Analysis of Gradient Descent Methods With Nondiminishing Bounded Errors

Arunselvan Ramaswamy and Shalabh Bhatnagar

***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) nondiminishing 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 asymptotically 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 and Solodov, Bertsekas and Tsitsiklis, and Tadić and Doucet. Using our framework, a simple yet effective implementation of GD using simultaneous perturbation stochastic approximations, 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*—Differential inclusions (DIs), gradient descent (GD) methods, nondiminishing errors, stability and convergence, stochastic approximation algorithms.**

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 (GD) method is often employed to find such a minimum:

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

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* [ϵ_n in (2)]. This is particularly true when the form of f or ∇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 [6, Ch. 4.4]. 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

$$x_{n+1} = x_n - \gamma_n(\nabla f(x_n) + \epsilon_n). \tag{2}$$

In the above equation ϵ_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: $J(x) = \int w\, d(F(w \mid x)) = E[W(x)]$, where $W(x)$ is the "waiting time" random variable with distribution $F(\cdot \mid x)$, 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, $\{F(\cdot \mid x) \mid x \in \mathbb{R}^d\}$, exactly. In such scenarios, one often works with approximate definitions of F, which, in turn, lead to approximate gradients, i.e., gradients with errors. More generally, 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 (see 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 the authors of [1]–[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 its analysis to work. Furthermore, 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

Manuscript received April 12, 2017; revised April 13, 2017 and August 14, 2017; accepted August 19, 2017. Date of publication August 24, 2017; date of current version April 24, 2018. This work was supported in part by Robert Bosch Center for Cyber-Physical Systems at the India Institute of Science, Bengaluru. Recommended by Associate Editor L. H. Lee. *(Corresponding author: Arunselvan Ramaswamy.)*

The authors are with the Department of Computer Science and Automation, Indian Institute of Science, Bengaluru 560012, India (e-mail: arunselvan@csa.iisc.ernet.in; shalabh@csa.iisc.ernet.in).

Digital Object Identifier 10.1109/TAC.2017.2744598

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] and [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, i.e., 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 nondiminishing errors, we show the following: There exists* $\epsilon(\delta) > 0$ *such that the iterates are stable and converge to the* δ*-neighborhood of the minimum set (*δ *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 $\{y_n\}_{n\geq 1}$ (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: 1) for each $x \in \mathbb{R}^n$, $H(x)$ is convex and compact; 2) *(pointwise boundedness)* for each $x \in \mathbb{R}^n$, $\sup_{w\in H(x)} \|w\| < K(1 + \|x\|)$ for some $K > 0$; and 3) H is *upper-semicontinuous*.

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

$$\dot{x} \in H(x) \tag{3}$$

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

$\Phi_t(x) = \{\mathbf{x}(t) \mid \mathbf{x} \in \sum, \mathbf{x}(0) = x\}$. 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 $\mathbf{x}$ with $\mathbf{x}(0) = x$ is given by $L(x) = \bigcap_{t\geq 0} \overline{\mathbf{x}([t, +\infty))}$.

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

Open and closed neighborhoods of a set: Let $x \in \mathbb{R}^d$ and $A \subseteq \mathbb{R}^d$, then $d(x, A) := \inf\{\|a - y\| \mid y \in A\}$. We define the δ*-open neighborhood* of A by $N^\delta(A) := \{x \mid d(x, A) < \delta\}$. The δ*-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, $B_r(0) = \{x \mid \|x\| < r\}$ and $\overline{B}_r(0) = \{x \mid \|x\| \leq r\}$.

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 exists n and $\Phi^1, \ldots, \Phi^n$ that are n solutions to the DI $\dot{x}(t) \in h(x(t))$, 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 $\Phi^i_{t_i}(x_i) \in N^\epsilon(x_{i+1})$ and $\Phi^i_{[0,t_i]}(x_i) \subset M$ for $1 \leq i \leq n$. The sequence $(x_1(= x), \ldots, x_{n+1}(= y))$ is called an (ϵ, 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 and fundamental neighborhood: $A \subseteq \mathbb{R}^d$ is *attracting* if it is compact and there exists a neighborhood U such that for any $\epsilon > 0$, $\exists\ T(\epsilon) \geq 0$ with $\Phi_{[T(\epsilon),+\infty)}(U) \subset N^\epsilon(A)$. 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 $B(A) = \{x \mid \omega_\Phi(x) \subset A\}$.

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

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 $\text{Limsup}_{n\to\infty} K_n := \{y \mid \underline{\lim}_{n\to\infty} d(y, K_n) = 0\}$.

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:

$$x_{n+1} = x_n - \gamma(n) g(x_n) \tag{4}$$

where $g(x_n) \in G(x_n)\ \forall\ n$ and $G(x) := \nabla f(x) + \overline{B}_\epsilon(0)$, $x \in \mathbb{R}^d$. In other words, the gradient estimate at stage n, $g(x_n)$, belongs to an ϵ-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 as follows.

