Lyapunov methods in nonsmooth optimization, Part II: persistently exciting finite differences

Andrew R. Teel[1]
ECE Department, University of California
Santa Barbara, CA 93106
teel@ece.ucsb.edu

Abstract

A recent converse Lyapunov theorem for differential inclusions is used to generate a class of finite difference algorithms for nonsmooth optimization. The algorithms rely on a proof of asymptotic stability for differential inclusions that contain persistently exciting signals and the ability to approximate these differential inclusions with finite differences. The notion of persistency of excitation that is used here generalizes that which is typically used in the identification and adaptive control literature.

1 Introduction

1.1 Background

The focus of this paper is a particular problem of unconstrained nonlinear programming for locally Lipschitz functions. We address the task of designing numerical algorithms that asymptotically determine a point that minimizes a locally Lipschitz function defined on Euclidean space without explicitly using the generalized gradient of the function. (We are motivated by extremum seeking problems in dynamical systems with nonsmooth readout maps; see [15].) For continuously differentiable functions, this problem has been addressed in at least two different ways. One approach has been labeled "derivative free optimization" and is described in [5] as "all methods which do not attempt to *directly* compute approximations to the unavailable derivative information." The results of [14], [16, 17], [25], and, more recently [7, 23, 24] are considered to be in this class. The other approach is to approximate the gradient by finite differences. This idea has received considerable attention in both a deterministic and stochastic setting. The initial stochastic results were reported in [11] and [2] and other important work can be found in, e.g., [12], [20] and the references therein. Deterministic results are nicely summarized in [6]. The main challenge of nonsmooth minimization by finite differences is that these don't always provide a good estimate of any element of the generalized gradient. See [13, Section 3.3] for a description of the "dangers and sins" of using finite differences in nonsmooth optimization.

[1]Research supported in part by NSF under grant ECS-9896140 and AFOSR under grant F49620-00-1-0106.

1.2 Contribution

In this paper we build upon the *Lyapunov method* of nonsmooth minimization developed in [21] which applies, most transparently, to locally Lipschitz functions that are regular. Locally Lipschitz functions that are convex or strictly differentiable are special cases of regular functions. (A precise definition is given in the next section.) For the gradient-free nonsmooth optimization problem, our main observation is that, while multiple finite differences cannot be used to approximate an element of the generalized gradient, a single finite difference can be used to approximate the inner product of a probing vector with some element of the generalized gradient. We will show that if this probing vector is persistently exciting, in a sense that generalizes the notion of persistency of excitation used in identification and adaptive control (see, for example, [1], [10, p. 177] and the references therein), then enough information about the generalized gradient is gathered over time to produce convergence to the minimizer.

Single finite differences have been used in a stochastic setting for thrice continuously differentiable functions in [20]. It was shown in [20] that the method proposed there, called "simultaneous perturbation stochastic approximation (SPSA)", tends to produce convergence results superior, in total number of function evaluations required, to algorithms that use multiple finite differences. An idea for extending SPSA to nonsmooth functions was given in [9], however no rigorous statements are made and it is not clear that the idea works in general. (See Section 6 below.)

Our proof of convergence to the minimizer is accomplished by first establishing asymptotic stability of the minimizer for a related differential inclusion, then using recent results on converse Lyapunov functions for asymptotically stable differential inclusions to deduce the existence of a descent function for the numerical algorithm that approximates the differential inclusion.

Our paper is organized as follows: In Section 2 we collect our main definitions and establish notation, which largely repeats [21]. In Section 3 we present results on asymptotic stability of differential and difference inclusions related to minimization problems. In Section 4 we develop the notion of persistency of excitation that we will use. In Section 5 we consider the problem

of minimizing locally Lipschitz, regular functions. We first construct a differential inclusion, using the generalized gradient of the locally Lipschitz function and a generic set of persistently exciting positive semidefinite matrices, that exhibits convergence to the minimum of the function. Then we show that this inclusion can be approximated by function evaluation differences. The combination of this result with the results in Section 3 forms the core idea behind the construction of our minimization algorithms. For locally Lipschitz, regular functions this combination yields a generic class of persistently exciting finite difference minimization algorithms. The necessity of our generalized persistency of excitation condition is discussed in Section 6.

2 Definitions

A function $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is said to belong to **class-$\mathcal{K}_\infty$** if it is continuous, zero at zero, strictly increasing and unbounded. Given a closed set $\mathcal{A} \subset \mathbb{R}^n$ and a vector $x \in \mathbb{R}^n$, we define $|x|_{\mathcal{A}} := \inf_{z \in \mathcal{A}} |x - z|$.

