Lyapunov methods in nonsmooth optimization, Part I: Quasi-Newton algorithms for Lipschitz, regular functions

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 large class of algorithms for nonsmooth optimization. Particular attention is given to quasi-Newton algorithms for the minimization of locally Lipschitz, regular functions.

1 Introduction

1.1 Background

The focus of this paper is unconstrained nonlinear programming for locally Lipschitz functions. We address the task of designing numerical algorithms that asymptotically determine a point that globally minimizes a locally Lipschitz function defined on Euclidean space. For continuously differentiable functions, this problem and its solutions have reached a very mature state, which is summarized in many excellent textbooks (see, for example, [2],[11]). The nonsmooth optimization problem is more recent. Serious attention was first given to it in the 1960's and, over the years, many authors have addressed the problem by imposing various extra assumptions, beyond Lipschitz continuity, on the function to be minimized, e.g., convexity, quasidifferentiability, semismoothness, regularity, etc., to facilitate proving the convergence of their proposed algorithms. Some books on the initial developments in the field include [5], [7], [13] and [21]. We also direct the reader's attention to [1], [8], [9], [10], [12], [14], [15], [16], [17], [19], and the references therein. Nonsmooth optimization remains an active area of research.

The main aspect of nonsmooth minimization that makes it a challenging problem is that not every vector belonging to the function's generalized gradient (the object that naturally generalizes the gradient for a continuously differentiable function) provides a descent (or non-ascent) direction for the function. The two main techniques used to solve unconstrained nonsmooth optimization problems are the *subgradient method*, which applies to convex functions, and *bundle methods*, which can handle both convex and nonconvex functions. The subgradient method is the topic of the monographs [5] and [21]. Bundle methods are the main subject of the books [7] and [13].

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

The basic idea exploited in the subgradient method for nonsmooth minimization is that, while elements of the generalized gradient of the function do not necessarily provide descent directions for the function, the negative of these elements are descent directions for the Euclidean distance to the minimizer. Thus, while moving in opposite the direction of elements in the generalized gradient does not cause the value of the function to monotonically decrease, it does cause the distance to the minimizer to decrease monotonically to zero.

The basic idea exploited in bundle methods for nonsmooth minimization is that the negative of at least one element of the generalized gradient of the function is a descent direction for the function and information at nearby points can be accumulated, or bundled, to eventually find such a direction. Bundle algorithms comprise "null steps", where it is determined that the current gradient information does not provide a descent direction and gradient information from a nearby point is used to help find a descent direction, and "serious steps" which are taken when the current gradient information provides a descent direction.

When applied to continuously differentiable functions, the initial versions of the subgradient and bundle methods produce algorithms that reduce to the method of steepest descent. More recently, these algorithms have been adapted to allow quasi-Newton type algorithms. The subgradient method with space dilation discussed in [21, Chapter 3] has this aim. Recent bundle method algorithms that allow quasi-Newton type flexibility can be found in [12], [16], and [17].

1.2 Contribution

In this paper we describe what we call the *Lyapunov method* of nonsmooth minimization. It applies, most transparently, to locally Lipschitz functions that are regular. Regular functions contain convex functions and strictly differentiable functions as special cases. (A precise definition is given in the next section.) It is a method that is formulated to directly include quasi-Newton type algorithms, and can be thought of as a natural extension of the subgradient method. The justification for this interpretation comes from our ability to show that regular functions generically admit smooth descent functions. By this we mean that, given

a family of linear transformations associated with a quasi-Newton type algorithm, there exists a smooth function that decreases along every direction in the set obtained by operating on elements of the generalized gradient with the family of linear transformations. These descent functions generalize the Euclidean distance to the minimizer which is a descent function for convex functions when using the method of steepest descent (where the linear transformation is $-I$.)

The existence of these descent functions comes from recent results on the existence of Lyapunov functions for differential inclusions satisfying certain basic conditions under the assumption of asymptotic stability. See [4] and [23]. We label our approach to nonsmooth optimization the "Lyapunov method" because we use our knowledge of the existence of Lyapunov functions for the relevant differential inclusions to produce conceptually simple quasi-Newton algorithms that minimize locally Lipschitz, regular functions.

