Analysis of gradient descent methods with non-diminishing, bounded errors

Arunselvan Ramaswamy*, arunselvan@csa.iisc.ernet.in
Shalabh Bhatnagar*, shalabh@csa.iisc.ernet.in
* Department of Computer Science and Automation, Indian Institute of Science, Bangalore - 560012, India.

Abstract—The main aim of this paper is to provide an analysis of gradient descent (GD) algorithms with gradient errors that do not necessarily vanish, asymptotically. In particular, sufficient conditions are presented for both stability (almost sure boundedness of the iterates) and convergence of GD with bounded, (possibly) non-diminishing gradient errors. In addition to ensuring stability, such an algorithm is shown to converge to a small neighborhood of the minimum set, which depends on the gradient errors. It is worth noting that the main result of this paper can be used to show that GD with asymmetrically vanishing errors indeed converges to the minimum set. The results presented herein are not only more general when compared to previous results, but our analysis of GD with errors is new to the literature to the best of our knowledge. Our work extends the contributions of Mangasarian & Solodov, Bertsekas & Tsitsiklis and Tadic & Doucet. Using our framework, a simple yet effective implementation of GD using simultaneous perturbation stochastic approximations (SPSA), with constant sensitivity parameters, is presented. Another important improvement over many previous results is that there are no ‘additional’ restrictions imposed on the step-sizes. In machine learning applications where step-sizes are related to learning rates, our assumptions, unlike those of other papers, do not affect these learning rates. Finally, we present experimental results to validate our theory.

Index Terms—Stochastic approximation algorithms, gradient descent methods, differential inclusions, non-diminishing errors, stability and convergence.

I. INTRODUCTION

Let us suppose that we are interested in finding a minimum (local/global) of a continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$. The following gradient descent method (GD) is often employed to find such a minimum:

$$x_{n+1} = x_n - \gamma_n \nabla f(x_n).$$

(1)

In the above equation, $\{\gamma_n\}_{n \geq 0}$ is the given step-size sequence and $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is a continuous map such that $\|\nabla f(x)\| \leq K(1 + \|x\|)$, $K > 0$ and $x \in \mathbb{R}^d$. GD is a popular tool to implement many machine learning algorithms. For example, the backpropagation algorithm for training neural networks employs GD due to its effectiveness and ease of implementation.

When implementing (1), one often uses gradient estimators such as Kiefer-wolfowitz estimator [8], simultaneous perturbation stochastic approximation (SPSA) [10], etc., to obtain estimates of the true gradient at each stage which in turn results in estimation errors $\{\epsilon_n\}$ in (2)). This is particularly true when the form of $f$ or $\nabla f$ is unknown. Previously in the literature, convergence of GD with errors was studied in [5]. However, their analysis required the errors to go to zero at the rate of the step-size (vanish asymptotically at a prescribed rate). Such assumptions are difficult to enforce and may adversely affect the learning rate when employed to implement machine learning algorithms, see Chapter 4.4 of [6]. In this paper, we present sufficient conditions for both stability (almost sure boundedness) and convergence (to a small neighborhood of the minimum set) of GD with bounded errors, for which the recursion is given by

$$x_{n+1} = x_n - \gamma_n (\nabla f(x_n) + \epsilon_n).$$

(2)

In the above equation $\epsilon_n$ is the estimation error at stage $n$ such that $\forall n \|\epsilon_n\| \leq \epsilon$ (a.s. in the case of stochastic errors) for a fixed $\epsilon > 0$ (positive real). As an example, consider the problem of estimating the average waiting time of a customer in a queue. The objective function $J$, for this problem, has the following form: $J(x) = \int w d F(w | x) = E[W(x)]$, where $W(x)$ is the “waiting time” random variable with distribution $F(\cdot | x)$, with $x$ being the underlying parameter (say the arrival or the service rate). In order to define $J$ at every $x$, one will need to know the entire family of distributions, $\{F(\cdot | x) \mid x \in \mathbb{R}^d\}$, exactly. In such scenarios, one often works with approximate definitions of $F$ which in turn lead to approximate gradients, i.e., gradients with errors. Moreover, the gradient errors could be inherent to the problem at hand or due to extraneous noise. In such cases, there is no reason to believe that these errors will vanish asymptotically. To the best of our knowledge, this is the first time an analysis is done for GD with biased/unbiased stochastic/deterministic errors that are not necessarily diminishing, and without imposing ‘additional’ restrictions on step-sizes over the usual standard assumptions, see (A2) in Section III-A.

Our assumptions, see Section III-A, not only guarantee stability but also guarantee convergence of the algorithm to a small neighborhood of the minimum set, where the neighborhood is a function of the gradient errors. If $\|\epsilon_n\| \to 0$ as $n \to \infty$, then it follows from our main result (Theorem 2) that the algorithm converges to an arbitrarily small neighborhood of the minimum set. In other words, the algorithm indeed converges to the minimum set. It may be noted that we do not impose any restrictions on the noise-sequence $\{\epsilon_n\}_{n \geq 0}$, except that almost surely for all $n \|\epsilon_n\| \leq \epsilon$ for some fixed $\epsilon > 0$. Our analysis uses techniques developed in the field of viability theory by [1], [2] and [3]. Experimental results supporting the analyses in this paper are presented in Section V.
A. Our contributions

(1) Previous literature such as [5] requires \(\|\epsilon_n\| \to 0\) as \(n \to \infty\) for it’s analysis to work. Further, both [5] and [9] provide conditions that guarantee one of two things (a) GD diverges almost surely or (b) converges to the minimum set almost surely. On the other hand, we only require \(\|\epsilon_n\| \leq \epsilon\) \(\forall\ n\), where \(\epsilon > 0\) is fixed a priori. Also, we present conditions under which GD with bounded errors is stable (bounded almost surely) and converges to an arbitrarily small neighborhood of the minimum set almost surely. Note that our analysis works regardless of whether or not \(\|\epsilon_n\|\) tends to zero. For more detailed comparisons with [5] and [9] see Section III-B.