A set-valued map $F(\cdot)$, a mapping from an open set $\mathcal{G}$ to subsets of $\mathbb{R}^n$, is said to satisfy the **basic conditions on $\mathcal{G}$** if, for each $x \in \mathcal{G}$, the set $F(x)$ is nonempty, compact and convex, and $F(\cdot)$ is upper semicontinuous, i.e., for each $x \in \mathcal{G}$ and each $\varepsilon > 0$ there exists $\delta > 0$ such that for each ξ such that $|\xi - x| \leq \delta$, we have

$$F(\xi) \subseteq F(x) + \varepsilon \overline{\mathcal{B}}$$

where $\overline{\mathcal{B}}$ denotes the closed unit ball in $\mathbb{R}^n$.

The next few definitions can be found in [3], for example. For a locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$, the **Clarke generalized directional derivative** of f at $x \in \mathbb{R}^n$ in the direction $v \in \mathbb{R}^n$, denoted $f^\circ(x, v)$, is defined as

$$f^\circ(x; v) := \limsup_{y \to x, t \to 0^+} \frac{f(y + tv) - f(y)}{t} . \qquad (1)$$

The **(Clarke) generalized gradient** of f at x, denoted $\partial f(x)$, is defined as

$$\partial f(x) := \{\xi \in \mathbb{R}^n : f^\circ(x; v) \geq \langle \xi, v \rangle \quad \forall v \in \mathbb{R}^n\} . \qquad (2)$$

The generalized gradient, which is a set-valued map in general, satisfies the basic conditions. A point $x \in \mathbb{R}^n$ is said to be a **stationary point** of f if $0 \in \partial f(x)$. A locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **regular** if, for all $x \in \mathbb{R}^n$ and all $v \in \mathbb{R}^n$, the usual one-sided directional derivative

$$f'(x; v) := \lim_{t \to 0^+} \frac{f(x + tv) - f(x)}{t} \qquad (3)$$

exists and equals $f^\circ(x; v)$. Locally Lipschitz functions that are strictly differentiable or convex are regular. Also, the sum and pointwise maximum of regular functions are regular. For more details see [3, Proposition 2.3.6].

We will use $\phi(\cdot, x)$ to denote an arbitrary **solution** of the differential inclusion $\dot{x} \in F(x)$, i.e., an absolutely continuous function satisfying $\phi(0, x) = x$ and whose derivative satisfies, for almost all t on its interval of definition, $\dot{\phi}(t, x) \in F(\phi(t, x))$. Whenever F satisfies the basic conditions on $\mathcal{G}$, there exists at least one solution for each $x \in \mathcal{G}$. (See, for example, [8, §7, Theorem 1].) We will let $\mathcal{S}(x)$ denote the set of **maximal solutions** starting at x, i.e., a solution defined on $[0, T)$ where either $T = +\infty$ or else the solution cannot be extended to a solution on a larger interval. If F satisfies the basic conditions on $\mathcal{G}$ then there exist maximal solutions for each $x \in \mathcal{G}$. (See, for example, [18, Propositions 1 and 2].)

For the differential inclusion $\dot{x} \in F(x)$, the compact set $\mathcal{A} \subset \mathbb{R}^n$ is **locally asymptotically stable** if

1. for each $\varepsilon > 0$ there exists $\delta > 0$ such that, for each $x \in \mathcal{A} + \delta\overline{\mathcal{B}}$, each solution $\phi \in \mathcal{S}(x)$ is defined for all $t \geq 0$ and satisfies $\phi(t, x) \in \mathcal{A} + \varepsilon\overline{\mathcal{B}} \qquad \forall t \geq 0$,
2. the set $\mathcal{G}$ of points $x \in \mathbb{R}^n$ such that each solution $\phi \in \mathcal{S}(x)$ is defined for all $t \geq 0$ and satisfies $\phi(t, x) \to \mathcal{A}$ as $t \to \infty$ contains a neighborhood of $\mathcal{A}$.

The set of points $\mathcal{G}$ that satisfies the second condition in the definition of local asymptotic stability is referred to as the **basin of attraction** for the set $\mathcal{A}$. If F satisfies the basic conditions on an open set containing $\mathcal{G}$ then $\mathcal{G}$ is an open set. (See [22, Proposition 3]; cf. [4, Proposition 2.2].) Similar notation and definitions apply to difference inclusions.

