Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 06–48

USING SIMPLEX GRADIENTS OF NONSMOOTH FUNCTIONS IN DIRECT SEARCH METHODS

A. L. CUSTÓDIO, J. E. DENNIS JR. AND L. N. VICENTE

ABSTRACT: It has been shown recently that the efficiency of direct search methods that use opportunistic polling in positive spanning directions can be improved significantly by reordering the poll directions according to descent indicators built from simplex gradients.

The purpose of this paper is twofold. First, we analyze the properties of simplex gradients of nonsmooth functions in the context of direct search methods like the Generalized Pattern Search (GPS) and the Mesh Adaptive Direct Search (MADS), for which there exists a convergence analysis in the nonsmooth setting. Our analysis does not require continuous differentiability and can be seen as an extension of the accuracy properties of simplex gradients known for smooth functions. Secondly, we test the use of simplex gradients when pattern search is applied to nonsmooth functions, confirming the merit of the poll ordering strategy for such problems.

KEYWORDS: Derivative free optimization, simplex gradients, poisedness, nonsmooth analysis, generalized pattern search methods, mesh adaptive direct search.
AMS SUBJECT CLASSIFICATION (2000): 65D05, 90C30, 90C56.

1. Introduction

Pattern search methods, and more generally, direct search methods, are directional methods that do not use derivatives. Thus, they can be applied to nonsmooth functions. The main goal of this paper is to analyze the properties of simplex gradients when direct search methods are applied to a nonsmooth function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. We are particularly interested in two classes of direct search methods of the directional type, for which convergence has been analyzed in the nonsmooth setting, namely generalized pattern search (GPS) and mesh adaptive direct search (MADS) (see Audet and Dennis [1, 2, 3]). Other classes of direct search methods have been developed and analyzed, and we refer the reader to the surveys in [15, 20].

Simplex gradients are basically the first order coefficients of polynomial interpolation or regression models, which, in turn, are used in derivative-free

Received October 26, 2006.
Support for this work was provided by Centro de Matemática da Universidade de Coimbra, by FCT under grant POCI/MAT/59442/2004, and by Centro de Matemática e Aplicações da Universidade Nova da Lisboa.

trust region methods. However, simplex gradients can also serve as directions for search or orientation, as suggested by Mifflin [19]. Bortz and Kelley [4] used simplex gradients as search directions in their implicit filtering method. In the context of the Nelder-Mead simplex-based direct search algorithm, Kelley [13] used the simplex gradient norm in a sufficient decrease type condition to detect stagnation, and the simplex gradient signs to orient the simplex restarts. More recently, Custódio and Vicente [8] suggested several procedures to improve the efficiency of pattern search methods using simplex derivatives. In particular, they showed that when opportunistic polling is employed, i.e., polling is terminated at an iteration as soon as a better point is found, then ordering the poll directions according to a negative simplex gradient can lead to a significant reduction in the overall number of function evaluations.

This paper focuses on the unconstrained case and is structured as follows. In Section 2 we revise the basic smooth case properties of simplex gradients. The properties of simplex gradients of nonsmooth functions are stated and proved in Section 3 for a general application of direct search methods, using the concepts of refining subsequence and refining direction. The use of simplex gradients in direct search methods based on positive spanning sets is discussed in Section 4. We confirm in Section 5 that, in particular, it is possible for both GPS and MADS to identify sample sets as specified in Section 3. We report numerical results in Section 6 for a set of nonsmooth problems, confirming that ordering the poll directions according to a negative simplex gradient leads to significant reductions in the overall number of function evaluations, as it was observed in [8] for smooth problems.

2. Simplex gradients

Consider a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ and a finite set of sampling points. When the sample set is poised for linear interpolation or regression, simplex gradients are defined as the gradients of the corresponding models. Depending on the number of points available, simplex gradients can be computed in determined or underdetermined forms (corresponding to linear interpolation models) or in overdetermined forms (corresponding to linear regression models).

In the determined case, let us assume that we have a sample set with $n+1$ affinely independent points $\{y^0, y^1, \ldots, y^n\}$. Set $S = [\, y^1 - y^0 \, \cdots \, y^n - y^0 \,]$

and $\delta = [\, f(y^1) - f(y^0) \cdots f(y^n) - f(y^0)\,]^\top$. The simplex gradient $\nabla_s f(y^0)$ computed at y^0 is calculated as $\nabla_s f(y^0) = S^{-\top}\delta$.