(2) The analyses presented herein will go through even when the gradient errors are “asymptotically bounded” almost surely. In other words, \(\|\epsilon_n\| \leq \epsilon\) for all \(n \geq N\) almost surely. Here \(N\) may be sample path dependent.

(3) Previously, convergence analysis of GD required severe restrictions on the step-size, see [5], [10]. However, in our paper we do not impose any such restrictions on the step-size. See Section III-B (specifically points 1 and 3) for more details.

(4) Informally, the main result of our paper, Theorem 2, states the following. One wishes to simulate GD with gradient errors that are not guaranteed to vanish over time. As a consequence of allowing non-diminishing errors, we show the following: There exists \(\epsilon(\delta) > 0\) such that the iterates are stable and converge to the \(\delta\)-neighborhood of the minimum set \((\delta\ being\ chosen\ by\ the\ simulator)\) as long as \(\|\epsilon_n\| \leq \epsilon(\delta) \ \forall\ n\).

(5) In Section IV-B we discuss how our framework can be exploited to undertake convenient yet effective implementations of GD. Specifically, we present an implementation using SPSA, although other implementations can be similarly undertaken.

II. DEFINITIONS USED IN THIS PAPER

[Minimum set of a function] This set consists of all global and local minima of the given function.

[Upper-semicontinuous map] We say that \(H\) is upper-semicontinuous, if given sequences \(\{x_n\}_{n \geq 1} \) (in \(\mathbb{R}^n\)) and \(\{y_n\}_{n \geq 2} \) (in \(\mathbb{R}^m\)) with \(x_n \to x\), \(y_n \to y\) and \(y_n \in H(x_n)\), \(n \geq 1\), then \(y \in H(x)\).

[Marchaud Map] A set-valued map \(H: \mathbb{R}^n \to \{\text{subsets of } \mathbb{R}^m\}\) is called Marchaud if it satisfies the following properties: (i) for each \(x \in \mathbb{R}^n\), \(H(x)\) is convex and compact; (ii) (point-wise boundedness) for each \(x \in \mathbb{R}^n\),

\[
\sup_{w \in H(x)} \|w\| \leq K(1 + \|x\|) \text{ for some } K > 0; \text{ (iii) } H \text{ is upper-semicontinuous.}
\]

Let \(H\) be a Marchaud map on \(\mathbb{R}^d\). The differential inclusion (DI) given by

\[
\dot{x} \in H(x)
\]

is guaranteed to have at least one solution that is absolutely continuous. The reader is referred to [1] for more details. We say that \(x \in \sum\) if \(x\) is an absolutely continuous map that satisfies (3). The set-valued semiflow \(\Phi\) associated with (3) is defined on \([0, +\infty) \times \mathbb{R}^d\) as:

\[
\Phi_t(x) = \{x(t) \mid x \in \sum, x(0) = x\}.
\]

Let \(B \times M \subset [0, +\infty) \times \mathbb{R}^d\) and define \(\Phi_B(M) = \bigcup_{t \in B, x \in M} \Phi_t(x)\).

[Limit set of a solution] The limit set of a solution \(x\) with \(x(0) = x\) is given by \(L(x) = \bigcap_{t \geq 0} x[t, +\infty)\).

[Invariant set] \(M \subset \mathbb{R}^d\) is invariant if for every \(x \in M\) there exists a trajectory, \(x \in \sum\), entirely in \(M\) with \(x(0) = x\), \(x(t) \in M\), for all \(t \geq 0\).

[Open and closed neighborhoods of a set] Let \(x \in \mathbb{R}^d\) and \(A \subset \mathbb{R}^d\), then \(d(x, A) := \inf\{\|a - y\| \mid y \in A\}\). We define the \(\delta\)-open neighborhood of \(A\) by \(N^\delta(A) := \{x \mid d(x, A) < \delta\}\). The \(\delta\)-closed neighborhood of \(A\) is defined by \(\overline{N^\delta(A)} := \{x \mid d(x, A) \leq \delta\}\).

[\(B_r(0)\) and \(\overline{B}_r(0)\)] The open ball of radius \(r\) around the origin is represented by \(B_r(0)\), while the closed ball is represented by \(\overline{B}_r(0)\). In other words, \(B_r(0) = \{x \mid \|x\| < \delta\}\) and \(\overline{B}_r(0) = \{x \mid \|x\| \leq \delta\}\).

[Internally chain transitive set] \(M \subset \mathbb{R}^d\) is said to be internally chain transitive if \(M\) is compact and for every \(x, y \in M\), \(\epsilon > 0\) and \(T > 0\) we have the following: There exist \(n\) and \(\Phi^1, \ldots, \Phi^n\) that are \(n\) solutions to the differential inclusion \(\dot{x}(t) \in h(x(t))\), points \(x_1 = x, \ldots, x_n+1 = y\) \(\in M\) and \(n\) real numbers \(t_1, t_2, \ldots, t_n\) greater than \(T\) such that:

\[
\Phi_{t_1}(x_1) = x_n \in N^\epsilon(x_{i+1})\text{ and }\Phi_{[0, t_i]}(x_i) \subset M\text{ for }1 \leq i \leq n.
\]

The sequence \((x_1 = x), \ldots, x_n+1 = y\) is called an \((\epsilon, T)\) chain in \(M\) from \(x\) to \(y\). If the above property only holds for all \(x = y\), then \(M\) is called chain recurrent.