3 Inclusion results related to minimization

We now recall some results stated in [21] that follow from [22]. These results will invoke the following assumption:

Assumption 1 *The set-valued map F satisfies the basic conditions on $\mathcal{G}$ and, for $\dot{x} \in F(x)$, the compact set $\mathcal{A}$ is asymptotically stable with basin of attraction $\mathcal{G}$.*

Theorem 1 *Let Assumption 1 hold. Then there exists a continuous function $\delta : \mathcal{G} \to \mathbb{R}_{\geq 0}$ that is positive on $\mathcal{G} \backslash \mathcal{A}$ so that, for the differential inclusion*

$$\dot{x} \in \overline{\text{co}} F(x + \delta(x)\overline{\mathcal{B}}) + \delta(x)\overline{\mathcal{B}} , \qquad (4)$$

the compact set $\mathcal{A}$ is asymptotically stable with basin of attraction $\mathcal{G}$.

Corollary 1 *Let Assumption 1 hold and let $\mathcal{C}$ and $\mathcal{D}$ be arbitrary compact subsets of $\mathcal{G}$ such that $\mathcal{A}$ is a strict subset of $\mathcal{D}$. Then there exists $\varepsilon > 0$ and a compact set $\mathcal{A}_\varepsilon$ that is a strict subset of $\mathcal{D}$ such that, for the differential inclusion*

$$\dot{x} \in \overline{\text{co}} F(x + \varepsilon\overline{\mathcal{B}}) + \varepsilon\overline{\mathcal{B}} =: F_\varepsilon(x) , \qquad (5)$$

the set $\mathcal{A}_\varepsilon$ is asymptotically stable with basin of attraction containing $\mathcal{C}$.

Theorem 2 *Let Assumption 1 hold and let $\mathcal{C}$ and $\mathcal{D}$ be arbitrary compact subsets of $\mathcal{G}$ such that $\mathcal{A}$ is a strict*

subset of $\mathcal{D}$. Then there exists $\tau^* > 0$ and a compact set $\mathcal{A}_\varepsilon$ that is a strict subset of $\mathcal{D}$ such that, for each $\tau \in (0, \tau^*)$, the difference inclusion

$$x_{k+1} \in x_k + \tau F(x_k) \tag{6}$$

is such that the set $\mathcal{A}_\varepsilon$ is locally asymptotically stable with basin of attraction containing the set $\mathcal{C}$.

4 Generalized persistency of excitation (GPE)

Our upcoming result on asymptotic stability for a differential inclusion related to nonsmooth optimization by finite differences will rely on the notion of persistency of excitation given in the following definition:[1]

Definition 1 (GPE) *A bounded, measurable function $\omega : [0, 2\pi] \to \mathbb{R}^{n\times n}$ is said to be* generalized persistently exciting *with respect to a function $f(\cdot)$ on a set $\mathcal{H}$ if, for every $x \in \mathcal{H}$ that is not a stationary point for f, there exists $\kappa > 0$ such that, for each measurable function $z(x, \cdot) : [0, 2\pi] \to \partial f(x)$,*

$$\int_0^{2\pi} z^T(x,\theta)\omega(\theta)z(x,\theta)d\theta \geq \kappa \ . \tag{7}$$

If $\partial f(x) = \{\nabla f(x)\}$, i.e., $\partial f(x)$ is a singleton, then the measurable function $z(x, \cdot)$ in the definition must necessarily be the constant vector $\nabla f(x)$. In this case the left-hand side of (7) becomes

$$\begin{aligned}\int_0^{2\pi} & z^T(x,\theta)\omega(\theta)z(x,\theta)d\theta \\ & = (\nabla f(x))^T \left(\int_0^{2\pi} \omega(\theta)d\theta\right) \nabla f(x) \ .\end{aligned} \tag{8}$$

Since the definition only concerns nonstationary points, i.e., points where $|\nabla f(x)| \neq 0$, it follows that (7) holds if and only if there exists $\widetilde{\kappa} > 0$ such that

$$\int_0^{2\pi} \omega(\theta)d\theta \geq \widetilde{\kappa} I \ . \tag{9}$$

This is the standard persistency of excitation condition in identification and adaptive control. In the planar case it holds, for example, when, with the definitions

$$e_1 := \begin{bmatrix} 1 \\ 0 \end{bmatrix} \ , \ R(\vartheta) := \begin{bmatrix} \cos(\vartheta) & -\sin(\vartheta) \\ \sin(\vartheta) & \cos(\vartheta) \end{bmatrix} \tag{10}$$

we have

$$\begin{aligned}\omega(\theta) = \Bigg\{ & \sum_{i=1}^{2} \lambda_i R\left(\frac{(i-1)\pi}{2}\right) e_1 e_1^T R\left(\frac{(i-1)\pi}{2}\right)^T \\ & \lambda_i \geq 0 \ , \ \sum_{i=1}^{2} \lambda_i = 1 \ , \\ & \lambda_i = 1 \ \forall \theta \in ((i-1)\pi, i\pi) \Bigg\} \ .\end{aligned} \tag{11}$$

[1] For simplicity we restrict our attention to persistently exciting functions that are 2π-periodic in t and independent of the state. These conditions can easily be relaxed.

