)\left(\mathbb{E}[f(\theta_k) - f(\theta^*)] - \frac{aL\sigma^2}{2\mu}\right).
\end{aligned} \tag{5.31}$$

Since $a\mu < \frac{\mu}{L} < 1$, a repeated application of the above inequality leads to the following bound:

$$\mathbb{E}[f(\theta_n) - f(\theta^*)] \leq \frac{aL\sigma^2}{2\mu} + (1 - a\mu)^{n-1} \left(f(\theta_1) - f(\theta^*) - \frac{aL\sigma^2}{2\mu} \right). \tag{5.32}$$

The claim follows. □

Remark 5.3. Taking limits as $k \to \infty$ in (5.27), we obtain

$$\mathbb{E}[f(\theta_k) - f(\theta^*)] \to \frac{aL\sigma^2}{2\mu} \text{ as } k \to \infty. \tag{5.33}$$

The result above implies that a constant stepsize SG 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 stepsize.

Theorem 5.6. Let f be a μ-strongly convex function. Assume A5.1. Then, the SG algorithm governed by (5.26) and with $a(k) = \frac{c}{k+1}$ s.t. $\frac{1}{\mu} < c \leq L$, satisfies

$$\mathbb{E}[f(\theta_n) - f(\theta^*)] \leq \frac{1}{n+1} \max \left\{ \frac{c^2 L\sigma^2}{2(cm-1)}, 2\left(f(\theta_1) - f(\theta^*)\right) \right\}. \tag{5.34}$$

Proof. We prove by induction. The base case holds trivially. Assuming the claim holds for n, we show that it holds for $n+1$.

From (5.6), 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 stepsize $a(k)$ and letting $K = \max\left\{\frac{c^2 L\sigma^2}{2(cm-1)}, 2\left(f(\theta_1) - f(\theta^*)\right)\right\}$, we obtain

$$\begin{aligned}\mathbb{E}[f(\theta_{n+1})] - f(\theta^*)] &\leq \left(1 - \frac{cm}{n+1}\right)\frac{K}{n+1} + \frac{c^2 L\sigma^2}{2(n+1)^2} \\ &= \frac{Kn}{(n+1)^2} - \frac{(cm-1)K}{(n+1)^2}) + \frac{c^2 L\sigma^2}{2(n+1)^2} \\ &\leq \frac{K}{n+2},\end{aligned}$$

where the final inequality used the following fact:

$$-\frac{(cm-1)K}{(n+1)^2}) + \frac{c^2 L\sigma^2}{2(n+1)^2} \leq 0.$$

The inequality above holds by the definition of K, and simple algebra to infer $\frac{Kn}{(n+1)^2} \leq \frac{K}{n+2}$.

The claim follows. □

Remark 5.4. In contrast to the constant stepsize case handled previously, with a diminishing stepsize, we have a bound that vanishes as $n \to \infty$. However, the stepsize choice requires the knowledge of the strong convexity parameter μ, while the constant stepsize case in Theorem 5.5 did not assume such information.

5.3.2 SG with biased gradient information

As before, we consider the update iteration in (5.26). Unlike the previous section, where we assumed unbiased gradient estimates (i.e., the condition A5.1 holds), here the estimate $\widehat{\nabla} f(\theta_n)$ is a biased approximation to the gradient of the objective function f at θ_n.

The biased gradient estimate can be decomposed as follows:

$$\widehat{\nabla} f(\theta_n) = \nabla f(\theta_n) + \beta_n + \eta_n, \text{ where} \tag{5.35}$$

$$\beta_n = E(\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n) - \nabla f(\theta_n),$$
$$\eta_n = \widehat{\nabla} f(\theta_n) - E(\widehat{\nabla} f(\theta_n) \mid \mathcal{F}_n),$$

where $\mathcal{F}_n$ is an increasing sequence of σ-fields generated by $\theta_i, i \leq n$, for all n. In the above, β_n is the bias in the gradient estimate and $\eta(n)$ is a martingale difference.

Using a simultaneous perturbation-based gradient estimate implies $\beta_n = O(\delta_n^2)$, where δ_n is the perturbation constant used in forming the estimate (see Chapter 3 for several examples). While the bias goes down as δ_n^2, the variance of the gradient estimate scales inversely with δ_n^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 (5.26) with inputs from a biased gradient oracle that satisfies the aforementioned assumptions.

Proposition 5.2. Suppose the objective function f is L-smooth (as defined in 3.1), and assumptions A5.2–A5.3 hold. Then, we have

$$\begin{aligned}\mathbb{E}\,\|\theta_{n+1} - \theta^*\|^2 \leq & \underbrace{2\exp(-2\mu\Gamma(n))\,\|\theta_0 - \theta^*\|^2}_{\textbf{initial error}} \\ & \underbrace{+\,2\sum_{k=1}^{n} a_k^2 \exp(-2\mu(\Gamma(n) - \Gamma_k))c_1^2\delta_k^4 +}_{\textbf{bias error}} \\ & \underbrace{\sum_{k=1}^{n} a_k^2 \exp(-2\mu(\Gamma(n) - \Gamma_k))c_2\delta_k^{-2}}_{\textbf{sampling error}}, \end{aligned} \tag{5.36}$$

where $\Gamma(k) := \sum_{i=1}^{k} a_i$.

Proof. Let $z_n = \theta_n - \theta^*$ denote the error at time instant n of the algorithm (5.26). Using $\nabla f(\theta^*) = 0$, we have

$$\left(\int_0^1 \nabla^2 f(\theta^* + \lambda(\theta_n - \theta^*))d\lambda\right) z_n = \nabla f(\theta_n).$$

Using the fact above, we arrive at a recursion for z_n from (5.35). Letting

$J_n := \int_0^1 \nabla^2 f(\theta^* + \lambda(\theta_n - \theta^*)d\lambda$, we have

$$\begin{aligned} z_{n+1} =& (I - a(n)J_n)z_n - a(n)\left(\beta_n + \eta(n)\right) \\ =& \Pi_n z_0 - \sum_{k=1}^{n} a(k)\Pi_n \Pi_k^{-1}(\beta_k + \eta(k)), \end{aligned}$$

where $\Pi_n := \prod_{k=1}^{n} (I - a(k)J_k)$.

By Jensen's inequality, we obtain

$$\begin{aligned} &\left(\mathbb{E}_n \left\|z_{n+1}\right\|\right)^2 \leq \mathbb{E}_n(\langle z_n, z_n\rangle) \\ &= \mathbb{E}_n \left(\left\|\Pi_n z_0\right\|^2 + \left\|\sum_{k=1}^{n} a(k)\Pi_n\Pi_k^{-1}\beta_k\right\|^2 + \left\|\sum_{k=1}^{n} a(k)\Pi_n\Pi_k^{-1}\eta(k)\right\|^2 \right. \qquad (5.37) \\ &\quad - \left\langle \Pi_n z_0, \sum_{k=1}^{n} a(k)\Pi_n\Pi_k^{-1}\beta_k \right\rangle - \left\langle \Pi_n z_0, \sum_{k=1}^{n} a(k)\Pi_n\Pi_k^{-1}\eta(k) \right\rangle \\ &\quad \left. - \left\langle \sum_{k=1}^{n} a(k)\Pi_n\Pi_k^{-1}\beta_k, \sum_{k=1}^{n} a(k)\Pi_n\Pi_k^{-1}\eta(k) \right\rangle \right) \\ &\leq 2\left\|\Pi_n z_0\right\|^2 + 2\sum_{k=1}^{n} a(k)^2 \left\|\Pi_n\Pi_k^{-1}\right\|^2 c_1^2\delta_k^4 \\ &\quad + \sum_{k=1}^{n} a(k)^2 \left\|\Pi_n\Pi_k^{-1}\right\|^2 \mathbb{E}\left\|\eta(k)\right\|^2 \qquad (5.38) \end{aligned}$$

For the last inequality, we have used the following facts: (i) $\eta(k)$ is a martingale difference in order to the last two two cross terms; (ii) $\beta_k \leq c_1\delta_k^2$ from A5.3; and (iii) Cauchy-Schwarz inequality for the first cross term.

Now, we bound each of the square term in (5.38) separately. Since the objective is strongly convex, we have that $\|I - a(n)J_n\| \leq \exp(-\mu a(n))$. Hence,

$$\left\|\Pi_n\Pi_k^{-1}\right\|_2 = \left\|\prod_{j=k+1}^{n} (I - a_j J_j)\right\|_2$$

$$\leq \prod_{j=k+1}^{n} \|(1-a_j\mu)I - a_j(J_j - \mu I)\|_2$$
$$\leq \prod_{j=k+1}^{n} \|(1-a_j\mu)I\|_2 \leq \prod_{j=k+1}^{n} (1-a_j\mu)$$
$$\leq \exp\left(-\mu(\Gamma(n)-\Gamma(k))\right). \tag{5.39}$$

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_n and (5.39) into (5.38). □

By specializing the result in the proposition above, we derive a non-asymptotic bound of the order $O(1/\sqrt{n})$.

Theorem 5.7. (**Biased gradients and strongly convex objective**) Let $a(k) = c/k$ and $\delta_k = \delta_0/k^\delta$. Then,

$$\mathbb{E}\,\|\theta_n - \theta^*\| \leq \frac{\sqrt{2}\,\|\theta_0 - \theta^*\|}{n^{\mu c}} + \frac{\sqrt{2}cc_1\delta_0^2}{\sqrt{2\mu c - 4\delta - 1}} n^{-\frac{1}{2}-2\delta}$$
$$+ \frac{\sqrt{c_2}c}{\delta_0\sqrt{2\mu c + 2\delta - 1}} n^{\delta-\frac{1}{2}}.$$

Remark 5.5. Choosing $\delta = 0$, one can recover the optimal rate of the order $O\left(n^{-1/2}\right)$ for simultaneous perturbation schemes. Contrast this with the $O\left(n^{-1/3}\right)$ obtained for the non-convex and convex cases in the previous sections.

Proof. Bounding a sum by an integral, we obtain

$$\exp(-\mu\Gamma(n)) \leq \exp(-\mu c \ln n) \leq n^{-\mu c}.$$

Plugging $a(k) = c/k$ and $\delta_k = \delta_0/k^\delta$ into the bias error term in (5.36), we obtain

$$\sum_{k=1}^{n} a^2(k)\exp(-2\mu(\gamma(n)-\gamma_k))c_1^2\delta_k^4 \leq \sum_{k=1}^{n} \frac{c^2}{k^2} n^{-2\mu c} k^{2\mu c} c_1^2 \frac{\delta_0^4}{n^{4\delta}}$$
$$\leq c^2 n^{-2\mu c} c_1^2 \delta_0^4 \sum_{k=1}^{n} k^{2\mu c - 4\delta - 2}$$

$$\leq \frac{c^2 c_1^2 \delta_0^4}{(2\mu c - 4\delta - 1)} n^{-1-4\delta}.$$

Along similar lines, the sampling error term in (5.36) can be upper-bounded as follows:

$$\sum_{k=1}^{n} a^2(k) \exp(-2\mu(\Gamma(n) - \Gamma_k)) \frac{c_2}{\delta_k^2} \leq \frac{c^2 c_2}{\delta_0^2 (2\mu c - 4\delta - 1)} n^{-1+2\delta}.$$

□

5.4 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: $\theta \in \mathbb{R}^N$, perturbation constant $\delta > 0$.

Output: a gradient estimate $\widehat{\nabla} f(\theta) \in \mathbb{R}^N$ that satisfies

(a) $\| \mathbb{E}_\xi \left[\widehat{\nabla} f(\theta) \right] - \nabla f(\theta) \| \leq C_1 \delta^2$,

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

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

$$\Delta_m = \mathbb{E} \left[f(\hat{\theta}_m) \right] - \inf_{\theta \in \mathcal{K}} f(\theta), \tag{5.40}$$

where $\mathcal{K} \subset \mathbb{R}^N$ 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

$$\Delta^{\mathcal{A}}_{\mathcal{F},m}(C_1, C_2) = \sup_{f\in\mathcal{F}} \sup_{\gamma\in\Gamma_1(f,C_1,C_2)} \Delta^{\mathcal{A}}_m(f,\gamma)\,, \tag{5.41}$$

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.

Theorem 5.8. Let $m > 0$ be an integer, $p, q > 0$, $C_1, C_2 > 0$, $\mathcal{K} \subset \mathbb{R}^N$ convex, closed, with $[+1, -1]^N \subset \mathcal{K}$. Then, for any algorithm that observes m random elements from a **(O1)** oracle, the minimax error satisfies the following bound:

$$\Delta^*_{\mathcal{F},m}(C_1, C_2) \geq K_1 \sqrt{N}\, C_1^{\frac{2}{3}} C_2^{\frac{1}{3}} m^{-\frac{1}{3}},$$

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 N-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:

$$\widehat{\nabla} f(\theta, \delta) = \overline{\gamma}(\theta, \delta) + \xi \tag{5.42}$$

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[2]. For $v \in \{\pm 1\}$, let

$$f_v(\theta) := \epsilon\,(\theta - v) + 2\epsilon^2 \ln\left(1 + e^{-\frac{\theta - v}{\epsilon}}\right)\,,\ x \in \mathcal{K}\,. \tag{5.43}$$

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.

[2]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}).