A1) $G(x) := \nabla f(x) + \overline{B}_\epsilon(0)$ for some fixed $\epsilon > 0$. ∇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 $\sup_n \gamma(n) \leq 1$.

Note that G is an upper-semicontinuous map since ∇f is continuous and pointwise bounded. For each $c \geq 1$, we define $G_c(x) := \{y/c \mid y \in G(cx)\}$. Define $G_\infty(x) := \overline{co}\,\langle \text{Limsup}_{c\to\infty} G_c(x)\rangle$; see Section II for the definition of Limsup. Given $S \subseteq \mathbb{R}^d$, the convex closure of S, denoted by $\overline{\text{co}}\langle S\rangle$, is the closure of the convex hull of S. It is worth noting that $\text{Limsup}_{c\to\infty} G_c(x)$ is nonempty for every $x \in \mathbb{R}^d$. Furthermore, we show that G_∞ is a Marchaud map in Lemma 1. In other words, $\dot{x}(t) \in -G_\infty(x(t))$ has at least one solution that is absolutely continuous; see [1]. Here, $-G_\infty(x(t))$ is used to denote the set $\{-g \mid g \in G_\infty(x(t))\}$.

A3) $\dot{x}(t) \in -G_\infty(x(t))$ 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 $\sup_{x\in\mathcal{A}}\|x\| < a$. Let us fix the following sequence of real numbers: $\sup_{x\in\mathcal{A}}\|x\| = \delta_1 < \delta_2 < \delta_3 < \delta_4 < a$.

A4) Let $c_n \geq 1$ be an increasing sequence of integers such that $c_n \uparrow \infty$ as $n \to \infty$. Furthermore, 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 $\dot{x}(t) \in -G_\infty(x(t))$ *is sufficient to guarantee that A3 holds. Furthermore, A4 is satisfied when* ∇f *is Lipschitz continuous.*

Lemma 1: G_∞ is a Marchaud map.

Proof: From the definition of G_∞ and G, we have that $G_\infty(x)$ is convex, compact, and $\sup_{y\in G(x)}\|y\| \le K(1+\|x\|)$ for every $x \in \mathbb{R}^d$. It is left to show that G_∞ is an upper-semicontinuous map. Let $x_n \to x$, $y_n \to y$, and $y_n \in G_\infty(x_n)$, for all $n \ge 1$. 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 $\sup_{z\in G_\infty(x)} f(z) \le \alpha - \epsilon$ and $f(y) \ge \alpha + \epsilon$, for some $\alpha \in \mathbb{R}$ and $\epsilon > 0$. Since $y_n \to y$, there exists $N > 0$ such that for all $n \ge N$, $f(y_n) \ge \alpha + \frac{\epsilon}{2}$. In other words, $G_\infty(x) \cap [f \ge \alpha + \frac{\epsilon}{2}] \ne \phi$ for all $n \ge N$. We use the notation $[f \ge a]$ to denote the set $\{x \mid f(x) \ge a\}$. For the sake of convenience, let us denote the set $\text{Limsup}_{c\to\infty} G_c(x)$ by $A(x)$, where $x \in \mathbb{R}^d$. We claim that $A(x_n) \cap [f \ge \alpha + \frac{\epsilon}{2}] \ne \phi$ for all $n \ge N$. We prove this claim later; for now, we assume that the claim is true and proceed. Pick $z_n \in A(x_n) \cap [f \ge \alpha + \frac{\epsilon}{2}]$ for each $n \ge N$. It can be shown that $\{z_n\}_{n\ge N}$ is norm bounded and, hence, contains a convergent subsequence $\{z_{n(k)}\}_{k\ge 1} \subseteq \{z_n\}_{n\ge N}$. Let $\lim_{k\to\infty} z_{n(k)} = z$. Since $z_{n(k)} \in \text{Limsup}_{c\to\infty}(G_c(x_{n(k)}))$, $\exists\ c_{n(k)} \in \mathbb{N}$ such that $\|w_{n(k)} - z_{n(k)}\| < \frac{1}{n(k)}$, where $w_{n(k)} \in G_{c_{n(k)}}(x_{n(k)})$. We choose the sequence $\{c_{n(k)}\}_{k\ge 1}$ such that $c_{n(k+1)} > c_{n(k)}$ for each $k \ge 1$.

We have the following: $c_{n(k)} \uparrow \infty$, $x_{n(k)} \to x$, $w_{n(k)} \to z$, and $w_{n(k)} \in G_{c_{n(k)}}(x_{n(k)})$, for all $k \ge 1$. It follows from assumption A4 that $z \in G_\infty(x)$. Since $z_{n(k)} \to z$ and $f(z_{n(k)}) \ge \alpha + \frac{\epsilon}{2}$ for each $k \ge 1$, we have that $f(z) \ge \alpha + \frac{\epsilon}{2}$. This contradicts the earlier conclusion that $\sup_{z\in h_\infty(x)} f(z) \le \alpha - \epsilon$.