The condition (9) is not sufficient to guarantee the generalized persistency of excitation condition given in Definition 1. For example, consider the function

$$f(x) = \max\{|x_1|, |x_2|\} \tag{12}$$

and the function ω defined by (11). It can be shown that the condition (7) is not satisfied at all points where $|x_1| = |x_2|$ (exactly the set of points where $\partial f(x)$ is not a singleton). For example, at points x where $x_1 = x_2 > 0$ we have

$$\partial f(x) = \left\{ \lambda_f R\left(\frac{\pi}{2}\right) e_1 + (1-\lambda_f) e_1 \qquad \lambda_f \in [0,1] \right\} \ . \tag{13}$$

If we make a selection $z(x, \theta) \in \partial f(x)$ so that $\lambda(\theta) = \lambda_f(\theta)$ then we get that $\omega(\theta)z(x, \theta) = 0$ for almost all $\theta \in [0, \pi]$ since

$$e_1^T R\left(\frac{\pi}{2}\right) e_1 = e_1^T R^T\left(\frac{\pi}{2}\right) e_1 = 0 \ , \tag{14}$$

and so (7) is not satisfied. On the other hand, if either the matrices that serve as the basis for ω or the coordinates used to produce f are rotated appropriately, e.g., by 45 degrees, then (7) is satisfied. (If both are rotated by the same amount, (7) will not be satisfied.) Again in the planar case, if it is known for $f(\cdot)$ that the cone defining the generalized gradient is limited in aperture to 90 degrees, as is the case for the function defined in (12), then (7) is guaranteed to be satisfied by taking $\omega(\theta)$ to be

$$\begin{aligned}\omega(\theta) = \Bigg\{ & \sum_{i=1}^{3} \lambda_i R\left(\frac{2(i-1)\pi}{3}\right) e_1 e_1^T R\left(\frac{2(i-1)\pi}{3}\right)^T \\ & \lambda_i \geq 0 \ , \ \sum_{i=1}^{3} \lambda_i = 1 \ , \\ & \lambda_i = 1 \ \ \forall \theta \in \left(\frac{2(i-1)\pi}{3}, \frac{2i\pi}{3}\right) \Bigg\} \ .\end{aligned} \tag{15}$$

This follows from the fact that, for each cone with 90 degrees aperture and at least one vector among

$$R\left(\frac{2(i-1)\pi}{3}\right) e_1 \quad i = 1, 2, 3 \ , \tag{16}$$

it is impossible to make a selection from the cone that is orthogonal to the vector.

Without any information about the aperture of the cone of the generalized gradient and again for the planar case, the choice

$$\omega(\theta) = R(\theta) e_1 e_1^T R(\theta) \tag{17}$$

is (generalized) persistently exciting. This idea easily generalizes to $\mathbb{R}^n$ since, for any $x \in \mathbb{R}^n$ that is not a stationary point of $f(\cdot)$ we can always find some direction in $\mathbb{R}^n$ that is not orthogonal to any element of $\partial f(x)$. This follows from the convexity of $\partial f(x)$.

5 Optimization by GPE finite differences

5.1 A preliminary inclusion result

Throughout this section we will assume the following for the function to be minimized (cf. the assumptions of [21, Proposition 1]):

Assumption 2 *Suppose*

1. $f : \mathbb{R}^n \to \mathbb{R}$ *is locally Lipschitz and regular,*
2. $\overline{f} \in \mathbb{R}$ *and* $\underline{f} \in (-\infty, \overline{f})$ *are such that*
   (a) $\{x \in \mathbb{R}^n : f(x) \leq \overline{f}\} =: \mathcal{C}_1$ *is compact,*
   (b) $\{x \in \mathbb{R}^n : f(x) \leq \underline{f}\} =: \mathcal{A}_1$ *is nonempty,*
   (c) $\{x \in \mathbb{R}^n : \underline{f} < f(x) \leq \overline{f}\} = \mathcal{C}_1 \backslash \mathcal{A}_1$ *contains no stationary points.*