When the number $q+1$ of points is not necessarily equal to $n+1$, simplex gradients can be also regarded as 'solutions' of the system

$$S^\top \nabla_s f(y^0) \;=\; \delta, \tag{1}$$

where $S = [\, y^1 - y^0 \cdots y^q - y^0\,]$ and $\delta = [\, f(y^1) - f(y^0) \cdots f(y^q) - f(y^0)\,]^\top$. For instance, when only $q+1 < n+1$ affinely independent points are available, a simplex gradient can be calculated as the minimum norm solution of the system (1).

Affine independence is not possible when $q > n$. In general, we say that a sample set is poised for a simplex gradient calculation (or for linear interpolation or regression) when S is full rank, *i.e.*, when $\mathrm{rank}(S) = \min\{n, q\}$. Thus, if we have a poised set with $q+1 > n+1$ points, one can compute a simplex gradient as the least squares solution of the system (1).

We can express the simplex gradient as $\nabla_s f(y^0) = V\Sigma^{-1}U^\top\delta/\Delta$ in any of the cases, where $U\Sigma V^\top$ is the reduced singular value decomposition (SVD) of $S^\top/\Delta$ and $\Delta = \max_{1\le i\le q}\|y^i - y^0\|$. Division by Δ is important to scale the points to an unitary ball centered at y^0.

For smooth functions, it is easy to derive bounds for the error between the simplex gradient and the true function gradient. The following result summarizes all the cases considered above (for proofs see [6] and [14]). The accuracy of these bounds is measured in terms of Δ. It is assumed that the gradient of f is Lipschitz continuous on a domain containing the smallest enclosing ball $B(y^0;\Delta) = \{y \in \mathbb{R}^n : \; \|y - y^0\| \le \Delta\}$ of the sample set, centered at y^0.

Theorem 1. *Let $\{y^0, y^1, \ldots, y^q\}$ be a poised sample set for a simplex gradient calculation in $\mathbb{R}^n$. Assume that ∇f is Lipschitz continuous in an open domain Ω containing $B(y^0;\Delta)$ with constant $\gamma_{\nabla f} > 0$. Then, the error of the simplex gradient at y^0, as an approximation to $\nabla f(y^0)$, satisfies*

$$\|\hat{V}^\top[\nabla f(y^0) - \nabla_s f(y^0)]\| \;\le\; \left(q^{\frac{1}{2}}\frac{\gamma_{\nabla f}}{2}\|\Sigma^{-1}\|\right)\Delta,$$

where $\hat{V} = I$ if $q \ge n$ and $\hat{V} = V$ if $q < n$.

In order to control the quality of the simplex gradient, it is therefore crucial to monitor the quality of the geometry of the sample set considered, in other words, the size of $\|\Sigma^{-1}\|$. Conn, Scheinberg, and Vicente in [7, 6] introduced the so-called notion of Λ–poisedness to measure the quality of sample sets, as well as algorithms to build or maintain Λ–poised sets. The definition of Λ–poisedness is omitted. For the purposes of this paper, we say that a poised set $\{y^0, y^1, \ldots, y^q\}$ is Λ–poised, for a given positive constant $\Lambda > 0$, if $\|\Sigma^{-1}\| \leq \Lambda$. A sequence of sample sets is Λ–poised if all the individual sample sets are.

3. Simplex gradients, refining subsequences, and nonsmooth functions

Let us start by recalling the definition of a *refining subsequence*, introduced first by Audet and Dennis in [1] in the context of GPS. This definition can be extended to any direct search algorithm that, at each iteration k, samples a poll set or a frame of the form $\{x_k + \alpha_k d : \ d \in D_k\}$, where D_k is a positive spanning set and $\alpha_k > 0$ is the mesh size or step size parameter.

A subsequence $\{x_k\}_{k\in\mathcal{K}}$ of the iterates generated by a direct search method is said to be a *refining subsequence* if two conditions are satisfied: (i) x_k is an unsuccessful iterate, meaning that $f(x_k) \leq f(x_k + \alpha_k d)$, for all $d \in D_k$; (ii) $\{\alpha_k\}_{k\in\mathcal{K}}$ converges to zero. A point x_k satisfying condition (i) is called a mesh local optimizer (in GPS) or a minimal frame center (in MADS). The analysis of direct search methods like GPS or MADS assumes that the sequence of iterates generated by the algorithms lie in compact sets. Hence, we can assume without loss of generality that a refining subsequence converges to a limit point.