It remains to prove that $A(x_n) \cap [f \ge \alpha + \frac{\epsilon}{2}] \ne \phi$ for all $n \ge N$. If this were not true, then $\exists\{m(k)\}_{k\ge 1} \subseteq \{n \ge N\}$ such that $A(x_{m(k)}) \subseteq [f < \alpha + \frac{\epsilon}{2}]$ for all k. It follows that $G_\infty(x_{m(k)}) = \overline{\text{co}}(A(x_{m(k)})) \subseteq [f \le \alpha + \frac{\epsilon}{2}]$ for each $k \ge 1$. Since $y_{n(k)} \to y$, $\exists N_1$ such that for all $n(k) \ge N_1$, $f(y_{n(k)}) \ge \alpha + \frac{3\epsilon}{4}$. This is a contradiction. ■

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: $\|\epsilon_n\| \le \gamma(n)(q + p\|\nabla f(x_n)\|) \forall\ n$, where $p, q > 0$. If the iterates are stable, then $\|\epsilon_n\| \to 0$. 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 $\|\epsilon_n\|$, which does not necessarily go to zero. Furthermore, *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 ($\epsilon_n = 0\ \forall\ n$). The main results of our paper, i.e., Theorems 1 and 2, show that if the GD under consideration satisfies A1–A4, then the iterates are stable (bounded almost surely). Furthermore, the algorithm is guaranteed to converge to a *given small neighborhood of the minimum set*, provided that 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 ∇f is assumed to be Lipschitz continuous by Bertsekas and Tsitsiklis [5]. This turns out to be sufficient (but not necessary) for A1 and 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 ϵ_n at stage n. It is assumed that $c_n \to 0$ and $\sum_{n\ge 0}\left(\frac{\gamma(n)}{c_n}\right)^2 < \infty$; see [10, A1, Sec. III]. 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: a) Our framework is more general and practical since the errors are not required to go to zero; b) we provide easily verifiable nonrestrictive set of assumptions that ensure almost sure boundedness and convergence of GD; and c) our assumptions A1–A4 do not affect the choice of step size.

5) Tadić and Doucet [11] showed that GD with bounded nondiminishing 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 $\epsilon \in [0, \infty)$, $m(A_{Q,\epsilon}) \le M_Q\,\epsilon^p$, where $A_{Q,\epsilon} = \{f(x) \mid x \in Q, \|f(x)\| \le \epsilon\}$ 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 f is 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 [7, Proposition 4].

Remark 1: Suppose the minimum set $\mathcal{M}$ 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 $\mathcal{M}$ almost surely; see [4]. On the other hand, suppose there are chain recurrent points outside $\mathcal{M}$; it may converge to this subset (of the chain recurrent set) outside $\mathcal{M}$. In Theorem 2, we will use the upper-semicontinuity of chain recurrent sets (see [3, Th. 3.1]) 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 that 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 its regular values (i.e., x for which $\nabla f(x) \ne 0$) are dense in $\mathbb{R}^d$, then the chain recurrent set of f is a subset of its minimum set; see [7, Proposition 4]. 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 $\overline{x}(t)$ for $t \in [0, \infty)$. First, define $t(0) := 0$ and $t(n) := \sum_{i=0}^{n-1}\gamma(i)$ for $n \ge 1$. Then, define $\overline{x}(t(n)) := x_n$, and for $t \in [t(n), t(n+1)]$, $\overline{x}(t)$ is the continuous linear interpolation of $\overline{x}(t_n)$ and $\overline{x}(t_{n+1})$. We also construct the following piecewise constant trajectory $\overline{g}(t)$, $t \ge 0$, as follows: $\overline{g}(t) := g(x_n)$ for $t \in [t(n), t(n+1))$, $n \ge 0$.

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 $\Phi_t(x_0) \in N^{\delta_2-\delta_1}(\mathcal{A})$ for $t \ge T(\delta_2 - \delta_1)$, where $\Phi_t(x_0)$ denotes solution to $\dot{x}(t) \in G_\infty(x(t))$ at time t with initial condition x_0 and $x_0 \in \overline{B}_a(0)$. 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 $T_n := \min\{t(m) : t(m) \ge T_{n-1} + T\}, n \ge 1$. Clearly, there exists a subsequence $\{t(m(n))\}_{n\ge 0}$ of $\{t(n)\}_{n\ge 0}$ such that $T_n = t(m(n))\ \forall\ n \ge 0$. 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, $\overline{B}_a(0)$. 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, $\hat{x}(t) := \frac{\overline{x}(t)}{r(n)}$, where $r(n) = \frac{\|\overline{x}(T_n)\|}{a} \vee 1$ (a is defined in A3. Also, let $\hat{x}(T_{n+1}^-) := \lim_{t \uparrow T_{n+1}} \hat{x}(t)$, $t \in [T_n, T_{n+1})$. The "rescaled g iterates" are given by $\hat{g}(t) := \frac{\overline{g}(t)}{r(n)}$.

Let $x^n(t)$, $t \in [0,T]$ be the solution (up to 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

$$x^n(t) \;=\; \hat{x}(T_n) - \int_0^t \hat{g}(T_n + z)\, dz. \tag{5}$$

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 $x^n(t) = \hat{x}(T_n + t)$, where $t \in [0,T]$.