We will also assume that we have a function ω, for simplicity assumed to be continuous, that is persistently exciting with respect to $f(\cdot)$ on $\mathcal{C}_1 \backslash \mathcal{A}_1$:

Assumption 3 *The function* $\omega : [0, 2\pi] \to \mathbb{R}^{n \times n}$ *is such that:*

1. *it is continuous,*
2. *each of its elements is symmetric, positive semidefinite,*
3. *it is persistently exciting with respect to* $f(\cdot)$ *on* $\mathcal{C}_1 \backslash \mathcal{A}_1$,
4. $\omega(0) = \omega(2\pi)$.

In the next proposition the function $P : \mathbb{R}^2 \setminus \{0\} \to [0, 2\pi)$ denotes the mapping from nonzero vectors in $\mathbb{R}^2$ to their angle in the plane, with increasing angle in the clockwise direction and zero identified with the vector $[0\ 1]^T$.

Proposition 1 *Let Assumptions 2 and 3 hold and let the interval* $[a, b]$ *on the real line be such that* $b \geq a > 0$. *Then Assumption 1 is satisfied with* $x = [x_1^T, x_2^T]^T$,

$$\begin{aligned} F(x) &:= \begin{bmatrix} -\omega\left(P(x_2)\right)\ \partial f(x_1) \\ \left([a,b] \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + (1 - |x_2|^2) \right) x_2 \end{bmatrix} \\ \mathcal{A} &:= \mathcal{A}_1 \times S^1 \ , \quad \mathcal{G} \supset \mathcal{C}_1 \times \mathbb{R}^2 \setminus \{0\} \ . \end{aligned} \tag{18}$$

Proof. (Compare with the proof of [21, Proposition 1].) We first define (r, φ) by

$$x_{21} = r \sin(\varphi) \ , \ x_{22} = r \cos(\varphi) \tag{19}$$

which is a smooth, invertible transformation on $\mathbb{R}^2 \setminus \{0\}$. We note that

$$\begin{aligned} \dot{x}_1 &\in -\omega(\varphi)\ \partial f(x_1) \\ \dot{r} &= -r(r^2 - 1) \\ \dot{\varphi} &\in [a, b] \ . \end{aligned} \tag{20}$$

We note that $\varphi(t)$ is a locally Lipschitz function of time that is invertible. Moreover, defining $V(x_2) := (r^2 - 1)^2$, we see that $V(\cdot)$ is smooth on $\mathbb{R}^2 \setminus \{0\}$ and

$$\dot{\overline{V(\phi_2(t,x))}} = -r^2(\phi_2(t,x)) V(\phi_2(t,x)) \ . \tag{21}$$

From these properties, to show that the set $\mathcal{A}$ defined in (18) is asymptotically stable with basin of attraction containing $\mathcal{C}_1 \times \mathbb{R}^2 \setminus \{0\}$, it is sufficient to show that the set $\mathcal{A}_1$ for the system[2]

$$\dot{x} \in -\frac{1}{[a,b]}\ \omega(t) \partial f(x) \tag{22}$$

is asymptotic stability with basin of attraction containing $\mathcal{C}_1$ (for all $t_\circ \in [0, 2\pi]$). For notational convenience, we will drop the $t_\circ$ dependence of the trajectories in what follows. Since $f(\cdot)$ is locally Lipschitz, it follows that $f(\phi(t,x))$ is absolutely continuous and $\dot{\overline{f(\phi(t,x))}}$ and $\dot{\overline{\phi(t,x)}}$ are defined for almost all t. We let $z(t) \in \partial f(\phi(t,x))$ and $\rho(t) \in [a, b]$ be measurable functions such that, for almost all t,

$$\frac{1}{\rho(t)} \omega(t) z(t) = -\dot{\overline{\phi(t,x)}} \ . \tag{23}$$

Then, using in succession i) definition of derivative, ii) regularity of f, and iii) equation (23), iv) $\rho(t) \leq b$, v) $\omega(t)$ symmetric, positive semidefinite, we have, for almost all t,

$$\begin{aligned} \dot{\overline{f(\phi(t,x))}} &= -f'(\phi(t,x), -\dot{\overline{\phi(t,x)}}) \\ &= -\max_{\xi \in \partial f(\phi(t,x))} \langle \xi, -\dot{\overline{\phi(t,x)}} \rangle \\ &\leq -\langle\ z(t), \frac{1}{\rho(t)} \omega(t) z(t)\ \rangle \\ &\leq -\frac{1}{b} z^T(t) \omega(t) z(t) \leq 0 \ . \end{aligned} \tag{24}$$