Note that when the function value cannot be calculated one can have $f(x_k + \alpha_k d) = +\infty$ for some $d \in D_k$. We must therefore assume that the poll points used in the simplex gradient calculations are such that $f(x_k + \alpha_k d) < +\infty$. To build appropriate simplex gradients at refining subsequences, we will also use the fact that D_k is a positive spanning set. However, we point out that the fact that the frame center is minimal ($f(x_k) \leq f(x_k + \alpha_k d)$, for all $d \in D_k$) is not needed in the analysis. Of importance to our analysis are the facts that a refining subsequence $\{x_k\}_{k\in\mathcal{K}}$ converges to x_* and that $\alpha_k \to 0$ for $k \in \mathcal{K}$.

Of relevance to us are also *refining directions* associated with refining subsequences. Refining directions are limits of the form $d_k/\|d_k\|$ for subsequences of $\mathcal{K}$. Refining directions are guaranteed to exist in GPS [1] and in MADS [3]. In our paper, we will assume for simplification and without loss of generality that $d_k/\|d_k\|$ converges for every refining subsequence considered.

Finally we will also use the fact that $\alpha_k \|d_k\| \to 0$ for $k \in \mathcal{K}$, which can be trivially guaranteed for GPS (since here D_k is contained in a positive spanning set D fixed for all k; see [1]) and also for MADS under further appropriate requirements on the frames (see [3, Definition 2.2]).

The global convergence results for pattern and direct search methods are obtained by analyzing the behavior of the generalized derivatives of f at the limit points of refining subsequences. Thus, it is natural to pay particular attention to simplex gradients calculated at iterates of refining subsequences. As we will see later, since these iterates are unsuccessful and positive bases have special geometrical properties, it is possible to calculate Λ–poised sample sets in a number of different ways, some of which have already been introduced by Custódio and Vicente [8]. For the time being, all we need is to assume that, given a refining subsequence, it is possible to identify a subset Z_k of the poll set, described as

$$Z_k \;=\; \{x_k + \alpha_k d :\; d \in E_k\} \;\subseteq\; \{x_k + \alpha_k d :\; d \in D_k\},$$

such that

$$Y_k \;=\; \{x_k\} \cup Z_k$$

is Λ–poised for $k \in \mathcal{K}$. Let $\mathcal{Z}_k$ denote the subset of the index set $\{1, \ldots, |D_k|\}$ which defines the poll points in Z_k (or the poll directions in E_k). The simplex gradient is calculated in an overdetermined form when $|\mathcal{Z}_k| \geq n+1$, and in a determined or underdetermined form when $|\mathcal{Z}_k| \leq n$.

First, we show that the subsequence of refining simplex gradients has a limit point. Let

$$\Delta_k \;=\; \max\{\|z - x_k\| :\; z \in Z_k\} \;=\; \alpha_k \max\{\|d_k^j\| :\; d_k^j \in E_k\},$$

$$\nabla_s f(x_k) \;=\; V_k \Sigma_k^{-1} U_k^\top \delta_k / \Delta_k, \quad \text{and} \quad S_k^\top / \Delta_k \;=\; U_k \Sigma_k V_k^\top,$$

where S_k is the matrix whose columns are $(x_k + \alpha_k d_k^j) - x_k = \alpha_k d_k^j$ and δ_k is the vector whose components are $f(x_k + \alpha_k d_k^j) - f(x_k)$, for all $d_k^j \in E_k$. For the result we need to assume that the number $|\mathcal{Z}_k|$ of elements used for the overdetermined simplex gradients remains uniformly bounded. If all D_k are positive bases, since these have a maximum number of $2n$ elements, we trivially get $|\mathcal{Z}_k| \leq 2n$. In general we need to assume, reasonably, that the number $|D_k|$ of elements of the positive spanning sets D_k is uniformly bounded.

Lemma 1. *Let $\{x_k\}_{k\in\mathcal{K}}$ be a refining subsequence converging to x_* such that $\{Y_k\}_{k\in\mathcal{K}}$ is Λ–poised. Let f be Lipschitz continuous near x_*. Then, the simplex gradient subsequence $\{\nabla_s f(x_k)\}_{k\in\mathcal{K}}$ has at least one limit point.*

Proof: Let Ω be a neighborhood of x_* where f is Lipschitz continuous, with Lipschitz constant γ_f. Since the sequence $\{x_k\}_{k\in\mathcal{K}}$ converges to x_*, the iterates x_k are in Ω for k sufficiently large. Thus, for all $i \in \mathcal{Z}_k$ and k sufficiently large,

$$\left|\left(\frac{\delta_k}{\Delta_k}\right)_i\right| \;\leq\; \frac{|f(x_k+\alpha_k d_k^i)-f(x_k)|}{\alpha_k \max\{\|d_k^j\| : d_k^j \in E_k\}} \;\leq\; \frac{\gamma_f \|d_k^i\|}{\max\{\|d_k^j\| : d_k^j \in E_k\}} \;\leq\; \gamma_f.$$