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:

$$\overline{\gamma}_+(\theta,\delta) = \begin{cases} f'_+(\theta) + \min(\epsilon, C_1\delta^2)\,, & \text{if } x < 0\,; \\ \min\left\{f'_+(\theta) + \min(\epsilon, C_1\delta^2), f'_-(\theta) - \min(\epsilon, C_1\delta^2)\right\}, & \text{else}\,, \end{cases} \tag{5.44}$$

and

$$\overline{\gamma}_-(\theta,\delta) = \begin{cases} f'_-(\theta) - \min(\epsilon, C_1\delta^2)\,, & \text{if } x > 0\,; \\ \max\left\{f'_-(\theta) - \min(\epsilon, C_1\delta^2), f'_+(\theta) + \min(\epsilon, C_1\delta^2)\right\}, & \text{else}\,. \end{cases} \tag{5.45}$$

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 $x \ge 0$, using that $f'_+(\theta) \le f'_-(\theta)$, we get

$$f'_+(\theta) - \min(\epsilon, C_1\delta^2) \le \overline{\gamma}_+(\theta,\delta) \le f'_+(\theta) + \min(\epsilon, C_1\delta^2),$$

showing $|\overline{\gamma}_+(\theta,\delta) - f'_+(\theta)| \le C_1\delta^2$. 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,

$$f_v(\theta) - f_v(\theta_v^*) \geq \frac{\epsilon}{2} \mathbb{I}\{\theta v < 0\}\,. \tag{5.46}$$

Given the above definitions and (5.46), by Yao's principle, the minimax error (5.41) is lower bounded by

$$\Delta_m^* \geq \inf_{\mathcal{A}} \mathbb{E}[f_V(\hat{X}_m) - \inf_{x \in X} f_V(\theta)] \geq \inf_{\mathcal{A}} \frac{\epsilon}{2} \mathbb{P}(\hat{X}_m V < 0)\,, \tag{5.47}$$

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 (5.42) 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 $\Omega = \mathbb{R}^n \times \{-1, 1\}$, 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

$$
\begin{aligned}
p_{\mathcal{A},\gamma}(g_{1:n}, v) &= \frac{1}{2}\Big(p_{\mathcal{A},\gamma}(g_m \mid g_{1:m-1}) \cdot \ldots \cdot p_{\mathcal{A},\gamma}(g_{m-1} \mid g_{1:m-2}) \cdot \ldots \cdot p_{\mathcal{A},\gamma}(g_1)\Big) \\
&= \frac{1}{2}\Big(p_{\mathcal{N}}(g_m - \overline{\gamma}(\theta_m^{\mathcal{A}}(g_{1:m-1}), \delta_m^{\mathcal{A}}(g_{1:m-1})), c_2(\delta_m^{\mathcal{A}}(g_{1:m-1}))) \cdot \ldots \cdot \\
&\qquad p_{\mathcal{N}}(g_1 - \overline{\gamma}(\theta_1^{\mathcal{A}}, \delta_1^{\mathcal{A}}), c_2(\delta_1^{\mathcal{A}}))\Big),
\end{aligned}
$$

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 (5.47) 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 $V(g_{1:n}, v) = v$.[3] Define $\mathbb{P}_+(\cdot) := \mathbb{P}(\cdot \mid V = 1)$, $\mathbb{P}_-(\cdot) := \mathbb{P}(\cdot \mid V = -1)$. From (5.47), we obtain

$$
\begin{aligned}
\Delta_m^* &\geq \inf_{\mathcal{A}} \frac{\epsilon}{4}\left(\mathbb{P}_+(\hat{X}_m < 0) + \mathbb{P}_-(\hat{X}_m > 0)\right), && (5.48)\\
&\geq \inf_{\mathcal{A}} \frac{\epsilon}{4}\left(1 - \|\mathbb{P}_+ - \mathbb{P}_-\|_{\mathrm{TV}}\right), && (5.49)\\
&\geq \inf_{\mathcal{A}} \frac{\epsilon}{4}\left(1 - \left(\frac{1}{2} D_{\mathrm{kl}}\left(P_+ \| P_-\right)\right)^{\frac{1}{2}}\right), && (5.50)
\end{aligned}
$$

where (5.48) uses the definitions of $\mathbb{P}_+$ and $\mathbb{P}_-$, $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance, (5.49) follows from its definition, while (5.50) follows from Pinsker's inequality. It remains to upper bound $D_{\mathrm{kl}}\left(P_+ \| P_-\right)$.

[3] Here, we are slightly abusing the notation as $\mathbb{P}$ depends on $\mathcal{A}$, but the dependence 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.

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

$$D_{\mathrm{kl}}\left(P_+ \| P_-\right) = \sum_{t=1}^{m} \int_{\mathbb{R}^{t-1}} D_{\mathrm{kl}}\left(P_+^t(\cdot \mid g_{1:t-1}) \| P_-^t(\cdot \mid g_{1:t-1})\right) N P_+^t(g_{1:t-1}). \tag{5.51}$$

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

$$D_{\mathrm{kl}}\left(P_+^t(\cdot \mid g_{1:t-1}) \| P_-^t(\cdot \mid g_{1:t-1})\right) = \frac{(\overline{\gamma}_+(\theta_t^{\mathcal{A}}, \delta_t^{\mathcal{A}}) - \overline{\gamma}_-(\theta_t^{\mathcal{A}}, \delta_t^{\mathcal{A}}))^2}{2c_2(\delta_t^{\mathcal{A}})},$$

as the KL-divergence between normal distributions $\mathsf{N}(\mu_1, \sigma^2)$ and $\mathsf{N}(\mu_2, \sigma^2)$ is equal to $\dfrac{(\mu_1 - \mu_2)^2}{2\sigma^2}$.

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,

$$\begin{aligned}
|\gamma_+(\theta, \delta) - \gamma_-(\theta, \delta)| &= \gamma_-(\theta, \delta) - \gamma_+(\theta, \delta) \\
&< \sup_x \gamma_-(\theta, \delta) - \inf_x \gamma_+(\theta, \delta) \\
&= \lim_{x \to \infty} \gamma_-(\theta, \delta) - \lim_{x \to -\infty} \gamma_+(\theta, \delta) \\
&= \epsilon - \epsilon \wedge C_1\delta^2 - (-\epsilon + \epsilon \wedge C_1\delta^2) \\
&= 2\epsilon - 2\epsilon \wedge C_1\delta^2 \\
&\leq 2(\epsilon - C_1\delta^2)^+,
\end{aligned} \tag{5.52}$$

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),

$$D_{\mathrm{kl}}\left(P_+^t(\cdot \mid g_{1:t-1}) \| P_-^t(\cdot \mid g_{1:t-1})\right) < \frac{2\{(\epsilon - C_1(\delta_t^{\mathcal{A}})^2)^+\}^2 (\delta_t^{\mathcal{A}})^2}{C_2} \tag{5.53}$$

$$\leq \sup_{\delta>0} \frac{2\{(\epsilon - C_1\delta^2)^+\}^2\,\delta^2}{C_2}\,, \qquad (5.54)$$

where inequality (5.53) follows from (5.52). 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

$$\delta_* = \left(\frac{2\epsilon}{6C_1}\right)^{1/2}. \qquad (5.55)$$

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 (5.54) into (5.51) and using this last observation we obtain

$$D_{\mathrm{kl}}\left(P_+\|P_-\right) \leq \frac{2m}{C_2}\left(\epsilon - C_1\delta_*^2\right)^2\delta_*^2\,. \qquad (5.56)$$

Note that the above bound holds uniformly over all algorithms $\mathcal{A}$. Substituting the above bound into (5.50), we obtain

$$\Delta_m^* \geq \frac{\epsilon}{4}\left(1 - \sqrt{m}\frac{(\epsilon - C_1\delta_*^2)\delta_*}{\sqrt{C_2}}\right) = \frac{\epsilon}{4}\left(1 - \sqrt{m}K_1\epsilon^{\frac{3}{2}}\right), \qquad (5.57)$$

where $K_1 = \dfrac{4}{6\sqrt{C_2}}\left(\dfrac{2}{6C_1}\right)^{\frac{1}{2}}$.

By choosing $\epsilon = \left(\dfrac{2}{5\sqrt{m}K_1}\right)^{\frac{2}{3}}$, we see that

$$\Delta_m^* \geq \frac{9}{20}\left(\frac{1}{25}\right)^{1/3} C_1^{1/3}C_2^{1/3}m^{-1/3}\,. \qquad (5.58)$$

Generalization to N dimensions: To prove the N-dimensional result, we introduce a new device which allows us to relate the minimax error of the N-dimensional problem to that of the 1-dimensional problem. The main idea is to use separable N-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 $1 \le i \le N$, 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 $\mathcal{K} = \times_{i=1}^N \mathcal{K}_i$ and for $v = (v_1, \dots, v_d) \in V^N$, let $f_v : \mathcal{K} \to \mathbb{R}$ be defined by

$$f_v(\theta) = \sum_{i=1}^N f_{v_i}^{(i)}(\theta_i), \qquad \theta \in \mathcal{K}\,. \tag{5.59}$$

Without the loss of generality, we assume that $\inf_{\theta_i \in \mathcal{K}_i} f_{v_i}^{(i)}(\theta_i) = 0$, and hence $\inf_{x \in \times_{i=1}^N \mathcal{K}_i} f_v(\theta) = 0$, 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 N-dimensional *separable* oracle γ_v as follows: The oracle is obtained from "composing" the N one-dimensional oracles, $(\gamma_{v_i}^{(i)})_i$. In particular, the ith component of the response of γ_v given the history of queries $(\theta_t, \delta_t, \dots, \theta_1, \delta_1) \in (\mathcal{K} \times [0,1))^t$ is defined as the response of $\gamma_{v_i}^{(i)}$ given the history of queries $(\theta_{t,i}, \delta_t, \dots, \theta_{1,i}, \delta_1) \in (\mathcal{K}_i \times [0,1))^t$. 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 $\Gamma_{\mathrm{sep}}(f_v, c_1, c_2)$ the set of separable oracles for the function f_v defined above. We also define $\mathcal{F}_{\mathrm{sep}} = \{f : f(\theta) = \sum_{i=1}^N f_{v_i}^{(i)}(\theta_i), x \in \mathcal{K}, v_i \in V_i\}$, the set of componentwise separable functions. Note that when $\|\cdot\| = \|\cdot\|_2$ is used in the definition of type-I oracles then $\Gamma_{\mathrm{sep}}(f_v, C_1/\sqrt{N}, C_2/N) \subset \Gamma(f_v, C_1, C_2)$.

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

$$L^{\mathcal{A}}(f, \gamma) = \int f(\theta) F_{\mathcal{A},\gamma}(Nx)\,.$$

Note that this definition applies both in the one and the N-dimensional cases. For $v \in V^N$, we introduce the abbreviation

$$L^{\mathcal{A}}(v) = L^{\mathcal{A}}(f_v, \gamma_v)\,.$$

We also define

$$\tilde{L}_i^{\mathcal{A}}(v) = \int f_{v_i}^{(i)}(\theta_i) F_{\mathcal{A},\gamma_v}(Nx)$$

so that

$$L^{\mathcal{A}}(v) = \sum_{i=1}^{N} \tilde{L}_i^{\mathcal{A}}(v)\,.$$

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^N, 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 N-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 N-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 5.9. Let $(f_v)_{v\in V^N}, f_v \in \mathcal{F}_{\text{sep}}$, $(\gamma_v)_{v\in V^N}, \gamma_v \in \Gamma_{\text{sep}}(f_v, c_1, c_2)$ be separable for some arbitrary functions c_1, c_2, and let $\mathcal{A}$ be any N-dimensional algorithm. Then there exist N one-dimensional algorithms, $\mathcal{A}_i^*$, $1 \leq i \leq N$ (using only one-dimensional oracles), such that

$$\max_{v\in V} L^{\mathcal{A}}(v) \geq \max_{v_1\in V_1} L_1^{\mathcal{A}_1^*}(v_1) + \cdots + \max_{v_d\in V_d} L_d^{\mathcal{A}_d^*}(v_d)\,. \tag{5.60}$$

Now, let

$$\mathcal{F}^{(i)} = \{f_{v_i} \,:\, v_i \in V\}, \qquad i = 1, \ldots, N\,.$$

The next result follows easily from the previous lemma:

Lemma 5.10. Let $\|\cdot\| = \|\cdot\|_2$ in the definition of the type-I oracles. Then, we have that

$$\Delta^*_{\mathcal{F}_{\text{sep}},n}(c_1, c_2) \geq \sum_{i=1}^{N} \Delta^*_{\mathcal{F}^{(i)},n}(c_1/\sqrt{N}, c_2/N)\,.$$

Let $\mathcal{K} \subset \mathbb{R}^N$, such that $\times_i \mathcal{K}_i \subset \mathcal{K}$, $\{\pm 1\} \subset \mathcal{K}_i \subset \mathbb{R}$, $\mathcal{F}_N = \mathcal{F}_{L,0}(\mathcal{K})$, where recall that $L \geq 1/2$. For any $1 \leq i \leq N$, $\theta_i \in \mathcal{K}_i$,

$$f^{(i)}_{v_i}(\theta_i) := \epsilon\,(\theta_i - v_i) + 2\epsilon^2 \ln\left(1 + e^{-\frac{\theta_i - v_i}{\epsilon}}\right)\,. \tag{5.61}$$

i.e., $f^{(i)}_{v_i}$ is like in the one-dimensional lower bound proof (cf. equation 5.43). Note that $f_v \in \mathcal{F}_N$ since f_v is separable, so its Hessian is diagonal and from our earlier calculation we know that $0 \leq \frac{\partial^2}{\partial \theta_i^2} f^{(i)}_{v_i}(\theta_i) \leq 1/2$. Let $\Delta^{(N)*}_m$ denote the minimax error $\Delta^*_{\mathcal{F}_N,n}\left(C_1\delta^2, \frac{C_2}{\delta^2}\right)$ for the N-dimensional family of functions $\mathcal{F}_N$. Let $\mathcal{F}^{(i)} = \{f^{(i)}_{-1}, f^{(i)}_{+1}\}$. As it was noted above, $f_v \in \mathcal{F}_N$ for any $v \in \{\pm 1\}^N$. Hence, by Lemma 5.10,

$$\Delta^{(N)*}_m \geq \sum_{i=1}^{N} \Delta^*_{\mathcal{F}^{(i)},m}\left(\frac{C_1}{\sqrt{N}}\,\delta^2, \frac{C_2}{N}\delta^{-2}\right)\,. \tag{5.62}$$

Plugging the lower bound derived in (5.58) for the one-dimensional setting into the bound in (5.62), we obtain a $\sqrt{N}$-times bigger lower bound for the N-dimensional case. In particular, we obtain

$$\Delta^{(N)*}_m \geq \frac{9}{10}\left(\frac{C_1 C_2}{25}\right)^{1/3} \sqrt{N} m^{-1/3}.$$

□

5.5 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.

5.4 The presentation of the lower bound is based on the results in (Hu *et al.*, 2016).

6

Hessian estimation

Recall that a stochastic Newton algorithm would update as follows:

$$\theta_{n+1} = \theta_n - a_n \left(\overline{H}_n\right)^{-1} \widehat{\nabla} f(\theta_n), \tag{6.1}$$

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 (1.13), 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., $\sum_{k=1}^{n} \widehat{H}_k$, with $\widehat{H}_n$ denoting the Hessian estimate formed using a certain number of function measurements in iteration n of (6.1). Alternatively, one can employ stochastic approximation with a more general stepsize 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 (1.13) 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 $O(N^2)$ 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 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 $\frac{df(\theta)}{d\theta}$ 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:

$$\frac{d^2 f(\theta)}{d\theta^2} \approx \frac{\left(\frac{f(\theta+\delta+\delta) - f(\theta+\delta-\delta)}{2\delta} \right) - \left(\frac{f(\theta-\delta+\delta) - f(\theta-\delta-\delta)}{2\delta} \right)}{2\delta} \tag{6.3}$$

As before, using Taylor series expansions, it can 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 $4N^2$ 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 noisy case as follows: Suppose we have the following function measurements: For any $i, j \in \{1, \ldots, d\}$,

$$y_1 = f(\theta + \delta e_i + \delta e_j) + \xi_{1ij}, y_2 = f(\theta + \delta e_i - \delta e_j) + \xi_{2ij}, \tag{6.5}$$

$$y_3 = f(\theta - \delta e_i + \delta e_j) + \xi_{3ij} \text{ and } y_4 = f(\theta - \delta e_i - \delta e_j) + \xi_{4ij}. \tag{6.6}$$

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

$$\widehat{H}_{ij} = \left(\frac{y_1 - y_2 - y_3 + y_4}{4\delta^2} \right), \forall i, j \tag{6.7}$$

We analyze the bias of the estimator defined above, under the following assumptions:

A6.1. f is four-times differentiable[1] with $\left| \nabla^4_{i_1,i_2,i_3,i_4} f(\theta) \right| < \infty$, for $i_1, i_2, i_3, i_4 = 1, \ldots, N$ and for all $\theta \in \mathbb{R}^d$.

