APPROXIMATING SUBDIFFERENTIALS BY RANDOM SAMPLING OF GRADIENTS

J.V. Burke[*], A.S. Lewis[†], and M.L. Overton[‡]

December 11, 2001

Key words: nonsmooth analysis, Clarke subdifferential, generalized gradient, bundle method, stochastic gradient, eigenvalue optimization, spectral abscissa

AMS 2000 Subject Classification:
Primary: 90C46, 49K40
Secondary: 65K05, 15A42

Abstract

Many interesting real functions on Euclidean space are differentiable almost everywhere. All Lipschitz functions have this property, but so, for example, does the spectral abscissa of a matrix (a non-Lipschitz function). In practice, the gradient is often easy to compute. We investigate to what extent we can approximate the Clarke subdifferential of such a function at some point by calculating the convex hull of some gradients sampled at random nearby points.

[*]Department of Mathematics, University of Washington, Seattle, WA 98195, U.S.A. Email: burke@math.washington.edu. Research supported by NSF.

[†]Department of Combinatorics & Optimization, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. Email: aslewis@math.uwaterloo.ca. Research supported by NSERC.

[‡]Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, U.S.A. Email: overton@cs.nyu.edu. Research supported by NSF, DOE.

1 Introduction

Over the last quarter century there has been remarkable progress in the theoretical analysis of nonsmooth functions $f : \mathbf{R}^n \to \mathbf{R}$, primarily motivated by optimization. Clarke's introduction of his *generalized gradient* in 1973 (summarized in his seminal book [7]) pioneered a rapid development, recently presented in detail in [8] and [20].

Computational methods for nonsmooth optimization have also developed rapidly, with many interesting applications. For a recent look, see [19], which focuses on mechanical applications, or [17], which concentrates on optimal control. Nonsmooth optimization algorithms such as the *subgradient methods* outlined in [21], or the *bundle methods* described in [2, 15], typically assume a locally Lipschitz function, and at each iterate x_k compute only one element of the generalized gradient (or *subdifferential*). Even for convex functions it was recognized very early that an exact computation of the entire subdifferential was generally impractical [23]. Good overviews are [14] and [12].

In certain specially structured problems (such as nonlinear minimax), algorithms that can exploit the structure of the entire subdifferential are possible. An interesting example is a remarkable early paper on eigenvalue optimization [9] that uses the special structure of the function f to compute an *approximation* of the entire subdifferential, an avenue pursued further by Overton et al. (see [16] for a survey). In general-purpose nonsmooth optimization algorithms, however, the subdifferential *set* really only appears as a theoretical tool.

So, to what extent can we really "do" general nonsmooth analysis? Without assuming any particular structure for our function f, what might a general purpose algorithm learn about the subdifferential? Our aim in this paper is to approach these questions via random sampling of gradients at nearby points. *Stochastic gradient* algorithms have been analyzed in [10, 21], for example, but again the aim was to analyze algorithms working with a single subgradient at each iteration rather than to approximate the subdifferential.

Our starting point is to assume our function f is differentiable almost everywhere. By Rademacher's theorem, this is true for all locally Lipschitz functions, but there are interesting non-Lipschitz functions that have this property, including all "directionally Lipschitz" functions (in the sense of [20])—see [3], and also the *spectral radius* and *spectral abscissa* of a complex square matrix (respectively the largest modulus and real part of the eigenvalues), regarded as functions of the real and imaginary parts of its entries.

Indeed, it is hard to imagine a continuous function arising in a concrete setting that is not differentiable almost everywhere.

Our philosophy is to suppose that, wherever the function f is differentiable, *the gradient is cheap to compute*. For example, the spectral abscissa is differentiable at any matrix having a unique eigenvalue whose real part equals the spectral abscissa: the gradient is just vu^*, where u and v are corresponding right and left eigenvectors with $u^*v = 1$ (and furthermore, many such gradients can be computed in parallel). On the other hand, computing the subdifferential of the spectral abscissa at a general matrix is much harder, requiring some knowledge of the Jordan form of the underlying matrix [5]. Even for the much easier example of the maximum eigenvalue function on the space of Hermitian matrices, this convex function is easy to differentiate whenever the maximum eigenvalue has multiplicity one, but calculating the subdifferential in general requires a complete orthonormal set of corresponding eigenvectors.

What can we say in general? If the function f is locally Lipschitz around the point $\bar{x} \in \mathbf{R}^n$ then the Clarke subdifferential is given by

$$\partial_C f(\bar{x}) = \operatorname{conv}\{\lim_r \nabla f(x_r) : x_r \to \bar{x},\ x_r \in Q\}, \tag{1.1}$$

where conv denotes the convex hull operation and Q is *any* full-measure subset of a neighbourhood of $\bar{x}$ consisting of points where f is differentiable (see [7]: for the most part we follow the notation in the book of Rockafellar and Wets [20], where ∂_C is written $\bar{\partial}$). It is easy to see, in this Lipschitz setting, the relationship

$$\partial_C f(\bar{x}) = \bigcap_{\delta > 0} I_\delta, \tag{1.2}$$

where

$$I_\delta = \operatorname{cl}\operatorname{conv}\left(\nabla f(Q \cap (\bar{x} + \delta B))\right)$$

and B denotes the open unit ball in $\mathbf{R}^n$. This suggests that if we sample random points $x_1, x_2, \ldots, x_k \in \mathbf{R}^n$ near $\bar{x}$ and consider the set

$$C_k = \operatorname{conv}\{\nabla f(x_i) : i = 1, 2, \ldots, k\}, \tag{1.3}$$

then we might hope that C_k approximates $\partial_C f(\bar{x})$.

It is reasonably straightforward to see that this approximation does indeed work, in a suitable stochastic sense, for locally Lipschitz functions. In the non-Lipschitz case too there are some positive results. However, we

present some simple non-Lipschitz examples that reveal the difficulties of approximating the subdifferential in this manner.

In outline, we first show that the set C_k converges almost surely to the set I_δ defined above. In the Lipschitz case this proves the desired approximation, by equation (1.2). In the non-Lipschitz case we show that, under reasonable conditions, the set I_δ is still an *outer* approximation to the Clarke subdifferential, but examples show that it may be much too large.

The theory relating the Clarke subdifferential with limits of convex combinations of gradients at nearby points can also be considered in the light of "mollifiers": see [20, Thm 9.67, Eq 9(38) and p. 420] for details in the Lipschitz case and further references. Our approach here does not use mollifier theory.

2 The sampling framework

We consider a continuous function $f : \mathbf{R}^n \to \mathbf{R}$ that is differentiable almost everywhere. The gradient map $\nabla f : \mathbf{R}^n \to \mathbf{R}^n$ is Lebesgue measurable. To see this, fix any direction $w \in \mathbf{R}^n$ and consider the sequence of continuous functions

$$x \in \mathbf{R}^n \mapsto r(f(x + r^{-1}w) - f(x)), \quad \text{for } r = 1, 2, \ldots.$$

Then the lim sup of this sequence of functions is measurable and agrees with $w^T \nabla f(x)$ almost everywhere. (In fact, to see this we just need to assume f is measurable.)

Given real $\delta > 0$ (the *sampling radius*) and a point $\bar{x} \in \mathbf{R}^n$, we fix a sample space $\Omega = \bar{x} + \delta B$ with an associated probability measure, absolutely continuous with respect to Lebesgue measure λ on $\mathbf{R}^n$. We assume the corresponding density θ, is strictly positive almost everywhere on Ω. (We refer to [6] or [13], for example, for probabilistic terminology.) Thus θ is an integrable function satisfying $\int_\Omega \theta \, d\lambda = 1$ and $\theta > 0$ a.e. For example, we could choose $\theta \equiv \lambda(\Omega)^{-1}$.