From these inequalities, we get

$$\|\nabla_s f(x_k)\| \;=\; \left\|V_k \Sigma_k^{-1} U_k^\top \frac{\delta_k}{\Delta_k}\right\| \;\leq\; \|\Sigma_k^{-1}\| \sqrt{|\mathcal{Z}_k|}\,\gamma_f \;\leq\; \|\Sigma_k^{-1}\|\,\sqrt{|D_k|}\,\gamma_f.$$

Thus, since $\|\Sigma_k^{-1}\| \leq \Lambda$ for all $k \in \mathcal{K}$, we conclude that $\{\nabla_s f(x_k)\}_{k\in\mathcal{K}}$ is bounded, from which the statement of the theorem follows trivially. ∎

The next step is to study, in the nonsmooth context, the properties of a limit point identified in Lemma 1 for subsequences of simplex gradients constructed at refining subsequences. For this purpose, we will make use of Clarke's nonsmooth analysis [5]. Next we summarize the results we need for locally Lipschitz functions.

Let f be Lipschitz continuous near x_*. The Clarke generalized directional derivative of f computed at x_* in the direction v is the limit

$$f^\circ(x_*; v) \;=\; \limsup_{\substack{y \to x_* \\ t \downarrow 0}} \frac{f(y+tv)-f(y)}{t}.$$

Since f is Lipschitz continuous near x_*, this limit is well defined and so is the generalized subdifferential (or subgradient)

$$\partial f(x_*) \;=\; \{s \in \mathbb{R}^n :\; f^\circ(x_*; v) \geq v^\top s,\; \forall v \in \mathbb{R}^n\}.$$

Moreover,

$$f^\circ(x_*; v) \;=\; \max\{v^\top s :\; s \in \partial f(x_*)\}. \tag{2}$$

The Clarke generalized subdifferential is a nonempty convex cone and, as set-valued mapping, is closed and locally bounded (see [5]). The mean value theorem can be formulated for locally Lipschitz functions using the Clarke generalized subdifferential. In fact, if x and y are points in $\mathbb{R}^n$ and if f is

Lipschitz continuous on an open set containing the line segment $[x, y]$, then there exists a point z in (x, y) such that

$$f(y) - f(x) \;=\; \nabla f(z)^\top (y - x), \tag{3}$$

for some $\nabla f(z) \in \partial f(z)$.

A function is strictly differentiable at x_* if and only if is Lipschitz continuous near x_* and there exists a vector $\nabla f(x_*)$ such that

$$\lim_{\substack{x \to x_* \\ t \downarrow 0}} \frac{f(x + tv) - f(x)}{t} \;=\; \nabla f(x_*)^\top v, \quad \forall v \in \mathbb{R}^n.$$

In this case, the Clarke generalized subdifferential reduces to a singleton $\partial f(x_*) = \{\nabla f(x_*)\}$.

3.1. The Lipschitz continuous case. The first case we consider is when $|\mathcal{Z}_k| \leq n$, in other words, when simplex gradients are determined or underdetermined. This case is not of great interest since underdetermined simplex gradients do not capture the appropriate geometrical properties of positive spanning sets. In the limit case $|\mathcal{Z}_k| = 1$, we are dealing with approximations to one-sided directional derivatives.

Theorem 2. *Let $\{x_k\}_{k \in \mathcal{K}}$ be a refining subsequence converging to x_* for which*

$$\lim_{\substack{k \to +\infty \\ k \in \mathcal{K}}} \frac{d_k}{\|d_k\|} \;=\; v \quad \textit{and} \quad \lim_{\substack{k \to +\infty \\ k \in \mathcal{K}}} \alpha_k \|d_k\| \;=\; 0.$$

Moreover, assume that $\{Y_k\}_{k \in \mathcal{K}}$ is Λ–poised, $|\mathcal{Z}_k| \leq n$ for all $k \in \mathcal{K}$, and $d_k \in E_k$ is a direction used in the computation of $\nabla_s f(x_k)$ for all $k \in \mathcal{K}$.