Integrating this inequality and using that $\mathcal{A}_1$ is a sublevel set of $f(\cdot)$ establishes stability of the set $\mathcal{A}_1$. We next establish convergence to the set $\mathcal{A}_1$ from the set $\mathcal{C}_1$. Since $\mathcal{A}_1$ is stable and $\mathcal{C}_1$ is compact and forward invariant (because it is a sublevel set of $f(\cdot)$), if $\phi(t,x)$ does not converge to $\mathcal{A}_1$ then $\phi(t,x)$ has an accumulation point $x^* \in \mathcal{C}_1 \backslash \mathcal{A}_1$. Using the persistency of excitation condition and the upper semicontinuity of $\partial f(\cdot)$, there exists $\delta > 0$ such that

$$\begin{aligned} |\phi(t,x) - x^*| \leq \delta \quad \forall t \in [s, s + 2\pi] \quad \Longrightarrow \\ \int_s^{s+2\pi} z^T(t) \omega(t) z(t) dt \geq \frac{\kappa(x^*)}{2} > 0 \ . \end{aligned} \tag{25}$$

On the other hand, by integrating (24) we get that

$$\int_0^\infty z^T(t) \omega(t) z(t) dt \leq b f(x) \tag{26}$$

[2] In inclusion (22), we are using $\omega(t) := \omega(t \mod 2\pi)$.

from which it follows that

$$\lim_{s\to\infty}\int_s^{s+2\pi} z^T\omega(t)z(t)dt = 0\ . \tag{27}$$

Using the fact that $\omega(\cdot)$ is bounded, and the fact that $\omega(t)$ is symmetric for each t, it follows from (23) that there exists $\sigma > 0$ such that for each $\varepsilon > 0$

$$\begin{aligned}\dot{\overline{|\phi(t,x)|}} &\leq \frac{1}{b}|\omega(t)z(t)| \\ &\leq \varepsilon + \frac{\sigma}{\varepsilon b^2}z^T(t)\omega(t)z(t)\ .\end{aligned} \tag{28}$$

From this and (27) it follows that there exists a positive real number s^* such that, for all $s \geq s^*$,

$$|\phi(t,x) - \phi(s,x)| \leq \frac{\delta}{2} \qquad \forall t \in [s, s+2\pi]\ . \tag{29}$$

Now, let s_i be a sequence of times with $s_{i+1} - 2\pi \geq s_i \geq s^*$ for all i and such that

$$|\phi(s_i,x) - x^*| \leq \frac{\delta}{2}\ . \tag{30}$$

Such a sequence of times exists since x^* is an accumulation point for $\phi(\cdot,x)$. It follows from combining (29) and (30) that $|\phi(t,x) - x^*| \leq \delta$ for all $t \in [s_i, s_i + 2\pi]$. In turn, it follows from (25) that the integral on the left-hand side of (26) is unbounded. This contradicts (26) and so establishes convergence to $\mathcal{A}_1$ from $\mathcal{C}_1$. ■

5.2 A finite difference optimization algorithm

The algorithm presented in this section relies on the mean-value theorem for locally Lipschitz functions (see, e.g., [3, Theorem 2.3.7]) a special case of which states the following:

Lemma 1 *If $f(\cdot)$ is locally Lipschitz then for each $\lambda_1, \lambda_2 \in \mathbb{R}$ and $v \in \mathbb{R}^n$, there exists $\rho \in [0,1]$ such that*

$$\begin{aligned}&f(x+\lambda_1 v) - f(x+\lambda_2 v) \in \\ &\qquad (\lambda_1 - \lambda_2)v^T\partial f\left(x + (\rho\lambda_1 + (1-\rho)\lambda_2)v\right)\ .\end{aligned} \tag{31}$$

We remark that, for all $\rho \in [0,1]$,

$$\begin{aligned}&\partial f\left(x + (\rho\lambda_1 + (1-\rho)\lambda_2)v\right) \\ &\qquad \subseteq \partial f\left(x + \max\{|\lambda_1|, |\lambda_2|\}\,|v|\overline{\mathcal{B}}\right)\end{aligned} \tag{32}$$

so that if $\max\{|\lambda_1|, |\lambda_2|\}\,|v| \leq \varepsilon$ then

$$f(x+\lambda_1 v) - f(x+\lambda_2 v) \in (\lambda_1 - \lambda_2)v^T\partial f\left(x + \varepsilon\overline{\mathcal{B}}\right)\ . \tag{33}$$

Finite difference minimization algorithm: Let ω satisfy Assumption 3. Let the interval $[a,b]$ satisfy the assumption of Proposition 1. Let ε and τ^* come from the combination of Proposition 1, Corollary 1 and Theorem 2.