Our paper is organized as follows: In Section 2 we collect our main definitions and establish notation. In Section 3 we present results on asymptotic stability of differential and difference inclusions related to minimization problems. In particular, in Section 3.1 we consider locally Lipschitz, regular functions to be minimized and construct a differential inclusion, using the generalized gradient of the locally Lipschitz function and a generic set of positive definite matrices, that exhibits asymptotic convergence to a (set of) minimizer(s). Section 3.2 points out that perturbations of this differential inclusion also exhibit nice convergence properties. Then, in Section 3.3, it is shown that the discrete-time Euler approximation (difference inclusion) of a globally asymptotically stable differential inclusion is semiglobally, practically asymptotically stable. The combination of this result, which is made possible by recent advances in the theory of converse Lyapunov functions for differential inclusions, with the results like in Section 3.1 forms the core idea behind the construction of our minimization algorithms. For locally Lipschitz, regular functions this combination yields a generic class of quasi-Newton minimization algorithms. In Section 4 we address adjusting the step size of our algorithms to achieve faster convergence to the minimum and to provide implementable stopping conditions.

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 $\mathbb{R}^n$ to subsets of $\mathbb{R}^n$, is said to satisfy the **basic conditions** if, for each $x \in \mathbb{R}^n$, the set $F(x)$ is nonempty, compact and convex, and $F(\cdot)$ is upper semicontinuous, i.e., for each $x \in \mathbb{R}^n$ 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 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, there exists at least one solution for each $x \in \mathbb{R}^n$. (See, for example, [6, §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 then there exist maximal solutions for each $x \in \mathbb{R}^n$. (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 then $\mathcal{G}$ is an open set. (See [23, Proposition 3]; cf. [4, Proposition 2.2].) Similar notation and definitions apply to difference inclusions.

3 Inclusions related to minimization

The results of Sections 3.2 and 3.3 will assume:

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

3.1 Relation of Assumption 1 to minimization

We give an example of a differential inclusion $\dot{x} \in F(x)$ and sets $\mathcal{A}$ and $\mathcal{G}$ that satisfy Assumption 1 that will be useful for the design of minimization algorithms for locally Lipschitz, regular functions:

Proposition 1 *Let*

1. $f : \mathbb{R}^n \to \mathbb{R}$ *be locally Lipschitz and regular,*
2. $\overline{f} \in \mathbb{R}$ *and* $\underline{f} \in (-\infty, \overline{f})$ *be such that*
   (a) $\left\{x \in \mathbb{R}^n : f(x) \leq \overline{f}\right\} =: \mathcal{C}$ *is compact,*
   (b) $\left\{x \in \mathbb{R}^n : f(x) \leq \underline{f}\right\} =: \mathcal{A}$ *is nonempty,*
3. *the set valued map* $\mathcal{M}$ *from* $\mathbb{R}^n$ *to subsets of* $\mathbb{R}^{n\times n}$ *is nonempty, compact and convex for each* x*, is upper semicontinuous, and, for each* $x \in \mathcal{C}\backslash\mathcal{A}$,

$$\min_{\zeta \in \overline{\text{co}}\{\mathcal{M}(x)\partial f(x)\}} \max_{\xi \in \partial f(x)} \langle \xi, \zeta \rangle > 0 \ . \tag{4}$$

Then Assumption 1 holds with $\mathcal{A}$ *defined above,* $\mathcal{G} \supset \mathcal{C}$ *and* $F(x) := -\overline{\text{co}}\,\{\mathcal{M}(x)\ \partial f(x)\}$.

Remark 3.1 It is easy to see that if $\overline{f}$ can be taken to be arbitrary then $\mathcal{G} = \mathbb{R}^n$. If $\underline{f} = \min_x f(x)$ then the result states that the set of minimizers is asymptotically stable. ■

Remark 3.2 If each element of $\mathcal{M}(x)$ is positive definite, x is a nonstationary point, and $\partial f(x)$ or $\mathcal{M}(x)$ is a singleton, then the condition (4) is automatically satisfied. This is a consequence of the fact that, since both $f(x)$ and $\mathcal{M}(x)$ are compact and convex, when either set is a singleton we have

$$\overline{\text{co}}\,\{\mathcal{M}(x)\partial f(x)\} = \mathcal{M}(x)\partial f(x) \ .$$

Then the facts that $0 \notin \partial f(x)$ and $\mathcal{M}(x)$ is a compact set of positive definite matrices yields (4). ■

