Nonlinear Analysis 75 (2012) 1282–1289

Contents lists available at SciVerse ScienceDirect

Nonlinear Analysis

journal homepage: www.elsevier.com/locate/na

On Nesterov's nonsmooth Chebyshev–Rosenbrock functions

Mert Gürbüzbalaban *, Michael L. Overton

Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012, USA

ARTICLE INFO

Communicated by Ravi Agarwal

Keywords:
Nonsmooth optimization
Optimization algorithms
Variational analysis

ABSTRACT

We discuss two nonsmooth functions on $\mathbf{R}^n$ introduced by Nesterov. We show that the first variant is partly smooth in the sense of Lewis and that its only stationary point is the global minimizer. In contrast, we show that the second variant has 2^{n-1} Clarke stationary points, none of them local minimizers except the global minimizer, but also that its only Mordukhovich stationary point is the global minimizer. Nonsmooth optimization algorithms from multiple starting points generate iterates that approximate all 2^{n-1} Clarke stationary points, not only the global minimizer, but it remains an open question as to whether the nonminimizing Clarke stationary points are actually points of attraction for optimization algorithms.

Published by Elsevier Ltd

1. Introduction

In 2008, Nesterov [1] introduced the following *smooth* (differentiable, in fact polynomial) function on $\mathbf{R}^n$:

$$\tilde{f}(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1}(x_{i+1} - 2x_i^2 + 1)^2.$$

The only stationary point of f is the global minimizer $x^* = [1, 1, \ldots, 1]^T$. Consider the point $\hat{x} = [-1, 1, 1, \ldots, 1]^T$ and the manifold

$$\mathcal{M} = \{x : x_{i+1} - 2x_i^2 + 1 = 0,\ i = 1, \ldots, n-1\}$$

which contains both x^* and $\hat{x}$. For $x \in \mathcal{M}$,

$$x_{i+1} = 2x_i^2 - 1 = T_2(x_i) = T_{2^i}(x_1), \quad i = 1, \ldots, n-1,$$

where $T_k(x)$ denotes the kth Chebyshev polynomial of the first kind [2, Section 2.4].

The function $\tilde{f}$ is the sum of a quadratic term and a nonnegative sum whose zero set is the manifold $\mathcal{M}$. Minimizing $\tilde{f}$ is equivalent to minimizing the first quadratic term on $\mathcal{M}$. Standard optimization methods, such as Newton's method and the BFGS quasi-Newton method, when applied to minimize $\tilde{f}$ and initiated at $\hat{x}$, generate iterates that, as in the well known Rosenbrock example [3] and its extensions [4], approximately "track" $\mathcal{M}$ to approach the minimizer. The iterates do not track $\mathcal{M}$ exactly, but because they typically follow this highly oscillatory manifold fairly closely, the tracking process requires many iterations. To move from $\hat{x}$ to x^* along $\mathcal{M}$ *exactly* would require x_n to trace the graph of the 2^{n-1}th Chebyshev polynomial, which has $2^{n-1} - 1$ extrema in $(-1, 1)$, as x_1 increases from -1 to 1. Hence, $\tilde{f}$ is a challenging test problem for optimization methods.

[*] Corresponding author.
E-mail addresses: mert@cims.nyu.edu (M. Gürbüzbalaban), overton@cs.nyu.edu (M.L. Overton).

0362-546X/$ – see front matter. Published by Elsevier Ltd
doi:10.1016/j.na.2011.07.062

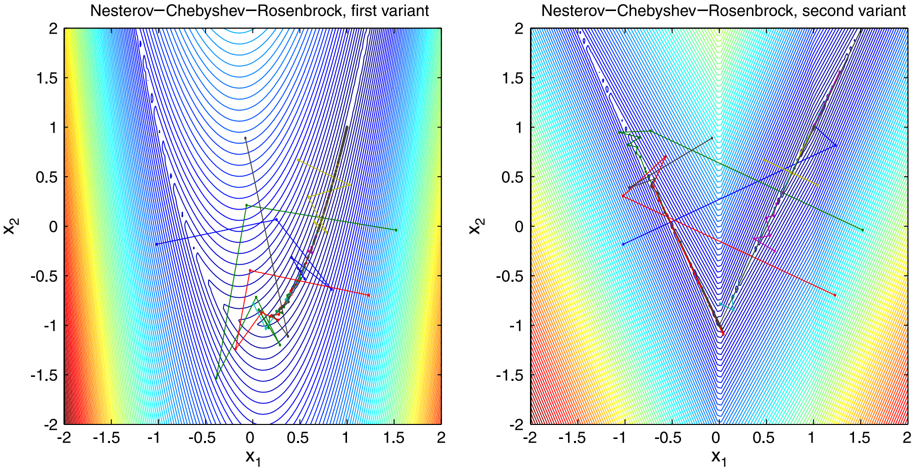

Fig. 1. Contour plots for Nesterov's first (left) and second (right) non-smooth Chebyshev–Rosenbrock functions $\hat{f}$ and f respectively, with $n = 2$. Points connected by line segments show the iterates generated by the BFGS method (see Section 3) initialized at 7 different randomly generated starting points (iterates plotted later may overwrite those plotted earlier). For the first variant $\hat{f}$, convergence always takes place to the only Clarke stationary point: the global minimizer $x^* = [1, 1]^T$. For the second variant f, some runs of BFGS generate iterates that approximate the nonminimizing Clarke stationary point $[0, -1]^T$ while others converge to the minimizer $[1, 1]^T$.