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 equicontinuous and pointwise bounded. Furthermore, 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 $\{\hat{x}(T_n + t), t \in [0,T] : n \geq 0\}$ is of the form $x(t) = x(0) + \int_0^t g_\infty(s)\; ds$, where $y : [0,T] \to \mathbb{R}^d$ is a measurable function and $g_\infty(t) \in G_\infty(x(t))$, $t \in [0,T]$.

Proof: For $t \geq 0$, define $[t] := \max\{t(k) \mid t(k) \leq t\}$. Observe that for any $t \in [T_n, T_{n+1})$, we have $\hat{g}(t) \in G_{r(n)}(\hat{x}([t]))$ and $\|\hat{g}(t)\| \leq K\,(1 + \|\hat{x}([t])\|)$, since $G_{r(n)}$ 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 $\sup_{t \in [0,\infty)} \|\hat{g}(t)\| < \infty$ a.s.

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

$$x^l(t) = x^l(0) - \int_0^t \hat{g}(t(m(l) + z))\, dz,\; t \;\in\; [0,T]$$

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 $\hat{g}(T_l + \cdot) \to g_\infty(\cdot)$ weakly in $L_2\left([0,T], \mathbb{R}^d\right)$, there exists $\{l(k)\} \subseteq \{l\}$ such that

$$\frac{1}{N} \sum_{k=1}^{N} \hat{g}(T_{l(k)} + \cdot) \to g_\infty(\cdot) \text{ strongly in } L_2\left([0,T], \mathbb{R}^d\right).$$

Furthermore, there exists $\{N(m)\} \subseteq \{N\}$ such that

$$\frac{1}{N(m)} \sum_{k=1}^{N(m)} \hat{g}(T_{l(k)} + \cdot) \to g_\infty(\cdot) \text{ a.e. on } [0,T].$$

Let us fix $t_0 \in \{t \mid \frac{1}{N(m)} \sum_{k=1}^{N(m)} \hat{g}(T_{l(k)} + t) \to g_\infty(t),\; t \in [0,T]\}$; then

$$\lim_{N(m) \to \infty} \frac{1}{N(m)} \sum_{k=1}^{N(m)} \hat{g}(T_{l(k)} + t_0) = g_\infty(t_0).$$

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 $\lim_{l(k) \to \infty} d\left(\hat{g}(T_{l(k)} + t_0), G_\infty(x(t_0))\right) = 0$. Suppose this is not true and $\exists\, \epsilon > 0$ and $\{n(k)\} \subseteq \{l(k)\}$ such that $d\left(\hat{g}(T_{n(k)} + t_0), G_\infty(x(t_0))\right) > \epsilon$. 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 $\lim_{k \to \infty} \hat{g}(T_{n(k)} + t_0) = g_0$, for some $g_0 \in \mathbb{R}^d$. Since $\hat{g}(T_{n(k)} + t_0) \in G_{r(n(k))}(\hat{x}([T_{n(k)} + t_0]))$ and $\lim_{k \to \infty} \hat{x}([T_{n(k)} + t_0]) = x(t_0)$, 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 $\{\hat{x}(T_k + t), t \in [0,T]\}_{\{k\} \subseteq \{n\}}$ in $C([0,T], \mathbb{R}^d)$, conditioned on $r(k) \uparrow \infty$, is of the form $x(t) = x(0) - \int_0^t g_\infty(z)\, dz$, 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 ϵ_n is stochastic then $g_\infty(\cdot)$ is a random variable). Recall that $\sup_{x \in \mathcal{A}} \|x\| = \delta_1 < \delta_2 < \delta_3 < \delta_4 < a$ (see the sentence following A3 in Section III-A). The following is an immediate corollary of Lemma 3.

Corollary 1: $\exists\, 1 < R_0 < \infty$ such that $\forall\, r(l) > R_0$, $\|\hat{x}(T_l + \cdot) - x(\cdot)\| < \delta_3 - \delta_2$, where $\{l\} \subseteq \mathbb{N}$ and $x(\cdot)$ is a solution (up to time T) of $\dot{x}(t) \in -G_\infty(x(t))$ such that $\|x(0)\| \leq 1$. The form of $x(\cdot)$ is as given by Lemma 3.

Proof: Assume to the contrary that $\exists\, r(l) \uparrow \infty$ such that $\hat{x}(T_l + \cdot)$ is at least $\delta_3 - \delta_2$ away from any solution to the DI. It follows from Lemma 3 that there exists a subsequence of $\{\hat{x}(T_l + t), 0 \leq t \leq T \;:\; l \subseteq \mathbb{N}\}$ guaranteed to converge, in $C([0,T], \mathbb{R}^d)$, to a solution of $\dot{x}(t) \in -G_\infty(x(t))$ such that $\|x(0)\| \leq 1$. This is a contradiction. ■

Remark 2: It is worth noting that R_0 may be sample path dependent. Since $T = T(\delta_2 - \delta_1) + 1$, we get $\|\hat{x}([T_l + T])\| < \delta_3$ for all T_l such that $\|\overline{x}(T_l)\| (= r(l)) > R_0$.