Proof. By construction, F satisfies the basic conditions. It remains to establish asymptotic stability of the set $\mathcal{A}$ with basin of attraction containing $\mathcal{C}$. Since $f(\cdot)$ is locally Lipschitz, it follows that $f(\phi(t,x))$ is absolutely continuous and for all t such that $\dot{\overline{f(\phi(t,x))}}$ and $\dot{\overline{\phi(t,x)}}$ are defined we have, using in succession i) the definition of derivative, ii) regularity of f, iii) definition of solution, iv) the condition (4),

$$\begin{aligned}
\dot{\overline{f(\phi(t,x))}} &= -f'(\phi(t,x), -\dot{\overline{\phi(t,x)}}) \\
&= -\max_{\xi \in \partial f(\phi(t,x))} \langle \xi, -\dot{\overline{\phi(t,x)}} \rangle \\
&\leq -\min_{\zeta \in \overline{\text{co}}\{\mathcal{M}\partial f(\phi(t,x))\}} \max_{\xi \in \partial f(\phi(t,x))} \langle \xi, \zeta \rangle \\
&< 0 \ \text{ when } \phi(t,x) \in \mathcal{C}\backslash\mathcal{A} \ .
\end{aligned} \tag{5}$$

Asymptotic stability of the set $\mathcal{A}$ with basin of attraction containing $\mathcal{C}$ is then a consequence of integrating (5), and exploiting the facts that $\mathcal{C}$ and $\mathcal{A}$ are compact and that ∂f satisfies the basic conditions. ■

Remark 3.3 The arguments used to get (5) have also been used in [20] in the development of Lyapunov stability theory for differential equations with discontinuous right-hand side. Similar arguments also appear to be used in [24]. ■

3.2 Robustness of asymptotic stability

We now recall results on robustness of asymptotic stability for differential inclusions, which indicate that the generalized gradient can be replaced in Proposition 1 by an approximation of the generalized gradient.

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*

$$\dot{x} \in \overline{\text{co}}F(x + \delta(x)\overline{\mathcal{B}}) + \delta(x)\overline{\mathcal{B}} \ , \tag{6}$$

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

Proof. Follows from the combination of [23, Propositions 2, 3 and Theorem 3]. ■

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 exist* $\varepsilon > 0$ *and a compact set* $\mathcal{A}_\varepsilon$ *that is a strict subset of* $\mathcal{D}$ *such that, for the differential inclusion*

$$\dot{x} \in \overline{\text{co}}F(x + \varepsilon\overline{\mathcal{B}}) + \varepsilon\overline{\mathcal{B}} =: F_\varepsilon(x) \ , \tag{7}$$

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

Proof. Let the continuous function $\delta : \mathcal{G} \to \mathbb{R}_{\geq 0}$ come from the conclusion of Theorem 1; in particular, the differential inclusion

$$\dot{x} \in \overline{\text{co}}F(x + \delta(x)\overline{\mathcal{B}}) + \delta(x)\overline{\mathcal{B}} =: F_{\delta(x)}(x) \tag{8}$$

is such that the set $\mathcal{A}$ is asymptotically stable with basin of attraction $\mathcal{G}$. Let $\omega : \mathcal{G} \to \mathbb{R}_{\geq 0}$ be a continuous function such that $\omega(x) = 0$ if and only if $x \in \mathcal{A}$ and $\omega(x) \to \infty$ as x approaches the boundary of $\mathcal{G}$ or x tends toward infinity. (In [23], such functions are called proper indicator functions for $\mathcal{A}$ on $\mathcal{G}$.) Using the main results from [23], (cf. [3]), there exists a smooth function function $V : \mathcal{G} \to \mathbb{R}_{\geq 0}$ and class-$\mathcal{K}_\infty$ functions α_1 and α_2 such that

$$\alpha_1(\omega(x)) \leq V(x) \leq \alpha_2(\omega(x)) \tag{9}$$

and

$$\max_{w \in F_{\delta(x)}(x)} \langle \nabla V(x), w \rangle \leq -V(x) \ . \tag{10}$$

From (9) and the properties of $\omega(\cdot)$, we can find $\mu > 0$ sufficiently large so that

$$\mathcal{C} \subseteq \{x \in \mathcal{G} : V(x) \leq \mu\} =: \mathcal{H} \tag{11}$$

and $\rho \in (0, \mu)$ and $\widetilde{\varepsilon} > 0$ sufficiently small so that

$$\Big(\mathcal{N} := \{x \in \mathcal{G} : V(x) \leq \rho\}\Big) + \widetilde{\varepsilon}\overline{\mathcal{B}} \subseteq \mathcal{D} \ . \tag{12}$$