Nesterov also introduced two *nonsmooth* variants of $\tilde{f}$, the first being

$$\hat{f}(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|. \tag{1}$$

A contour plot of this function when $n = 2$ is shown on the left side of Fig. 1. Again, the unique global minimizer is x^*. Like $\tilde{f}$, the function $\hat{f}$ is the sum of a quadratic term and a nonnegative sum whose zero set is the manifold $\mathcal{M}$, so, as previously, minimizing $\hat{f}$ is equivalent to minimizing the first quadratic term on $\mathcal{M}$, but unlike $\tilde{f}$, the function $\hat{f}$ is not differentiable at points in $\mathcal{M}$. However, as we show in the next section, $\hat{f}$ is *partly smooth* with respect to $\mathcal{M}$, in the sense of [5], at points in $\mathcal{M}$. It follows that, like $\tilde{f}$, the function $\hat{f}$ has only one stationary point — the global minimizer x^* — where by stationary point we mean both in the sense of Clarke and of Mordukhovich.

The second nonsmooth variant is

$$f(x) = \frac{1}{4}|x_1 - 1| + \sum_{i=1}^{n-1} |x_{i+1} - 2|x_i| + 1|. \tag{2}$$

Again, the unique global minimizer is x^*. Consider the set

$$S = \{x : x_{i+1} - 2|x_i| + 1 = 0,\ i = 1, \ldots, n-1\}. \tag{3}$$

Minimizing f is equivalent to minimizing its first term on S. Like $\mathcal{M}$, the set S is highly oscillatory, but it has "corners": it is not a manifold around any point x where any of the components $x_1, \ldots, x_{n-1}$ vanishes. For example, consider the case $n = 2$, for which a contour plot is shown on the right side of Fig. 1. It is easy to verify that the point $[0, -1]^T$ is Clarke stationary (zero is in the convex hull of gradient limits at the point), but not a local minimizer ($[1, 2]^T$ is a direction of linear descent from $[0, -1]^T$). We will show in the next section that, in fact, f has 2^{n-1} Clarke stationary points, that the only local minimizer is the global minimizer x^*, and furthermore that the only stationary point in the sense of Mordukhovich is x^*.

In the next section, we define stationarity in both senses and present the main results. In Section 3, we report on numerical experiments showing the behavior of nonsmooth minimization algorithms on these functions.

2. Main results

Before stating our main results, we will need the following well-known definitions. The Clarke subdifferential or generalized gradient [6] of a locally Lipschitz function on a finite-dimensional space can be defined as follows [7, Theorem 6.2.5]. Let ∇ denote gradient.

Definition 1 (*Clarke Subdifferential*). Consider a function $\phi : \mathbf{R}^n \to \mathbf{R}$ and a point $x \in \mathbf{R}^n$, and assume that ϕ is locally Lipschitz around x. Let $\mathcal{G} \subset \mathbf{R}^n$ be the set of all points where ϕ is differentiable, and $A \subset \mathbf{R}^n$ be an arbitrary set with measure zero. Then the Clarke subdifferential of ϕ at x is

$$\partial^C \phi(x) = \operatorname{conv} \left\{ \lim_{m\to\infty} \nabla\phi(x^m) : x^m \to x,\ x_m \in \mathcal{G},\ x^m \notin A \right\}. \tag{4}$$

Note that by Rademacher's Theorem [8], locally Lipschitz functions are differentiable almost everywhere so A can be chosen as the set of points at which ϕ is not differentiable.

As expounded in [9], the Mordukhovich [10] subdifferential is defined as follows.

Definition 2 (*Mordukhovich Subdifferential*)**.** Consider a function $\phi : \mathbf{R}^n \to \mathbf{R}$ and a point $x \in \mathbf{R}^n$. A vector $v \in \mathbf{R}^n$ is a *regular subgradient* of ϕ at x (written $v \in \hat{\partial}\phi(x)$) if

$$\liminf_{\substack{z \to x \\ z \neq x}} \frac{\phi(z) - \phi(x) - \langle v, z - x \rangle}{|z - x|} \geq 0,$$

where $\langle \cdot\,,\ \cdot \rangle$ is the usual inner product on $\mathbf{R}^n$. A vector $v \in \mathbf{R}^n$ is a *Mordukhovich subgradient* of ϕ at x (written $v \in \partial^M \phi(x)$) if there exist sequences x^m and v^m in $\mathbf{R}^n$ satisfying

$$\begin{aligned} & x^m \to x \\ & \phi(x^m) \to \phi(x) \\ & v^m \in \hat{\partial}\phi(x^m) \\ & v^m \to v. \end{aligned}$$