With this probability space, we now consider a sequence of independent trials with outcomes $x_i \in \Omega$ for $i = 1, 2, \ldots$. Our assumptions on the density θ guarantee that for each trial the outcome x_i lies outside any fixed set of Lebesgue measure zero almost surely, and that x_i lies in any fixed nonempty open subset of Ω with a strictly positive probability that is independent of the trial number i.

From the outcomes x_i we construct a sequence of gradients $G_i = \nabla f(x_i)$. Since ∇f is measurable, $G_1, G_2, \ldots$ is a sequence of independent, identically distributed random vectors. Each random vector G_i corresponds to an induced probability measure: the measure of any Borel set $A \subset \mathbf{R}^n$ is just

$$\int_{(\nabla f)^{-1}(A)} \theta \, d\lambda.$$

Corresponding to this sequence of random vectors, we define a corresponding increasing sequence of closed convex random sets

$$C_k = \operatorname{conv} \{G_1, G_2, \ldots, G_k\} \quad (k = 1, 2, \ldots).$$

Our aim is to compare C_k with the Clarke subdifferential of f at $\bar{x}$. We call k the *sample size*.

We will not need any general discussion of random sets. The events we consider are measurable subsets of the infinite product space Ω^∞, typically of the form

$$\{(G_1, G_2, \ldots) \in S\}$$

for some closed set S: the probability of such an event is just its measure with respect to the product measure associated with the density θ.

Under reasonable conditions, the sets C_k converge to the closed convex hull of the image of the neighbourhood $\bar{x} + \delta B$ under the gradient map ∇f. This is the content of the following result. Thus a central question of this paper is how well this convex hull captures the Clarke subdifferential of f at $\bar{x}$.

Theorem 2.1 (limiting approximation) *Consider a continuous function $f : \mathbf{R}^n \to \mathbf{R}$ that is continuously differentiable on an open set $Q \subset \bar{x} + \delta B$ of full measure. Then*

$$\operatorname{cl} \bigcup_{k=1}^{\infty} C_k = \operatorname{cl} \operatorname{conv} \nabla f(Q) \quad \textit{almost surely,}$$

and for any direction $w \in \mathbf{R}^n$ we have

$$\max w^T C_k \uparrow \sup w^T \nabla f(Q) \quad \textit{as } k \to \infty, \textit{ almost surely.}$$

Proof Suppose a vector g lies in $\nabla f(Q)$. Given any real $\epsilon > 0$, since ∇f is continuous on Q, the set

$$\{x \in Q : \|\nabla f(x) - g\| < \epsilon\}$$

is open and nonempty. Hence for each index i we have

$$\mathrm{pr}\,\{\|G_i - g\| < \epsilon\} \quad > \quad 0,$$

so

$$g \in \bigcup_k C_k + \epsilon B \ \text{ almost surely.}$$

Since ϵ was arbitrary, we deduce

$$g \in \mathrm{cl} \bigcup_k C_k \ \text{ almost surely.}$$

Now choose a countable dense subset $\{g_1, g_2, \ldots\}$ of $\nabla f(Q)$, and observe, by the above,

$$g_r \in \mathrm{cl} \bigcup_k C_k \ \ (r = 1, 2, \ldots) \ \text{ almost surely.}$$

Then taking closed convex hulls shows

$$\mathrm{cl}\,\mathrm{conv}\,\nabla f(Q) \subset \mathrm{cl} \bigcup_k C_k \ \text{ almost surely.}$$

On the other hand, since Q has full measure, each x_i lies in Q almost surely, so G_i lies in $\mathrm{cl}\,\mathrm{conv}\,\nabla f(Q)$ almost surely. Hence each set C_k is contained in $\mathrm{cl}\,\mathrm{conv}\,\nabla f(Q)$ almost surely, so we have

$$\mathrm{cl}\,\mathrm{conv}\,\nabla f(Q) \supset \bigcup_k C_k \ \text{ almost surely,}$$

and the first equation follows by taking closures. The final claim then follows easily. ♣

(The continuity assumption on f is in fact superfluous, since the other assumptions imply that f is measurable.)

Our assumption on the continuous differentiability of f is stronger than we need for most of our paper. However, it seems reasonable in practice. For example, since the matrices with distinct eigenvalues form an open set of full measure, this assumption holds for the spectral abscissa. Corollary 5.11 provides a variant of the result where the continuous differentiability of f is dropped in favour of local Lipschitzness.

3 Sampling gradients: the Lipschitz case

Let us suppose first that the function f is locally Lipschitz around the point $\bar{x}$. In this case the analysis is reasonably straightforward.

Theorem 3.1 (inner approximation) *If f is locally Lipschitz around $\bar{x}$ then for any real $\epsilon > 0$ we have, for any sufficiently small sampling radius,*

$$C_k \subset \partial_C f(\bar{x}) + \epsilon B \text{ for } k = 1, 2, \ldots, \quad \text{almost surely,}$$

and so

$$\operatorname{cl} \bigcup_{k=1}^{\infty} C_k \subset \partial_C f(\bar{x}) + \epsilon \operatorname{cl} B \quad \text{almost surely.}$$

Proof The Clarke subdifferential is upper semicontinuous at $\bar{x}$, so there exists a radius $\delta > 0$ such that

$$\partial_C f(\bar{x} + \delta B) \subset \partial_C f(\bar{x}) + \epsilon B.$$

But C_k is almost surely contained in the left hand side. The result therefore follows. ♣

Example 7.1 (overestimating the subdifferential) shows how this result can fail for non-Lipschitz functions.

For the opposite inclusion we use the following lemma. We define the *regular subderivative* (or *Clarke directional derivative*) at $\bar{x}$ as the finite sublinear function $\hat{d}f(\bar{x}) : \mathbf{R}^n \to \mathbf{R}$ given by

$$\hat{d}f(\bar{x})(w) = \max w^T \partial_C f(\bar{x}).$$

Lemma 3.2 *For all real $\epsilon > 0$, directions $w \in \mathbf{R}^n$, and indices $i = 1, 2, \ldots$, we have*

$$\operatorname{pr}\{G_i^T w > \hat{d}f(\bar{x})(w) - \epsilon\} \quad > \quad 0.$$

Proof By our assumptions on the probability density θ, it suffices to show that the measurable set

$$\{x \in \bar{x} + \delta B : \nabla f(x)^T w > \hat{d}f(\bar{x})(w) - \epsilon\}$$

has strictly positive Lebesgue measure. Suppose this fails, so $\nabla f(x)^T w \leq \hat{d}f(\bar{x})(w) - \epsilon$ for all points x in Q, a subset of $\bar{x} + \delta B$ of full measure. Using

our definition of the Clarke subdifferential (1.1), we can choose a sequence $\{x^r\}$ in Q approaching $\bar{x}$ such that $\nabla f(x^r)^T w \to \hat{d}f(\bar{x})(w)$, and this is a contradiction. ♣

Theorem 3.3 (outer approximation) *If f is locally Lipschitz around $\bar{x}$ then for any sufficiently small sampling radius and any real $\epsilon > 0$ we have*

$$\text{pr}\,\{\partial_C f(\bar{x}) \subset C_k + \epsilon B\} \quad \to \quad 1 \text{ as } k \to \infty.$$