A6.2. $\mathbb{E}\left[\xi_{kij} | \theta \right] = 0$ for $k = 1, \ldots, 4, i, j = 1, \ldots, N$.

The four-times continuously differentiability assumption on f in A6.1 allows Taylor series expansions, while A6.2 ensures the noise factors $\xi_i, i = 1, \ldots, 4$ vanish in the bias analysis. Under A6.1–A6.2, we have

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

The final equality can be arrived at using Taylor series expansions followed by straightforward simplifications.

[1] Here $\nabla^4 f(\theta) = \dfrac{\partial^4 f(\theta)}{\partial \theta^\top \partial \theta^\top \partial \theta^\top \partial \theta^\top}$ 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 $i_1, i_2, i_3, i_4 = 1, \ldots, N$.

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 $\Delta^{-1} = (1/\Delta_1, \ldots, 1/\Delta_N)^T$.

What remains to be specified is the specification of the gradient estimates for input parameters $\theta + \delta\Delta$. For forming this estimate, we use the simultaneous perturbation trick again, i.e.,

$$G(\theta \pm \delta\Delta) = \hat{\Delta}^{-1} \frac{y(\theta \pm \delta\Delta + \delta\hat{\Delta}) - y(\theta \pm \delta\Delta)}{\delta},$$

where $\hat{\Delta}$ are another independent set of perturbations having same distribution as Δ,

$$y(\theta + \delta\Delta + \delta\hat{\Delta}) = f(\theta + \delta\Delta + \delta\hat{\Delta}) + \xi_1,$$
$$y(\theta - \delta\Delta + \delta\hat{\Delta}) = f(\theta - \delta\Delta + \delta\hat{\Delta}) + \xi_2,$$
$$y(\theta + \delta\Delta) = f(\theta + \delta\Delta) + \xi_3, \text{ and } y(\theta - \delta\Delta) = f(\theta - \delta\Delta) + \xi_4.$$

For the bias bound of the Hessian estimator defined in (6.8), we require the following assumption on the noise elements.

A6.3. Given θ, $\{\xi_k, k = 1, \ldots, 4\}$ is independent of Δ. In addition, $\mathbb{E}[\xi_k | \theta] = 0$ for $k = 1, \ldots, 4$.

Lemma 6.1. Assume A6.1 and A6.3. Then, for any $i, j \in \{1, \ldots, N\}$, 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 $\nabla^2_{ij} f(\cdot)$ 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

$$f(\theta\pm\delta\Delta+\delta\hat{\Delta}) = f(\theta\pm\delta\Delta)+\delta\sum_{k=1}^{N}\hat{\Delta}_k\nabla_k f(\theta\pm\delta\Delta) + \frac{1}{2}\delta^2\sum_{k=1}^{N}\sum_{l=1}^{N}\hat{\Delta}_k\nabla^2_{k,l}f(\theta\pm\delta\Delta)\hat{\Delta}_l+O(\delta^3).$$

Using (6.9) and the expansion above, we have

$$E\left[\widehat{H}_{ij}\middle|\theta\right] = E\Bigg[\frac{\nabla_i f(\theta+\delta\Delta)-\nabla_i f(\theta-\delta\Delta)}{2\delta\Delta_j}+\sum_{k\neq i}\frac{\hat{\Delta}_k}{\hat{\Delta}_i}\frac{\nabla_k f(\theta+\delta\Delta)-\nabla_k f(\theta-\delta\Delta)}{2\delta\Delta_j} + \delta\sum_{k=1}^{N}\sum_{l=1}^{N}\frac{\hat{\Delta}_k(\nabla^2_{k,l}f(\theta+\delta\Delta)-\nabla^2_{k,l}f(\theta-\delta\Delta))\hat{\Delta}_l}{4\delta\Delta_j\hat{\Delta}_i}+O(\delta^2)\ \Big|\ \theta\Bigg] \tag{6.10}$$

Expanding $\nabla_i f(\theta\pm\delta\Delta)$ around $\nabla_i f(\theta)$, we obtain

$$\frac{\nabla_i f(\theta+\delta\Delta)-\nabla_i f(\theta-\delta\Delta)}{2\delta\Delta_j} = \nabla^2_{i,j}f(\theta)+\sum_{l\neq j}\frac{\Delta_l}{\Delta_j}\nabla^2_{i,j}f(\theta)+O(\delta^3).$$

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:

$$\delta\sum_{k=1}^{N}\sum_{l=1}^{N}\frac{\hat{\Delta}_k(\nabla^2_{k,l}f(\theta+\delta\Delta)-\nabla^2_{k,l}f(\theta-\delta\Delta))\hat{\Delta}_l}{4\delta\Delta_j\hat{\Delta}_i} = \delta\sum_{k=1}^{N}\sum_{l=1}^{N}\sum_{m=1}^{N}\frac{\hat{\Delta}_k\Delta(m)\nabla^3_{k,l,m}f(\theta)\hat{\Delta}_l}{2\hat{\Delta}_i\Delta_j}+O(\delta^2).$$

In the above, we used the following equality:

$$\frac{\nabla^2_{k,l} f(\theta + \delta\Delta) - \nabla^2_{k,l} f(\theta - \delta\Delta)}{4\delta\Delta_j} = \sum_{m=1}^{N} \frac{\Delta(m)\nabla^3_{k,l,m} f(\theta)}{2\Delta_j} + O(\delta^2)$$

Using the simplified forms for each of the terms on the RHS of (6.10), we have

$$\begin{aligned}
E\left[\widehat{H}_{ij} \middle| \theta\right] &= E\Bigg[\nabla^2_{i,j} f(\theta) + \sum_{l\neq i} \frac{\Delta_l}{\Delta_i} \nabla^2_{i,l} f(\theta) + \sum_{k\neq j} \frac{\hat{\Delta}_k}{\hat{\Delta}_j} \nabla^2_{k,i} f(\theta) \\
&\quad + \sum_{k\neq i} \sum_{l\neq i} \frac{\hat{\Delta}_k}{\hat{\Delta}_i} \frac{\Delta_l}{\Delta_j} \nabla^2_{k,l} f(\theta) + \delta \sum_{k,l,m=1}^{N} \frac{\hat{\Delta}_k \Delta(m) \nabla^3_{k,l,m} f(\theta) \hat{\Delta}_l}{2\hat{\Delta}_i \Delta_j} + O(\delta^2) \mid \theta\Bigg] \\
&= \nabla^2_{i,j} f(\theta) + \sum_{l\neq j} E\left[\frac{\Delta_l}{\Delta_j} \mid \theta\right] \nabla^2_{i,l} f(\theta) + \sum_{k\neq i} E\left[\frac{\hat{\Delta}_k}{\hat{\Delta}_i} \mid \theta\right] \nabla^2_{k,i} f(\theta) \\
&\quad + \sum_{k\neq i} \sum_{l\neq j} E\left[\frac{\hat{\Delta}_k}{\hat{\Delta}_i} \frac{\Delta_l}{\Delta_j} \,\middle|\, \theta\right] \nabla^2_{k,l} f(\theta) \\
&\qquad + \delta \sum_{k=1}^{N} \sum_{l=1}^{N} \sum_{m=1}^{N} E\left[\frac{\hat{\Delta}_k \hat{\Delta}_l \Delta(m)}{2\hat{\Delta}_j \Delta_i} \,\middle|\, \theta\right] \nabla^3_{k,l,m} f(\theta) + O(\delta^2).
\end{aligned}$$

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 requires for forming the Hessian estimate is brought down to three. This scheme was proposed by Bhatnagar and Prashanth, 2015b, and can be motivated by using the following balanced approximation to the second derivative in the case of a scalar parameter:

$$\begin{aligned}
\frac{d^2 f(\theta)}{d\theta^2} &\approx \frac{\left(\frac{f(\theta+\delta) - f(\theta)}{\delta}\right) - \left(\frac{f(\theta) - f(\theta-\delta)}{\delta}\right)}{\delta} \\
&= \left(\frac{f(\theta+\delta) + f(\theta-\delta) - 2f(\theta)}{\delta^2}\right). \qquad (6.11)
\end{aligned}$$

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:

$$\widehat{H}_{ij} = \left(\frac{y^+ + y^- - 2y}{\delta^2 \Delta_i \hat{\Delta}_j}\right), \forall i, j. \tag{6.12}$$

For the noise elements to vanish in the bias analysis of the estimator above, we make the following assumption.

A6.4. Given θ, $\{\xi, \xi^+, \xi^-\}$ is independent of $\Delta, \hat{\Delta}$. In addition, $\mathbb{E}\left[\xi^+ + \xi^- - 2\xi \middle| \theta\right] = 0$.

Lemma 6.2. Assume A6.1 and A6.4. Then, for any $i, j \in \{1, \ldots, N\}$, 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 $i, j \in \{1, \ldots, N\}$, $i \neq j$. 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

$$\begin{aligned}
\frac{\hat{f}(\theta, \Delta, \hat{\Delta})}{\delta^2 \triangle_i \hat{\triangle}_j} &= \frac{(\Delta + \hat{\Delta})^T \nabla^2 f(\theta)(\Delta + \hat{\Delta})}{\triangle_i \hat{\triangle}_j} + O(\delta^2) \\
&= \sum_{l=1}^{N}\sum_{m=1}^{N} \frac{\Delta_l \nabla^2_{lm} f(\theta) \Delta_m}{\Delta_i \hat{\Delta}_j} + 2\sum_{l=1}^{N}\sum_{m=1}^{N} \frac{\Delta_l \nabla^2_{lm} f(\theta) \hat{\Delta}_m}{\Delta_i \hat{\Delta}_j} \\
&\quad + \sum_{l=1}^{N}\sum_{m=1}^{N} \frac{\hat{\Delta}_l \nabla^2_{lm} f(\theta) \hat{\Delta}_m}{\Delta_i \hat{\Delta}_j} + O(\delta^2).
\end{aligned}$$

It is now easy to see that

$$E\left[\sum_{l=1}^{N}\sum_{m=1}^{N} \frac{\Delta_l \nabla^2_{lm} f(\theta) \Delta_m}{\Delta_i \hat{\Delta}_j} \mid \theta\right] = E\left[\sum_{l=1}^{N}\sum_{m=1}^{N} \frac{\hat{\Delta}_l \nabla^2_{lm} f(\theta) \hat{\Delta}_m}{\Delta_i \hat{\Delta}_j} \mid \theta\right] = 0 \text{ a.s.}$$

and $E\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\frac{\Delta_l\nabla^2_{lm}f(\theta)\hat{\Delta}_m}{\Delta_i\hat{\Delta}_j} \mid \theta\right] = \nabla^2_{i,j}f(\theta)$ a.s.

Thus,

$$\mathbb{E}\left[\frac{\hat{f}(\theta,\Delta,\hat{\Delta})}{\delta^2\triangle_i\hat{\triangle}_j} \mid \theta\right] = 2\nabla^2_{i,j}f(\theta) + O(\delta^2).$$

The case when $i = j$, $i, j \in \{1, \ldots, N\}$ follows in a similar manner. The claim follows after observing that

$$\mathbb{E}\left[\widehat{H}_{ij}\middle|\,\theta\right] = \mathbb{E}\left[\frac{\hat{f}(\theta,\Delta,\hat{\Delta})}{\delta^2\triangle_i\hat{\triangle}_j} \mid \theta\right].$$

The equality above holds since the noise elements $\xi^{\pm}, \xi$ satisfy A6.9. □

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.

A6.5. The function $f : \mathbb{R}^d \to \mathbb{R}$ is two-times continuously differentiable with a bounded third derivative.

As with gradient SF, we begin by taking a convolution of the objective function Hessian with a multi-variate Gaussian density functional. Through a double integration by parts argument, 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 $\theta, \Delta' \in \mathcal{R}^d$.

A6.6. 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 1SF Hessian estimator is then the following:

$$\hat{H}(\theta) = \frac{(\Delta\Delta^T - I)}{\delta^2} y^+. \tag{6.14}$$

The reason for having this form for the Hessian estimator will become evident in what follows.

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

$$\nabla_{\Delta'}(\Delta' G_\delta(\Delta')) =$$

$$\begin{bmatrix} \nabla_{\Delta'_1}(\Delta'_1 G_\delta(\Delta')) & \nabla_{\Delta'_2}(\Delta'_1 G_\delta(\Delta')) & \cdots & \nabla_{\Delta'_d}(\Delta'_1 G_\delta(\Delta')) \\ \nabla_{\Delta'_1}(\Delta'_2 G_\delta(\Delta')) & \nabla_{\Delta'_2}(\Delta'_2 G_\delta(\Delta')) & \cdots & \nabla_{\Delta'_d}(\Delta'_2 G_\delta(\Delta')) \\ \cdots & \cdots & \cdots & \cdots \\ \nabla_{\Delta'_1}(\Delta'_d G_\delta(\Delta')) & \nabla_{\Delta'_2}(\Delta'_d G_\delta(\Delta')) & \cdots & \nabla_{\Delta'_d}(\Delta'_d G_\delta(\Delta')) \end{bmatrix}$$

$$= \begin{bmatrix} \left(1 - \frac{\Delta'^2_1}{\delta^2}\right) & -\frac{\Delta'_1 \Delta'_2}{\delta^2} & \cdots & -\frac{\Delta'_1 \Delta'_d}{\delta^2} \\ -\frac{\Delta'_2 \Delta'_1}{\delta^2} & \left(1 - \frac{\Delta'^2_2}{\delta^2}\right) & \cdots & -\frac{\Delta'_2 \Delta'_d}{\delta^2} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{\Delta'_d \Delta'_1}{\delta^2} & -\frac{\Delta'_d \Delta'_2}{\delta^2} & \cdots & \left(1 - \frac{\Delta'^2_d}{\delta^2}\right) \end{bmatrix} G_\delta(\Delta')$$

$$= \left(I - \frac{\Delta' \Delta'^T}{\delta^2}\right) G_\delta(\Delta') \triangleq \check{H}(\Delta') G_\delta(\Delta').$$

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 $\Delta \triangleq \Delta'/\delta$. 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

$$\bar{I}(\Delta) \triangleq (\Delta\Delta^T - I). \tag{6.18}$$

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 A6.5-A6.6, we have that

$$\| E[\hat{H}(\theta)|\theta] - \nabla^2 f(\theta) \| \leq O(\delta) \to 0, \text{ as } \delta \to 0.$$

Proof. From the definition of $\hat{H}(\theta)$,

$$E[\hat{H}(\theta)|\theta)] = \frac{1}{\delta^2} E[\bar{I}(\Delta)(f(\theta + \delta\Delta) + \xi^+)|\theta]$$

$$= D^2_{\delta,1} f(\theta) + \frac{1}{\delta^2} E[(I - \Delta\Delta^T)\xi^+|\theta].$$