We say that ϕ is *Clarke stationary* at x if $0 \in \partial^C \phi(x)$. Similarly, ϕ is *Mordukhovich stationary* at x if $0 \in \partial^M \phi(x)$. For a locally Lipschitz function ϕ, we have [9, Theorem 8.49]

$$\partial^C \phi(x) = \operatorname{conv} \{\partial^M \phi(x)\}. \tag{5}$$

The following simple example illustrates equation (5).

Example 1. For $g(x) = |x_1| - |x_2|, x \in \mathbf{R}^2$, explicit formulas for the Clarke and Mordukhovich subdifferentials can be derived at $x = [0, 0]^T$. Using Definitions 1 and 2, a straightforward computation leads to

$$\partial^C g([0, 0]^T) = [-1, 1] \times [-1, 1] \quad \text{and} \quad \partial^M g([0, 0]^T) = [-1, 1] \times \{-1, 1\},$$

where the former subdifferential is the convex hull of the latter one.

We will need the concept of regularity (also known as subdifferential regularity or Clarke regularity) [9], which can be characterized for locally Lipschitz functions as follows [11, Theorem 6.10].

Definition 3 (*Regularity*)**.** A locally Lipschitz function $\phi : \mathbf{R}^n \to \mathbf{R}$ is regular at a point x if and only if its ordinary directional derivative satisfies

$$\phi'(x; d) = \limsup_{z \to x} \ \langle \nabla \phi(z), d \rangle$$

for every direction $d \in \mathbf{R}^n$.

One consequence of regularity of ϕ at a point x is that $\partial^C \phi(x) = \partial^M \phi(x)$ [12, Proposition 4.1(iii)] and another is that the Clarke stationarity condition $0 \in \partial^C \phi(x)$ is equivalent to the first-order optimality condition $\phi'(x, d) \geq 0$ for all directions d [13, Section 14.1].

A property that will be central in our analysis is *partial smoothness* [5].

Definition 4. A function ϕ is *partly smooth* at x relative to a manifold $\mathcal{X}$ containing x if

1. its restriction to $\mathcal{X}$, denoted by $\phi_{|\mathcal{X}}$, is twice continuously differentiable at x,
2. at every point close to $x \in \mathcal{X}$, the function ϕ is regular and has a subgradient,
3. par $\{\partial^M \phi(x)\}$, the subspace parallel to the affine hull of the subdifferential of ϕ at x, is the normal subspace to $\mathcal{X}$ at x, and
4. the subdifferential map $\partial^M \phi$ is continuous at x relative to $\mathcal{X}$.

We illustrate the definition by proving that $\hat{f}$ is partly smooth.

Lemma 1. *Nesterov's first nonsmooth Chebyshev–Rosenbrock function* $\hat{f}$*, defined in* (1)*, is partly smooth with respect to* $\mathcal{M}$ *at all points in* $\mathcal{M}$*.*

Proof. For each $i \in \{1, \ldots, n-1\}$, consider the function $h_i(x) = |x_{i+1} - 2x_i^2 + 1|$ and the manifold $\mathcal{M}_i := \{x : H_i(x) := x_{i+1} - 2x_i^2 + 1 = 0\}$. By the chain rule [9, Proposition 10.5], h_i is globally regular as a composition of two regular functions and we have

$$\partial^M h_i(x) = \nabla H_i(x) \left[\left\{ \partial^M |\cdot| \right\} (x_{i+1} - 2x_i^2 + 1) \right]$$

for any $x \in \mathbf{R}^n$. We have the normal space $N_{\mathcal{M}_i} = \text{Range}(\nabla H_i(x))$ [9, Example 6.8] which is clearly parallel to the subdifferential $\partial^M h_i(x)$. Since $h_{i|\mathcal{M}_i} = 0$ is smooth and $\partial^M h_i$ is continuous at x relative to $\mathcal{M}_i$, it follows from Definition 4 that h_i is partly smooth with respect to the manifold $\mathcal{M}_i$. We conclude from [5, Corollary 4.6 and 4.7] that $\hat{f}(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} h_i(x)$ is partly smooth with respect to the manifold $\mathcal{M} = \cap_{i=1}^{n-1} \mathcal{M}_i$ at all points in $\mathcal{M}$. □

It follows that $\hat{f}$ has only one stationary point.