[Attracting set & fundamental neighborhood] \(A \subset \mathbb{R}^d\) is attracting if it is compact and there exists a neighborhood \(U\) such that for any \(\epsilon > 0\), \(\exists T(\epsilon) \geq 0\) with \(\Phi_{T(\epsilon)}(\mathbb{R}^d(\epsilon) \cup U) \subset N^\epsilon(A)\). Such a \(U\) is called the fundamental neighborhood of \(A\).

[Attractor set] An attracting set that is also invariant is called an attractor set. The basin of attraction of \(A\) is given by

\[
B(A) = \{x \mid \omega(x) \subset A\}.
\]

[Lyapunov stable] The above set \(A\) is Lyapunov stable if for all \(\delta > 0\), \(\exists \epsilon > 0\) such that \(\Phi_{[0, +\infty)}(N^\delta(A)) \subset N^\epsilon(A)\).

[Upper-limit of a sequence of sets, \(\limsup\)] Let \(\{K_n\}_{n \geq 1}\) be a sequence of sets in \(\mathbb{R}^d\). The upper-limit of \(\{K_n\}_{n \geq 1}\) is given by:

\[
\limsup_{n \to \infty} K_n := \{y \mid \limsup_{n \to \infty} d(y, K_n) = 0\}.
\]

We may interpret that the lower-limit collects the limit points of \(\{K_n\}_{n \geq 1}\) while the upper-limit collects its accumulation points.

III. ASSUMPTIONS AND COMPARISON TO PREVIOUS LITERATURE

A. Assumptions

Recall that GD with bounded errors is given by the following recursion:

\[
x_{n+1} = x_n - \gamma(n)g(x_n), \quad (4)
\]

where \(g(x_n) \in G(x_n)\) \(\forall n\) and \(G(x) := \nabla f(x) + B_e(0), \quad x \in \mathbb{R}^d\). In other words, the gradient estimate at stage \(n\), \(g(x_n)\), belongs to an \(\epsilon\)-ball around the true gradient \(\nabla f(x_n)\) at stage \(n\). Note that (4) is consistent with (2) of Section I. Our assumptions, (A1)-(A4) are listed below.

(A1) \(G(x) := \nabla f(x) + B_e(0)\) for some fixed \(\epsilon > 0\). \(\nabla f\) is a continuous function such that \(\|\nabla f(x)\| \leq K(1 + \|x\|)\) for all \(x \in \mathbb{R}^d\), for some \(K > 0\).
(A2) \( \{\gamma(n)\}_{n \geq 0} \) is the step-size (learning rate) sequence such that: \( \gamma(n) > 0 \ \forall n \), \( \sum_{n \geq 0} \gamma(n) = \infty \) and \( \sum_{n \geq 0} \gamma(n)^2 < \infty \).

Without loss of generality we let \( \lim_{n \to \infty} \gamma(n) = 1 \).

Note that \( G \) is an upper-semicontinuous map since \( \nabla f \) is continuous and point-wise bounded. For each \( c \geq 1 \), we define \( G_c(x) \) := \( \{ y/e \ | \ y \in G(x) \} \). Define \( G(\infty)(x) := \overline{\text{conv}}(\text{Limsup}_{c \to \infty} G_c(x)) \), see Section II for the definition of Limsup. Given \( S \subseteq \mathbb{R}^d \), the convex closure of \( S \), is the convex closure of the convex hull of \( S \). It is worth noting that \( \text{Limsup}_{c \to \infty} G_c(x) \) is non-empty for every \( x \in \mathbb{R}^d \).

Further, we show that \( G(\infty) \) is a Marchaud map in Lemma 1. In other words, \( \hat{x}(t) \in -G(\infty)(x(t)) \) has at least one solution that is absolutely continuous, see [1]. Here \( -G(\infty)(x(t)) \) is used to denote the set \( \{ -g \ | \ g \in G(\infty)(x(t)) \} \).

(A3) \( \hat{x}(t) \in -G(\infty)(x(t)) \) has an attractor set \( A \) such that \( A \subseteq B_0(0) \) for some \( a > 0 \) and \( \overline{A}(0) \) is a fundamental neighborhood of \( A \).

Since \( A \subseteq B_0(0) \) is compact, we have that \( \sup_{x \in A} \|x\| < \alpha \).

Let us fix the following sequence of real numbers: \( \sup_{x \in A} \|x\| = \delta_1 < \delta_2 < \delta_3 < \delta_4 < \alpha \).

(A4) Let \( c_n \geq 1 \) be an increasing sequence of integers such that \( c_n \to \infty \) as \( n \to \infty \), let \( x_n \to x \) and \( y_n \to y \) as \( n \to \infty \), such that \( y_n \in G_{c_n}(x_n) \), \( \forall n \), then \( y \in G(\infty)(x) \).

It is worth noting that the existence of a global Lyapunov function for \( \hat{x}(t) \in -G(\infty)(x(t)) \) is sufficient to guarantee that (A3) holds. Further, (A4) is satisfied when \( \nabla f \) is Lipschitz continuous.

Lemma 1. \( G(\infty) \) is a Marchaud map.