The second term on the RHS equals zero in the light of Assumption A6.6. 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

$$\begin{aligned} D^2_{\delta,1} f(\theta) &= E\Big[\frac{1}{\delta^2}\bar{I}(\Delta)(f(\theta) + \delta\Delta^T \nabla f(\theta) \\ &\qquad + \frac{\delta^2}{2}\Delta^T \nabla^2 f(\theta)\Delta + o(\delta^2) \mid \theta\Big] \\ &= \frac{1}{\delta^2} E[\bar{I}(\Delta) f(\theta) \mid \theta] + \frac{1}{\delta} E[\bar{I}(\Delta)\Delta^T \nabla f(\theta) \mid \theta] \\ &\qquad + \frac{1}{2} E[\bar{I}(\Delta)\Delta^T \nabla^2 f(\theta)\Delta \mid \theta] + O(\delta). \end{aligned} \tag{6.19}$$

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

$$E[\bar{I}(\Delta)\Delta^T \nabla f(\theta) \mid \theta] =$$
$$\mathbb{E}\left[\begin{array}{cccc} (\Delta_1^2 - 1)\Delta^T \nabla f(\theta) & \Delta_1\Delta_2\Delta^T \nabla f(\theta) & \cdots & \Delta_1\Delta_N\Delta^T \nabla f(\theta) \\ \Delta_2\Delta_1\Delta^T \nabla f(\theta) & (\Delta_2^2 - 1)\Delta^T \nabla f(\theta) & \cdots & \Delta_2\Delta_N\Delta^T \nabla f(\theta) \\ \cdots & \cdots & \cdots & \cdots \\ \Delta_N\Delta_1\Delta^T \nabla f(\theta) & \Delta_N\Delta_2\Delta^T \nabla f(\theta) & \cdots & (\Delta_N^2 - 1)\Delta^T \nabla f(\theta) \end{array} \mid \theta\right]. \tag{6.20}$$

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

$$\frac{1}{2}\mathbb{E}[\bar{H}(\Delta)\Delta^T\nabla^2 f(\theta)\Delta \mid \theta] =$$

$$\frac{1}{2}\mathbb{E}\begin{bmatrix} (\Delta_1^2-1)\sum_{i,j=1}^N \nabla_{ij}f(\theta)\Delta_i\Delta_j & \cdots & \Delta_1\Delta_N\sum_{i,j=1}^N \nabla_{ij}f(\theta)\Delta_i\Delta_j \\ \Delta_2\Delta_1\sum_{i,j=1}^N \nabla_{ij}f(\theta)\Delta_i\Delta_j & \cdots & \Delta_2\Delta_N\sum_{i,j=1}^N \nabla_{ij}f(\theta)\Delta_i\Delta_j \\ \cdots & \cdots & \cdots \\ \Delta_N\Delta_1\sum_{i,j=1}^N \nabla_{ij}f(\theta)\Delta_i\Delta_j & \cdots & (\Delta_N^2-1)\sum_{i,j=1}^N \nabla_{ij}f(\theta)\Delta_i\Delta_j \end{bmatrix} \mid \theta \Big]. \quad (6.21)$$

Consider now the term corresponding to the first row and first column above. Note that

$$\begin{aligned} &\mathbb{E}[(\Delta_1^2-1)\sum_{i,j=1}^N \nabla_{ij}f(\theta)\Delta_i\Delta_j \mid \theta] \\ &= \mathbb{E}[\Delta_1^2\sum_{i,j=1}^N \nabla_{ij}f(\theta)\Delta_i\Delta_j \mid \theta] - \mathbb{E}[\sum_{i,j=1}^N \nabla_{ij}f(\theta)\Delta_i\Delta_j \mid \theta]. \end{aligned} \quad (6.22)$$

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 $-\sum_{i=1}^N \nabla_{ii}f(\theta)$. Adding the above two terms, one obtains

$$\mathbb{E}[(\Delta_1^2-1)\sum_{i,j=1}^N \nabla_{ij}f(\theta)\Delta_i\Delta_j \mid \theta] = 2\nabla_{11}f(\theta).$$

Consider now the term in the first row and second column of the matrix

in (6.21). Note that

$$\begin{aligned}
&\mathbb{E}[\Delta_1\Delta_2 \sum_{i,j=1}^{N} \nabla_{ij} f(\theta)\Delta_i\Delta_j \mid \theta] \\
&= 2\mathbb{E}[\Delta_1^2\Delta_2^2\nabla_{12} f(\theta) \mid \theta] + \mathbb{E}[\sum_{(i,j)\notin\{(1,2),(2,1)\}} \Delta_1\Delta_2\Delta_i\Delta_j\nabla_{ij} f(\theta) \mid \theta] \\
&= 2\nabla_{12} f(\theta).
\end{aligned}$$

Proceeding in a similar manner, it is easy to verify that the (i,j)th term $(i,j \in \{1,\ldots,N\})$ in the matrix in (6.21) equals $2\nabla_{ij} f(\theta)$. Substituting the above back in (6.21), one obtains

$$\frac{1}{2}\mathbb{E}[\bar{I}(\Delta)\Delta^T\nabla^2 f(\theta)\Delta] = \nabla^2 f(\theta).$$

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

$$D^2_{\delta,2} f(\theta) = E\left[\frac{1}{2\delta^2}\bar{I}(\Delta)(f(\theta+\delta\Delta) + f(\theta-\delta\Delta)) \mid \theta\right].$$

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:

$$\hat{H}(\theta) = \frac{(\Delta\Delta^T - I)}{2\delta^2}(y^+ + y^-). \tag{6.23}$$

A6.7. The function $f : \mathbb{R}^d \to \mathbb{R}$ is three-times continuously differentiable with a bounded fourth derivative.

A6.8. 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 A6.7-A6.8, we have that

$$\| E[\hat{H}(\theta)|\theta] - \nabla^2 f(\theta) \| \leq O(\delta^2) \to 0 \text{ as } \delta \to 0.$$

Proof. From (6.23), note that

$$E[\hat{H}(\theta)|\theta)] = \frac{1}{2\delta^2} E[\bar{I}(\Delta)((f(\theta + \delta\Delta) + \xi^+) + (f(\theta - \delta\Delta) + \xi^-)|\theta]$$

$$= D^2_{\delta,2} f(\theta) + \frac{1}{2\delta^2} E[(I - \Delta\Delta^T)(\xi^+ + \xi^-)|\theta].$$

The second term on the RHS equals zero in the light of Assumption A6.8.

We now consider the first term on the RHS aboive. Using Taylor series expansions of $f(\theta + \delta\Delta)$ and $f(\theta - \delta\Delta)$ around θ, one obtains

$$f(\theta+\delta\Delta) = f(\theta)+\delta\Delta^T\nabla f(\theta)+\frac{\delta^2}{2}\Delta^T\nabla^2 f(\theta)\Delta+\frac{\delta^3}{6}\nabla^3 f(\theta)(\Delta\otimes\Delta\otimes\Delta)+O(\delta^4)$$

$$f(\theta-\delta\Delta) = f(\theta)-\delta\Delta^T\nabla f(\theta)+\frac{\delta^2}{2}\Delta^T\nabla^2 f(\theta)\Delta-\frac{\delta^3}{6}\nabla^3 f(\theta)(\Delta\otimes\Delta\otimes\Delta)+O(\delta^4).$$

From the foregoing, one obtains

$$D^2_{\delta,2} f(\theta) = E\left[\frac{1}{2\delta^2}\bar{I}(\Delta)\left(2f(\theta) + \delta^2\Delta^T\nabla^2 f(\theta)\Delta + O(\delta^4)\right) \mid \theta\right].$$

It has been shown in the proof of Proposition 6.2 that $\mathbb{E}[\bar{I}(\Delta)f(\theta) \mid \theta] = 0$ and $\frac{1}{2}\mathbb{E}[\bar{I}(\Delta)\Delta^T\nabla^2 f(\theta)\Delta \mid \theta] = \nabla^2 J(\theta)$, 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, $i = 1, \ldots, N$, 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 $i = 1, \ldots, N$, $\mathbb{E}\Delta_i = 0$, $\mathbb{E}(\Delta_i)^2 = 1 + \epsilon$ 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 an Hessian estimate $\widehat{H}$ that is not too far from the true Hessian $\nabla^2 f(\theta)$. Suppose we use the three measurements, together with a matrix M (to be specified later) to form $\widehat{H}$ as follows:

$$\widehat{H} = M\left(\frac{y^+ + y^- - 2y}{\delta^2}\right) \tag{6.25}$$

$$= M\left[\left(\frac{f(\theta+\delta\Delta) + f(\theta-\delta\Delta) - 2f(\theta)}{\delta^2}\right) + \left(\frac{\xi^+ + \xi^- - 2\xi}{\delta^2}\right)\right]$$

$$= M\left(\Delta^\intercal \nabla^2 f(\theta)\Delta + O(\delta^2) + \left(\frac{\xi^+ + \xi^- - 2\xi}{\delta^2}\right)\right). \tag{6.26}$$

Taking expectations on both sides above, 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:

$$M = \begin{bmatrix} \frac{1}{\kappa}\left((\Delta^1)^2 - (1+\epsilon)\right) & \cdots & \frac{1}{2(1+\epsilon)^2}\Delta^1\Delta^d \\ \frac{1}{2(1+\epsilon)^2}\Delta^2\Delta^1 & \cdots & \frac{1}{2(1+\epsilon)^2}\Delta^2\Delta^d \\ \cdots & \cdots & \cdots \\ \frac{1}{2(1+\epsilon)^2}\Delta^d\Delta^1 & \cdots & \frac{1}{\kappa}\left((\Delta^d)^2 - (1+\epsilon)\right) \end{bmatrix}, \tag{6.27}$$

where $\kappa = \tau\left(1 - \dfrac{(1+\epsilon)^2}{\tau}\right)$ and $\tau = E(\Delta^i)^4 = \dfrac{(1+\epsilon)(1+(1+\epsilon)^3)}{(2+\epsilon)}$, for any $i = 1, \ldots, N$.

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).

A6.9. Given θ, $\{\xi, \xi^+, \xi^-\}$ is independent of Δ. In addition, $\mathbb{E}\left[\xi^+ + \xi^- - 2\xi \middle| \theta\right] = 0$.

Lemma 6.3. (**Bias in Hessian estimate**) Assume A6.1 and A6.9. Then, $\widehat{H}$ defined according to (6.27) satisfies the following bound for any $i, j = 1, \ldots, N$,

$$\left|\mathbb{E}\left[\widehat{H}(i,j)\middle|\theta\right] - \nabla^2_{ij} f(\theta)\right| = O(\delta^2). \tag{6.28}$$

From the above lemma, it is evident that the bias in the Hessian estimate above is of the same order as that of the estimators in the previous sections.

Proof. By a Taylor's series expansion, we obtain

$$\begin{aligned} f(\theta \pm \delta\Delta) &= f(\theta) \pm \delta\Delta^\mathsf{T}\nabla f(\theta) + \frac{\delta^2}{2}\Delta^\mathsf{T}\nabla^2 f(\theta)\Delta \\ &\pm \frac{\delta^3}{6}\nabla^3 f(\theta)(\Delta\otimes\Delta\otimes\Delta) + \frac{\delta^4}{24}\nabla^4 f(\tilde{\theta}^+)(\Delta\otimes\Delta\otimes\Delta\otimes\Delta). \end{aligned}$$

Hence,

$$\begin{aligned} &\frac{f(\theta+\delta\Delta) + f(\theta-\delta\Delta) - 2f(\theta)}{\delta^2} \\ =&\Delta^\mathsf{T}\nabla^2 f(\theta)\Delta + O(\delta^2) \\ =&\sum_{i=1}^{d}\sum_{j=1}^{d}\Delta^i\Delta^j\nabla^2_{ij}f(\theta) + O(\delta^2) \\ =&\sum_{i=1}^{d}(\Delta^i)^2\nabla^2_{ii}f(\theta) + 2\sum_{i=1}^{d-1}\sum_{j=i+1}^{d}\Delta^i\Delta^j\nabla^2_{ij}f(\theta) + O(\delta^2). \end{aligned}$$

Now, taking the conditional expectation of the Hessian estimate $\widehat{H}$ and observing that $\mathbb{E}[\xi^+ + \xi^- - 2\xi \mid \theta] = 0$ by A6.9, we obtain the following:

$$\mathbb{E}[\widehat{H} \mid \theta] = \mathbb{E}\left[M\left(\sum_{i=1}^{d-1}(\Delta^i)^2\nabla^2_{ii}f(\theta)\right.\right.$$

$$+2\sum_{i=1}^{d}\sum_{j=i+1}^{d}\Delta^i\Delta^j\nabla^2_{ij}f(\theta)+O(\delta^2)\Bigg)\Bigg|\theta\Bigg]. \tag{6.29}$$

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 $i = 1, \ldots, N$. Consider the lth diagonal term inside the conditional expectation in (6.29):

$$\begin{aligned}
&\frac{1}{\tau(1-\frac{(1+\epsilon)^2}{\tau})}\mathbb{E}\left(\left((\Delta_l)^2-(1+\epsilon)\right)\left(\sum_{i=1}^{N}(\Delta_i)^2\nabla^2_{ii}f(\theta)\right.\right.\\
&\qquad\left.\left.+2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\Delta_i\Delta_j\nabla^2_{ij}f(\theta)\right)\Bigg|\theta\right)\\
=&\frac{1}{\tau(1-\frac{(1+\epsilon)^2}{\tau})}\mathbb{E}\left((\Delta_l)^2\sum_{i=1}^{N}(\Delta_i)^2\nabla^2_{ii}f(\theta)\Big|\theta\right)\\
&-\frac{(1+\epsilon)}{\tau(1-\frac{(1+\epsilon)^2}{\tau})}\mathbb{E}\left(\sum_{i=1}^{N}(\Delta_i)^2\nabla^2_{ii}f(\theta)\Big|\theta\right)
\end{aligned} \tag{6.30}$$

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 $\mathbb{E}\left((d_n^l)^2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\Delta_i\Delta_j\nabla^2_{ij}f(\theta)\Big|\theta\right) = 0$ and $\mathbb{E}\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\Delta_i\Delta_j\nabla^2_{ij}f(\theta)\Big|\theta\right) = 0$. 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) be simplified as follows:

$$\frac{1}{\tau(1-\frac{(1+\epsilon)^2}{\tau})}\mathbb{E}\left((\Delta_l)^2\sum_{i=1}^{N}(\Delta_i)^2\nabla^2_{ii}f(\theta)\Big|\theta\right)$$

$$
\begin{aligned}
&= \frac{1}{\tau(1-\frac{(1+\epsilon)^2}{\tau})} \mathbb{E}\left((\Delta_l)^4 \nabla^2_{ll} f(\theta) + \sum_{i=1, i\neq l}^{N} (\Delta_l)^2 (\Delta_i)^2 \nabla^2_{ii} f(\theta) \right) \\
&= \frac{1}{(1-\frac{(1+\epsilon)^2}{\tau})} \left(\nabla^2_{ll} f(\theta) + \frac{(1+\epsilon)^2}{\tau} \sum_{i=1, i\neq l}^{N} \nabla^2_{ii} f(\theta) \right). \qquad (6.31)
\end{aligned}
$$