Theorem 1. *Nesterov's first nonsmooth Chebyshev–Rosenbrock function $\hat{f}$ is Clarke stationary or Mordukhovich stationary only at the unique global minimizer $x^* = [1, 1, \ldots, 1]^T$.*

Proof. If $x \notin \mathcal{M}$, then $\hat{f}$ is smooth and nonstationary at x as the partial derivative of $\hat{f}$ with respect to x_n is ± 1. On the other hand, if $x \in \mathcal{M}$, then the restricted function $\hat{f}_{|\mathcal{M}} = \frac{1}{4}(x_1 - 1)^2$ is smooth and has a critical point only at the global minimizer x^*. If $x \in \mathcal{M}$ and $x \neq x^*$, it follows from [5, Proposition 2.4] that $0 \notin \text{aff}\, \partial^M \hat{f}(x)$. Thus, $0 \notin \partial^M \hat{f}(x)$. By regularity, we have $\partial^C \hat{f}(x) = \partial^M \hat{f}(x)$ and the result follows. □

The main results of the paper concern Nesterov's second nonsmooth example. For this we will need the usual sign function:

$$\text{sign}(x) = \begin{cases} 1 & : x > 0, \\ 0 & : x = 0, \\ -1 & : x < 0. \end{cases}$$

We start by stating a simple lemma.

Lemma 2. *Let S be defined as in (3). There are $2^{n-1} - 1$ points in S such that $x_j = 0$ for some $j < n$. Let $\bar{x}$ be such a point. Then $\bar{x}_i$ takes non-integer values between -1 and 1 for $i < j$, $\bar{x}_i = 0$ for $i = j$, $\bar{x}_i = -1$ for $i = j + 1$ and $\bar{x}_i = 1$ if $n \geq i > j + 1$. In particular, $\bar{x}_1 < 1$ (with $\bar{x}_1 = 0$ if $j = 1$).*

Proof. For $j < n$ fixed, it is easy to see that there are 2^{j-1} points in S such that $x_j = 0$. Summing over j, we obtain $2^{n-1} - 1 = \sum_{j=1}^{n-1} 2^{j-1}$ points. The rest of the proof is straightforward. □

Theorem 2. *Nesterov's second nonsmooth Chebyshev–Rosenbrock function f, defined in (2), is Clarke stationary at the $2^{n-1} - 1$ points in the set S with a vanishing x_j for some $j < n$.*

Proof. Let $\bar{x}$ be such a point. Then, using Lemma 2, we see that around $\bar{x}$ the function $\frac{|x_1 - 1|}{4}$ is equal to $\frac{1 - x_1}{4}$ and furthermore $\bar{x}_i \neq 0$ if $i \neq j$. These observations allow us to write f in a simpler form eliminating most of the absolute values. We first prove the case $j = n - 1$. We will show that in an arbitrarily small neighborhood of $\bar{x}$ the gradient vector (if defined) can take arbitrary signs in each coordinate. This will ensure that $0 \in \partial^C f(\bar{x})$ by (4).

Around $\bar{x}$, the function $f(x)$ may be rewritten as

$$f(x) = \frac{1 - x_1}{4} + |x_2 + 2c_1 x_1 + 1| + \cdots + |x_{n-1} + 2c_{n-2} x_{n-2} + 1| + |x_n - 2|x_{n-1}| + 1| \tag{6}$$

where $c_i = -\text{sign}(\bar{x}_i)$, $i = 1, 2, \ldots, n - 2$, depend only on the point $\bar{x}$ and are fixed, (note that $\bar{x}_i \neq 0$ for $i < j = n - 1$). Since $\bar{x} \in S$ and $\bar{x}_{n-1} = 0$, all the absolute value terms appearing in (6) are equal to 0 at $\bar{x}$. By the continuity of f at $\bar{x}$, it is possible to find points x arbitrarily close to $\bar{x}$ such that each of the absolute value terms evaluated at x has arbitrary sign and at those points

$$\nabla f(x) = \left[-\frac{1}{4} + 2c_1 d_1, d_1 + 2c_2 d_2, \ldots, d_{n-2} + 2c_{n-1} d_{n-1}, d_{n-1} \right]^T$$

where $c_{n-1} := -\text{sign}(x_{n-1})$ and each of $d_1, d_2, \ldots, d_{n-1}$ can be chosen to be $+1$ or -1 as desired. Hence, it is possible to have $\nabla f(x)$ in any of the 2^n quadrants of $\mathbb{R}^n$. Consequently, 0 lies in the convex combination of these gradient vectors and we conclude from (4) that $0 \in \partial^C f(\bar{x})$.