Now define $\varepsilon := \min_{x \in \mathcal{H}\backslash\mathcal{N}} \delta(x)$. It follows from (10) that $x \in \mathcal{H}\backslash\mathcal{N}$ implies $\max_{w \in F_\varepsilon(x)} \langle \nabla V(x), w\rangle \leq -V(x)$.
It follows, for $\dot{x} \in F_\varepsilon(x)$, that the set $\mathcal{N}$ is asymptotically stable with basin of attraction containing $\mathcal{H}$. ■

Remark 3.4 The result of Corollary 1 allows the use of ε-generalized gradients, defined for $\varepsilon \geq 0$ as

$$\partial f_\varepsilon(x) := \overline{\text{co}}\left(\bigcup_{z \in \{x\}+\varepsilon\overline{\mathcal{B}}} \partial f(z)\right) + \varepsilon\overline{\mathcal{B}} \ , \tag{13}$$

to achieve convergence to an arbitrarily small neighborhood of the optimum. These and related ε-subgradients/subdifferentials are commonly used in bundle methods for nonsmooth optimization as a means of guaranteeing ε-decrease away from the minimizer when using line searches. (For a nice discussion see [10, Section 6.1].) ■

3.3 From differential to difference inclusions

The following result, which relates asymptotic stability for a differential inclusion to asymptotic stability for the corresponding discrete-time Euler approximation (difference inclusion) (cf. [24]), is the main tool for the minimization algorithms we propose.

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*

$$x_{k+1} \in x_k + \tau F(x_k) \tag{14}$$

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

Proof. Let $\omega : \mathcal{G} \to \mathbb{R}_{\geq 0}$ be a continuous function such that $\omega(x) = 0$ if and only if $x \in \mathcal{A}$ and $\omega(x) \to \infty$ as x approaches the boundary of $\mathcal{G}$ or x tends toward infinity. Using the main results from [23], (cf. [3]), there exists a smooth function $V : \mathcal{G} \to \mathbb{R}_{\geq 0}$ and class-$\mathcal{K}_\infty$ functions α_1 and α_2 such that (9) holds and

$$\max_{w \in F(x)} \langle \nabla V(x), w\rangle \leq -V(x) \ . \tag{15}$$

Like in the proof of Corollary 1, we can find $\mu > 0$, define $\mathcal{H}$ as in (11), find $\rho \in (0, \mu)$ and $\widetilde{\varepsilon} > 0$ sufficiently small so that (12), where $\mathcal{N}$ is defined, holds. Since $\mathcal{H}$ is compact and $F(\cdot)$ satisfies the basic conditions,

$$M := \sup\{|w| : x \in \mathcal{H}, w \in F(x)\} \tag{16}$$

is finite. Let $\tau_1^* > 0$ be such that $\mathcal{H} + \tau_1^* M\overline{\mathcal{B}} \subset \mathcal{G}$ and let L be a Lipschitz constant for the smooth function $\nabla V(\cdot)$ on the set $\mathcal{H} + \tau_1^* M\overline{\mathcal{B}}$. Define $\tau_2^* := \dfrac{\rho}{2LM^2}$ and $\tau^* := \min\{1, \tau_1^*, \tau_2^*\}$. Let $x \in \mathcal{H}$, $w \in F(x)$ and $\tau \in (0, \tau^*)$. It follows that the line segment connecting x to $x + \tau w$ belongs to $\mathcal{H} + \tau M\overline{\mathcal{B}} \subseteq \mathcal{H} + \tau_1^* M\overline{\mathcal{B}}$. Now, using the classical Mean Value Theorem, the condition (15), and the definitions of L, M and τ_2^*, we have the existence of $\lambda \in [0, 1]$ such that

$$\begin{aligned} V(x + \tau w) - V(x) &= \tau\langle \nabla V(x + \lambda\tau w), w\rangle \\ &\leq -\tau V(x) + \tau\langle \nabla V(x + \lambda\tau w) - \nabla V(x), w\rangle \\ &\leq -\tau V(x) + \tau^2 L|w|^2 \leq -\tau(V(x) - \tau LM^2) \\ &\leq -\tau\left(V(x) - \frac{\rho}{2}\right) \ . \end{aligned} \tag{17}$$