Proof Denote the unit sphere in $\mathbf{R}^n$ by S and the Lipschitz constant of f on $\bar{x} + \delta B$ by L, and choose points $w_1, w_2, \ldots, w_m$ in S such that

$$S \subset \bigcup_{j=1}^{m} \Big(w_j + \frac{\epsilon}{3L} B\Big).$$

By Lemma 3.2 we know, for each index $i = 1, 2, \ldots$ and $j = 1, 2, \ldots, m$, the probability

$$\text{pr}\,\{G_i^T w_j > \hat{d}f(\bar{x})(w_j) - \epsilon/3\}$$

is strictly positive, and independent of i. Hence, for each j we have

$$\text{pr}\Big\{ \max_{1 \le i \le k} G_i^T w_j > \hat{d}f(\bar{x})(w_j) - \epsilon/3 \Big\} \quad \to \quad 1 \text{ as } k \to \infty,$$

so

$$\text{pr}\Big\{ \max_{1 \le i \le k} G_i^T w_j > \hat{d}f(\bar{x})(w_j) - \epsilon/3 \text{ for each } j \Big\} \quad \to \quad 1 \text{ as } k \to \infty,$$

or in other words

$$\text{pr}\Big\{ \max w_j^T C_k > \hat{d}f(\bar{x})(w_j) - \epsilon/3 \text{ for each } j \Big\} \quad \to \quad 1 \text{ as } k \to \infty.$$

Now notice that the functions $w \in \mathbf{R}^n \mapsto \max w^T C_k$ and $\hat{d}f(\bar{x})(\cdot)$ both have Lipschitz constant L, so by our choice of the w_j's, the inequalities

$$\max w_j^T C_k > \hat{d}f(\bar{x})(w_j) - \epsilon/3 \text{ for } j = 1, 2, \ldots, m$$

imply

$$\max w^T C_k > \hat{d}f(\bar{x})(w) - \epsilon \text{ for all } w \in S,$$

which in turn implies

$$\partial_C f(\bar{x}) \subset C_k + \epsilon B.$$

The result now follows. ♣

Such results clearly may fail for non-Lipschitz functions since the set C_k is always bounded whereas the subdifferential may be unbounded.

Corollary 3.4 *If f is locally Lipschitz around $\bar{x}$ then for any sufficiently small sampling radius we have*

$$\partial_C f(\bar{x}) \subset \mathrm{cl} \bigcup_{k=1}^{\infty} C_k \quad \textit{almost surely.}$$

Proof By Theorem 3.3, $\partial_C f(\bar{x}) \subset \epsilon B + \cup_k C_k$, almost surely, for any real $\epsilon > 0$, and the result follows. ♣

We extend this result in Theorem 5.13.

In summary, in the locally Lipschitz case, Theorem 3.1 (inner approximation) says that providing we sample gradients close to $\bar{x}$, the sets C_k will not overestimate the subdifferential $\partial_C f(\bar{x})$ too badly, while Theorem 3.3 (outer approximation) says that, as we increase our sample size, the probability of underestimating the subdifferential shrinks to zero.

4 Non-Lipschitz analysis

The variational analysis of non-Lipschitz functions is more subtle than the Lipschitz case. In this section we summarize the notions we use. At the risk of slight notational extravagance, we introduce a new subdifferential-like object, which we call the "convex-stable subdifferential". We make no attempt to study its properties as a subdifferential, but rather observe how it arises naturally in our gradient-sampling framework, and how it compares with the Clarke subdifferential. We refer throughout to [20].

We suppose, as before, that the function $f : \mathbf{R}^n \to \mathbf{R}$ is continuous. The *regular subdifferential* of f at a point $x \in \mathbf{R}^n$ is the set of vectors $y \in \mathbf{R}^n$ satisfying

$$f(\bar{x} + z) \geq f(\bar{x}) + y^T z + o(z) \quad \text{for small } z \in \mathbf{R}^n.$$

We denote this closed convex set $\hat{\partial} f(\bar{x})$, and we define the *subdifferential* of f at $\bar{x}$ by

$$\partial f(\bar{x}) = \bigcap_{\delta > 0} \mathrm{cl}\,(\hat{\partial} f(\bar{x} + \delta B)).$$

Thus y lies in $\partial f(x)$ if and only if there are sequences $x_r \to x$ and $y_r \to y$ with $y_r \in \hat{\partial} f(x_r)$ for all r. This object has become fundamental in modern variational analysis. When f is locally Lipschitz we have $\partial_C f(x) = \operatorname{conv} \partial f(x)$ [20, Thm 9.61].

Part of the subtlety of non-Lipschitz analysis arises from *horizon* behaviour. The *horizon cone* of a nonempty set $C \subset \mathbf{R}^n$ is the closed cone

$$C^\infty = \{\lim_r t_r y_r : t_r \downarrow 0,\ y_r \in C\},$$

and we define $\emptyset^\infty = \{0\}$. Thus a set is bounded exactly when its horizon cone is $\{0\}$. We call a cone $K \subset \mathbf{R}^n$ *pointed* when $K \cap -K = \{0\}$. The *horizon subdifferential* of a continuous function $f : \mathbf{R}^n \to \mathbf{R}$ at a point $x \in \mathbf{R}^n$ is the closed cone

$$\partial^\infty f(x) = \{\lim_r t_r y_r : t_r \downarrow 0,\ y_r \in \hat{\partial} f(x_r),\ x_r \to x\} \cup \{0\}.$$

It is easy to check $\partial^\infty f(x) = \{0\}$ if f is locally Lipschitz. The polar cone of the horizon subdifferential is the closed cone

$$\partial^\infty f(x)^* = \{w : w^T y \leq 0 \text{ for all } y \in \partial^\infty f(x)\},$$

and the *regular subderivative* of f at x is the sublinear function $\hat{d} f(x) : \mathbf{R}^n \to [-\infty, +\infty])$ defined by

$$\hat{d} f(x)(w) = \begin{cases} \sup w^T \partial f(x) & \text{if } w \in \partial^\infty f(x)^* \\ +\infty & \text{otherwise} \end{cases}$$

[20, Ex 8.23]. We can then define the *Clarke subdifferential* of f at x as the closed convex set

$$\partial_C f(x) = \{y : w^T y \leq \hat{d} f(x)(w) \text{ for all } w \in \mathbf{R}^n\}$$

[20, Thm 8.49]. These definitions agree with our previous notions in the Lipschitz case. We call the point x *Clarke-critical* if $0 \in \partial_C f(x)$.

Since we interpret $\sup \emptyset = -\infty$, we see from our definitions that $\partial f(x) = \emptyset$ if and only if $\partial_C f(x) = \emptyset$ (cf. [20, Thm 8.49]). Assuming $\partial f(x)$ is nonempty, we call f *regular* at x if

$$\partial f(x) = \hat{\partial} f(x) \neq \emptyset \quad \text{and} \quad \hat{\partial} f(x)^\infty = \partial^\infty f(x)$$

[20, Cor 8.11]. In the next result we collect some useful representations of the Clarke subdifferential.

Proposition 4.1 (Clarke subdifferential) *For any continuous function $f : \mathbf{R}^n \to \mathbf{R}$, the Clarke subdifferential at x has the following representations:*

(i) $\partial_C f(x) = \mathrm{cl}\,(\mathrm{conv}\,\partial f(x) + \mathrm{conv}\,\partial^\infty f(x))$;

(ii) $\partial_C f(x) = \mathrm{conv}\,\partial f(x) + \mathrm{conv}\,\partial^\infty f(x)$ *providing* $\partial^\infty f(x)$ *is pointed;*

(iii) $\partial_C f(x) = \partial f(x)$ *providing* f *is regular at* x.

Proof (i) This result appears for example in [18, p.58, Prop. 2.6]: for convenience, we outline a short proof. If $\partial f(x) = \emptyset$ then both sides of the equation are empty. Hence we can assume $\partial f(x) \neq \emptyset$, so both sides are nonempty closed convex sets. It suffices to show the corresponding support functions coincide. On the left hand side we obtain $\hat{d}f(x)(w)$, and on the right,

$$\sup w^T \partial f(x) + \sup w^T \partial^\infty f(x).$$

These two functions of the vector w clearly coincide by our definition of the regular subderivative. Indeed, for all $w \in \partial^\infty f(x)^*$ the second term in the above sum is zero, while for all other w it is $+\infty$.