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 $\sup_n \|r(n)\| < \infty$ a.s. Once we show that the iterates are stable, we use the main results of Benaïm, Hofbauer, and Sorin to conclude that the iterates converge to a closed, connected, internally chain transitive and invariant set of $\dot{x}(t) \in G(x(t))$.

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

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

We have $\|x(T)\| < \delta_2$ since $x(\cdot)$ is a solution, up to time T, to the DI given by $\dot{x}(t) \in G_\infty(x(t))$ 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, $\sup_{t \geq 0} \|\hat{x}(t)\| \leq K_w < \infty$, where K_w could be sample path dependent. Now, we observe that there exists N such that all of the following happen:

1) $m(l) \geq N \Rightarrow r(l) \;>\; R_0$ (since $r(l) \uparrow \infty$);

2) $m(l) \geq N \Rightarrow \|\hat{x}([T_l + T])\| < \delta_3$ (since $r(l) > R_0$ and Remark 2);
3) $n \geq N \Rightarrow \gamma(n) < \frac{\delta_4 - \delta_3}{K(1+K_\omega)}$ (since $\gamma(n) \to 0$).

We have $\sup_{x \in \mathcal{A}} \|x\| = \delta_1 < \delta_2 < \delta_3 < \delta_4 < a$ (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 $\hat{x}(T_{n+1}^-) = \lim_{t \uparrow t(m(n+1))} \hat{x}(t), t \in [T_n, T_{n+1})$ 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

$$\|\hat{g}(t(m(l)+k))\| \leq K(1 + \|\hat{x}(t(m(l)+k)\|) \leq K(1+K_\omega). \tag{6}$$

Hence

$$\|\hat{x}(T_{l+1}^-)\| \leq \|\hat{x}(t(m(l)+k))\| + \gamma(m(l)+k)K(1+K_\omega).$$

In other words, $\|\hat{x}(T_{l+1}^-)\| < \delta_4$. Furthermore

$$\frac{\|\overline{x}(T_{l+1})\|}{\|\overline{x}(T_l)\|} = \frac{\|\hat{x}(T_{l+1}^-)\|}{\|\hat{x}(T_l)\|} < \frac{\delta_4}{a} < 1. \tag{7}$$

It follows from (7) that $\|\overline{x}(T_{n+1})\| < \frac{\delta_4}{a}\|\overline{x}(T_n)\|$ if $\|\overline{x}(T_n)\| > R_0$. From Corollary 1 and the aforementioned, we get that the trajectory falls at an exponential rate till it enters $\overline{B}_{R_0}(0)$. Let $t \leq T_l, t \in [T_n, T_{n+1})$ and $n+1 \leq l$, be the last time that $\overline{x}(t)$ jumps from within $\overline{B}_{R_0}(0)$ to the outside of the ball. It follows that $\|\overline{x}(T_{n+1})\| \geq \|\overline{x}(T_l)\|$. Since $r(l) \uparrow \infty, \overline{x}(t)$ would be forced to make larger and larger jumps within an interval of length $T+1$. This leads to a contradiction since the maximum jump size within any fixed time interval can be bounded using the *Gronwall* inequality. Thus, the iterates are shown to be stable.

It now follows from [2, Th. 3.6 and Lemma 3.8] that the iterates converge almost surely to a closed, connected, internally chain transitive and invariant set of $\dot{x}(t) \in G(x(t))$. ■

Now that the GD with nondiminishing bounded errors, given by (4), is shown to be stable (bounded a.s.), we proceed to show that these iterates in fact converge to an arbitrarily small neighborhood of the minimum set. The proof uses [3, Th. 3.1] that we state below. First, we make a minor comment on the limiting set of GD with errors.

Recall from Remark 1 that the chain recurrent set of $\dot{x}(t) = -\nabla f(x(t))$ is a subset of $\mathcal{M}$, where $\mathcal{M}$ is the minimum set of f. We consider two cases: a) $\mathcal{M}$ is the unique global attractor of $\dot{x}(t) = -\nabla f(x(t))$; and b) $\mathcal{M}$ comprises of multiple local attractors. Suppose we are in case a; it can be shown that any compact neighborhood, $\mathcal{M} \subseteq \mathcal{K} \subset \mathbb{R}^d$, is a fundamental neighborhood of $\mathcal{M}$. It follows from Theorem 1 that the iterates are bounded almost surely. In other words, $\overline{x}(t) \in \mathcal{K}_0, \forall\, t \geq 0$, for some compact set $\mathcal{K}_0$, that could be sample path dependent, such that $\mathcal{M} \subseteq \mathcal{K}_0$. In this case, GD with errors is expected to converge to a small neighborhood of $\mathcal{M}$. Suppose we are in case b; we need to consider $\mathcal{M}' \subseteq \mathcal{M}$ such that the aforementioned $\mathcal{K}_0$ is a fundamental neighborhood of it. In this case, GD with errors is expected to converge to a small neighborhood of $\mathcal{M}'$.