Proof. From the definition of \( G(\infty) \) and \( G \) we have that \( G(\infty)(x) \) is convex, compact and \( \sup_{y \in G(x)} \|y\| \leq K(1 + \|x\|) \) for every \( x \in \mathbb{R}^d \). It is left to show that \( G(\infty) \) is an upper-semicontinuous map. Let \( x_n \to x \), \( y_n \to y \) and \( y_n \in G(\infty)(x_n) \), for all \( n \geq 1 \). We need to show that \( y \in G(\infty)(x) \). We present a proof by contradiction. Since \( G(\infty)(x) \) is convex and compact, \( y \notin G(\infty)(x) \) implies that there exists a linear functional on \( \mathbb{R}^d \), say \( f \), such that \( \sup_{x \in G(\infty)(x)} f(x) \leq 0 \). Let \( x_n \to x \) and \( y_n \to y \) as \( n \to \infty \), such that \( y_n \in G_{c_n}(x_n) \), \( \forall n \), then \( y \in G(\infty)(x) \).

B. Relevance of our results

(1) Gradient algorithms with errors have been previously studied by Bertsekas and Tsitsiklis [5]. They impose the following restriction on the estimation errors: \( \|e_n\| \leq \gamma(n)(q+ \phi)\|\nabla f(x(n))\| \) \( \forall n, \) where \( q, \phi > 0 \). If the iterates are stable then \( \|e_n\| \to 0 \). In order to satisfy the aforementioned assumption the choice of step-size may be restricted, thereby affecting the learning rate (when used within the framework of a learning algorithm). In this paper we analyze the more general and practical case of bounded \( \|e_n\| \) which does not necessarily go to zero. Further none of the assumptions used in our paper impose further restrictions on the step-size, other than standard requirements, see (A2).

(2) The main result of Bertsekas and Tsitsiklis [5] states that the GD with errors either diverges almost surely or converges to the minimum set almost surely. An older study by Mangasarian and Solodov [9] shows the exact same result as [5] but for GD without estimation errors (\( e_n = 0 \) \( \forall n \)). The main results of our paper, Theorems 1 & 2 show that if the GD under consideration satisfies (A1)-(A4) then the iterates are stable (bounded almost surely). Further, the algorithm is guaranteed to converge to a given small neighborhood of the minimum set provided the estimation errors are bounded by a constant that is a function of the neighborhood size. To summarize, under the more restrictive setting of [5] and [9] the GD is not guaranteed to be stable, see the aforementioned references, while the assumptions used in our paper are less restrictive and guarantee stability under the more general setting of bounded error GD. It may also be noted that \( \nabla f \) is assumed to be Lipschitz continuous by [5]. This turns out to be sufficient (but not necessary) for (A1) & (A4) to be satisfied.

(3) The analysis of Spall [10] can be used to analyze a variant of GD that uses SPSA as the gradient estimator. Spall introduces a gradient sensitivity parameter \( c_n \) in order to control the estimation error \( e_n \) at stage \( n \). It is assumed that \( c_n \to 0 \) and \( \sum_{n \geq 0} \left( \frac{c_n}{n} \right) < \infty \), see A1, Section III, [10]. Again, this restricts the choice of step-size and affects the learning rate. In this setting our analysis works for the more practical scenario where \( c_n = c \) for all \( n \) i.e., a constant, see Section IV-B.

(4) The important advancements of this paper are the following: (i) Our framework is more general and practical since the errors are not required to go to zero; (ii) We provide
easily verifiable, non-restrictive set of assumptions that ensure almost sure boundedness and convergence of GD and (iii) Our assumptions (A1)-(A4) do not affect the choice of step-size.

(5) Tadić and Doucet [11] showed that GD with bounded non-diminishing errors converges to a small neighborhood of the minimum set. They make the following key assumption: (A) There exists \( p \in (0, 1) \), such that for every compact set \( Q \subset \mathbb{R}^d \) and every \( \epsilon \in [0, \infty) \), \( m(A_{Q, \epsilon}) \leq M_Q \epsilon^p \), where \( A_{Q, \epsilon} = \{ f(x) \mid x \in Q, \| f(x) \| \leq \epsilon \} \) and \( M_Q \in [1, \infty) \).

Note that \( m(A) \) is the Lebesgue measure of the set \( A \subset \mathbb{R}^d \). The above assumption holds if \( d = d_0 \) times differentiable, where \( d < d_0 < \infty \), see [11] for details. In comparison, we only require that the chain recurrent set of \( f \) be a subset of its minimum set. One sufficient condition for this is given in Proposition 4 of Hurley [7].

Remark 1. Suppose the minimum set \( M \) of \( f \), contains the chain recurrent set of \( \dot{x}(t) = -\nabla f(x(t)) \), then it can be shown that GD without errors (\( \epsilon = 0 \) in (4)) will converge to \( M \) almost surely, see [4]. On the other hand suppose there are chain recurrent points outside \( M \), it may converge to this subset (of the chain recurrent set) outside \( M \). In Theorem 2, we will use the upper-semicontinuity of chain recurrent sets (Theorem 3.1 of Benaim, Hofbauer and Sorin [3]), to show that GD with errors will converge to a small neighborhood of the limiting set of the “corresponding GD without errors.” In other words, GD with errors converges to a small neighborhood of the minimum set provided the corresponding GD without errors converges to the minimum set. This will trivially happen if the chain recurrent set of \( \dot{x}(t) = -\nabla f(x(t)) \) is a subset of the minimum set of \( f \), which we implicitly assume is true. Suppose GD without errors does not converge to the minimum set, then it is reasonable to expect that GD with errors may not converge to a small neighborhood of the minimum set.