The case $j < n - 1$ is handled similarly. For a choice of x around $\bar{x}$, we get

$$\nabla f(x) = \left[-\frac{1}{4} + 2c_1 d_1, d_1 + 2c_2 d_2, \ldots, d_{j-1} + 2c_j d_j, d_j + 2d_{j+1}, d_{j+1} - 2d_{j+2}, \ldots, d_{n-1} \right]^T$$

where $c_i = -\text{sign}(x_i)$, $i = 1, 2, \ldots, j-1$, are fixed (when $j > 1$) and $c_j = -\text{sign}(x_j)$, $d_1, d_2, \ldots, d_{n-1}$ are free parameters to choose from $\{-1, 1\}$. Suppose $d_j = d_j^0$, $d_{j+1} = d_{j+1}^0, \ldots, d_{n-1} = d_{n-1}^0$ are fixed. By choosing $c_j, d_1, \ldots, d_{j-1}$ appropriately, the signs of the first j components of $\nabla f(x)$ vector can be chosen to be positive or negative. Thus, by convexity,

$$\left[0, \ldots, 0, d_j^0 + 2d_{j+1}^0, d_{j+1}^0 - 2d_{j+2}^0, \ldots, d_{n-1}^0\right]^T \in \partial^C f(\bar{x}).$$

Choosing $d_j = -d_j^0, d_{j+1} = -d_{j+1}^0, \ldots, d_{n-1} = -d_{n-1}^0$, we have

$$\left[0, \ldots, 0, -(d_j^0 + 2d_{j+1}^0), -(d_{j+1}^0 - 2d_{j+2}^0), \ldots, -d_{n-1}^0\right]^T \in \partial^C f(\bar{x}).$$

and so by convexity $0 \in \partial^C f(\bar{x})$, completing the proof. □

The following theorem characterizes all the stationary points of f in the sense of both subdifferentials.

Theorem 3. *Nesterov's second nonsmooth Chebyshev–Rosenbrock function f is Mordukhovich stationary only at the global minimizer $x^* = [1, 1, \ldots, 1]^T$. Furthermore, f is Clarke stationary only at x^* and the $2^{n-1} - 1$ points in S with a vanishing x_j for some $j < n$. None of the Clarke stationary points of f except the global minimizer are local minimizers of f and there exists a direction of linear descent from each of these points.*

Proof. If $x \notin S$, f is smooth at x and we have $0 \notin \partial^M f(x) = \partial^C f(x) = \{\nabla f(x)\}$ since the partial derivative of f with respect to x_n at x is ± 1.

When $x = x^* \in S$, we have $0 \in \hat{\partial} f(x) \subset \partial^M f(x) \subset \partial f^C(x)$. If $x \in S, x \neq x^*$ ($x_1 \neq 1$) and $x_j \neq 0$ for $j = 1, 2, \ldots, n-1$, then the set S is a manifold around x. The function f is partly smooth with respect to S at x, with $f_{|S}(x) = \frac{|1-x_1|}{4}$, the restriction of f to S, smooth around x, and x is not a critical point of $f_{|S}$. It follows from [5, Proposition 2.4] that $0 \notin \text{aff}\,\{\partial^M f(x)\}$. This implies directly that $0 \notin \partial^M f(x)$ and $0 \notin \partial^C f(x) = \text{conv}\,\{\partial^M f(x)\}$, using (5).

The remaining case is when $x \in S$ is such that $x_j = 0$ for some $j < n$. We have $x_1 < 1$. Let $\delta > 0$ be small and x^δ be the unique point near x such that $x^\delta \in S$ and $x_1^\delta = x_1 + \delta$. It follows from the definition of S that $x^\delta = x + \delta v$ where v is a fixed vector independent of $\delta > 0$ for δ sufficiently small. Since $f_{|S}(x) = \frac{1-x_1}{4}$ around x, we have $f(x^\delta) = f(x + \delta v) = f(x) - \frac{1}{4}\delta < f(x)$ which shows that v is a direction of linear descent. Furthermore, we have $0 \notin \hat{\partial} f(x)$ since the existence of the descent direction at $\bar{x}$ implies

$$\liminf_{\substack{z \to \bar{x} \\ z \neq x}} \frac{f(z) - f(x)}{|z - x|} \leq \liminf_{\delta \downarrow 0} \frac{f(x + \delta v) - f(x)}{\delta |v|} = -\frac{1}{4|v|} < 0.$$

We want to prove that $0 \notin \partial^M f(x)$. This requires an investigation of the regular subdifferential $\hat{\partial} f(y)$ for y near x. Let y be a point near x, $y \neq x$. We have $x_j = 0$, so we distinguish two cases: $y \notin S$, $\{y \in S \text{ and } y_j \neq 0\}$. (If $y \in S$ and $y_j = 0$, then, for y to be near x, we would need $y = x$.)