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:

$$
\begin{aligned}
&\frac{(1+\epsilon)}{\tau(1-\frac{(1+\epsilon)^2}{\tau})} \mathbb{E}\left(\sum_{i=1}^{N} (\Delta_i)^2 \nabla^2_{ii} f(\theta) \middle| \theta \right) \\
&= \frac{(1+\epsilon)}{\tau(1-\frac{(1+\epsilon)^2}{\tau})} \sum_{i=1}^{N} \mathbb{E}\left[(\Delta_i)^2 \right] \nabla^2_{ii} f(\theta) \\
&= \frac{(1+\epsilon)^2}{\tau(1-\frac{(1+\epsilon)^2}{\tau})} \sum_{i=1}^{N} \nabla^2_{ii} f(\theta). \qquad (6.32)
\end{aligned}
$$

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

$$
\begin{aligned}
&\frac{1}{2(1+\epsilon)^2} \mathbb{E}\left[\Delta_k \Delta_l \left(\sum_{i=1}^{N} (\Delta_i)^2 \nabla^2_{ii} f(\theta) + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \Delta_i \Delta_j \nabla^2_{ij} f(\theta) \right) \middle| \theta \right] \\
&= \frac{1}{2(1+\epsilon)^2} \sum_{i=1}^{N} \mathbb{E}\left(\Delta_k \Delta_l (\Delta_i)^2 \right) \nabla^2_{ii} f(\theta) \\
&\quad + \frac{1}{(1+\epsilon)^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbb{E}\left(\Delta_k \Delta_l \Delta_i \Delta_j \right) \nabla^2_{ij} f(\theta) \qquad (6.33) \\
&= \nabla^2_{kl} f(\theta).
\end{aligned}
$$

Note that the first term on the RHS of (6.33) equals zero since $k \neq l$. The claim follows. □

6.6 Summary

Property → Hessian estimate ↓	# measurements	Bias
FDSA (6.7)	$4N^2$	$O(\delta^2)$
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	$O(\delta^2)$

6.7 Bibliographic remarks

In (Fabian, 1971), the author analyzes a finite differences Hessian estimation scheme with $O(N^2)$ 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, 2015b) 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, 2022). 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, 2022, 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.

7

Asymptotic analysis of stochastic Newton algorithms

To be updated.

8

Applications to reinforcement learning

8.1 REINFORCE with an SPSA Gradient Estimate

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 $X_n, n \geq 0$ take values in a set S called the state-space. Let $A(s)$ denote the set of 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 is say s and a feasible action a is chosen, the next state seen is s' with a probability $p(s'|s,a) \stackrel{\triangle}{=} P(S_{n+1} = s' \mid S_n = s, A_n = a)$, $\forall n$. 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 n. Many times, instead of calling $\pi = \{\mu, \mu, \mu, \ldots\}$ a stationary policy, we simply refer to the function μ 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 state. Thus, $S = \{1, 2, \ldots, p, t\}$ denotes the state space here.

In this chapter, we are concerned with the stochastic shortest path problem, see Bertsekas, 2012, where under any policy there is a positive probability of hitting the goal or 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.

Under a given policy π, define

$$V_\pi(s) = E_\pi \left[\sum_{k=0}^{T} g(s_k, \mu_k(s_k), s_{k+1}) \mid s_0 = s \right],$$

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 $V^*(i), i \in S$ where

$$V^*(i) = \min_{\pi \in \Pi} V_\pi(i) = V_{\pi^*}(i), \ i \in S.$$

Here π^* denotes the optimal policy, i.e., the one that minimizes $V_\pi(i)$ 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.

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 i, 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

$$V^*(s) = \min_{a \in A(s)} \sum_{j=1}^{p} p(j \mid s, a)(g(s, a, j) + V^*(j)), \tag{8.1}$$

$s = 1, \ldots, p$. 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 $\phi_i : S \to P(A)$. In other words, given a state s, a randomized policy would provide a distribution $\phi_i(s) = (\phi_i(s, a), a \in A(s))$ for the action to be chosen in the ith stage. 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 case 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, $G_k = \sum_{j=k}^{T-1} g_k$ where $g_k \equiv g(s_k, a_k, s_{k+1})$. Note that if all actions

are chosen according to a policy ϕ, then the value function (under ϕ) would be

$$V_\phi(s) = E_\phi[G_k \mid S_k = s]. \tag{8.2}$$

We consider here a class of stationary randomized policies that are parameterized by a parameter $\theta = (\theta_1, \ldots, \theta_d)^T \in C \subset \mathbb{R}^d$ 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 discussed previously, 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 $x = (x_1, \ldots, x_d)^T \in \mathcal{R}^d$ 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 according to $\Gamma(x) = (\Gamma_1(x_1), \ldots, \Gamma_N(x_N))^T$, where the individual operators $\Gamma_i : \mathcal{R} \to \mathcal{R}$ are specified by $\Gamma_i(x_i) = \min(a_{i,\max}, \max(a_{i,\min}, x))$, $i = 1, \ldots, d$. Also, let $\mathcal{C}(C)$ denote the space of all continuous functions from C to $\mathcal{R}^d$.

Let $\theta(n)$ and $\Gamma(\theta(n) + \delta\Delta(n)), n \geq 0$ be two parameter sequences where $\theta(n) = (\theta_1(n), \ldots, \theta_d(n))^T \in \mathcal{R}^d$, $\delta > 0$ is a small constant and $\Delta(n) = (\Delta_1(n), \ldots, \Delta_d(n))^T, n \geq 0$, and where $\Delta_i(n), i = 1, \ldots, d, n \geq 0$ are 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. It is easy to see that $\Gamma(\theta(n) + \delta\Delta(n)) \in C, \forall n$. Moreover, $\theta(n) \in C, \forall n$ from the algorithm 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+}). 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 $\phi_{\Gamma(\theta(n)+\delta\Delta(n))}$. The initial states in the two trajectories are kept the same, i.e., $s^n = s^{n+}$, 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,

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

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 $\mathcal{F}_n \stackrel{\triangle}{=} \sigma(\theta(m), m \leq n, \Delta(m), \chi^m, \chi^{m+}, m < n), n \geq 1$ be a sequence of increasing sigma fields and with $\mathcal{F}_0 = \sigma(\theta(0))$. Let $M_n = (M^1_n, \ldots, M^d_n)^T$, $n \geq 0$. Here we let $\|\cdot\|$ denote the Euclidean norm.

Lemma 8.1. $(M_n, \mathcal{F}_n), n \geq 0$ 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,

$$E[G^n \mid \mathcal{G}_n] = \sum_s \nu(s) E[G^n \mid s_0^n = s, \phi_{\theta(n)}] = \sum_s \nu(s) V_{\theta(n)}(s).$$

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

$$V_{\theta(n)+\delta\Delta(n)}(s_n) = V_{\theta(n)}(s_n) + \delta\Delta(n)^T \nabla V_{\theta(n)}(s_n)$$
$$+\frac{\delta^2}{2}\Delta(n)^T \nabla^2 V_{\theta(n)}(s_n)\Delta(n) + o(\delta^2).$$

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 $\eta(n)=M_{n+1}=\left(\frac{G_n^+-G_n}{\delta\Delta_i(n)}\right)-E\left[\left(\frac{G_n^+-G_n}{\delta\Delta_i(n)}\right)\mid\mathcal{F}_n\right]$ and $\beta(n)=(\beta_1(n),\ldots,\beta_d(n))$ with $\beta_i(n)=E\left[\left(\frac{G_n^+-G_n}{\delta\Delta_i(n)}\right)\mid\mathcal{F}_n\right]-\sum_s\nu(s)\nabla_i V_{\theta(n)}(s)$.

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 (see for instance Chapter 13 of Sutton and Barto, 2018 that $v_\theta(s)$ is differentiable in θ and satisfies

$$\nabla v_\theta(s) = \sum_{y \in S} \sum_{l=0}^{\infty} P_\theta^k(s, y) \sum_{a \in A(y)} \nabla \phi_\theta(a \mid y) q_\theta(y, a),$$

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 $q_\theta(y, a) = E_\theta[G_n \mid S_n = y, A_n = a]$ 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 denote the Lipschitz constant for the function $\nabla v_\theta(s)$. Then, for a given $\theta_0 \in C$,

$$\begin{aligned} \| \nabla v_\theta(s) \| - \| \nabla v_{\theta_0}(s) \| &\leq \| \nabla v_\theta(s) - \nabla v_{\theta_0}(s) \| \\ &\leq L_1^s \| \theta - \theta_0 \| \\ &\leq L_1^s \| \theta \| + L_1^s \| \theta_0 \|, \end{aligned}$$

where $L_1^s > 0$ is the Lipschitz constant of $\nabla v_\theta(s)$. Thus,

$$\| \nabla v_\theta(s) \| \leq \| \nabla v_{\theta_0}(s) \| + L_1^s \| \theta_0 \| + L_1^s \| \theta \| .$$

Let $K_s \stackrel{\triangle}{=} \| \nabla v_{\theta_0}(s) \| + L_1^s \| \theta_0 \|$. Let $K_1 \stackrel{\triangle}{=} \max(K_s, L_1^s, s \in S)$. 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

$$E[\| M_{n+1} \|^2 | \mathcal{F}_n] \leq \hat{L}(1+ \| \theta(n) \|^2),$$

for some constant $\hat{L} > 0$.

Proof. Note that

$$\| M_{n+1} \|^2 = \sum_{i=1}^{d} (M_{n+1}^i)^2$$

$$= \frac{(G^{n+} - G^n)^2}{\delta^2} + \frac{1}{\delta^2} \left(E \left[\frac{G^{n+} - G^n}{\Delta_i(n)} \mid \mathcal{F}_n \right] \right)^2$$

$$-2 \frac{G^{n+} - G^n}{\delta \Delta_i(n)} E \left[\frac{G^{n+} - G^n}{\delta \Delta_i(n)} \mid \mathcal{F}_n \right].$$

Thus,

$$E[\| M_{n+1} \|^2 | \mathcal{F}_n] = E \left[\frac{(G^{n+} - G^n)^2}{\delta^2} \mid \mathcal{F}_n \right] - \left(E \left[\frac{G^{n+} - G^n}{\delta \Delta_i(n)} \mid \mathcal{F}_n \right] \right)^2.$$

It now follows from Assumption A8.1 and the fact that all single-stage costs are bounded, that $E[\| M_{n+1} \|^2 | \mathcal{F}_n] \leq \check{K}$ 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

$$E[\| M_{n+1} \|^2 | \mathcal{F}_n] \leq \check{K} (1 + \| \theta(n) \|^2 .$$

□

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

$$\sum_{n=0}^{\infty} E[\| Z_{n+1} - Z_n \|^2 | \mathcal{F}_n] < \infty \text{ a.s.}$$

Note that

$$E[\| Z_{n+1} - Z_n \|^2 | \mathcal{F}_n] = a(n)^2 E[\| M_{n+1} \|^2 | \mathcal{F}_n].$$

Thus,

$$\sum_{n=0}^{\infty} E[\| Z_{n+1} - Z_n \|^2 | \mathcal{F}_n] = \sum_{n=0}^{\infty} a(n)^2 E[\| M_{n+1} \|^2 | \mathcal{F}_n]$$

$$\leq \check{K} \sum_{n=0}^{\infty} a(n)^2 (1 + \| \theta(n) \|^2),$$

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. □

Consider now the following ODE:

$$\dot{\theta}(t) = \bar{\Gamma}(-\sum_{s} \nu(s) \nabla V_\theta(s)). \tag{8.8}$$

where $\bar{\Gamma} : \mathcal{C}(C) \to \mathcal{C}(\mathcal{R}^d)$ is as defined in (2.22).

Let $H \stackrel{\triangle}{=} \{\theta | \bar{\Gamma}(-\sum_{s} \nu(s) \nabla V_\theta(s))\}$ denote the set of asymptotically stable attractors of (8.8). Let $H^\epsilon \stackrel{\triangle}{=} N^\epsilon(H) \cap C$ where $N^\epsilon(H) = \{\theta \mid \| \theta - \theta_0 \| < \epsilon, \theta_0 \in H\}$.

We now have the following result:

Theorem 8.5. Given $\epsilon > 0$, $\exists \delta_0 > 0$ such that $\forall \delta \in [0, \delta_0)$, the stochastic iterates $\theta(n)$ governed by (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). 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 now 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

$$\begin{aligned}\nabla F(\theta)^T \bar{\Gamma}(-\sum_s \nu(s)\nabla V_\theta(s)) &= (\sum_s \nu(s)\nabla_\theta V_\theta(s))^T \bar{\Gamma}(-\sum_s \nu(s)\nabla V_\theta(s)) \\ &\leq 0.\end{aligned}$$

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

$$\begin{aligned}\nabla F(\theta)^T \bar{\Gamma}(-\sum_s \nu(s)\nabla V_\theta(s)) &< 0 \text{ if } \theta \in H^c \cap C \\ &= 0 \text{ o.w.}\end{aligned}$$

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 **??**. □

8.2 Cubic-regularized policy Newton algorithm

To be done.

8.3 SPSA for risk-constrained MDPs

Appendices

A

ODEs and differential inclusions

A.1 Ordinary differential equations

In the following, we first discuss various forms of limit sets of ODEs before describing results on convergence of the stochastic approximation recursion (2.1) using the ODE (2.2).

A.1.1 Limit sets

We present here first some basic definitions on the limit sets of ODEs. We shall then present a couple of results, in particular, Theorem 2.3 on the convergence of an underlying stochastic approximation scheme. This result from Benaïm, 1996 is a generalization of the Kushner and Clark lemma (cf. Kushner and Clark, 1978).

We recall first the Gronwall inequality, 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 $C, K, T \geq 0$:

$$u(t) \leq C + K \int_0^t u(s)v(s)ds, \ \forall t \in [0, T].$$

Then it follows that for all $t \in [0, T]$,

$$u(t) \leq C \exp\left(K \int_0^T v(s)ds\right).$$

Consider now the ODE (2.2) 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 (2.2) 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.

The integral solution to the ODE (2.2) is obtained as

$$\theta(t) = \theta_0 + \int_0^t h(\theta(s))ds, \ t \geq 0.$$

If an ODE is well-posed, it has unique integral curves. The following theorem says that a sufficient condition for well-posedness of (2.2) is that the function h be Lipschitz continuous (see Theorem B.1 of Borkar, 2022 for a proof).

Theorem A.2. Suppose the function $h : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous. Then the ODE (2.2) is well-posed.

For the ODE (2.2), 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 (2.2). 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 (2.2) 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 (2.2) 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 (2.2) 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 (2.2) 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 (2.2), the set $\mathcal{L} \stackrel{\triangle}{=} \cap_{t \geq 0} \overline{\theta([t, \infty))}$ that comprises of the set of limit points of (2.2) is called the ω-limit set of (2.2).