(ii) See [20, Thm 8.49].

(iii) This part follows easily from part (i) (cf. [20, p. 337]). ♣

We remark that continuity of f in the above results could in fact be relaxed to lower semicontinuity.

An instructive example is the function

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ -\sqrt{x} & \text{if } x \geq 0. \end{cases} \tag{4.2}$$

An easy calculation shows

$$\hat{\partial} f(0) = \emptyset, \quad \partial f(0) = \{0\}, \quad \partial^\infty f(0) = \mathbf{R}_- = \partial_C f(0).$$

Note f is not regular at 0, and $\partial_C f(0) \neq \mathrm{cl}\,\mathrm{conv}\,\partial f(0)$.

The following notion of a *convex-stable subdifferential* will be helpful later: we define

$$\tilde{\partial} f(x) = \bigcap_{\delta > 0} \mathrm{cl}\,\mathrm{conv}\,(\hat{\partial} f(x + \delta B)).$$

For example, for the function f above we have $\tilde{\partial} f(0) = \mathbf{R}_-$. The next result shows that this subdifferential is at least as large as the Clarke subdifferential.

Proposition 4.3 (Clarke versus convex-stable subdifferential) *If the function $f : \mathbf{R}^n \to \mathbf{R}$ is continuous, then at any point $x \in \mathbf{R}^n$ we have*

$$\tilde{\partial} f(x) \supset \partial_C f(x).$$

Proof For any real $\delta > 0$ we need to show

$$\operatorname{cl}\operatorname{conv}(\hat{\partial} f(x + \delta B)) \supset \partial_C f(x).$$

Observe

$$\partial f(x) \subset \operatorname{cl}\operatorname{conv}(\hat{\partial} f(x + \delta B)) \quad \text{and} \quad \partial^\infty f(x) \subset (\operatorname{cl}\operatorname{conv}(\hat{\partial} f(x + \delta B)))^\infty.$$

We deduce

$$\begin{aligned} &\operatorname{conv}\partial f(x) + \operatorname{conv}\partial^\infty f(x) \\ &\quad\subset \operatorname{cl}\operatorname{conv}(\hat{\partial} f(x + \delta B)) + (\operatorname{cl}\operatorname{conv}(\hat{\partial} f(x + \delta B)))^\infty \\ &\quad= \operatorname{cl}\operatorname{conv}(\hat{\partial} f(x + \delta B)), \end{aligned}$$

by [20, Thm 3.6]. The desired inclusion now follows by taking closures and applying Proposition 4.1(i). ♣

Example 7.2 shows that the inclusion can be strict.

The next result shows that limits of convex combinations of "convex-stable subgradients" at nearby points must themselves be convex-stable subgradients.

Proposition 4.4 (stabilizing subdifferentials) *If the function $f : \mathbf{R}^n \to \mathbf{R}$ is continuous, then at any point $x \in \mathbf{R}^n$ we have*

$$\begin{aligned} \tilde{\partial} f(x) &= \bigcap_{\delta>0} \operatorname{cl}\operatorname{conv}(\hat{\partial} f(x + \delta B)) \\ &= \bigcap_{\delta>0} \operatorname{cl}\operatorname{conv}(\partial f(x + \delta B)) \\ &= \bigcap_{\delta>0} \operatorname{cl}\operatorname{conv}(\partial_C f(x + \delta B)) \\ &= \bigcap_{\delta>0} \operatorname{cl}\operatorname{conv}(\tilde{\partial} f(x + \delta B)) \end{aligned}$$

Proof Since each expression contains the previous one (using the previous result), it suffices to show that the last expression is contained in $\tilde{\partial} f(x)$. For this, we simply observe the inclusion

$$\operatorname{cl}\operatorname{conv}(\tilde{\partial} f(x+\delta' B)) \subset \operatorname{cl}\operatorname{conv}(\hat{\partial} f(x+\delta B))$$

whenever $0 < \delta' < \delta$. ♣

Finally we show, under a horizon condition, that limits of convex combinations of *Clarke* subgradients (and so, in particular, of gradients) at nearby points are themselves Clarke subgradients.

Theorem 4.5 (stability of Clarke subdifferential) *For any continuous function $f : \mathbf{R}^n \to \mathbf{R}$ and any point $x \in \mathbf{R}^n$, if $\partial^\infty f(x)$ is pointed then*

$$\tilde{\partial} f(x) = \partial_C f(x).$$

Proof By Proposition 4.3, we need to show that if

$$v \in \bigcap_{\delta>0} \operatorname{cl}\operatorname{conv}(\hat{\partial} f(x+\delta B))$$

then $v \in \partial_C f(x)$. Assume the above condition holds. Then by the theorem of Carathéodory, for each integer $j = 0, 1, \ldots, n$ there exist sequences $\{u_j^i\}$ and $\{v_j^i\}$ in $\mathbf{R}^n$ and $\{\lambda_j^i\}$ in $\mathbf{R}_+$ (indexed by $i \in \mathbf{N}$) satisfying

$$\begin{aligned}
\sum_{j=0}^{n} \lambda_j^i &= 1 \quad \text{for all } i, \\
\lim_{i\to\infty} u_j^i &= x \quad \text{for all } j, \\
v_j^i &\in \hat{\partial} f(u_j^i) \quad \text{for all } i, j, \quad \text{and} \\
\lim_{i\to\infty} \sum_{j=0}^{n} \lambda_j^i v_j^i &= v.
\end{aligned}$$

We claim that for each index j the sequence $\{\lambda_j^i v_j^i\}$ is bounded. If not, there is an index j' and a subsequence $I \subset \mathbf{N}$ such that

$$\lambda_{j'}^i \|v_{j'}^i\| = \max_j \lambda_j^i \|v_j^i\| \to +\infty \quad \text{as } i \to \infty \text{ in } I.$$

Call the left hand side σ^i. Since for each j the sequence $\{(\lambda_j^i/\sigma^i)v_j^i\}$ is bounded, we can assume (taking a further subsequence) that it converges to some vector $v_j \in \partial^\infty f(x)$ as $i \to \infty$ in I. Now notice

$$\|v_{j'}\| = 1 \quad \text{and} \quad \sum_{j=0}^{n} v_j = 0.$$

But since $\partial^\infty f(x)$ is pointed, so is its convex hull [20, Thm 8.49], and this is a contradiction.

Hence, as we claimed, for each index j the sequence $\{\lambda_j^i v_j^i\}$ is bounded, so we can assume (taking a subsequence) that it converges to some vector $w_j \in \mathbf{R}^n$, and furthermore that the bounded scalar sequence $\{\lambda_j^i\}$ converges to some scalar $\lambda_j \in \mathbf{R}_+$. We then have

$$\sum_{j=0}^{n} w_j = v \quad \text{and} \quad \sum_{j=0}^{n} \lambda_j = 1.$$

Define a (nonempty) index set $J = \{j : \lambda_j > 0\}$. For $j \in J$ we have

$$\lambda_j^{-1} w_j = \lim_{i \to \infty} v_j^i \in \partial f(x),$$

whereas for $j \notin J$ we have

$$w_j = \lim_{i \to \infty} \lambda_j^i v_j^i \in \partial^\infty f(x).$$

Hence

$$\begin{aligned} v &= \sum_{j \in J} \lambda_j (\lambda_j^{-1} w_j) + \sum_{j \notin J} w_j \\ &\in \operatorname{conv} \partial f(x) + \operatorname{conv} \partial^\infty f(x) \\ &= \partial_C f(x), \end{aligned}$$

by Proposition 4.1(ii). ♣

Functions with a pointed horizon subdifferential at a point are called *directionally Lipschitz* there [20].