It follows, for the system (14), that the sets $\mathcal{H}$ and $\mathcal{N}$ are forward invariant and, for each $x_\circ \in \mathcal{H}$, $V(x_k)$ reaches the set $\mathcal{N}$ is a finite number of steps. ■

The proof of the previous theorem leads to:

Theorem 3 *Let Assumption 1 hold. Then there exists a continuous function $\tau : \mathcal{G} \to \mathbb{R}_{\geq 0}$ that is positive on $\mathcal{G}\backslash\mathcal{A}$ such that, for the difference inclusion*

$$x_{k+1} \in x_k + \tau(x_k)F(x_k) \ , \tag{18}$$

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

Proof. Following the proof of Theorem 2 and using (9), the positive real numbers L and M can be expressed as continuous, positive, nondecreasing functions of $V(x)$. Then we take $\tau_1(\cdot)$ continuous such that $x + \tau_1(x)M(V(x))\overline{\mathcal{B}} \subset \mathcal{G}$ for all $x \in \mathcal{G}$,

$$\tau_2(x) = \frac{V(x)}{2L(V(x))M(V(x))^2} \ , \tag{19}$$

and $\tau(x) = \min\{\tau_1(x), \tau_2(x)\}$. ■

Remark 3.5 When $F(x) = -\partial f(x)$ where f is *convex*, it is sufficient to have (see [10, Lemma 5.1])

$$\tau(x) < 2\frac{f(x) - f(x^*)}{\sup\{|\xi|^2 : \xi \in \partial f(x)\}} \qquad \forall x \neq x^* \ . \tag{20}$$

■

4 Nonsmooth minimization algorithms

Combining Proposition 1 with Theorem 2, we generate a generic class of quasi-Newton minimization algorithms for a locally Lipschitz, regular functions. The

basic algorithm assumes that there is a subroutine available that, for each $x \in \mathbb{R}^n$, returns one (arbitrary) element of $\partial f_\varepsilon(x)$. The algorithm is:

Core algorithm: *Let $x_{k+1} = x_k - \tau_k M_k \xi_k$ where $\xi_k \in \partial f_\varepsilon(x_k)$, $M_k \in \mathcal{M}(x_k)$ (where $\mathcal{M}$ satisfies (4)), and $\tau_k > 0$ is chosen small enough to satisfy the conditions of Theorem 2 or the considerations given below.*

This algorithm is not a descent algorithm in general, i.e., we do not necessarily have $f(x_{k+1}) < f(x_k)$. We do have $V(x_{k+1}) < V(x_k)$ (away from the set of minimizers), where $V(x)$ is the Lyapunov function used to prove Theorem 2, but it is typically not available to the minimization algorithm. This is similar to the situation encountered when using the subgradient method of nonsmooth minimization for locally Lipschitz, convex functions. In that case, the Euclidean norm of the distance from x to the minimizer of the function is a Lyapunov function (see, e.g., [10, Lemma 5.1]), but since the minimizer is not available, the Lyapunov function is not available. Algorithms based on Theorem 2 take small steps at each iteration and typically converge to a small neighborhood of the minimum rather than to the minimum. Moreover, to guarantee convergence to a small neighborhood of the minimum, small step sizes are required. For these reasons, we will look at ways to adjust the step size τ_k to achieve faster convergence to the minimum and to develop stopping conditions.

4.1 Convergence and stopping conditions

The key observation needed for adjusting the step size of our core algorithm to achieve convergence to the minimum or for producing a reliable stopping condition is given in the next proposition. We are operating under the assumptions of Proposition 1 and we take $F(x) = -\overline{\text{co}}\{\mathcal{M}(x)\partial f(x)\}$ and, for simplicity, $\mathcal{A} = \{x^*\}$, and $\mathcal{G} = \mathbb{R}^n$, in the conclusion of Proposition 1. We let $0 < \varepsilon_1 < \varepsilon_2$, define

$$\mathcal{C} := \mathcal{A} + \varepsilon_2 \overline{\mathcal{B}} \ , \quad \mathcal{D} := \mathcal{A} + \varepsilon_1 \overline{\mathcal{B}} \tag{21}$$

and let Theorem 2 generate $\tau^* > 0$.