2. 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 (2.2) (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) $\eta_k = \eta$.

3. 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 (2.2) is chain transitive.

4. The flow Φ as above restricted to $\eta = \lambda$, for all $\lambda \in \mathbb{R}^d$ is called chain recurrent.

5. A compact invariant set $A \subset \mathcal{R}^d$ on which the flow Φ of the ODE (2.2) is chain recurrent (resp. chain transitive) is called an internally chain recurrent (resp. internally chain transitive) set for (2.2).

We now recall the following result from Benaïm, 1999 (cf. Proposition 5.3):

Lemma A.3. Let $A \subset \mathcal{R}^d$ be a compact invariant set for the ODE (2.2). 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 $\theta \in \mathcal{R}^d$ is an equilibrium of the ODE (2.2) if $\Phi_t(\theta) = \theta$, $\forall t$.

2. A compact invariant set $A \subset \mathbb{R}^d$ is said to be Lyapunov stable for the ODE (2.2) 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 $x \in \mathbb{R}^d$, $d(x, A) = \min_{\eta \in A} \|x - \eta\|$.

3. A set $A \subset \mathcal{R}^d$ is an attractor for (2.2) if A is nonempty, compact and invariant. Further, A has a positively invariant open neighborhood $M \subset \mathcal{R}^d$ such that $d(\Phi_t(\theta), A) \to 0$ as $t \to \infty$ uniformly in $\theta \in M$.

4. The largest open neighborhood M for an attractor A above is called the domain of attraction of A.

5. A compact invariant $A \subset \mathbb{R}^d$ is asymptotically stable for the ODE (2.2) if it is both Lyapunov stable and an attractor.

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 (2.2). Let $V : M \subset \mathbb{R}^d \to \mathbb{R}$ be a non-negative, continuously differentiable function. Suppose it satisfies the following conditions:

(i) $V(\theta) \to \infty$ as $\|\theta\| \to \partial M$ (the boundary of M).

(ii) Further,

$$\langle \nabla V(\theta), h(\theta) \rangle \begin{cases} < 0 \text{ if } \theta \in M \cap A^c \\ = 0 \text{ if } \theta \in A. \end{cases}$$

The function $h(\cdot)$ above is the driving vector field of the ODE (2.2). The asymptotic stability of A follows since $\dfrac{d}{dt} V(\theta(t) \leq 0$ with equality only for $\theta(t) \in A$.

We now recall the Lasalle Invariance Principle, see Theorem 2 of La Salle, 1966.

Theorem A.4 (Lasalle Invariance Principle)**.** Let $V(\cdot)$ as above be a Lyapunov function for the ODE (2.2). Then any trajectory $\theta(\cdot)$ of (2.2) 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 withe corresponding ODE is

$$\dot{\theta}(t) = -\nabla f(\theta(t)),\ \theta(0) = \theta_0. \tag{A.1}$$

Thus, here $h(\theta) = -\nabla f(\theta)$. Note that

$$\begin{aligned} \frac{df(\theta(t))}{dt} &= -\|\nabla f(\theta(t))\|^2 \\ &< 0 \quad \text{if } \nabla f(\theta) \neq 0 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

If $\lim_{\|\theta\|\to\infty} f(\theta) = \infty$, 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.1). We recall now Lemma 11.1 of Borkar, 2022.

Lemma A.5. The only invariant sets that can occur as w-limit sets for the ODE (A.1) are the subsets of $H \stackrel{\triangle}{=} \{\theta \in \mathbb{R}^d | \nabla f(\theta) = 0\}$.

Lasalle Invariance Principle, see Theorem A.4, in the case of gradient systems would say something similar as below.

Lemma A.6. Any trajectory $\theta(\cdot)$ of the ODE (A.1) 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 by Aubin and Frankowska, 1990 and Aubin and Cellina, 1984.

A.2.1 Set-valued maps

A set-valued map $x \mapsto J(x)$ is one where for any $x \in \mathbb{R}^d$, $J : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}^d\}$ and is specified via it's graph, i.e., $\text{Graph}(J) = \{(x, y) \mid y \in J(x)\}$. The domain ($\text{Dom}(J)$) and image ($\text{Im}(J)$) are respectively given by $\text{Dom}(J) = \{x \in \mathbb{R}^d \mid J(x) \neq \phi\}$ and $\text{Im}(J) = \cup_{x \in \mathbb{R}^d} J(x)$, respectively. The inverse J^{-1} of the set-valued map J (above) is also a set-valued map such that $x \in J^{-1}(y)$ if and only if $y \in J(x)$, viz., $(x, y) \in \text{Graph}(J)$.

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, $B_\epsilon(0) = \{x \in \mathbb{R}^d \mid \|x\| < \epsilon\}$ and $\overline{B}_\epsilon(0) = \{x \in \mathbb{R}^d \mid \|x\| \leq \epsilon\}$. For any set $A \subset \mathbb{R}^d$, for any $\delta > 0$, we call $N_\delta(A) = \{x \in \mathbb{R}^d \mid \|x - y\| < \delta, y \in A\}$ 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 $J : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}^d\}$ be a set-valued map.

Definition A.5 (Continuity of Set-Valued Maps). (i) J is said to be upper semi-continuous at a point $x \in \text{Dom}(J)$ if given sequences $\{x_n\}_{n \geq 1}$ (in $\mathbb{R}^d$) and $\{y_n\}_{n \geq 1}$ (in $\mathbb{R}^d$) with $x_n \to x$, $y_n \to y$ and $y_n \in J(x_n)$, $\forall n \geq 1$, we have $y \in J(x)$. We say that J is upper semi-continuous if it is upper semi-continuous at every $x \in \text{Dom}(J)$. In other words, $\text{Graph}(J)$ is closed.

(ii) J is said to be lower semi-continuous at a point $x \in \text{Dom}(J)$ if for any $y \in J(x)$, and any sequence of points $x_n \in \text{Dom}(J)$ converging to x, there exists a sequence of elements $y_n \in J(x_n) \to$

$y \in J(x)$. We say that J is lower semi-continuous if it is lower semi-continuous at every $x \in \mathrm{Dom}(J)$.

(iii) J is continuous at $x \in \mathrm{Dom}(J)$ 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 $x \in \mathrm{Dom}(J)$.

(iv) J is Lipschitz at $z \in \mathbb{R}^d$ if there exists $L > 0$ and $\epsilon > 0$ such that for all $x, y \in N_\epsilon(\{z\})$, we have that $J(x) \subset J(y) + L\|x - y\|B_1(0)$ where $U = \{w \in \mathbb{R}^d \mid \|w\| < 1\}$ is a unit ball around the origin in $\mathbb{R}^d$ or more compactly $J(x) \subset N_{L\|x-y\|}(y)$.

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 Map)**.** A set-valued map $J : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}^d\}$ is called a Peano map if it satisfies the following properties:

(i) For every $x \in \mathbb{R}^d$, $J(x)$ is convex and compact.

(ii) J is pointwise bounded for every $x \in \mathbb{R}^d$, i.e., for some $K > 0$ we have, $\sup_{w \in J(x)} \|w\| \leq K(1 + \|x\|)$.

(iii) J is upper semi-continuous.

The distance of a point $x \in \mathbb{R}^d$ to a set $A \subset \mathbb{R}^d$ 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 $J : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}^d\}$, we define the upper limit (Limsup) and lower limit (Liminf) of the sequence of sets $J(x_n)$ as follows:

(i) $\mathrm{Limsup}_{x_n \to x} J(x_n) = \{y \in \mathbb{R}^d \mid \liminf_{x_n \to x} d(y, J(x_n)) = 0\}$.

(ii) $\mathrm{Liminf}_{x_n \to x} J(x_n) = \{y \in \mathbb{R}^d \mid \lim_{x_n \to x} d(y, J(x_n)) = 0\}$.

Note that both Liminf and Limsup are closed sets. Liminf collects the limit points of $\{J(x_n)\}$ while Limsup collects it's accumulation points. Further, $\mathrm{Liminf}_{x_n \to x} J(x_n) \subset \overline{J(x)} \subset \mathrm{Limsup}_{x_n \to x} J(x_n)$.

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

$$\dot{x}(t) \in J(t, x(t)), \tag{A.2}$$

where $J : \mathbb{R} \times \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}^d\}$. We shall mainly be interested with the case where $J(t, x) \stackrel{\triangle}{=} J(x)$, i.e., there is no explicit time dependence of the set-valued map J. Thus, the DI in this case takes the form

$$\dot{x}(t) \in J(x(t)), \tag{A.3}$$

with $J : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}^d\}$. Any solution to (A.3) is viewed in the Caratheodory sense, i.e., as an absolutely continuous function satisfying (A.3) almost everywhere.

Definition A.8. (i) Let $K \subset \text{Dom}(J)$. 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.3) is said to be viable if for some closed subset K of $\text{Dom}(J)$, 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

$$C(x, K) \stackrel{\triangle}{=} \{y \in \mathbb{R}^d \mid \liminf_{k \to 0^+} \frac{d(x + ky, K)}{k} = 0\}.$$

(iv) We say that a set $K \subset \text{Dom}(J)$ is a viability domain of the set-valued map J if and only if for all $x \in K$, $J(x) \cap C(x, K) \neq \phi$.

Consider the case where $K = \{x\}$. Then the contingent cone to $\{x\}$ is given by $C(x, \{x\}) = \{y \mid \liminf_{k \to 0^+} \frac{d(x + ky, \{x\})}{k} = 0\} = \{0\}$. Then, from Definition A.8(iv), it follows that $K = \{x\}$ is a viability domain of J if and only if $J(x) \cap \{0\} \neq \phi$ or x is a stationary solution to the inclusion $0 \in J(x)$ implying that x is an equilibrium of J. 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.7. Consider a Peano map $J : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}^d\}$. Then the limit sets of the solutions to the DI (A.3) are closed viability domains. Further, the limit of a solution $x(t)$ to the DI (A.3) (if it exists), as $t \to \infty$, is an equilibrium of J.

Theorem A.8. Let $J : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}^d\}$ be a Peano map. If $K \subset \text{Dom}(J)$ is a compact viability domain and if $J(K)$ is convex, then there exists an equilibrium of J in K.

A.2.3 Limit Sets of Differential Inclusions

Recall that a solution to the DI (A.3) is an absolutely continuous mapping $\mathbf{x} : \mathbb{R} \to \mathbb{R}^d$ such that $\mathbf{x}(0) = x$ and $\dot{\mathbf{x}}(t) \in J(\mathbf{x}(t))$ for almost every $t \in \mathbb{R}$. The ω-limit set of a given solution $\mathbf{x}$ of the DI (A.3) 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 $\Phi_t(x) = \{\mathbf{x}(t) \mid \mathbf{x} \text{ is a solution to the DI (A.3) with } \mathbf{x}(0) = x\}$. Then $\{\Phi_t\}$ is the set-valued semi-flow associated with the DI (A.3). 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.3) 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 J) if $\forall x \in M$, $\exists$ a solution $\mathbf{x}$ to the DI (A.3) 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.3) 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.3) 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.3) 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) $\frac{dy(t)}{dt} - U(t) \in J^{\delta(t)}(y(t))$ for almost every $t > 0$, for some $\delta : [0, \infty) \to \mathbb{R}$ such that $\delta(t) \to 0$ as $t \to \infty$. Here $J^{\delta}(y) = \{x \in \mathbb{R}^d \mid \exists z \text{ s.t. } \|z - x\| < \delta, d(x, J(z)) < \delta\}$.

We now state a couple of important results (cf. Lemma 3.5 and Theorem 3.6 of Benaïm *et al.*, 2005).

Lemma A.9. Any internally chain transitive set for the DI (A.3) is invariant.

Theorem A.10. Let $\mathbf{z}$ be a bounded perturbed solution to the DI (A.3) 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.3).

Definition A.13 (Attracting and Lyapunov Stable Sets for a DI)**.** In relation to the DI (A.3), we have the following definitions:

(i) $M \subseteq \mathbb{R}^d$ 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 $\Phi_{[T(\epsilon),+\infty)}(U) \subset N^{\epsilon}(M)$. In other words, any DI trajectory initiated in U reaches the ϵ-neighborhood of M, $T(\epsilon)$ instants later and stays there forever subsequently.

(ii) The set U above is called the fundamental neighborhood of M.

(iii) An attracting set M that is also invariant is called an attractor set.

(iv) The basin of attraction of M is the set $B(M) = \{x \in \mathbb{R}^d \mid w_{\Phi}(x) \subset M\}$. In other words, this is the largest subset of $\mathbb{R}^d$ such that the DI initiated anywhere within this set has it's ω-limit set contained in M.

(v) The set M is said to be Lyapunov stable if for all $\delta > 0$, $\exists\ \epsilon > 0$ such that $\Phi_{[0,+\infty)}(N^{\epsilon}(M)) \subseteq N^{\delta}(M)$.

B

Martingales

B.1 Notions of convergence of random variables

Definition B.1 (Almost sure or with probability 1 convergence)**.** Let $\{X_m\}$ be the random variable set then, $X_m \to X$ almost surely or $X_m \to X$ with probability 1 as $X \to \infty$ 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.2 (Convergence in probability)**.** Let $\{X_m\}$ be the random variable set then, $X_m \xrightarrow{p} X$ if $\mathbb{P}\left[w|X_m(w) - X(w)| > \epsilon\right] = 0 \ \forall \ \epsilon > 0$, where $\mathbb{P}\left[w|X_m(w) - X(w)| > \epsilon\right]$ is usually written as $\mathbb{P}\left[|X_m(w) - X(w)| > \epsilon\right]$.

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.3 (L^2 or mean-squared convergence)**.** $X_m \xrightarrow{L^2} X$ 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.4 (Convergence in distribution)**.** $X_m \xrightarrow{d} X$ if $\mathbb{E}[f(X_m)] \to \mathbb{E}[f(X)]$ for all bounded continuous f.

The reader is referred to https://math.iisc.ac.in/~manju/PT/PT_bookform.pdf 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 https://www.ee.iitm.ac.in/~krishnaj/EE5110_files/notes/lecture28_Convergence.pdf.

In this book, we provide almost sure convergence guarantees for the well-known gradient-based zeroth-order optimization algorithms.

B.2 Martingales

A martingale is a stochastic process that is defined below.

Definition B.5. A sequence $Y = \{Y_n : n \geq 0\}$ is a *martingale* with respect to the sequence $X = \{X_n : n \geq 0\}$ if, for all $n \geq 0$,

- $\mathbb{E}[Y_n] < \infty$,
- $\mathbb{E}[Y_{n+1}|X_0, X_1, ..., X_n] = Y_n$.

Notice that

$$\begin{aligned}\mathbb{E}[Y_{n+2}|Y_1, Y_2, ..., Y_n] &= \mathbb{E}[\mathbb{E}[Y_{n+2}|Y_1, Y_2, ..., Y_{n+1}]|Y_1, Y_2, ..., Y_n] \\ &= \mathbb{E}[Y_{n+1}|Y_1, Y_2, ..., Y_n] = Y_n.\end{aligned}$$

Extending the argument, we have $\mathbb{E}[Y_{n+m}|Y_1, Y_2, ..., Y_n] = Y_n$, for any $m > 0$.