5 Sampling gradients: the non-Lipschitz case

We now return to our gradient-sampling framework in the case where the continuous function $f : \mathbf{R}^n \to \mathbf{R}$ may not be Lipschitz, but is nonetheless differentiable almost everywhere. We shall see that this case is more problematic, but we begin with some positive results.

Given the convex nature of our proposed approximation (1.3), we hope to approximate the Clarke subdifferential $\partial_C f(\bar{x})$. Our approximation is motivated by the relationship (1.2). To what extent does this relationship still hold if f is not Lipschitz? The result below, which is essentially a slight reworking of Clarke's original argument [7, p. 63], states a one-sided inclusion. It assumes f is *absolutely continuous on lines* near $\bar{x}$: that is, for any points u and v near $\bar{x}$, the function

$$t \in [0,1] \mapsto f(tu + (1-t)v) \tag{5.1}$$

is absolutely continuous. This is automatic for locally Lipschitz functions. It also holds for the spectral radius and abscissa: the space of matrices stratifies into submanifolds (according to Jordan structure) on each of which these functions are analytic [1], which shows that the function (5.1) is piecewise differentiable with piecewise monotonic derivative, and hence absolutely continuous [22, 4.50].

Theorem 5.2 (covering gradients) *Suppose, near the point $\bar{x} \in \mathbf{R}^n$, the function $f : \mathbf{R}^n \to \mathbf{R}$ is continuous, absolutely continuous on lines, and differentiable almost everywhere, and Q is a full-measure subset of a neighbourhood of $\bar{x}$ consisting of points where f is differentiable. Then the Clarke and convex-stable subdifferentials satisfy*

$$\partial_C f(\bar{x}) \subset \tilde{\partial} f(\bar{x}) = \bigcap_{\delta > 0} \operatorname{cl}\operatorname{conv}\left(\nabla f(Q \cap (\bar{x} + \delta B))\right).$$

Proof We first show, for any real $\delta > 0$,

$$\hat{\partial} f(\bar{x}) \subset \operatorname{cl}\operatorname{conv}\left(\nabla f(Q \cap (\bar{x} + \delta B))\right). \tag{5.3}$$

To see this, suppose a vector $y \in \mathbf{R}^n$ does not lie in the right hand side, so by separation there is a vector $z \in \mathbf{R}^n$ and real k such that $y^T z > k$ but

$$\nabla f(x)^T z \leq k \quad \text{for all } x \in Q \cap (\bar{x} + \delta B). \tag{5.4}$$

If $y \in \hat{\partial} f(\bar{x})$ then for small real t we have

$$f(\bar{x} + tz) \geq f(\bar{x}) + y^T tz + o(t),$$

so there exists $t \in (0, \delta/(2\|z\|))$ such that

$$f(\bar{x} + tz) > f(\bar{x}) + kt.$$

By continuity, for all points $w \in \mathbf{R}^n$ close to $\bar{x}$ we have

$$f(w + tz) > f(w) + kt. \tag{5.5}$$

By Fubini's theorem [22, 6.124], almost all w close to $\bar{x}$ satisfy

$$w + sz \in Q \quad \text{for almost all } s \in [0, t]. \tag{5.6}$$

Therefore we can choose w in $\bar{x} + (\delta/2)B$ satisfying both (5.5) and (5.6).

Now consider the function $g : [0, 1] \to \mathbf{R}$ defined by $g(s) = f(w + sz)$. By assumption, g is absolutely continuous and for almost all $s \in [0, t]$ we have

$$g'(s) = \nabla f(w + sz)^T z \leq k,$$

by (5.6) and (5.4). By the Fundamental Theorem of Calculus [22, 6.85] we deduce $g(t) \leq g(0) + kt$, which contradicts inequality (5.5). Hence $y \notin \hat{\partial} f(\bar{x})$, so we have proved the inclusion (5.3).

We now apply this inclusion at points in a neighbourhood of $\bar{x}$ and for suitable δ to obtain

$$\operatorname{cl}\operatorname{conv}(\hat{\partial} f(\bar{x} + \delta'' B)) \subset \operatorname{cl}\operatorname{conv}(\nabla f(Q \cap (\bar{x} + \delta' B))) \subset \operatorname{cl}\operatorname{conv}(\hat{\partial} f(\bar{x} + \delta' B))$$

whenever $0 < \delta'' < \delta'$, and the equality in the main result follows. The inclusion for the Clarke subdifferential is a consequence of Proposition 4.3 (Clarke versus convex-stable subdifferential). ♣

In passing, we note the analogy between the right hand side of the inclusion we have just proved, and Fillipov's notion of a generalized solution for differential equations with discontinuous right hand sides [11].

Returning to the sampling scheme described in Section 2, we can now generalize Lemma 3.2.

Lemma 5.7 *If the function f satisfies the assumptions of Theorem 5.2 then for all real $\epsilon > 0$, directions $w \in \mathbf{R}^n$, and indices $i = 1, 2, \ldots$ we have*

$$\operatorname{pr}\{w^T G_i > \hat{d}f(\bar{x})(w) - \epsilon\} \quad > \quad 0.$$

Proof It suffices to show that the measurable set

$$\{x \in \bar{x} + \delta B : w^T \nabla f(x) > \hat{d}f(\bar{x})(w) - \epsilon\}$$

has strictly positive Lebesgue measure. Suppose this fails, and define Q to be the set of points $x \in \bar{x} + \delta B$ where f is differentiable and

$$w^T \nabla f(x) \leq \hat{d}f(\bar{x})(w) - \epsilon.$$

Theorem 5.2 shows $\partial_C f(\bar{x}) \subset \operatorname{cl}\operatorname{conv} \nabla f(Q)$. Hence

$$\hat{d}f(\bar{x})(w) \leq \sup w^T \nabla f(Q) \leq \hat{d}f(\bar{x})(w) - \epsilon,$$

which is a contradiction. ♣

We deduce the following non-Lipschitz version of Theorem 3.3 (outer approximation), stating, loosely speaking, that our approximation to the subdifferential gives a good *upper* approximation to the regular subderivative.

Theorem 5.8 (upper approximation) *Suppose, near the point $\bar{x} \in \mathbf{R}^n$, the function $f : \mathbf{R}^n \to \mathbf{R}$ is continuous, absolutely continuous on lines, and differentiable almost everywhere. Then for any direction $w \in \mathbf{R}^n$ we have*

$$\lim_{k \to \infty} \max w^T C_k \geq \hat{d}f(\bar{x})(w) \quad \textit{almost surely.}$$

Proof This follows from Lemma 5.7. ♣

Example 7.1 shows that the inequality in the above result may be strict.

With more care we can gain a little more insight into this result. The following theorem parallels Theorem 2.1 (limiting approximation).

Theorem 5.9 (directional approximation) *Suppose, near a point $\bar{x} \in \mathbf{R}^n$, the function $f : \mathbf{R}^n \to \mathbf{R}$ is continuous, absolutely continuous on lines, and differentiable on a full measure subset $Q \subset \mathbf{R}^n$. Then, for any direction $w \in \mathbf{R}^n$ and sufficiently small sampling radius $\delta > 0$,*

$$\max w^T C_k \uparrow \sup w^T \nabla f(Q \cap (\bar{x} + \delta B)) \quad \textit{as } k \to \infty, \quad \textit{almost surely.}$$

Proof With probability one, each $x_i \in Q$, so $G_i \in \operatorname{cl}\operatorname{conv} \nabla f(Q \cap (\bar{x} + \delta B))$. Hence each set C_k is almost surely contained in $\operatorname{cl}\operatorname{conv} \nabla f(Q \cap (\bar{x} + \delta B))$, the right hand side is almost surely an upper bound.