Let $\tau_k \in (0, \tau^*)$. For $i = 1, 2$, let

$$\lambda_{i,k} \in \left[-\frac{\varepsilon}{M}, \frac{\varepsilon}{M}\right] \tag{34}$$

with $\lambda_{1,k} \neq \lambda_{2,k}$ for all k, let

$$v_k = \omega(\varphi_k) \tag{35}$$

and let

$$\varphi_{k+1} \in (\varphi_k + \tau_k[a,b]) \mod 2\pi\ . \tag{36}$$

Finally, let

$$\begin{aligned}&x_{k+1} - x_k = \\ &\quad -\frac{\tau_k}{\lambda_{1,k} - \lambda_{2,k}}v_k\left[f(x_k + \lambda_{1,k}v_k) - f(x_k + \lambda_{2,k}v_k)\right] \\ &\quad \in -\tau_k v_k v_k^T\partial f\left(x_k + \varepsilon\overline{\mathcal{B}}\right) \\ &\quad \subseteq -\tau_k\omega(\varphi_k)\partial f\left(x_k + \varepsilon\overline{\mathcal{B}}\right)\ .\end{aligned} \tag{37}$$

The convergence to a small neighborhood of the minimizer then follows from the results of this paper. Stopping conditions and conditions for the online reduction of τ_k and $\lambda_{i,k}$ follow the discussion in [21, Section 4].

6 On the necessity of GPE

Concerning the differential inclusion (22), suppose that $\omega(\cdot)$ satisfies every aspect of Assumption 3 except for being persistently exciting with respect to $f(\cdot)$ on $\mathcal{C}_1 \backslash \mathcal{A}_1$. It follows from the compactness of $\partial f(x)$ and the symmetry of $\omega(\theta)$ that there exist a point $x \in \mathcal{C}_1 \backslash \mathcal{A}_1$ and measurable selection $z : [0, 2\pi] \to \partial f(x)$ such that $\omega(\theta)z(\theta) = 0$ for all $\theta \in [0, 2\pi]$. Therefore $\phi(t, t_\circ, x) \equiv x$ is a solution to (22), i.e., $x \in \mathcal{C}_1 \backslash \mathcal{A}_1$ is a weak equilibrium point. So we see that the persistency of excitation condition is necessary for the conclusion of Proposition 1.

Despite this necessity for the differential inclusion result, it is not clear that the persistency of excitation condition is necessary, in general, for the finite difference algorithm to converge to a minimum. In particular, it may be sufficient to have a weaker persistency of excitation condition, like only requiring the *existence* of a measurable selection from $\partial f(x)$ that satisfies the integral condition. This is because the finite difference algorithm makes a single selection from (a neighborhood) of this set. On the other hand, there is no obvious mechanism for making the proper selection. We note that the use of probing vectors where each component is generated by a random choice from $\{-1, 1\}$ does not, in general, satisfy the (generalized) persistency of excitation condition. (It does satisfy the standard persistency of excitation condition.) This choice is advocated in the SPSA method of [20] for thrice continuously differential functions. Because it doesn't

satisfy the generalized persistency of excitation condition, it is not clear if SPSA with this choice of probing vectors will work for nonsmooth functions. (We conjecture that in the planar case it will work but there may be problems in higher dimensions; on the other hand, the idea of [9] seems to be to use SPSA in this manner for nonsmooth functions.) The results of the current paper do establish that using finite differences with probing vectors chosen randomly from a uniform distribution on the unit sphere works for nonsmooth functions. However, such distributions are ruled out by the assumptions of [20] (see [20, Footnote 1].)

7 Conclusion

We have shown how to generalize and use the notion of persistency of excitation from the fields of identification and adaptive control to make the method of finite differences applicable to the problem of nonsmooth optimization. In a sense, the method presented here can be seen to generalize the results on stochastic optimization by simultaneous perturbations presented in [20]. However, the discussion here has been carried out completely in a deterministic setting. This paper has presented the theoretical basis for a finite difference approach to nonsmooth optimization. Future work will include numerical examples that illustrate the theory.

References

[1] B.D.O. Anderson. Exponential stability of linear equations arising in adaptive identification. *IEEE Trans. Auto. Contr.*, vol. 22, pp. 83-88, 1977.