We are now ready to present [3, Th. 3.1]. The statement has been interpreted to the setting of this section for the sake of convenience.

([3, Th. 3.1]): Given $\delta > 0$, there exists $\epsilon(\delta) > 0$ such that the chain recurrent set of $\dot{x}(t) = -\nabla f(x(t)) + \overline{B}_r(0)$ is within the δ-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 $N^\delta(\mathcal{M})$, the δ-neighborhood of the minimum set of f, provided $\epsilon < \epsilon(\delta)$. Here, ϵ 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 $\dot{x}(t) \in -\left(\nabla f(x(t)) + \overline{B}_\epsilon(0)\right)$. It follows from *[3, Th. 3.1]* that (4) converges to a δ-neighborhood of the chain recurrent set, provided that $\epsilon < \epsilon(\delta)$. In other words, GD with errors converges to a small neighborhood of the minimum set, provided that GD without errors is guaranteed to converge to the minimum set. ■

B. Implementing GD Methods Using SPSA

Gradient estimators are often used in the implementation of GD methods such as SPSA [10]. When using SPSA, the update rule for the ith coordinate is given by

$$x_{n+1}^i = x_n^i - \gamma(n)\left(\frac{f(x_n + c_n\Delta_n) - f(x_n - c_n\Delta_n)}{2c_n\Delta_n^i}\right) \tag{8}$$

where $x_n = \left(x_n^1, \ldots, x_n^d\right)$ is the underlying parameter, $\Delta_n = \left(\Delta_n^1, \ldots, \Delta_n^d\right)$ is a sequence of perturbation random vectors such that Δ_n^i, $1 \leq i \leq d$, $n \geq 0$ are independent and identically distributed. It is common to assume Δ_n^i to be symmetric Bernoulli distributed, taking values ± 1 $w.p.$ $1/2$. The sensitivity parameter c_n is such that the following are assumed: $c_n \to 0$ as $n \to \infty$; $\sum_{n \geq 0}\left(\frac{\gamma(n)}{c_n}\right)^2 < \infty$, see [10, A1]. Furthermore, c_n needs to be chosen such that the estimation errors go to zero. This, in particular, could be difficult since the form of the function f is often unknown. One may need to run experiments to find each c_n. Also, smaller values of c_n in the initial iterates tends to blow up the variance, which, in turn, affects convergence. For these reasons, in practice, one often lets $c_n := c$ (a small constant) for all n. If we assume additionally that the second derivative of f is bounded, then it is easy to see that the estimation errors are bounded by $\epsilon(c)$ such that $\epsilon(c) \to 0$ as $c \to 0$. *Thus, keeping c_n fixed to c forces the estimation errors to be bounded at each stage. In other words, SPSA with a constant sensitivity parameter falls under the purview of the framework presented in this paper.* Also, it is worth noting that the iterates are assumed to be stable (bounded a.s.) in [10]. However, in our framework, stability is shown under verifiable conditions even when $c_n = c$, $n \geq 0$.

We arrive at the important question of how to choose this constant c in practice such that fixing $c_n := c$ we still get the following: 1) the iterates are stable and 2) GD implemented in this manner converges to the minimum set. Suppose that the simulator wants to ensure that the iterates converge to a δ-neighborhood of the minimum set, i.e., $N^\delta(\mathcal{M})$; then, it follows from Theorem 2 that there exists $\epsilon(\delta) > 0$ such that the GD converges to $N^\delta(\mathcal{M})$ provided the estimation error at each stage is bounded by $\epsilon(\delta)$. Now, c is chosen such that $\epsilon(c) \leq \epsilon(\delta)$. The simulation is carried out by fixing the sensitivity parameters to this c. As stated earlier, one may need to carry out experiments to find such a c. However, the advantage is that we only need to do this once before starting the simulation. Also, the iterates are guaranteed to be stable and converge to the δ-neighborhood of the minimum set, provided that A1–A4 are satisfied.

V. Experimental Results

The experiments presented in this section consider a quadratic objective function $f : \mathbb{R}^d \to \mathbb{R}$ with $f(x) := x^T Q x$, where Q is a positive-definite matrix. The origin is the unique global minimizer of