Example B.1. An i.i.d. sequence, say $\{X_i\}$, is not necessarily a martingale. However, if X_i are zero mean r.v.s, for all i, then $\{X_i\}$ is a martingale.

Definition B.6. Let $\mathcal{F}$ be a filtration of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and Y be a sequence of random variables which is adapted to $\mathcal{F}$. We call the pair $(Y, \mathcal{F}) = \{(Y_n, \mathcal{F}_n : n \geq 0)\}$ a *martingale* if, for all $n \geq 0$,

1. $\mathbb{E}[Y_n] < \infty$,
2. $\mathbb{E}[Y_{n+1}|\mathcal{F}_n] = Y_n$.

The former definition is retrieved by choosing $\mathcal{F}_n$ to be $\sigma(X_0, X_1, ..., X_n)$, which is the smallest σ-field with respect in which $X_0, X_1, ..., X_n$ are measurable. If Y is a martingale with respect to $\mathcal{F}$, then it is also a martingale with respect to $\mathcal{G}$ where $\mathcal{G}_n = \sigma(Y_0, Y_1, ..., Y_n)$.

Definition B.7. Let $\mathcal{F}$ be a filtration of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let Y be a sequence of random variables which is adapted to $\mathcal{F}$. We call the pair $(Y, \mathcal{F})$ is a *sub-martingale* if, for all n ≥ 0,

1. $\mathbb{E}(Y_n^+) < \infty$,
2. $\mathbb{E}(Y_{n+1}|\mathcal{F}_n) \geq Y_n$.

or a *super-martingale* if, for all $n \geq 0$,

1. $\mathbb{E}(Y_n^-) < \infty$,
2. $\mathbb{E}(Y_{n+1}|\mathcal{F}_n) \leq Y_n$.

Example B.2. Let $\{X_i\}$ be i.i.d, then,

1. If $\mathbb{E}[X_i] \geq 0$, then $\{X_i\}$ is a sub-martingale.
2. If $\mathbb{E}[X_i] \leq 0$, then $\{X_i\}$ is a super-martingale.

Let $X^+ = \max\{0, X\}$ and $X^- = -\min\{0, X\}$. Then, $X = X^+ - X^-$ and $|X| = X^+ + X^-$. Notice that Y is a martingale if and only if it is both a sub-martingale and a super-martingale. Also, Y is a sub-martingale if and only if $-Y$ is a super-martingale.

Example B.3. Let $\{Z_n\}$ be a martingale sequence, and $S_n = Z_n - Z_{n-1}$. Then,

$$\begin{aligned}\mathbb{E}[S_n|S_1, ..., S_{n-1}] &= \mathbb{E}[Z_n|S_1, ..., S_{n-1}] - \mathbb{E}[Z_{n-1}|S_1, ..., S_{n-1}] \\ &= Z_{n-1} - Z_{n-1} = 0.\end{aligned}$$

Further, $\mathbb{E}[S_n] = \mathbb{E}[Z_n] - \mathbb{E}[Z_{n-1}] < \infty$. Thus, the sequence $\{S_n\}$ is a martingale.

B.2.1 Applications

Mean Estimation

Consider a r.v. Y with mean μ and variance σ^2. Suppose we are given i.i.d samples $Y_1, Y_2 \dots Y_n$ 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

$$x_{n+1} = \frac{1}{n+1}\sum_{k=1}^{n+1} Y_{n+1} = \frac{n}{n+1}\left(\frac{1}{n}\sum_{k=1}^{n} Y_n\right) + \frac{1}{n+1}Y_{n+1}.$$

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 $\dfrac{1}{n+1}$, 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

$$\begin{aligned}x_{n+1} &= x_n + \alpha_n(Y_{n+1} - x_n) &\text{(B.2)}\\ &= x_n + \alpha_n[(\mu - x_n) + (Y_{n+1} - \mu)] &\text{(B.3)}\\ &= x_n + \alpha_n[(\mu - x_n) + w_{n+1}] &\text{(B.4)}\end{aligned}$$

where $w_{n+1} = Y_{n+1} - \mu$ is the noise factor. Notice that

$$\begin{aligned}\mathbb{E}[w_{n+1}|x_1, \dots x_n] &= \mathbb{E}[w_{n+1}|Y_1, \dots Y_n] \\ &= \mathbb{E}[Y_{n+1}|Y_1, \dots Y_n] - \mu \\ &= \mathbb{E}[Y_{n+1}] - \mu = 0.\end{aligned}$$

Hence, $\{w_n\}$ is a martingale difference sequence.

Urn model

Suppose we have an empty urn and add red or blue ball randomly in an iterative fashione. Let us define

$$Y_{n+1} = \begin{cases} 1, & \text{if (n+1)th ball is red} \\ 0, & \text{else} \end{cases}$$

$S_n = \sum_{k=1}^{n} Y_k$ denotes the total number of red balls. $x_n = \frac{S_n}{n}$ denotes the fraction of red balls. Then, we have

$$\begin{aligned}x_{n+1} &= \frac{1}{n+1} \sum_{i=1}^{n+1} Y_k \\ &= (1 - \frac{1}{n+1})x_n + \frac{1}{n+1} Y_{n+1} \\ &= x_n + \alpha_n(Y_{n+1} - x_n),\end{aligned}$$

where $\alpha_n = \frac{1}{n+1}$. Suppose the conditional probability that the next ball added at $(n+1)$, given the past, depends only on x_n, i.e.,

$$\mathbb{P}\left[Y_{n+1} = 1 | x_1 \dots x_n\right] = p(x_n).$$

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

$$\begin{aligned}\mathbb{E}(w_{n+1}|x_1, \dots x_n) &= \mathbb{E}(Y_{n+1} - p(x_n)|x_1, \dots x_n) \\ &= \mathbb{P}\left[Y_{n+1}|x_1, \dots x_n\right] - p(x_n) \\ &= p(x_n) - p(x_n) = 0.\end{aligned}$$

Therefore, $\{w_{n+1}\}$ is a martingale difference sequence.

SGD: A popular ML algorithm for training

Consider the following problem that is ubiquitous in machine learning applications involving training over a given dataset:

$$\min_x f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x). \tag{B.5}$$

It is common to assume that f_is are smooth and f is convex or strongly convex.

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

$$x_{n+1} = x_n - \alpha_n \left(\frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x_n)\right). \tag{B.6}$$

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:

$$x_{n+1} = x_n - \alpha_n \nabla f_{i_n}(x_n) \tag{B.7}$$

Rewriting the above update rule, we obtain

$$\begin{aligned} x_{n+1} &= x_n - \alpha_n \left(\frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x_n)) - \alpha_n(\nabla f_{i_n}(x_n) - \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x_n)\right) \\ &= x_n - \alpha_n \left(\frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x_n) + w_{n+1}\right), \end{aligned}$$

where $\{w_{n+1} = \nabla f_{i_n}(x_n) - \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x_n)\}$ is a martingale difference sequence because $\mathbb{E}[w_{n+1}|x_1, \ldots x_n] = 0$

Remark B.1. 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. In the next section, we state and prove the well-known maximal inequality for martingales. This inequality will be subsequently used 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.2.2 Maximal inequality

We state and prove *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. Let $A_0 = \Omega$, $A_k = \{|S_i| < \epsilon \forall i \leq k\}$, and let $B_k = A_{k-1} \cap \{|S_k| \geq \epsilon\}$ be the event that $|S_i| \geq \epsilon$ for the first time when $i = k$. Then,

$$A_k \cup \left(\bigcup_{i=1}^{k} B_i\right) = \Omega.$$

Therefore,

$$\mathbb{E}[S_n^2] = \sum_{i=1}^{n} \mathbb{E}[S_n^2 I_{B_i}] + \mathbb{E}[S_n^2 I_{A_n}] \geq \sum_{i=1}^{n} \mathbb{E}[S_n^2 I_{B_i}].$$

However,

$$\begin{aligned}\mathbb{E}[S_n^2 I_{B_i}] &= \mathbb{E}[(S_n - S_i + S_i)^2 I_{B_i}] \\ &= \underbrace{\mathbb{E}[(S_n - S_i)^2 I_{B_i}]}_{(I)} + \underbrace{2\mathbb{E}[(S_n - S_i)S_i I_{B_i}]}_{(II)} + \underbrace{\mathbb{E}[S_i^2 I_{B_i}]}_{(III)}.\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}[(S_n - S_i)S_i I_{B_i}] &= \mathbb{E}\left[S_i I_{B_i}\mathbb{E}[(S_n - S_i)|X_1, ..., X_i]\right] \\ &= 0,\end{aligned}$$

since B_i concerns $X_1, ..., X_i$ only, the inequality presented above becomes

$$\mathbb{E}[S_n^2] \geq \sum_{i=1}^{n} \epsilon^2 P(B_i) = \epsilon^2 P(\max_{1\leq i\leq n} |S_i| \geq \epsilon).$$

□

B.2.3 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 $m, n \geq 1$ 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 $C = \{\omega \mid S_n(\omega)$ 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,

$$C = \{\omega | \forall \epsilon > 0, \exists\, m \text{ such that } |S_{m+i}(\omega) - S_m(\omega)| < \epsilon\ \forall\, i \geq 1\}$$

$$= \bigcap_{\epsilon>0} \bigcup_{m\geq 1} \{|S_{m+i} - S_m| < \epsilon, \forall i \geq 1\}$$

$$C^{\mathsf{c}} = \bigcup_{\epsilon>0} \bigcap_{m\geq 1} \{|S_{m+i} - S_m| \geq \epsilon, \text{ for some } i \geq 1\}.$$

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

$$0 \leq \lim_{\epsilon \downarrow 0} P\left(\bigcap_m A_m(\epsilon)\right) \leq \lim_{\epsilon\downarrow 0} \lim_{m\to\infty} \mathbb{P}(A_m(\epsilon)).$$

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

$$\mathbb{P}(|Y_i \geq \epsilon \text{ for some } 1 \leq i \leq n) \leq \frac{1}{\epsilon^2}\mathbb{E}[Y_n]^2$$

$$\mathbb{P}(|S_{m+i} - S_m| \geq \epsilon, \text{ for some } 1 \leq i \leq n) \leq \frac{\mathbb{E}[S_{m+n} - S_m]^2}{\epsilon^2}$$

$$0 \leq \mathbb{P}(A_m(\epsilon)) \leq \frac{\mathbb{E}(S_{m+n} - S_m)^2}{\epsilon^2} = \frac{\mathbb{E}[S_{m+n}^2 + \mathbb{E}[S_m^2 - 2\mathbb{E}[S_{m+n}S_m]]]}{\epsilon^2}.$$

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

$\mathbb{P}(C^{\mathsf{c}} = 0 \text{ (or) } \mathbb{P}(C) = 1, \quad \text{i.e.}$
The sequence $\{S_n\}$ is Cauchy convergent $\exists\ S$ such that $S_n \underset{a.s}{\to} S$ 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 $\{X_n\}$ 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

$$\begin{aligned}
\mathbb{E}[(S_n - S)^2] &= \mathbb{E}[\liminf_{m\to\infty} (S_n - Sm)^2] && \text{(B.8)}\\
&\leq \liminf_{m\to\infty} \mathbb{E}[(S_n - S_m)^2] && \text{(Fatou's Lemma)}\\
&= M - \mathbb{E}[S_n^2] \overset{n\to\infty}{\longrightarrow} 0. && \text{(B.9)}
\end{aligned}$$

$\implies \mathbb{E}[(S_n - S)^2] \overset{n\to\infty}{\longrightarrow} 0$ or $S_n \overset{L^2}{\to} S$

To arrive at the equality in B.8, 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.9 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. □

C

Smoothness and Convexity

In this appendix, we discuss foundation of algorithms for non-linear smooth optimization problem which include Taylor's theorem and its applications, convex sets and convex/strongly-convex functions..

We are interested in finding a θ^* such that

$$\theta^* \in \arg\min_{x \in \mathcal{D}} f(\theta). \tag{C.1}$$

We have following relevant definitions for this minimization problem.

Definition C.1 (local minima)**.** A point $\theta^* \in \mathcal{D}$ is called local minima of f if there exists a neighbourhood $\mathcal{N}(\theta^*, \epsilon)$ of θ^* such that $f(\theta) \geq f(\theta^*)$ for all $x \in \mathcal{N}(\theta^*, \epsilon) \cap \mathcal{D}$.

Definition C.2 (global minima)**.** A point $\theta^* \in \mathcal{D}$ is called global minima of f if $f(\theta) \geq f(\theta^*)$ for all $x \in \mathcal{D}$.

Definition C.3 (strict local minima)**.** A point $\theta^* \in \mathcal{D}$ is called local minima of f if there exists a neighbourhood $\mathcal{N}(\theta^*, \epsilon)$ of θ^* such that $f(\theta) > f(\theta^*)$ for all $x \in \mathcal{N}(\theta^*, \epsilon) \cap \mathcal{D}$ with $x \neq \theta^*$.

C.1 Necessary conditions for local minima

Given a point $\theta^* \in \mathcal{D}$, how does one determine whether it is a local minima or not? The following results, which are standard in optimization literature, provide an answer to this question.

Theorem C.1 (First and second-order necessary conditions)**.** Let θ^* be a local minima of $f : \mathcal{D} \to \mathbb{R}$ and f is continuously differentiable. Then $\nabla f(\theta^*) = 0$.
Further if f is twice continuously differentiable, then $\nabla^2 f(\theta^*)$ is a positive semi-definite (p.s.d.) matrix.

Proof. Fix $s \in \mathbb{R}^N$. Recall that θ^* is a local minima. Then we have,

$$s^T \nabla f(\theta^*) = \lim_{\delta -> 0} \frac{f(\theta^* + \delta s) - f(\theta^*)}{\delta} \geq 0.$$

Similarly, we have,

$$-s^T \nabla f(\theta^*) \geq 0.$$

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

$$f(\theta^* + \delta s) - f(\theta^*) = \delta s^T \nabla f(\theta^*) + \frac{\delta^2}{2} s^T \nabla^2 f(\theta^*) s + o(\delta^3).$$

Since $\nabla f(\theta^*) = 0$, we have

$$0 \leq \frac{f(\theta^* + \delta s) - f(\theta^*)}{\delta^2} = \frac{1}{2} s^T \nabla^2 f(\theta^*) s + o(\delta).$$

Thus, as $\delta \to 0$, for all $s \in \mathbb{R}^N$, we have $s^T \nabla^2 f(\theta^*) s \geq 0$, implying $\nabla^2 f(\theta^*)$ is p.s.d. Hence proved. □

Example C.1. Consider $f(\theta) = \frac{1}{2} \theta^T A \theta - b^T \theta$. From the first-order necessary condition, we have
$\nabla f(\theta^*) = 0$ and $\nabla^2 f(\theta^*)$ is p.s.d., which is equivalent to $A\theta^* - b = 0$ and A is a psd.