We now claim, for all real $\epsilon > 0$ and indices $i = 1, 2, \ldots$,

$$\operatorname{pr}\{w^T G_i > \sup w^T \nabla f(Q \cap (\bar{x} + \delta B)) - \epsilon\} \quad > \quad 0. \tag{5.10}$$

To see this, choose a point $\tilde{x} \in Q \cap (\bar{x} + \delta B)$ satisfying

$$w^T \nabla f(\tilde{x}) > \sup w^T \nabla f(Q \cap (\bar{x} + \delta B)) - \frac{\epsilon}{2}.$$

Let $\tilde{\delta} = \delta - \|\tilde{x} - \bar{x}\| > 0$. Then, as in the proof of Lemma 5.7, the measurable set

$$\left\{x \in \tilde{x} + \tilde{\delta} B : w^T \nabla f(x) > w^T \nabla f(\tilde{x}) - \frac{\epsilon}{2}\right\}$$

has strictly positive Lebesgue measure, and our claim (5.10) follows. This proves the result. ♣

In the Lipschitz case this gives a variant of Theorem 2.1 (limiting approximation).

Corollary 5.11 (Lipschitz approximation) *Suppose the function f is locally Lipschitz around $\bar{x}$ and differentiable on a full-measure subset Q of a neighbourhood of $\bar{x}$. Then for any sufficiently small sampling radius $\delta > 0$,*

$$\operatorname{cl} \bigcup_{k=1}^{\infty} C_k = \operatorname{cl}\operatorname{conv} \nabla f(Q \cap (\bar{x} + \delta B)) \quad \textit{almost surely.}$$

Proof Denote the left and right hand side sets by C and D respectively. Clearly these closed convex sets satisfy $C \subset D$, almost surely. By Theorem 5.9 (directional approximation) we know the support functions agree at any given vector in $\mathbf{R}^n$, almost surely. Hence in fact they agree on any given countable dense subset of $\mathbf{R}^n$, almost surely. But D is bounded, whence so is C, so both support functions are continuous. Hence the support functions are identical, almost surely, and the result follows. ♣

A natural test for optimality, in our gradient sampling scheme, is to ask the question

$$0 \in C_k? \tag{5.12}$$

The next result relates this test to finding a Clarke-critical point.

Theorem 5.13 (detecting critical points) *Suppose, on a neighbourhood of $\bar{x}$,*

(i) f is locally Lipschitz, or

(ii) f is absolutely continuous on lines and continuously differentiable on a full-measure open subset.

Then

$$\partial_C f(\bar{x}) \subset \operatorname{cl} \bigcup_{k=1}^{\infty} C_k \quad \textit{almost surely.}$$

Consequently, dist$(0, C_k) \to 0$ almost surely whenever $\bar{x}$ is a Clarke-critical point of f, and furthermore

$$0 \in \operatorname{int} \partial_C f(\bar{x}) \quad \Rightarrow \quad 0 \in \operatorname{int} C_k \ \textit{eventually, almost surely.}$$

Proof Under either assumption we know

$$\operatorname{cl} \bigcup_{k=1}^{\infty} C_k = \operatorname{cl} \operatorname{conv} \nabla f(Q) \ \text{ almost surely,}$$

where Q is a full-measure subset of a neighbourhood of $\bar{x}$, using Theorem 2.1 (limiting approximation) or Corollary 5.11 (Lipschitz approximation). By Theorem 5.2 (covering gradients) we deduce the first inclusion. The result about Clarke-critical points now follows, and the last implication is a consequence of the fact that

$$\operatorname{int} \operatorname{cl} \bigcup_{k=1}^{\infty} C_k = \operatorname{int} \bigcup_{k=1}^{\infty} C_k = \bigcup_{k=1}^{\infty} \operatorname{int} C_k,$$

by convexity and nestedness. ♣

We have stated this result for the Clarke subdifferential, although obviously there is a completely parallel result replacing the Clarke subdifferential with the convex-stable subdifferential throughout.

What about the opposite inclusion in Theorem 5.2 (covering gradients)? To understand the issues here we need to consider horizon behaviour more carefully.

6 The Clarke subdifferential and the horizon condition

Consider once again a continuous function $f : \mathbf{R}^n \to \mathbf{R}$. If the horizon subdifferential $\partial^\infty f(\bar{x})$ is pointed (so f is directionally Lipschitz around $\bar{x}$) we know that f is differentiable almost everywhere near $\bar{x}$ [3]. We also know that the Clarke subdifferential coincides with the convex-stable subdifferential, by Theorem 4.5 (stability of Clarke subdifferential). Putting this fact together with Theorem 5.2 (covering gradients) leads to the following result, providing conditions under which convex combinations of gradients at nearby points give a good approximation of the Clarke subdifferential.

Corollary 6.1 (gradient-based approximation) *Suppose that, close to the point $\bar{x} \in \mathbf{R}^n$, the function $f : \mathbf{R}^n \to \mathbf{R}$ is continuous, and absolutely continuous on lines, with $\partial^\infty f(\bar{x})$ pointed. If Q is a full-measure subset of a neighbourhood of $\bar{x}$ consisting of points where f is differentiable, then*

$$\partial_C f(\bar{x}) = \bigcap_{\delta > 0} \operatorname{cl} \operatorname{conv} \nabla f(Q \cap (\bar{x} + \delta B)).$$

In passing, we remark that a directionally Lipschitz function even of one variable may not be absolutely continuous. For example, the "Lebesgue singular function" [22, Ex. 3.138] is continuous and nondecreasing on the interval $[0, 1]$, and hence directionally Lipschitz, but it is not absolutely continuous.

Our next result shows that, in searching for Clarke-critical points, providing our sampling radius is sufficiently small, the test $0 \in C_k$ will not generate a false positive, even approximately.

Corollary 6.2 (false positives) *Suppose, close to the point $\bar{x} \in \mathbf{R}^n$, the function $f : \mathbf{R}^n \to \mathbf{R}$ is continuous, with $\partial^\infty f(\bar{x})$ pointed. If $\bar{x}$ is not a Clarke-critical point of f then for any sufficiently small sampling radius we have*

$$\lim_{k \to \infty} \operatorname{dist}(0, C_k) > 0 \quad \textit{almost surely.}$$

Proof As above, we know f is differentiable almost everywhere near $\bar{x}$. Since $0 \notin \partial_C f(\bar{x})$, we know, for all small $\delta > 0$,

$$0 \notin \operatorname{cl} \operatorname{conv} (\hat{\partial} f(\bar{x} + \delta B)) \supset \operatorname{cl} \bigcup_{k=1}^{\infty} C_k \quad \text{almost surely,}$$

by Theorem 4.5 (stability of Clarke subdifferential). ♣

This suggests a conceptual algorithm for generating descent directions, outlined in the next result.

Corollary 6.3 (descent directions) *Suppose, near the point $\bar{x} \in \mathbf{R}^n$, the function $f : \mathbf{R}^n \to \mathbf{R}$ is continuous, and absolutely continuous on lines, with $\partial^\infty f(\bar{x})$ pointed. The sequence of random vectors*

$$Y_k = \textit{closest point to } 0 \textit{ in } C_k, \quad \textit{for } k = 1, 2, \ldots,$$

converges almost surely to a limit $Y \in \mathbf{R}^n$, and if $\bar{x}$ is not a Clarke-critical point of f and the sampling radius is small, then

$$\hat{d}f(\bar{x})(-Y) < 0 \quad \textit{almost surely.}$$

Proof As before, f is differentiable almost everywhere near $\bar{x}$. It is routine to check that the sequence Y_k converges to the closest point to 0 in the set $C = \operatorname{cl} \cup_k C_k$. (For example, the nonincreasing sequence of functions $\|\cdot\| + \delta_{C_k}$ epi-converges to $\|\cdot\| + \delta_C$, by [20, Prop 7.4], and is level-bounded, so the minimizers converge as required, by [20, Thm 7.33].)