Suppose \( f \) is continuously differentiable and it’s regular values (i.e., \( x \) for which \( \nabla f(x) \neq 0 \)) 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 [7]. We implicitly assume that an assumption of this kind is satisfied.

IV. PROOF OF STABILITY AND CONVERGENCE

We use (4) to construct the linearly interpolated trajectory, \( \pi(t) \) for \( t \in [0, \infty) \). First, define \( t(0) := 0 \) and \( t(n) := \sum_{i=0}^{n-1} \gamma(i) \) for \( n \geq 1 \). Then, define \( \pi(t(n)) := x_n \) and for \( t \in [t(n), t(n+1)] \), \( \pi(t) \) is the continuous linear interpolation of \( \pi(t_n) \) and \( \pi(t_{n+1}) \). We also construct the following piecewise-constant trajectory \( \overline{\pi}(t) \), \( t \geq 0 \) as follows: \( \overline{\pi}(t) := g(x_n) \) for \( t \in [t(n), t(n+1)] \), \( n \geq 0 \).

We need to divide time, \( 0, \infty \), into intervals of length \( T \), where \( T = T(\delta_2 - \delta_1) + 1 \). Note that \( T(\delta_2 - \delta_1) \) is such that \( \Phi_t(x_0) \in N_{\delta_2 - \delta_1}(A) \) for \( t \geq T(\delta_2 - \delta_1) \), where \( \Phi_t(x_0) \) denotes solution to \( \dot{x}(t) \in G(x(t)) \) at time \( t \) with initial condition \( x_0 \) and \( x_0 \in \text{int}(A) \). Note that \( T(\delta_2 - \delta_1) \) is independent of the initial condition \( x_0 \), see Section II for more details. Dividing time is done as follows: define \( T_0 := 0 \) and \( T_n := \min \{ t(m) : t(m) \geq T_{n-1} + T \}, n \geq 1 \). Clearly, there exists a subsequence \( \{ t(m(n)) \}_{n \geq 0} \) such that \( T_n = t(m(n)) \) \( \forall n \geq 0 \). In what follows we use \( t(m(n)) \) and \( T_n \) interchangeably.

To show stability, we use a projective scheme where the iterates are projected periodically, with period \( T \), onto the closed ball of radius \( a \) around the origin, \( B_a(0) \). Here, the radius \( a \) is given by (A3). This projective scheme gives rise to the following rescaled trajectories \( \hat{x}(\cdot) \) and \( \hat{g}(\cdot) \). First, we construct \( \hat{x}(t) \), \( t \geq 0 \). Let \( t \in [T_n, T_{n+1}) \) for some \( n \geq 0 \), then \( \hat{x}(t) := \frac{\pi(t)}{r(n)} \), where \( r(n) := T(T_n \in [\| \pi(T_n) \| / a \) \( \forall (a \) is defined in (A3)). Also, let \( \hat{x}(T_n^{-}) := \lim_{t \to T_n^{+}} \hat{x}(t) \), \( t \in [T_n, T_{n+1}) \). The ‘rescaled g iterates’ are given by \( \hat{g}(\cdot) := \frac{\pi(t)}{r(n)} \).

Let \( x_n(t), t \in [0, T] \) be the solution (upto time \( T \)) to \( \dot{x}(n)(t) = -\hat{g}(T_n(t)) \), with the initial condition \( x_n(0) = \hat{x}(T_n) \), recall the definition of \( \hat{g}(\cdot) \) from the beginning of Section IV. Clearly, we have

\[
x_n(t) = \hat{x}(T_n) - \int_0^t \hat{g}(T_n + z) dz.
\]

We begin with a simple lemma which essentially claims that \( \{ x_n(t), 0 \leq t \leq T \mid n \geq 0 \} = \{ \hat{x}(T_n(t)), 0 \leq t \leq T \mid n \geq 0 \} \).

The proof is a direct consequence of the definition of \( \hat{g} \) and is hence omitted.

Lemma 2. For all \( n \geq 0 \), we have \( x_n(t) = \hat{x}(T_n + t), t \in [0, T] \).

It directly follows from Lemma 2 that \( \{ x_n(t), t \in [0, T] \mid n \geq 0 \} = \{ \hat{x}(T_n(t)), t \in [0, T] \mid n \geq 0 \} \) in other words, the two families of \( T \)-length trajectories, \( \{ x_n(t), t \in [0, T] \mid n \geq 0 \} \) and \( \{ \hat{x}(T_n(t)), t \in [0, T] \mid n \geq 0 \} \), are really one and the same. When viewed as a subset of \( C([0, T], \mathbb{R}^d), \{ x_n(t), t \in [0, T] \mid n \geq 0 \} \) is equi-continuous and point-wise bounded. Further, from the Arzela-Ascoli theorem we conclude that it is relatively compact. In other words, \( \{ \hat{x}(T_n + t), t \in [0, T] \mid n \geq 0 \} \) is relatively compact in \( C([0, T], \mathbb{R}^d) \).

Lemma 3. Let \( r(n) \uparrow \infty \), then any limit point of \( \{ \hat{x}(T_n + t), t \in [0, T] \mid n \geq 0 \} \) is of the form \( x(t) = x(0) + \int_0^t g_{\infty}(s) ds \), where \( y : [0, T] \rightarrow \mathbb{R}^d \) is a measurable function and \( g_{\infty}(t) \in G_{\infty}(x(t)) \), \( t \in [0, T] \).