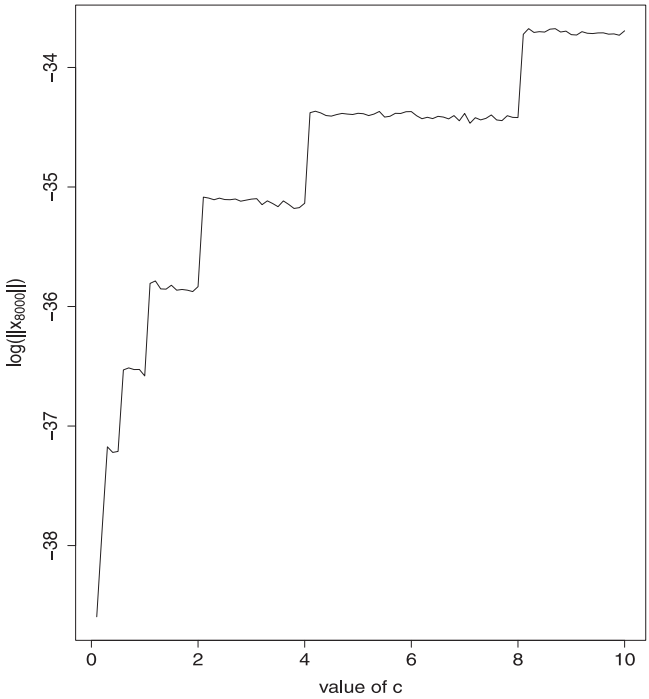

Fig. 1. Average performance variation of 20 independent simulation runs as a function of the sensitivity parameter c.

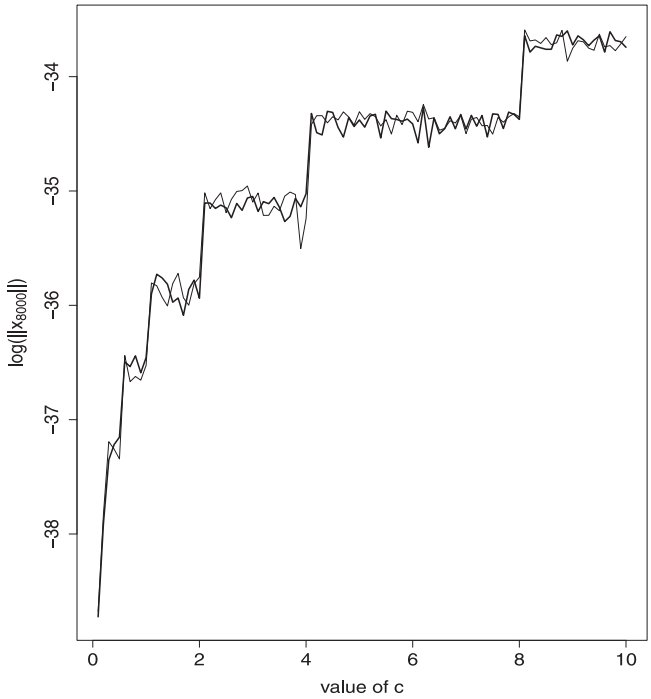

Fig. 2. Two sample runs.

f. On the other hand, if one were to conduct these experiments using f with multiple local minima, then their results would expected to be similar.

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: $a(n) = \frac{1}{(n \mod 800)+100}$, $n \geq 1$.

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 Σ 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 Figs. 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 ith 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:

$$x_{n+1} = x_n + 1/n\,(Qx_n + \underline{\epsilon}) \tag{9}$$

where the starting point x_0 was randomly chosen and dimension $d = 10$, the matrix Q was a randomly generated positive-definite matrix (Q is generated as explained before), and $\underline{\epsilon} = (\epsilon/\sqrt{d} \ldots \epsilon/\sqrt{d})$ is the constant noise vector added at each stage and $\epsilon \in \mathbb{R}$. Since Q is positive definite, we expect (9) to converge to the origin when $\epsilon = 0$ in the noise vector. A natural question to ask is the following: If a "small" noise vector is 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 $\epsilon \in \mathbb{R}$. 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

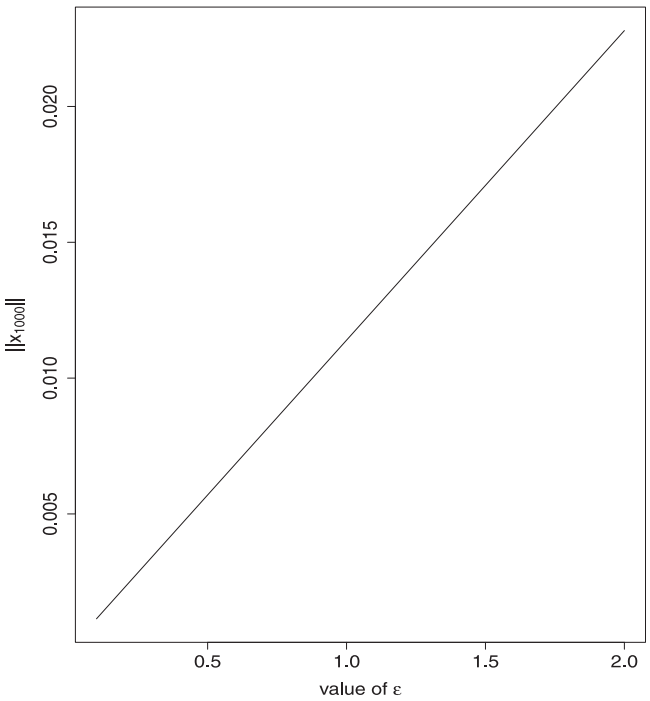

Fig. 3. Average performance variation of 20 independent simulation runs as a function of the neighborhood parameter ϵ.