Since $0 \notin \operatorname{cl} \cup_k C_k$, by Corollary 6.2 (false positives) we deduce

$$0 < \inf Y^T \bigcup_{k=1}^{\infty} C_k = \lim_{k \to \infty} \min Y^T C_k \leq -\hat{d}f(\bar{x})(-Y),$$

by Theorem 5.8 (upper approximation). ♣

Notice, as with Theorem 5.13 (detecting critical points), we have stated the above two corollaries for the Clarke subdifferential, although there are analogous results for the convex-stable subdifferential needing no pointedness assumption.

At least when the function f is Lipschitz near $\bar{x}$, the above result is reassuring. When $0 \notin \partial_C f(\bar{x})$ and we pick a small sampling radius, any approximation $\bar{Y}$ close to Y will satisfy

$$\limsup_{x \to \bar{x},\ t \downarrow 0} \frac{f(x - t\bar{Y}) - f(x)}{t} = \hat{d}f(\bar{x})(-\bar{Y}) < 0,$$

by [20, Ex 9.15] and the continuity of $\hat{d}f(\bar{x})$. Thus $-\bar{Y}$ is a descent direction which is stable with respect to small perturbations, both to itself and to the base point $\bar{x}$.

On the other hand, when f is not Lipschitz around $\bar{x}$, the subdifferential $\partial_C f(\bar{x})$ is unbounded [20, Thm 9.13], so the choice of descent direction may be highly sensitive under perturbation.

7 Examples

We end with some simple examples illustrating the delicate features of non-Lipschitz optimization in this framework. The functions we consider here even satisfy the strong assumption of regularity for all points x near the point of interest $\bar{x}$ (so in particular the Clarke subdifferential coincides locally with the subdifferential).

Example 7.1 (overestimating subderivatives) We show here how, for certain directions $w \in \mathbf{R}^n$, the estimate $\max w^T C_k$ may be much larger than the regular subderivative $\hat{d}f(\bar{x})(w)$, even when f is regular near $\bar{x}$ and satisfies all the assumptions of Corollary 6.3 (descent directions).

We define a function $f : \mathbf{R}^2 \to \mathbf{R}$ by

$$f(x) = \sqrt{(\|x\| - 1)^+} \ ,$$

where, for real u, we define $u^+ = \max\{u, 0\}$. A direct calculation shows

$$\hat{\partial} f(x) = \partial f(x) = \partial_C f(x) = \begin{cases} \{0\} & (\|x\| < 1) \\ \left\{ \left(2\|x\| \sqrt{(\|x\| - 1)^+} \right)^{-1} x \right\} & (\|x\| > 1) \\ \mathbf{R}_+ x & (\|x\| = 1), \end{cases}$$

and

$$\hat{\partial} f(x)^\infty = \partial^\infty f(x) = \begin{cases} \{0\} & (\|x\| \neq 1) \\ \mathbf{R}_+ x & (\|x\| = 1). \end{cases}$$

Thus f is everywhere regular and satisfies all the assumptions of Corollary 6.3 at any point: furthermore, it is continuously differentiable on the full-measure open set $\{x : \|x\| \neq 1\}$.

Consider the point $\bar{x} = (1, 0)$. Our calculations are slightly easier if, rather than a circular neighbourhood, we consider, in polar coordinates, the neighbourhood

$$N \quad = \quad \{x = (r, \theta) : |\theta| < \delta, \ |r - 1| < \delta\}.$$

Let $Q = \{x \in N : \|x\| \neq 1\}$, an open, full-measure subset of N. Then, essentially by Theorem 2.1 (limiting approximation), if we sample our points x_i from N we obtain, almost surely,

$$\begin{aligned} \operatorname{cl} \bigcup_{k=1}^{\infty} C_k &= \operatorname{cl} \operatorname{conv} \nabla f(Q) \\ &= \operatorname{cl} \operatorname{conv} \left(\{0\} \cup \{(r,\theta) : |\theta| < \delta,\ 2r\sqrt{\delta} > 1\}\right) \\ &= \{(r,\theta) : |\theta| \leq \delta\}. \end{aligned}$$

Since

$$\partial_C f(\bar{x}) = \{(r,\theta) : \theta = 0\},$$

the conclusion of Theorem 3.1 (inner approximation) fails.

We see from this that our approximations may give overestimates for regular subderivatives: returning to Cartesian coordinates, if $w = (0,1)$ then $\hat{d}f(\bar{x})(w) = 0$, whereas $\lim_{k\to\infty} w^T C_k = \infty$ almost surely.

Notice finally that, if we denote the spectral abscissa of a matrix by α, then we can write our function in the following form:

$$f(x_1, x_2) = \alpha \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ x_1 - 1 & x_2 & 0 & 0 \\ x_2 & -x_1 - 1 & 0 & 0 \end{bmatrix}.$$

Example 7.2 (The horizon condition) This example shows the importance of the horizon condition in Section 6. We define $f : \mathbf{R}^2 \to \mathbf{R}$ by

$$f(x) = \sqrt{|\|x\| - 1|}.$$

A direct calculation shows f is continuously differentiable on the full-measure open set $\{x : \|x\| \neq 0, 1\}$,

$$\hat{\partial} f(x) = \partial f(x) = \partial_C f(x) = \begin{cases} \left\{ \frac{\operatorname{sgn}(\|x\|-1)}{2\|x\|\sqrt{|\|x\|-1|}} x \right\} & (\|x\| \neq 0, 1) \\ \mathbf{R}x & (\|x\| = 1) \end{cases}$$

(where, for nonzero real u we define $\operatorname{sgn} u = u/|u|$), and

$$\hat{\partial} f(x)^\infty = \partial^\infty f(x) = \begin{cases} \{0\} & (\|x\| \neq 0, 1) \\ \mathbf{R}x & (\|x\| = 1). \end{cases}$$

Thus f is regular everywhere except at the origin (where we could easily smooth it, if so desired), and it satisfies all the assumptions of Corollary 6.3 at any nonzero point x, except that $\partial f(\bar{x})$ has nonpointed horizon cone when $\|\bar{x}\| = 1$.

We claim, for any sampling radius $\delta > 0$, at any point $\bar{x}$ with $\|\bar{x}\| = 1$ we have

$$(7.3)\quad \tilde{\partial} f(\bar{x}) = \operatorname{cl}\operatorname{conv}(\partial f(\bar{x} + \delta B)) = \operatorname{cl}\operatorname{conv}\nabla f(Q \cap (\bar{x} + \delta B)) = \mathbf{R}^2,$$

(where Q is any full measure subset of a neighbourhood of $\bar{x}$ consisting of points where f is differentiable). That is, the conclusions of Theorem 4.5 (stability of Clarke subdifferential) and Corollary 6.1 (gradient-based approximation) both fail badly.

To see this, without loss of generality choose $\bar{x} = (0, 1)$. As before, our calculations are slightly easier if we use the neighbourhood N (along with the subset Q) and the sampling scheme of the previous example. We then obtain, essentially by Theorem 2.1 (limiting approximation) again, almost surely,

$$\begin{aligned}
\operatorname{cl} \bigcup_{k=1}^{\infty} C_k &= \operatorname{cl}\operatorname{conv}\nabla f(Q) \\
&= \operatorname{cl}\operatorname{conv}\{(r,\theta) : \theta \in (-\delta, \delta) \cup (\pi - \delta, \pi + \delta),\ 2r\sqrt{\delta} > 1\} \\
&= \mathbf{R}^2.
\end{aligned}$$

We deduce equations (7.3) easily, and our observation follows.