We have the following cases:

- If A is not psd, then f has no local minima

- If A is p.s.d., then f is convex and any θ^* solving $A\theta^* - b = 0$ is a global minima.
- if A is p.d., then f has a unique global minima given by $\theta^* = A^{-1}b$.
- What happens in the case where A is psd and singular?

C.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 C.2. Let $f : \mathbb{R}^N \to \mathbb{R}$ be a continuously differentiable function. Given $x, p \in \mathbb{R}^N$, we have

$$f(\theta + p) = f(\theta) + \int_0^1 \nabla f(\theta + \alpha p)^T p d\alpha, \text{ and} \tag{C.2}$$

$$f(\theta + p) = f(\theta) + \nabla f(\theta + \alpha p)^T p. \tag{C.3}$$

The expression in (C.2) is the "integral form" and the one in (C.3) is the "mean-value form" of Taylor's theorem.
If f is twice continuously differentiable, we have

$$\nabla f(\theta + p) = \nabla f(\theta) + \int_0^1 \nabla^2 f(\theta + \alpha p) p d\alpha, \text{ and}$$

$$f(\theta + p) = f(\theta) + \nabla f(\theta)^T p + \frac{1}{2} p^T \nabla^2 f(\theta + \alpha p) p,$$

for some $\alpha \in (0, 1)$.

A consequence of (C.2) is that for a continuously differentiable f at x, we have

$$f(\theta + p) = f(\theta) + \nabla f(\theta)^T p + o(||p||).$$

Definition C.4 (smooth function)**.** A function $f : \mathcal{D}(\subset \mathbb{R}^N) \to \mathbb{R}$ is said to be L-smooth if for all $x, y \in \mathcal{D}$, the following condition holds:

$$||\nabla f(\theta) - \nabla f(y)|| \leq L||x - y||. \tag{C.4}$$

The three results below provide useful characterizations of L-smooth functions.

Lemma C.3. Let $f : \mathcal{D}(\subset \mathbb{R}^N) \to \mathbb{R}$ be a L-smooth function. Then for any $x, y \in \mathcal{D}$, we have the following:

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}||y - x||^2. \tag{C.5}$$

Lemma C.4. Suppose $f : \mathcal{D}(\subset \mathbb{R}^N) \to \mathbb{R}$ is twice continuously differentiable function. Then,
(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 C.5. Suppose f is twice continuously differentiable on $\mathbb{R}^N$. Then if f is L-smooth, we have $\nabla^2 f(\theta) \preceq L\mathbb{I}$ for all x.
Conversely, if $-L\mathbb{I} \preceq \nabla^2 f(\theta) \preceq L\mathbb{I}$, then f is L-smooth

C.3 Sufficient conditions for local minima

Theorem C.6 (Sufficient Conditions for Smooth Unconstrained Optimization)**.** Suppose that f is twice continuously differentiable and that, for some θ^*, we have $\nabla f(\theta^*) = 0$, and $\nabla^2 f(\theta^*)$ is positive definite. Then θ^* is a strict local minimizer of $\min_{\theta \in \mathbb{R}^N} f(\theta)$

Proof. We use formula (4.6) from Taylor's theorem. Define a radius ρ sufficiently small and positive such that the eigenvalues of $\nabla^2 f(\theta^* + \gamma p)$ are bounded below by some positive number ϵ, for all $p \in \mathbb{R}^N$ with $\|p\| < \rho$, and all $\gamma \in (0, 1)$. (Because $\nabla^2 f$ is positive definite at θ^* and continuous, and because the eigenvalues of a matrix are continuous functions of the elements of a matrix, it is possible to choose $\rho > 0$ and $\epsilon > 0$ with these properties.) By setting $\theta = \theta^*$ in (4.6), we have for some $\gamma \in (0, 1)$

$$\begin{aligned} f(\theta^* + p) &= f(\theta^*) + \nabla f(\theta^*)^T p + \frac{1}{2} p^T \nabla^2 f(\theta^* + \alpha p) p \\ &\geq f(\theta^*) + \epsilon \|p\|^2, \quad \text{for all p with } \|p\| < \rho. \end{aligned}$$

Thus, by setting $N = \theta^* + p \| \|p\| < \rho$, we have found a neighborhood of θ^* such that $f(\theta) > f(\theta^*)$ for all $\theta \in \mathcal{N}$ with $\theta \neq \theta^*$, hence satisfying the conditions for a strict local minimizer. □

C.4 Convex Sets and Functions

Definition C.5. A set $A \subset \mathbb{R}^d$ is a convex set if $\forall x, y \in A$ and for all $\lambda \in [0, 1]$ it satisfies:

$$\lambda\theta + (1 - \lambda)y \in A. \tag{C.6}$$

Example C.2. Hyperplanes: $\mathrm{H} = \{\theta \mid A^T\theta = b\}$ is a convex set.

Example C.3. Halfspaces: $\mathrm{H} = \{\theta \mid A^T\theta \leq b\}$ is a convex set.

Example C.4. Euclidean Balls: $B = \{\theta \mid \|x\|_n \leq 1\}$ is a convex set.

Definition C.6. $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function if its domain Ω is convex and it satisfies the following condition for all $x, y \in \Omega$ and $\lambda \in [0, 1]$:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(\theta) + (1 - \lambda)f(y). \tag{C.7}$$

Note that a function f is concave if $-f$ is convex.

Lemma C.7. Suppose f is convex. Then,

1. Any local minima is a global minima.
2. The set of all global minima is convex.

Theorem C.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,

$$f(x + \alpha(y - x)) = f(\theta) + \alpha\nabla f(x)^T(y - x) + o(\alpha) \leq (1 - \alpha)f(x) + \alpha f(y).$$

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + o(1).$$

when $\alpha \downarrow 0$, $o(1)$ term vanishes, and we obtain

$$f(y) \geq f(x) + \nabla f(x)^T (y - x).$$

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 C.9. Suppose $f : \mathbb{R}^N \to \mathbb{R}$ is twice differentiable over an open domain. Then the following are equivalent f is convex $f(y) \geq f(\theta) + \nabla f(\theta)^T (y - x)$, for all $x, y \in \mathcal{D}$

1. f is convex;
2. $f(y) \geq f(\theta) + \nabla f(\theta)^T (y - x), \forall x, y \in \mathcal{D}$;
3. $\nabla^2 f(\theta) \succeq 0$, for all $\theta \in \mathcal{D}$.

Proof. We prove $(i) \Leftrightarrow (ii)$ then $(ii) \Leftrightarrow (iii)$.

$(i) \Rightarrow (ii)$ If f is convex, by definition

$$f(\lambda y + (1 - \lambda)x) \leq \lambda f(y) + (1 - \lambda) f(\theta), \forall \lambda \in [0, 1], x, y \in \text{dom}(f)$$

After rewriting, we have

$$f(\theta + \lambda(y - x)) \leq f(\theta) + \lambda(f(y) - f(\theta))$$
$$\Rightarrow f(y) - f(\theta) \geq \frac{f(\theta + \lambda(y - x)) - f(\theta)}{\lambda}, \forall \lambda \in (0, 1]$$

As $\lambda \downarrow 0$, we get

$$f(y) - f(\theta) \geq \nabla f^T(\theta)(y - x) \tag{C.8}$$

$(ii) \Rightarrow (i)$ Suppose (C.8) holds $\forall x, y \in \text{dom}(f)$. Take any $x, y \in \text{dom}(f)$ and let

$$z = \lambda\theta + (1 - \lambda)y$$

We have

$$f(\theta) \geq f(z) + \nabla f^T(z)(x - z) \tag{C.9}$$
$$f(y) \geq f(z) + \nabla f^T(z)(y - z) \tag{C.10}$$

Multiplying (C.9) by λ, (C.10) by $(1 - \lambda)$ and adding, we get

$$\begin{aligned}\lambda f(\theta) + (1 - \lambda) f(y) &\geq f(z) + \nabla f^T(z)(\lambda\theta + (1 - \lambda)y - z) \\ &= f(z) \\ &= f(\lambda\theta + (1 - \lambda)y).\end{aligned}$$

$(ii) \Leftrightarrow (iii)$ 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

$$f(y) \geq f(\theta) + f'(\theta)(y - x) \tag{C.11}$$
$$\text{and } f(\theta) \geq f(y) + f'(y)(x - y) \tag{C.12}$$
$$\Rightarrow f'(\theta)(y - x) \leq f(y) - f(\theta) \leq f'(y)(y - x)$$

using (C.11) then (C.12). Dividing LHS and RHS by $(y - x)^2$ gives

$$\frac{f'(y) - f'(\theta)}{y - x} \geq 0, \forall x, y, x \neq y$$

As we let $y \to x$, we get

$$f''(\theta) \geq 0, \forall x \in \text{dom}(f)$$

$(iii) \Rightarrow (ii)$*(uni-variate case)* Suppose $f''(\theta) \geq 0, \forall x \in \text{dom}(f)$. By the mean value version of Taylor's theorem we have

$$f(y) = f(\theta) + f'(\theta)(y - x) + \frac{1}{2} f''(z)(y - x)^2, \text{ for some } z \in [x, y].$$
$$\Rightarrow f(y) \geq f(\theta) + f'(\theta)(y - x).$$

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 iff

$$g''(\alpha) = v^T \nabla^2 f(x_0 + \alpha v) v \geq 0$$

$\forall x_0 \in \text{dom}(f), \forall v \in \mathbb{R}^n$ and $\forall \alpha$ s.t. $x_0 + \alpha v \in \text{dom}(f)$. Hence, f is convex iff $\nabla^2 f(\theta) \succeq 0$ for all $x \in \text{dom}(f)$. □

C.5 Strongly Convex Functions

Definition C.7. A function $f : \mathbb{R}^N \to \mathbb{R}$ is said to be m-strongly convex ($m > 0$) if for all $x, y \in \mathbb{R}^N$,them:

$$f((1-\lambda)x+\lambda y) \leq (1-\lambda)f(\theta)+\lambda f(y)+\frac{m}{2}(1-\lambda)\left\|y-x\right\|_2^2 \quad \text{(C.13)}$$

Theorem C.10. Suppose f is continuously differentiable and m-strongly convex, then

$$f(y) \geq f(\theta) + \nabla f(\theta)^T(y-x) + \frac{m}{2}\left\|y-x\right\|_2^2$$

Lemma C.11. Suppose that f is twice-continuously differentiable on $\mathbb{R}^N$. Then f has modulus of convexity m if and only if $\nabla f(\theta) \succeq mI$ for all x.

Proof. For any $x, y \in \mathbb{R}^N$ and $\alpha > 0$, we have from Taylor's theorem that

$$f(x+\alpha u) = f(\theta) + \alpha \nabla f(\theta)^T u + \frac{1}{2}\alpha^2 u^T \nabla^2 f(\theta + \gamma\alpha u)u,$$

for some $\gamma \in (0,1)$.

From the strong convexity property, we have

$$f(x+\alpha u) \geq f(\theta) + \alpha \nabla f(\theta)^T u + \frac{m}{2}\alpha^2 \left\|u\right\|^2$$

By comparing the two equations above, we obtain

$$u^T \nabla^2 f(x+\gamma\alpha u)u \geq m\left\|u\right\|^2$$

By taking $\alpha \downarrow 0$, we obtain

$$u^T \nabla^2 f(\theta)u \geq m\left\|u\right\|^2.$$

Hence, we have

$$\nabla^2 f(\theta) \succeq mI. \quad \text{(C.14)}$$

□

References

A.Dvoretzky. (1956). "On stochastic approximation". *Proc. Third Berkeley Symp. Math. Stat. and Prob.* 1: 39–55.

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.

Artzner, P., F. Delbaen, J. Eber, and D. Heath. (1999). "Coherent measures of risk". *Mathematical Finance*. 9(3): 203–228.

Asmussen, S. and P. W. Glynn. (2007). *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. (2022). "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.

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".

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. A. Prashanth. (2015a). "Simultaneous perturbation Newton algorithms for simulation optimization". *Journal of Optimization Theory and Applications*. 164(2): 621–643.

Bhatnagar, S. and L. Prashanth. (2015b). "Simultaneous perturbation Newton algorithms for simulation optimization". *Journal of Optimization Theory and Applications*. 164(2): 621–643.

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. (2003). "Multiscale chaotic SPSA and smoothed functional algorithms for simulation optimization". *Simulation*. 79(10): 568–580.

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. 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. (2016). "Multiscale Q-learning with linear function approximation". *Discrete Event Dynamic Systems*. 26: 477–509.

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.

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.

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.

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, S., A. Devraj, A. Busic, and S. Meyn. (2020). "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.

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.

Dunkel, J. and S. Weber. (2010). "Stochastic root finding and efficient estimation of convex risk measures". *Operations Research*. 58(5): 1505–1521.

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.

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.

Ge, R., F. Huang, C. Jin, and Y. Yuan. (2015). "Escaping From Saddle Points – Online Stochastic Gradient for Tensor Decomposition". *Conference of Learning Theory*.

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.

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.

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.

Hurley, M. (1995). "Chain recurrence, semiflows, and gradients". *Journal of Dynamics and Differential Equations*: 437–456.

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.

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.

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.

Kiefer, J. and J. Wolfowitz. (1952). "Stochastic estimation of the maximum of a regression function". *Ann. Math. Statist.* 23: 462–466.

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.

La Salle, J. P. (1966). "An invariance principle in the theory of stability". *Tech. rep.*

Lasalle, J. (1966). "An Invariance Principle in the Theory of Stability". *Tech. rep.* No. TR-66-1. Center for Dynamical Systems, Brown University.

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

Meyn, S. (2022). *Control systems and reinforcement learning*. Cambridge University Press.

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.

Nesterov, Y. and V. Spokoiny. (2017). "Random gradient-free minimization of convex functions". *Foundations of Computational Mathematics*. 17(2): 527–566.

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., 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. 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., 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.

Prashanth, L., 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. and M. Ghavamzadeh. (2016). "Variance-constrained actor-critic algorithms for discounted and average reward MDPs". *Machine Learning*. 105: 367–417.

Ramaswamy, A. and S. Bhatnagar. (2016). "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.

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.

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.

Spall, J. C. (2000). "Adaptive stochastic approximation by the simultaneous perturbation method". *IEEE Trans. Autom. Contr.* 45: 1839–1853.

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. (2005). *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons.

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.

Sutton, R. S. and A. W. Barto. (2018). *Reinforcement Learning, 2'nd Edition*. MIT Press.

Swain, J. (2017). "Simulation Software Survey-Simulation: new and improved reality show". *OR/MS Today*. 44(5): 38–49.

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. Prashanth. (2021). "Smoothed functional-based gradient algorithms for off-policy reinforcement learning: A non-asymptotic viewpoint". *Systems & Control Letters*. 155: 104988.

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.

Yao, A. C. C. (1977). "Probabilistic computations: Toward a unified measure of complexity". In: *FOCS*. 222–227.