Proposition 2 *For each $\tau \in (0, \tau^*)$, there exists a positive integer ℓ such that, for all sequences starting in $\mathcal{C}$ and satisfying $x_{k+1} \in x_k + \tau F(x_k)$, if $f(x_{k+\ell}) \geq f(x_k)$ then $x_{k+\ell} \in \mathcal{D}$.*

Proof. Following the proof of Theorem 2, we take $\omega(x) = f(x) - f(x^*)$ which, from the assumptions of Proposition 1, is a proper indicator function for $\{x^*\}$ on $\mathbb{R}^n$. We let $V(x)$, $\alpha_1(s)$, $\alpha_2(s)$, μ and ρ be as in the proof of Theorem 2. We define

$$\gamma := \max_{s \in [\rho, \mu]} \frac{\alpha_2 \circ \alpha_1^{-1}(s) - \frac{\rho}{2}}{s - \frac{\rho}{2}} \tag{22}$$

and we let ℓ be a positive integer such that

$$e^{\tau \ell} > \gamma \ . \tag{23}$$

It follows from the proof of Theorem 2 that

$$V(x_{k+\ell}) - \frac{\rho}{2} \leq e^{-\tau \ell} \left(V(x_k) - \frac{\rho}{2} \right) \ . \tag{24}$$

In particular, since $x_0 \in \mathcal{C}$,

$$V(x_{k+\ell}) \leq \mu \qquad \forall k \ . \tag{25}$$

Suppose $f(x_{k+\ell}) \geq f(x_k)$, which implies

$$\alpha_2 \circ \alpha_1^{-1}(V(x_{k+\ell})) \geq V(x_k) \ , \tag{26}$$

and $x_{k+\ell} \notin \mathcal{D}$, which implies

$$V(x_{k+\ell}) \geq \rho \ . \tag{27}$$

Combining (24) and (26), we have

$$\alpha_2 \circ \alpha_1^{-1}(V(x_{k+\ell})) \geq e^{\tau \ell} \left(V(x_{k+\ell}) - \frac{\rho}{2} \right) + \frac{\rho}{2} \tag{28}$$

or, using (22), (25) and (27),

$$e^{\tau \ell} \leq \frac{\alpha_2 \circ \alpha_1^{-1}(V(x_{k+\ell})) - \frac{\rho}{2}}{V(x_{k+\ell}) - \frac{\rho}{2}} \leq \gamma \ . \tag{29}$$

This contradiction of (23) proves the result. ■

Remark 4.1 It follows from the proof (in particular, (22) and (23)) that, in the case where $\alpha_2 \circ \alpha_1^{-1}(s) = s$, we can take $\ell = 1$. ■

4.1.1 Convergence: With Proposition 2, we can use ℓ and stored information about $f(x_k)$ to determine when convergence to the minimum is being impeded by τ being too large. Upon detection of this we can, in principle, decrease ε_1 by a small factor in the definition of $\mathcal{D}$, compute a new τ^* from Theorem 2, pick $\tau \in (0, \tau^*)$ and then calculate a new ℓ using Proposition 2 and then return to checking whether convergence to the minimum is being impeded by τ being too large. In this way, τ will converge to zero and x_k will converge to x^*.

4.1.2 Stopping condition: The discussion about convergence to the minimum leads to implementable stopping conditions. For example, it is natural to set a sufficiently small threshold for ε_1 and when ε_1 passes below this threshold and the blocking condition with respect to ℓ has been detected then the algorithm may terminate.

4.2 Incorporating line searches

In the case where line searches for the minimum of the function in the selected direction $M_k \xi_k$ are permitted, they may be used to speed up convergence. One way to prove convergence while taking large steps, perhaps corresponding to a step size returned by a line search subroutine, for algorithms that do not guarantee descent at every step is to make sure that the Lyapunov function decreases as a result of the large step. This

can be guaranteed by making sure that the decrease in the function to be minimized is sufficiently large. We emphasize that the lower bound on the line search parameter suggested by the preceding theory should be respected so that the disastrous behavior described in [10, Section 6.1] is explicitly avoided.