Proof. For \( t \geq 0 \), define \( [l] := \max \{ t(k) \mid t(k) \leq t \} \). Observe that for any \( t \in [T_n, T_{n+1}) \), we have \( \hat{g}(\pi(t)) = g_{\infty}(\pi(t)) \) and \( \| \hat{g}(\pi(t)) \| \leq K (1 + \| \pi(t) \|) \), since \( G_r(x_0) \) is a Marchaud map. Since \( \hat{x}(\cdot) \) is the rescaled trajectory obtained by periodically projecting the original iterates onto a compact set, it follows that \( \hat{x}(\cdot) \) is bounded a.s. i.e., \( \sup_{t \in [0, \infty)} \| \hat{x}(t) \| < \infty \) a.s. It now follows from the observation made earlier that \( \sup_{t \in [0, \infty)} \| \hat{g}(\pi(t)) \| < \infty \) a.s.

Thus, we may deduce that there exists a subsequence of \( \mathbb{N} \), say \( \{ l \} \subset \{ n \} \), such that \( \hat{x}(T_n(l)) \rightarrow x(\cdot) \) in \( C([0, T], \mathbb{R}^d) \) and \( \hat{g}(l) \rightarrow g_{\infty}(\cdot) \) weakly in \( L_2([0, T], \mathbb{R}^d) \). From Lemma 2 it follows that \( x(l) \rightarrow x(\cdot) \) in \( C([0, T], \mathbb{R}^d) \). Letting \( r(l) \uparrow \infty \) in

\[
x(l) = x(0) - \int_0^t \hat{g}(r(l) + z) dz, t \in [0, T],
\]
we get \( x(t) = x(0) - \int_0^t g_\infty(z)dz \) for \( t \in [0, T] \). Since \( \| \hat{x}(T_n) \| \leq 1 \) we have \( \| x(0) \| \leq 1 \).

Since \( g(T_l + \cdot) \rightarrow g_\infty(\cdot) \) weakly in \( L_2([0, T], \mathbb{R}^d) \), there exists \( \{l(k)\} \subseteq \{l\} \) such that
\[
\frac{1}{N} \sum_{k=1}^{N} \hat{g}(T_l(k) + \cdot) \rightarrow g_\infty(\cdot) \text{ strongly in } L_2([0, T], \mathbb{R}^d).
\]

Further, there exists \( \{N(m)\} \subseteq \{N\} \) such that
\[
\frac{1}{N(m)} \sum_{k=1}^{N(m)} \hat{g}(T_l(k) + t) \rightarrow g_\infty(t), \ t \in [0, T].
\]

Let us fix \( t_0 \in \{ t \mid \frac{1}{N(m)} \sum_{k=1}^{N(m)} \hat{g}(T_l(k) + t_0) \rightarrow g_\infty(t_0), \ t \in [0, T] \} \), then
\[
\lim_{N(m) \rightarrow \infty} \frac{1}{N(m)} \sum_{k=1}^{N(m)} \hat{g}(T_l(k) + t_0) = g_\infty(t_0).
\]

Since \( G_\infty(x(t_0)) \) is convex and compact (Proposition 1), to show that \( g_\infty(t_0) \in G_\infty(x(t_0)) \) it is enough to show \( d(\hat{g}(T_l(k) + t_0), G_\infty(x(t_0))) = 0 \). Suppose this is not true and \( \exists \epsilon > 0 \) and \( \{n(k)\} \subseteq \{l(k)\} \) such that \( d(\hat{g}(T_n(k) + t_0), G_\infty(x(t_0))) > \epsilon \). Since \( \{\hat{g}(T_n(k) + t_0)\}_{k \geq 1} \) is norm bounded, it follows that there is a convergent subsequence. For convenience, assume \( k \rightarrow \infty \hat{g}(T_n(k) + t_0) = g_0 \), for some \( g_0 \in \mathbb{R}^d \). Since \( \hat{g}(T_n(k) + t_0) \in G_{r_\infty}(\hat{x}(T_n(k) + t_0)) \) and \( \lim_{k \rightarrow \infty} \hat{x}(T_n(k) + t_0) = x(t_0) \), it follows from assumption (A4) that \( g_0 \in G_\infty(x(t_0)) \). This leads to a contradiction.

Note that in the statement of Lemma 3 we can replace \( r(n) \uparrow \infty \) by \( r(k) \uparrow \infty \), where \( \{r(k)\} \) is a subsequence of \( \{r(n)\} \).
Specifically we can conclude that any limit point of \( \{\hat{x}(T_k + t), t \in [0, T]\}_{k \in \mathbb{N}} \) in \( C([0, T], \mathbb{R}^d) \) conditioned on \( r(k) \uparrow \infty \), is of the form \( \hat{x}(t) = x(0) - \int_0^t g_\infty(z)dz \), where \( g_\infty(t) \in G_\infty(x(t)) \) for \( t \in [0, T] \). It should be noted that \( g_\infty(\cdot) \) may be sample path dependent (if \( c_\infty \) is stochastic then \( g_\infty(\cdot) \) is a random variable). Recall that \( \sup_{x \in A} \| x \| = \delta_1 < \delta_2 < \delta_3 < \delta_4 < a \) (see the sentence following (A3) in Section III-A).

**A. Main Results**

We are now ready to prove the two main results of this paper. We begin by showing that (4) is stable (bounded a.s.). In other words, we show that \( \sup \| r(n) \| < \infty \) a.s. Once we show that the iterates are stable we use the main results of Benaim, Hofbauer and Sorin to conclude that the iterates converge to a closed, connected, internally chain transitive and invariant set of \( \hat{x}(t) \in G(x(t)) \).