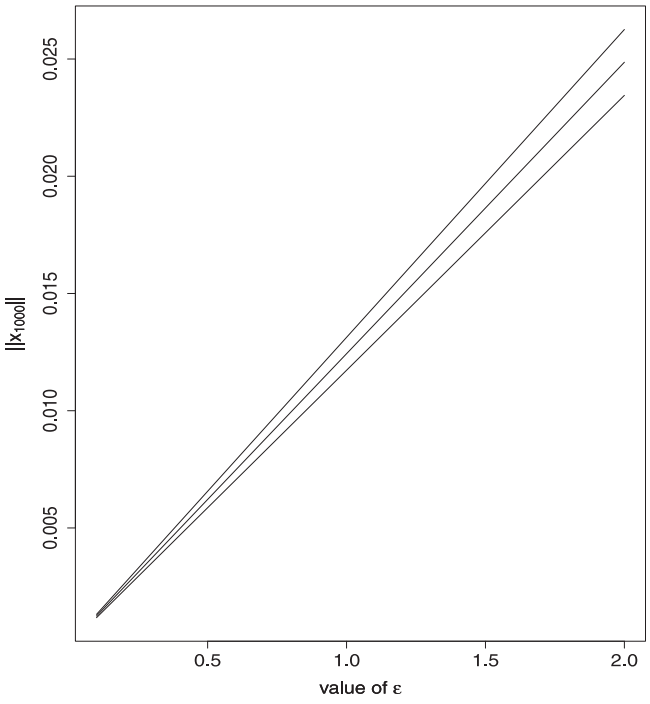

Fig. 4. Three sample runs.

forces the iterates to diverge. As in the first experiment, we expect the iterates to be farther from the origin for larger values of ϵ. This is evidenced by the plots in Figs. 3 and 4. As before, Fig. 3 shows the average performance of 20 independent simulation runs (for each $\underline{\epsilon}$), and Fig. 4 shows three of these sample runs. The x-axis in Figs. 3 and 4 represents values of the ϵ parameter in (9) that varies from 0.1 to 2, i.e., $\|\underline{\epsilon}\|$ varies from 0.1 to 2 in steps of 0.01. The y-axis in Fig. 3 represents the average distance of the iterate from the origin after 1000 iterations, i.e., $1/20\sum_{i=1}^{20}\|x_{1000}^i\|$, where x_{1000}^i is the iterate value after 1000 iterations from the ith run. The y-axis in Fig. 3 represents $\|x_{1000}^i\|$. For ϵ close to 0, the iterate (after 1000 iterations) is within $B_{0.0003}(0)$, while for ϵ close to 2, the iterate (after 1000 iterations) is only within $B_{0.1}(0)$.

VI. Extensions and Conclusion

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) nondiminishing 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) nondiminishing errors. To see this, define $G(x) := H(x)^{-1}\nabla f(x) + \overline{B}_\epsilon(0)$ in $(A1)$; G_∞ changes accordingly. Here, $H(x)$ (assumed positive definite) denotes the Hessian evaluated at x. Theorems 1 and 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 nondiminishing errors. More generally, it would be interesting to derive convergence rates of stochastic approximation algorithms with set-valued mean fields.

References

[1] J. Aubin and A. Cellina, *Differential Inclusions: Set-Valued Maps and Viability Theory*. New York, NY, USA: Springer, 1984.
[2] M. Benaïm, J. Hofbauer, and S. Sorin, "Stochastic approximations and differential inclusions," *SIAM J. Control Optim.*, vol. 44, pp. 328–348, 2005.
[3] M. Benaïm, J. Hofbauer, and S. Sorin, "Perturbations of set-valued dynamical systems, with applications to game theory," *Dyn. Games Appl.*, vol. 2, no. 2, pp. 195–205, 2012.
[4] M. Benam, "A dynamical system approach to stochastic approximations," *SIAM J. Control Optim.*, vol. 34, no. 2, pp. 437–472, 1996.
[5] D. P. Bertsekas and J. N. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM J. Optim.*, vol. 10, no. 3, pp. 627–642, 2000.
[6] S. S. Haykin, *Neural Networks and Learning Machines*, vol. 3. Upper Saddle River, NJ, USA: Pearson Education, 2009.
[7] M. Hurley, "Chain recurrence, semiflows, and gradients," *J. Dyn. Differ. Equ.*, vol. 7, no. 3, pp. 437–456, 1995.
[8] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Ann. Math. Statist.*, vol. 23, no. 3, pp. 462–466, 1952.
[9] O. L. Mangasarian and M. V. Solodov, "Serial and parallel backpropagation convergence via nonmonotone perturbed minimization," *Optim. Methods Softw.*, vol. 4, no. 2, pp. 103–116, 1994.
[10] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Trans. Automat. Control*, vol. 37, no. 3, pp. 332–341, Mar. 1992.
[11] V. B. Tadić and A. Doucet, "Asymptotic bias of stochastic gradient search," in *Proc. 50th IEEE Conf. Decision Control Eur. Control Conf.*, 2011, pp. 722–727.