1. $y \notin S$: $\nabla f(y)$ exists, we have $\hat{\partial} f(y) = \{\nabla f(y)\}$ and the n-th coordinate of $\nabla f(y)$ is ± 1. This shows that there exists no sequence $y^m \to x$ such that $y^m \notin S$ for all m with $\hat{\partial} f(y^m) = \{\nabla f(y^m)\} \ni v^m \to 0$.
2. $y \in S$ and $y_j \neq 0$: We have, for y sufficiently close to x, that $y_k \neq 0$ for $k = 1, \ldots, n$ and

$$S = \{x : F_i(x) = 0, \quad i = 1, \ldots, n-1\}$$

where $F_i(x) = x_{i+1} - 2|x_i| + 1$ is smooth at y. Hence, S is a manifold around y and it is easy to see that f is partly smooth at y with respect to S. The restricted function $f_{|S}(x) = \frac{1-x_1}{4}$ is smooth at y and since $y_1 < 1$, y is not a critical point of f, so from [5, Proposition 2.4] we conclude that $0 \notin \text{aff}\,\{\partial^M f(y)\}$ which leads to $0 \notin \hat{\partial} f(y)$. Furthermore, by [5, Proposition 2.2], we have

$$\hat{\partial} f(y) \subset \nabla g(y) + N_S(y) \tag{7}$$

where $g(x) = \frac{1-x_1}{4}$ and $N_S(y)$ is the normal space to S at y. The normal space to S at y coincides with the normal cone to S at y so by [9, Example. 6.8]

$$N_S(y) = \text{Range}(\nabla F)$$

where

$$\nabla F(y)^T = \left[\frac{\partial F_i}{\partial x_j}(y)\right]_{i,j=1}^{n-1,n} \in \mathbf{R}^{(n-1)\times n}$$

is the Jacobian matrix. We have

$$\nabla F(y) = \begin{bmatrix} -2\,\text{sign}(y_1) & & & \\ 1 & -2\,\text{sign}(y_2) & & \\ & 1 & \ddots & \\ & & \ddots & -2\,\text{sign}(y_{n-1}) \\ & & & 1 \end{bmatrix} \in \mathbf{R}^{n\times(n-1)} \tag{8}$$

and $\nabla g(y) = [-1/4, 0, \dots, 0]^T$. From (7) and (8), we see that $0 \in \hat{\partial} f(y)$ is possible only if $[1/4, 0, \dots, 0]^T \in N_S(y)$. A straightforward calculation shows that this is impossible. We conclude that $0 \notin \hat{\partial} f(y)$. The next step is to investigate the possible limits of $v^m \in \hat{\partial} f(y^m)$ as the sequence $y^m \in S$ approaches x. Let y^m be a sequence such that $y^m \to x$, $y^m \in S$ and $y^m \neq x$ for all m (this implies $y^m_j \neq 0$ for all m as before). Without loss of generality, assume $y^m_j > 0$ for all m. For fixed $k \in \{1, 2, \dots, n\}$, the quantity $\operatorname{sign}(y^m_k)$ does not depend on m and is nonzero. Thus, $G := \nabla F(y^m)$ does not depend on m. Let $v \in \mathbf{R}^n$ be such that $\hat{\partial} f(y^m) \ni v^m \to v$. From (7), we have $v^m = [-1/4, 0, \dots, 0]^T + Gc^m$ for some $c^m \in \mathbf{R}^{n-1}$. Since $v^m \to v$ and G has full rank, we have $c^m \to c \in \mathbf{R}^{n-1}$ and $v = [-1/4, 0, \dots, 0]^T + Gc$. As previously, $v = 0$ is impossible.

We conclude that $0 \notin \partial^M f(x)$. Since we already know from Theorem 2 that $0 \in \partial^C f(x)$, this completes the proof of the theorem. □

It follows immediately from Theorem 3 that f is not regular at the $2^{n-1} - 1$ non-locally-minimizing Clarke stationary points of f: see the comments after Definition 3.

3. Numerical experiments

Nesterov has observed that Newton's method with an inexact line search, when applied to minimize the *smooth* function $\tilde{f}$ initiated at $\hat{x}$, takes many iterations to reduce the value of the function below a small tolerance ϵ. Indeed, the number of iterations is typically exponential in n, although quadratic convergence is observed eventually if the method is run for long enough. Our experimental results are mainly obtained using the BFGS quasi-Newton algorithm with a line search based on the Armijo and "Wolfe" conditions, a well-known method generally used to optimize *smooth* functions [4]. However, as explained in [14], BFGS with the same line search is surprisingly effective for nonsmooth functions too. For the results reported below, we used a publicly available MATLAB implementation.[1]

For smooth but nonconvex functions such as $\tilde{f}$, there is no theorem known that guarantees that the BFGS iterates will converge to a stationary point, and pathological counterexamples have been constructed [15,16], although, unlike $\tilde{f}$, these are not analytic. However, it is widely accepted that BFGS generally produces sequences converging to local minimizers of smooth, nonconvex functions [17], so it is not surprising that this is the case for $\tilde{f}$, with superlinear convergence to x^* in the limit. As with Newton's method, many iterations are required. For $n = 8$, starting at $\hat{x}$ and with the initial inverse Hessian approximation H set to the identity matrix I, the BFGS method requires about 6700 iterations to reduce $\tilde{f}$ below 10^{-15}, and for $n = 10$, nearly 50,000 iterations are needed.