**Theorem 1.** Under assumptions (A1) – (A4), the iterates given by (4) are stable i.e., \( \sup \| x(n) \| < \infty \) a.s. Further, they converge to a closed, connected, internally chain transitive and invariant set of \( \hat{x}(t) \in G(x(t)) \).

**Proof.** First, we show that the iterates are stable. To do this we start by assuming the negation i.e., \( P(\sup r(n) = \infty) > 0 \). Clearly, there exists \( \{l\} \subseteq \{n\} \) such that \( r(l) \uparrow \infty \). Recall that \( T_l = t(m(l)) \) and that \( [T_l + T] = \max(t(k) + t(l) \leq T_l + T) \).

We have \( \| x(T_l) \| < \delta_2 \) since \( x(\cdot) \) is a solution, up to time \( T_l \), to the DI given by \( \hat{x}(t) \in G_{\infty}(x(t)) \) and \( T = T(\delta_2 - \delta_1) + 1 \). Since the rescaled trajectory is obtained by projecting onto a compact set, it follows that the trajectory is bounded. In other words, \( \sup \| \hat{x}(t) \| \leq K_w < \infty \), where \( K_w \) could be sample path dependent. Now, we observe that there exists \( N \) such that all of the following happen:

(i) \( m(l) \geq N \implies r(l) > R_0 \), [since \( r(l) \uparrow \infty \)]

(ii) \( m(l) \geq N \implies \| \hat{x}(T_l + T) \| < \delta_3 \). [since \( r(l) > R_0 \) and Remark 2]

(iii) \( n \geq N \implies \gamma(n) < \frac{\delta_4 - \delta_3}{K_w(K_w + 1)} \). [since \( \gamma(n) \rightarrow 0 \)]

We have \( \sup_{x \in A} \| x \| = \delta_1 < \delta_2 < \delta_3 < \delta_4 < a \) (see the sentence following (A3) in Section III-A for more details). Let \( m(l) \geq N \) and \( T_{l+1} = t(m(l + 1)) = t(m(l) + k + 1) \) for some \( k > 0 \). If \( T_l + T \neq T_{l+1} \) then \( t(m(l) + k) = T_l + T \), else if \( T_l + T = T_{l+1} \) then \( t(m(l) + k + 1) = T_l + T \). We proceed assuming that \( T_l + T \neq T_{l+1} \) since the other case can be identically analyzed. Recall that \( \hat{x}(T_{l+1}) = \lim_{n \rightarrow \infty} x(n + 1) \), \( t \in [T_n, T_{n+1}] \) and \( n \geq 0 \). Then,
\[
\hat{x}(T_{l+1}) = \hat{x}(t(m(l) + k)) - \gamma(m(l) + k)\hat{g}(t(m(l) + k)).
\]

Taking norms on both sides we get
\[
\| \hat{x}(T_{l+1}) \| \leq \| \hat{x}(t(m(l) + k)) \| + \gamma(m(l) + k)\| \hat{g}(t(m(l) + k)) \|.
\]

As a consequence of the choice of \( N \) we get:
\[
\| \hat{g}(t(m(l) + k)) \| \leq K (1 + \| \hat{g}(t(m(l) + k)) \|) \leq K (1 + K_w).
\]

Hence,
\[
\| \hat{x}(T_{l+1}) \| \leq \| \hat{x}(t(m(l) + k)) \| + \gamma(m(l) + k) K (1 + K_w).
\]

In other words, \( \| \hat{x}(T_{l+1}) \| < \delta_4 \). Further,
\[
\frac{\| \hat{x}(T_{l+1}) \|}{\| \hat{x}(T_l) \|} = \frac{\| \hat{x}(T_{l+1}) \|}{\| \hat{x}(T_l) \|} < \frac{\delta_4}{a} < 1.
\]

It follows from (7) that \( \| \hat{x}(T_{l+1}) \| < \frac{\delta_4}{a} \| \hat{x}(T_l) \| \) if \( \| \hat{x}(T_{l+1}) \| > R_0 \). From Corollary 1 and the aforementioned we get that the trajectory falls at an exponential rate till it enters...
Theorem 3.1 [3] Given $\delta > 0$, there exists $\epsilon(\delta) > 0$ such that the chain recurrent set of $\dot{x}(t) = -\nabla f(x(t)) + \mathcal{B}_n(0)$ is within the $\delta$-open neighborhood of the chain recurrent set of $\dot{x}(t) = -\nabla f(x(t))$ for all $r \leq \epsilon(\delta)$.

Theorem 2. Given $\delta > 0$, there exists $\epsilon(\delta) > 0$ such that the GD with bounded errors given by (4) converges to $N^\delta(M)$, the $\delta$-neighborhood of the minimum set of $f$, provided $\epsilon < \epsilon(\delta)$. Here $\epsilon$ is the bound for estimation errors from assumption (A1).

Proof. As stated in Remark 1, the chain recurrent set of $\dot{x}(t) = -\nabla f(x(t))$ is assumed to be a subset of the minimum set of $f$. Note that the iterates given by (4) track a solution to $\dot{x}(t) = -\nabla f(x(t)) + \mathcal{B}_n(0)$. It follows from Theorem 3.1 [3] that (4) converge to a $\delta$-neighborhood of the chain recurrent set provided $\epsilon < \epsilon(\delta)$. In other words, GD with errors converges to a small neighborhood of the minimum set provided GD without errors is guaranteed to converge to the minimum set.
A. Exp.1: SPSA with constant sensitivity parameters (SPSA-C)

First we consider SPSA with constant sensitivity parameters to find the minimum set of \( f \). This scheme is given by (8) but with \( c_n = c \) for all \( n \), and we refer to it as SPSA-C.