Returning to Cartesian coordinates, consider the new function $\tilde{f}(x) = f(x) - x_2$. Clearly $0 \notin \partial_C \tilde{f}(1, 0)$; that is, $(1, 0)$ is not a Clarke-critical point of $\tilde{f}$. Indeed, tracing around the unit circle from the point $(1, 0)$ causes the value of $\tilde{f}$ to decrease locally at linear rate. However, for any sampling radius $\delta > 0$, for this new function

$$0 \in \operatorname{int} C_k \quad \text{eventually, almost surely}$$

(using the same convexity argument as in the proof of Theorem 5.13 (detecting critical points)), so our optimality test will always give a false positive for large enough sample size: Corollaries 6.2 (False positives) and 6.3 (Descent directions) both fail.

Example 7.4 (Large samples) Our last example shows that even though Theorem 5.13 guarantees that we can, in some sense, detect Clarke-critical points, we may require a large sample size.

Consider the function $f : \mathbf{R}^2 \to \mathbf{R}$ defined by

$$f(x_1, x_2) = 2\max\{|x_1|, \sqrt{|x_2|}\} + x_1.$$

A calculation shows f is everywhere regular, with

$$\partial f(0,0) = [-1, 3] \times \mathbf{R},$$

so $0 \in \operatorname{int} \hat{\partial} f(0,0)$: the origin is a "sharp" local minimizer, since, for example $f(x) \geq \|x\|/2 \geq 0 = f(0)$ for all small x.

We know that f is continuously differentiable on the full-measure open set $\{x : x_2 \neq \pm x_1^2\}$ and absolutely continuous on lines, so Theorem 5.13 guarantees

$$0 \in \operatorname{int} C_k \quad \text{eventually, almost surely.} \tag{7.5}$$

To make our calculations slightly easier, let us take our points x_i uniformly distributed on $[-\delta, \delta]^2$. Note $C_k \subset [1,3] \times \mathbf{R}$ unless $|(x_i)_2| < (x_i)_1^2$ and $x_i < 0$ for some index $i \in \{1, 2, \ldots, k\}$. The probability of this event is $\delta/6$ for each i, so $C_k \subset [1,3] \times \mathbf{R}$ with probability $(1 - \delta/6)^k$.

Suppose we choose our sample size as $\lfloor 1/\delta \rfloor$ (the largest integer less than $1/\delta$). Our argument shows, despite (7.5), we have

$$\operatorname{pr}\{\operatorname{dist}(0, C_{\lfloor 1/\delta \rfloor}) \geq 1\} \quad \to \quad e^{-1/6} \text{ as } \delta \downarrow 0.$$

A similar calculation shows that to ensure the distance of C_k from the origin is less than one with any given strictly positive probability, we need a sample size growing like $1/\delta$. In summary, we need a very large sample size to secure $0 \in C_k$.

Denoting the spectral abscissa by α, we can write our function in the form

$$f(x_1, x_2) = \alpha \begin{bmatrix} 3x_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -x_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & x_1 & 2 & 0 & 0 \\ 0 & 0 & 2x_2 & x_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & x_1 & 2 \\ 0 & 0 & 0 & 0 & -2x_2 & x_1 \end{bmatrix}$$

Notice that at the minimum the corresponding matrix is *derogatory*: the multiple zero eigenvalue appears in several Jordan blocks. Such solutions are, in a well-defined sense, atypical in spectral abscissa minimization. The minimizing matrix for the problem at the end of Example 7.1 is, by contrast, nonderogatory. It is unclear whether the phenomena exhibited in Examples 7.2 and 7.4 can occur in typical spectral abscissa minimization problems. For a discussion of what we mean by "typical", see [4].

Acknowledgements We thank J.M. Borwein, F.H. Clarke, B. Mordukhovich, R.T. Rockafellar and R.J.-B. Wets for helpful suggestions.

References

[1] V.I. Arnold. On matrices depending on parameters. *Russian Mathematical Surveys*, 26:29–43, 1971.

[2] M.L. Balinski and P. Wolfe, editors. *Nondifferentiable Optimization*, Mathematical Programming Study 3, 1975.

[3] J.M. Borwein, J.V. Burke, and A.S. Lewis. Differentiability of cone-monotone functions on separable Banach space. Technical report, Simon Fraser University, 2001.

[4] J.V. Burke, A.S. Lewis, and M.L. Overton. Optimal stability and eigenvalue multiplicity. *Foundations of Computational Mathematics*, 1:205–225, 2001.

[5] J.V. Burke and M.L. Overton. Variational analysis of non-Lipschitz spectral functions. *Mathematical Programming*, 90:317–352, 2001.

[6] K.L. Chung. *A Course in Probability Theory.* Academic Press, New York, second edition, 1974.

[7] F.H. Clarke. *Optimization and Nonsmooth Analysis.* Wiley, New York, 1983. Republished as Vol. 5, Classics in Applied Mathematics, SIAM, 1990.

[8] F.H. Clarke, Yu.S. Ledyaev, R.J. Stern, and P.R. Wolenski. *Nonsmooth Analysis and Control Theory.* Springer-Verlag, New York, 1998.

[9] J. Cullum, W.E. Donath, and P. Wolfe. The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices. *Mathematical Programming Study*, 3:35–55, 1975.

[10] Yu.M. Ermoliev. Methods of nondifferentiable and stochastic optimization and their applications. In E.A. Nurminski, editor, *Progress in Nondifferentiable Optimization*, pages 5–27, Laxenburg, Austria, 1982. International Institute for Applied Systems Analysis.

[11] A.F. Filippov. *Differential Equations with Discontinuous Righthand Sides*. Kluwer, Norwell, MA, 1988.

[12] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer-Verlag, Berlin, 1993.

[13] J.F.C. Kingman and S.J. Taylor. *Introduction to Measure and Probability*. Cambridge University Press, Cambridge, 1966.

[14] K.C. Kiwiel. *Minimization Methods for Non-Differentiable Functions*, volume 1133 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1985.

[15] C. Lemaréchal. Nondifferentiable optimization. In G.L. Nemhauser, A.H.G. Rinnooy Kan, and M.J. Todd, editors, *Optimization*, volume 1 of *Handbooks in Operations Research and Management Science*, pages 529–572. North-Holland, Amsterdam, 1989.

[16] A.S. Lewis and M.L. Overton. Eigenvalue optimization. *Acta Numerica*, 5:149–190, 1996.

[17] M.M. Makela and P. Neittaanmaki. *Nonsmooth Optimization*. World Scientific, Singapore, 1992.

[18] B.S. Mordukhovich. *Approximation Methods in Problems of Optimization and Control*. Nauka, Moscow, 1988. 2nd English edition to be published by Wiley.

[19] J. Outrata, M. Kocvara, and J. Zowe. *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*. Kluwer, Dordrecht, 1998.

[20] R.T. Rockafellar and R.J.-B. Wets. *Variational Analysis*. Springer, Berlin, 1998.

[21] N.Z. Shor. *Minimization Methods for Non-Differentiable Functions*. Springer-Verlag, Berlin, 1985.

[22] K.R. Stromberg. *An Introduction to Classical Real Analysis*. Wadsworth, Belmont, CA, 1981.

[23] P. Wolfe. A method of conjugate subgradients for minimizing nondifferentiable functions. *Mathematical Programming Study*, 3:145–173, 1975.