For nonsmooth functions, there is no general convergence theory for the BFGS method, but as discussed in [14], when applied to locally Lipschitz functions the method seems to always generate sequences of function values converging linearly to Clarke stationary values, and our experiments confirm this observation for small n for both nonsmooth functions studied in this paper. To apply BFGS to Nesterov's first nonsmooth variant $\hat{f}$, we cannot use $\hat{x}$ for the initial point as the method immediately breaks down, $\hat{f}$ being nondifferentiable at $\hat{x}$. Instead, we initialize x randomly, retaining the identity matrix for initializing H. The left panel of Fig. 1 shows the iterates generated by BFGS for the case $n = 2$ using 7 random starting points: all sequences of iterates converge to the global minimizer $x^* = [1, 1]^T$. However, the accuracy to which BFGS can minimize $\hat{f}$ drops rapidly as n increases. Because of the difficulty of the problem combined with the limited machine precision, the method breaks down, that is the line search fails to return a point satisfying the Armijo and Wolfe conditions, at an iterate x that is close to $\mathcal{M}$ but not very near x^*. When the calculations are carried out to higher precision, more accurate results are obtained [18]. For example, for $n = 4$, using standard IEEE "double" precision (about 16 decimal digits), from most starting points, BFGS reduces $\hat{f}$ to final values ranging from 10^{-3} to 10^{-2}, while using "double double" precision (about 32 decimal digits), from the same starting points, the final values that are obtained range from 10^{-4} to 10^{-3}.

For Nesterov's second nonsmooth variant f, we find that BFGS generates iterates approximating Clarke stationary points, but not necessarily the minimizer x^*. The iterates for the case $n = 2$, again for 7 randomly generated starting points, are shown in the right panel of Fig. 1. Most of the runs converge to the minimizer $[1, 1]^T$, but some terminate near the Clarke stationary point $[0, -1]^T$. For $n \leq 6$, given enough randomly generated starting points, BFGS finds, that is approximates well, all 2^{n-1} Clarke stationary points. The left and right panels of Fig. 2 plot final values of f found by 1000 runs of BFGS starting with random x and $H = I$, sorted into increasing order, for the cases $n = 5$ and $n = 6$ respectively. Most runs find either the minimizer or one of the $2^{n-1} - 1$ nonminimizing Clarke stationary points, although a few runs break down away from these points. For $n = 7$, the method usually breaks down without finding any Clarke stationary point, presumably because of the limitations of machine precision.

Experiments with the gradient sampling algorithm [19] and Kiwiel's bundle code [20] give similar results. Both of these methods have well established convergence theories ensuring convergence to Clarke stationary points. However, it remains an open question whether the nonminimizing Clarke stationary points are points of attraction for any of these algorithms. For small n, the computations usually terminate near Clarke stationary points, because eventually rounding error prevents the

[1] http://www.cs.nyu.edu/overton/software/hanso.

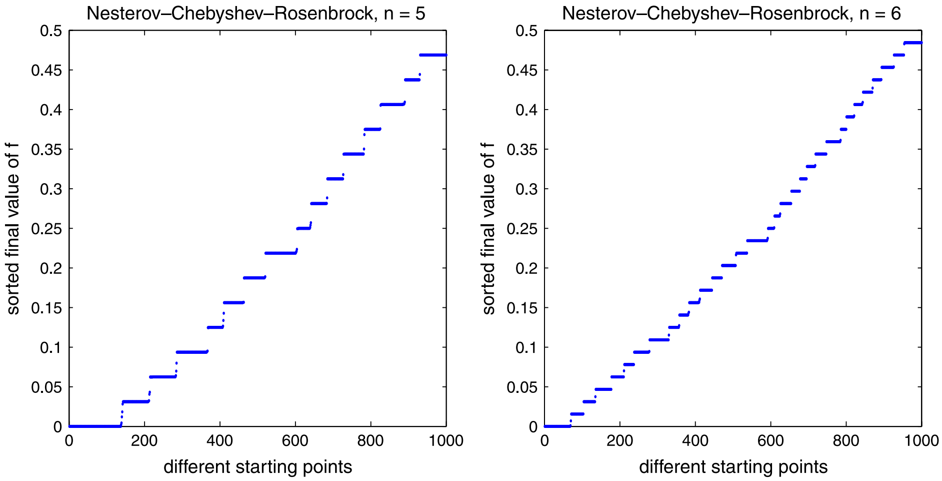

Fig. 2. Left: sorted final values of f for 1000 randomly generated starting points, when $n = 5$: BFGS finds all 16 Clarke stationary points. Right: same with $n = 6$: BFGS finds all 32 Clarke stationary points.

method from obtaining a lower point in the line search. But this does not establish whether, in exact arithmetic, the methods would actually generate sequences converging to the nonminimizing Clarke stationary points. Indeed, experiments in [18] suggest that the higher the precision used, the more likely BFGS is to move away from the neighborhood of a nonminimizing Clarke stationary point and eventually find a lower one, perhaps the minimizer.