Parameter settings:
(1) The positive definite matrix \( Q \) and the starting point \( x_0 \) were randomly chosen. (2) The dimension \( d = 10 \). The number of iterations of SPSA-C was 8000. (3) \( c \) was varied from 0.1 to 10. For each value of \( c \), SPSA-C was run for 8000 iterations and \( \|x_{8000}\| \) was recorded. Since origin is the unique global minimizer of \( f \), \( \|x_{8000}\| \) records the distance of the iterate after 8000 iterations from the origin. (4) For \( 0 \leq n \leq 7999 \), we chose the following step-size sequence: \( a(n) = \frac{1}{(n \mod 8000)+100} \), \( n \geq 1 \). This step-size sequence seems to expedite the convergence of the iterates to the minimum set. We were able to use this sequence since our framework does not impose extra restrictions on step-sizes, unlike [10].

Since we keep the sensitivity parameters fixed the implementation was greatly simplified. Based on the theory presented in this paper, for larger values of \( c \) one expects the iterates to be farther from the origin than for smaller values of \( c \). This theory is corroborated by the experiment illustrated in Fig. 1. Note that to generate \( Q \) we first randomly generate a column-orthonormal matrix \( U \) and let \( Q := U \Sigma U^T \), where \( \Sigma \) is a diagonal matrix with strictly positive entries. To generate \( U \), we sample its entries independently from a Gaussian distribution and then apply Gram-Schmidt orthogonalization to the columns. Fig. 1 shows the average performance of 20 independent simulation runs (for each \( c \) ) of the experiment, where \( Q \) and \( x_0 \) were randomly chosen for each run; Fig. 2 shows two sample runs. In Fig. 1 and 2 the \( x \)-axis represents the values of \( c \) ranging from 0.1 to 10 in steps of 0.01. The \( y \)-axis in Fig. 1 represents the logarithm of the average of corresponding distances from the origin after 8000 iterations i.e., \( \log\left(1/20 \sum_{i=1}^{20} \|x_{8000}^i\|\right) \), where \( x_{8000}^i \) is the iterate-value after 8000 runs from the \( i \)-th simulation. The \( y \)-axis in Fig. 2 represents the logarithm of the corresponding distances from the origin after 8000 iterations i.e., \( \log(\|x_{8000}\|) \). Note that for \( c \) close to 0, \( x_{8000} \in B_{e^{-38}}(0) \) while for \( c \) close to 10, \( x_{8000} \in B_{e^{-32}}(0) \) only. Also note that the graph has a series of “steep rises” followed by “plateaus”. These indicate that for values of \( c \) within the same plateau the iterate converges to the same neighborhood of the origin. As stated earlier for larger values of \( c \) the iterates are farther from the origin than for smaller values of \( c \).

B. Exp.2: GD with constant gradient errors

For the second experiment we ran the following recursion for 1000 iterations:

\[ x_{n+1} = x_n + \frac{1}{n} (Q x_n + \epsilon) \]

where \( \epsilon \) is the constant noise-vector added at each stage. \( \epsilon \) is a “small” noise-vector added at each stage does the iterate sequence still converge to a small neighborhood of the origin or do the iterates diverge? It can be verified that (9) satisfies (A1)-(A4) of Section III-A for any \( \epsilon \). Hence it follows from Theorem 1 that the iterates are stable and do not diverge. In other words, the addition of such a noise does not accumulate and force the iterates to diverge. As in the first experiment we expect the iterates to be farther from the origin for larger values of \( \epsilon \). This is evidenced by the plots in Fig. 3 and 4. As before,
VI. Extensions and conclusions

In this paper we have provided sufficient conditions for stability and convergence (to a small neighborhood of the minimum set) of GD with bounded and (possibly) non-diminishing errors. To the best of our knowledge our analysis of GD with errors is new to the literature. In addition to being easily verifiable, the assumptions presented herein do not affect the choice of step-size. Finally, experimental results presented in Section V are seen to validate the theory. An important step in the analysis of ‘GD with errors’ is to show stability (almost sure boundedness) of the iterates. It is worth noting that this step is not straightforward even in the case of asymptotically vanishing errors, i.e., $\epsilon_n \to 0$ as $n \to \infty$.

An extension to our main results is the introduction of an additional martingale noise term $M_{n+1}$ at stage $n$. Our results will continue to hold provided $\sum_{n\geq 0} \gamma(n)M_{n+1} < \infty$ a.s. Another extension is to analyze implementations of GD using Newton’s method with bounded, (possibly) non-diminishing errors. To see this, define $G(x) := H(x)^{-1}\nabla f(x) + B_0(0)$ in (A1); $G_{\infty}$ changes accordingly. Here $H(x)$ (assumed positive definite) denotes the Hessian evaluated at $x$. Theorems 1 & 2 hold under this new definition of $G$ and appropriate modifications of (A1) – (A4). Our analysis is valid in situations where the function $f$ is not differentiable at some points, however, the error in the gradient estimate at any stage is bounded. An interesting future direction will be to derive convergence rates of gradient schemes with non-diminishing errors. More generally, it would be interesting to derive convergence rates of stochastic approximation algorithms with set-valued mean-fields.

REFERENCES