[2] J.R. Blum. Multidimensional stochastic approximation methods. *Ann. Math. Stat.*, vol. 25, pp. 737–744, 1954.

[3] F.H. Clarke. *Optimization and nonsmooth analysis.* SIAM, Philadelphia, 1990.

[4] F.H. Clarke, Y.S. Ledyaev and R.J. Stern. Asymptotic stability and smooth Lyapunov functions. J. of Diff. Eqs., vol. 149, no. 1, pp. 69–114, 1998.

[5] A.R. Conn, K. Scheinberg, and Ph.L. Toint. Recent progress in unconstrained nonlinear optimization without derivatives. *Math. Prog.*, 79 (1997) 397–414.

[6] J.E. Dennis and R.B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[7] J.E. Dennis, and V. Torczon. Direct search methods on parallel machines. *SIAM J. Opt.*, **1**:4:448-474, 1991.

[8] A.F. Filippov. *Differential Equations with Discontinuous Righthand Sides.* Kluwer Academic Pub., 1988.

[9] L. Gerencsér, G. Kozmann, and Z. Vàgò. SPSA for non-smooth optimization with application in ECG analysis. *Proceedings of the 37th IEEE CDC*, Tampa, FL, December 1998, pp. 3907–3908.

[10] P.A. Ioannou and J. Sun. *Robust adaptive control.* PTR Prentice-Hall, Upper Saddle River, NJ, 1996.

[11] J. Kiefer and J. Wolfowitz. Stochastic estimation of a regression function. *Ann. Math. Stat.*, vol. 23, pp. 462–466, 1952.

[12] H.J. Kushner and D.S. Clark. *Stochastic approximation methods for constrained and unconstrained systems*, Springer-Verlag, New York, 1978.

[13] C. Lemaréchal. Nondifferentiable optimization, in Handbooks in Op. Res. and Man. Sci., Vol. 1, *Optimization*, G.L. Nemhauser, A.H.G. Rinnooy Kan, and M.J. Todd, eds., North-Holland, Amsterdam, 1989.

[14] J.A. Nelder and R. Mead. A simplex method for function minimization. *Computer J.*, 7 (1965) 308–313.

[15] D. Popovic and A.R. Teel. Solving smooth and nonsmooth multivariable extremum seeking problems by the methods of nonlinear programming. Submitted.

[16] M.J.D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer J.*, 17 (1964) 155–162.

[17] M.J.D. Powell. A new algorithm for unconstrained optimization. In: J.B. Rosen, O.L. Mangasarian and K. Ritter, eds., *Nonlinear Programming*, Academic Press, New York, 1970.

[18] E.P. Ryan. Discontinuous feedback and universal adaptive stabilization. in "Control of Uncertain Systems", D. Hinrichsen and B. Martensson, eds., Birkhauser, Boston, 1990, pp. 245–258.

[19] D. Shevitz and B. Paden. Lyapunov stability theory of nonsmooth systems. *IEEE Trans. on Auto. Contr.* vol. 39, no. 8, Sept. 1994, pp. 1910–1914.

[20] J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE TAC* vol. 37, 1992, pp. 332-341.

[21] A.R. Teel. Lyapunov methods in nonsmooth optimization, Part I: Quasi-Newton algorithms for Lipschitz, regular functions. *39th IEEE CDC*, Sydney.

[22] A.R. Teel and L. Praly. A smooth Lyapunov function from a class-$\mathcal{KL}$ estimate involving two positive semidefinite functions. *ESAIM: Cont., Opt., and Cal. of Var.*, vol. 5, 2000, pp. 313–368. See also "Results on converse Lyapunov functions from class-$\mathcal{KL}$ estimates", In *Proceedings of the 38th IEEE Conf. on Decision and Control*, Phoenix, AZ, December 1999, pp. 2545–2550.

[23] V. Torczon. On the convergence of the multidirectional search algorithm. *SIAM J. Opt.*, **1**:123–145, 1991.

[24] V. Torczon. On the convergence of pattern search algorithms. *SIAM J. Opt.*, **7**:1-25, 1997.

[25] D. Winfield. Function minimization by interpolation in a data table. *J. of the Inst. of Math. and its Applications*, 12 (1973) 339–347.

[26] S.K. Zavriev and A.G. Perevozchikov. Attraction of trajectories of finite-difference inclusions and stability of numerical methods of stochastic nonsmooth optimization. Sov. Phys. Dokl. **35**(8), Aug. 1990, 709-711.