With Proposition 1's assumptions, $\exists \alpha_1, \alpha_2 \in \mathcal{K}_\infty$ s.t.

$$\alpha_1\left(f(x)-f(x^*)\right) \leq V(x) \leq \alpha_2\left(f(x)-f(x^*)\right) \ . \tag{30}$$

We note that in the case where $f(\cdot)$ is C^1, we can take α_1 and α_2 so that $\alpha_2^{-1} \circ \alpha_1(s) = s$. To guarantee that $V(x_{k+1}) \leq V(x_k)$ it is sufficient to have

$$\alpha_2\left(f(x_{k+1})-f(x^*)\right) \leq \alpha_1\left(f(x_k)-f(x^*)\right) \ , \tag{31}$$

equivalently,

$$f(x_{k+1}) \leq \alpha_2^{-1} \circ \alpha_1\left(f(x_k)-f(x^*)\right) + f(x^*) \ . \tag{32}$$

We note that the closer $\alpha_2^{-1} \circ \alpha_1$ is to the identity, the smaller the decrease in f is needed to permit moving x_{k+1} to the location returned by the line search subroutine. In the case where f is C^1 so that $\alpha_2^{-1} \circ \alpha_1(s) = s$, line searches are permitted at every step and our algorithm simplifies to standard quasi-Newton minimization algorithms with line search.

5 Conclusion

We have presented a methodology for solving nonsmooth optimization problems for locally Lipschitz, regular functions. Most of the ideas remain at a conceptual level and still need to be demonstrated on examples. This is the topic of future work. A companion paper [22] discusses the application of the methodology to the case of optimization when only function calls are available and gradient information is not available.

References

[1] Ya. I. Alber, A.N. Iusem, and M.V. Solodov. On the projected subgradient method for nonsmooth convex optimization in a Hilbert space. *Mathematical Programming*, 81 (1998), 23–35.

[2] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 1995.

[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] V.F. Dem'yanov and L.V. Vasil'ev. *Nondifferentiable optimization*. Optimization Software, Inc., 1985.

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

[7] K.C. Kiwiel. *Methods of descent for nondifferentiable optimization*. Springer-Verlag, Berlin, 1985.

[8] K.C.. Kiwiel. An ellipsoid trust region bundle method for nonsmooth convex optimization. *SIAM J. Control and Optimization*, **27** (1989), pp. 737–757.

[9] C. Lemaréchal, J.-J. Strodiot and A. Bihain. On a bundle algorithm for nonsmooth optimization, in Nonlinear Programming 4, O.L. Mangasarian, R.R. Meyer, and S.M. Robinson, eds., Academic Press, NY, 1981.

[10] 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.

[11] D.G. Luenberger. *Linear and nonlinear programming*, 2nd ed. Addison-Wesley, Reading, MA, 1984.

[12] L. Lukšan and J. Vlček. A bundle-Newton method for nonsmooth unconstrained minimization. *Mathematical Programming*, 83 (1998) 373–391.

[13] M.M. Mäkelä and P. Naittaanmäki. *Nonsmooth optimization: analysis and algorithms with applications to optimal control*. World Scientific, Singapore, 1992.

[14] R. Mifflin. An algorithm for constrained optimization with semismooth functions. *Mathematics of Operations Research*. Vol. 2, no. 2, May 1977, pp. 191–207.

[15] R. Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM J. Control and Optimization*. Vol. 15, no. 6, November 1977, pp. 959–972.

[16] R. Mifflin, D. Sun and L. Qi. Quasi-Newton bundle-type methods for nondifferentiable convex optimization. *SIAM J. Opt.* Vol. 8, no. 2, 1998, 583–603.

[17] L. Qi and X. Chen. A preconditioning proximal Newton method for nondifferentiable convex optimization. *Mathematical Programming*, 76 (1997) 411–429.

[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] H. Schramm and J. Zowe. A version of the bundle idea for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results. *SIAM J. Opt.* Vol. 2, no. 1, February 1992, pp. 121–152.

[20] 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.

[21] N.Z. Shor. *Minimization methods for nondifferentiable functions*. Springer-Verlag, Berlin, 1985.

[22] A.R. Teel. Lyapunov methods in nonsmooth optimization, Part II: persistently exciting gradient-free optimization. *39th IEEE CDC*, Sydney.

[23] A.R. Teel and L. Praly. A smooth Lyapunov function from a class-$\mathcal{KL}$ estimate involving two positive semidefinite functions. *ESAIM: Cont., Opt. & 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.

[24] 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.