Another observation is the difficulty of finding descent directions from the nonminimizing Clarke stationary points using random search. Although we know that such descent directions exist by Theorem 3, numerical experiments show that finding a descent direction by random search typically needs exponentially many trials. For example, when $n = 5$, usually 100,000 random trials do not suffice to find a descent direction. This illustrates the difficulty faced by an optimization method in moving away from these points.

4. Conclusion

Nesterov's Chebyshev–Rosenbrock functions provide very interesting examples for optimization, both in theory and in practice. Specifically, the smooth function $\tilde{f}$, the first nonsmooth function $\hat{f}$ and the second nonsmooth function f are very challenging nonconvex instances of smooth functions, partly smooth functions and non-regular functions respectively. As far as we know, Nesterov's function f is the first documented case for which methods for nonsmooth optimization result in the approximation of Clarke stationary points from which there exist directions of linear descent. This observation is primarily due to Kiwiel [20]. Furthermore, since all first-order nonsmooth optimization methods, including bundle methods [21], the gradient sampling method [19] and the BFGS method [14], are based on sampling gradient or subgradient information, the results given here for f suggest that limitation of convergence results to Clarke stationary points may be unavoidable, in the sense that one may not in general be able to expect stronger results such as convergence only to Mordukhovich stationary points. Nonetheless, it remains an open question as to whether the nonminimizing Clarke stationary points of f are actually points of attraction for methods using exact arithmetic.

Acknowledgments

We thank Krzysztof Kiwiel, Adrian Lewis and Yuri Nesterov for various helpful comments and for their interest in this work. The work of both authors was supported in part by the National Science Foundation under Grant DMS-1016325.

References

[1] Y. Nesterov, (2008), private communication.
[2] Gabor Szegö, Orthogonal Polynomials. American Mathematical Society, New York, 1939. American Mathematical Society Colloquium Publications, vol. 23.
[3] P.E. Gill, W. Murray, M.H. Wright, Practical Optimization, Academic Press, New York and London, 1981.
[4] J. Nocedal, S.J. Wright, Nonlinear Optimization, second ed., Springer, New York, 2006.
[5] A.S. Lewis, Active sets, nonsmoothness and sensitivity, SIAM J. Optim. 13 (2003) 702–725.
[6] F.H. Clarke, Optimization and Nonsmooth Analysis, John Wiley, New York, 1983, Reprinted by SIAM, Philadelphia, 1990.
[7] J.M. Borwein, A.S. Lewis, Convex Analysis and Nonlinear Optimization: Theory and Examples, second ed., Springer, New York, 2005.
[8] L.C. Evans, R.F. Gariepy, Measure Theory and Fine Properties of Functions, in: Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1992.
[9] R.T. Rockafellar, R.J.B. Wets, Variational Analysis, Springer, New York, 1998.
[10] B.S. Mordukhovich, Maximum principle in the problem of time optimal response with nonsmooth constraints, J. Appl. Math. Mech. 40 (1976) 960–969.

[11] A.S. Lewis, Eigenvalues and nonsmooth optimization, in: Foundations of Computational Mathematics, Santander 2005, in: London Math. Soc. Lecture Note Ser., vol. 331, Cambridge University Press, Cambridge, 2006, pp. 208–229.
[12] J.V. Burke, A.S. Lewis, M.L. Overton, Approximating subdifferentials by random sampling of gradients, Math. Oper. Res. 27 (2002) 567–584.
[13] W.-Y. Sun, Y.-X. Yuan, Optimization Theory and Methods, in: Springer Optimization and Its Applications, vol. 1, Springer, New York, 2006, Nonlinear programming.
[14] A.S. Lewis, M.L. Overton, Nonsmooth optimization via quasi-Newton methods, In revision for Math. Programming (submitted for publication).
[15] Y.-H. Dai, Convergence properties of the BFGS algorithm, SIAM J. Optim. 13 (2002) 693–701.
[16] W.F. Mascarenhas, The BFGS method with exact line searches fails for non-convex objective functions, Math. Program. 99 (2004) 49–61.
[17] D.-H. Li, M. Fukushima, On the global convergence of the BFGS method for nonconvex unconstrained optimization problems, SIAM J. Optim. 11 (2001) 1054–1064.
[18] A. Kaku, Implementation of high precision arithmetic in the BFGS method for nonsmooth optimization. Master's thesis, NYU, Jan 2011. URL: http://www.cs.nyu.edu/overton/mstheses/kaku/msthesis.pdf.
[19] J.V. Burke, A.S. Lewis, M.L. Overton, A robust gradient sampling algorithm for nonsmooth, nonconvex optimization, SIAM J. Optim. 15 (2005) 751–779.
[20] K.C. Kiwiel, (2008), private communication.
[21] K.C. Kiwiel, Methods of Descent for Nondifferentiable Optimization, in: Lecture Notes in Mathematics, vol. 1133, Springer-Verlag, Berlin and New York, 1985.