Let f be Lipschitz continuous near x_*. Then, $\{\nabla_s f(x_k)\}_{k \in \mathcal{K}}$ has a limit point $\nabla_s f(x_k) \to \nabla_s f_*$, $k \in \mathcal{L} \subseteq \mathcal{K}$, such that

$$f^\circ(x_*; v) \;\geq\; \nabla_s f_*^\top v.$$

Proof: The proof is omitted since it is basically the proof of the next theorem, which is in same way more general. In fact, (4) below reduces to $\delta_k = S_k^\top \nabla_s f(x_k)$ and the proof of Theorem 3 applies here trivially. ∎

Let us consider now the more interesting case where $|\mathcal{Z}_k| \geq n + 1$ (overdetermined simplex gradients). From the definition of simplex gradient, we have

$$\delta_k \;=\; S_k^\top \nabla_s f(x_k) + (I - S_k^\top (S_k S_k^\top)^{-1} S_k) \delta_k, \tag{4}$$

where $R_k = (I - S_k^\top (S_k S_k^\top)^{-1} S_k)$ is a projector onto the null space of S_k. For convenience, we will denote the rows of R_k by $(r_k^i)^\top$, $i \in \mathcal{Z}_k$. In this subsection we analyze the case where f is Lipschitz continuous near x_*.

Theorem 3. *Let $\{x_k\}_{k \in \mathcal{K}}$ be a refining subsequence converging to x_* for which*

$$\lim_{\substack{k \to +\infty \\ k \in \mathcal{K}}} \frac{d_k}{\|d_k\|} \;=\; v \quad \text{and} \quad \lim_{\substack{k \to +\infty \\ k \in \mathcal{K}}} \alpha_k \|d_k\| \;=\; 0. \tag{5}$$

Moreover, assume that $\{Y_k\}_{k \in \mathcal{K}}$ is Λ–poised, $|\mathcal{Z}_k| \geq n+1$ for all $k \in \mathcal{K}$, and $d_k \in E_k$ is a direction used in the computation of $\nabla_s f(x_k)$ for all $k \in \mathcal{K}$.

Let f be Lipschitz continuous near x_*. Then, $\{\nabla_s f(x_k)\}_{k \in \mathcal{K}}$ has a limit point $\nabla_s f(x_k) \to \nabla_s f_*$, $k \in \mathcal{L} \subseteq \mathcal{K}$, such that

$$f^\circ(x_*; v) \;\geq\; \nabla_s f_*^\top v + \limsup_{\substack{k \to +\infty \\ k \in \mathcal{L}}} \left(r_k^{i_k}\right)^\top \left(\frac{\delta_k}{\alpha_k \|d_k\|}\right), \tag{6}$$

where i_k is the index in $\mathcal{Z}_k$ for which $d_k = d_k^{i_k} \in E_k$.

Proof: From Lemma 1, there exists a subsequence $\mathcal{L} \subseteq \mathcal{K}$ such that $\nabla_s f(x_k) \to \nabla_s f_*$ for $k \in \mathcal{L}$. Now, we express the i_k-th row in (4) as

$$\frac{f(x_k + \alpha_k d_k) - f(x_k)}{\alpha_k} \;=\; \nabla_s f(x_k)^\top d_k + \frac{1}{\alpha_k} (r_k^{i_k})^\top (\delta_k).$$

From the basic properties of the generalized derivatives of locally Lipschitz functions (see, for instance (2)), one can easily see that

$$f^\circ(x_*; v) \;=\; f^\circ\left(x_*; \lim_{\substack{k \to +\infty \\ k \in \mathcal{K}}} \frac{d_k}{\|d_k\|}\right) = \lim_{\substack{k \to +\infty \\ k \in \mathcal{K}}} f^\circ\left(x_*; \frac{d_k}{\|d_k\|}\right).$$

Since $\alpha_k \|d_k\| \to 0$ for $k \in \mathcal{K}$,

$$\begin{aligned} f^\circ(x_*; v) \;&\geq\; \limsup_{\substack{k \to +\infty \\ k \in \mathcal{K}}} \frac{f(x_k + \alpha_k \|d_k\| \frac{d_k}{\|d_k\|}) - f(x_k)}{\alpha_k \|d_k\|} \\ &=\; \limsup_{\substack{k \to +\infty \\ k \in \mathcal{K}}} \frac{f(x_k + \alpha_k \, d_k) - f(x_k)}{\alpha_k \|d_k\|} \end{aligned}$$

$$
\begin{aligned}
&= \limsup_{\substack{k \to +\infty \\ k \in \mathcal{K}}} \left\{ \frac{\nabla_s f(x_k)^\top d_k}{\|d_k\|} + (r_k^{i_k})^\top \left(\frac{\delta_k}{\alpha_k \|d_k\|} \right) \right\} \\
&\geq \nabla_s f_*^\top v + \limsup_{\substack{k \to +\infty \\ k \in \mathcal{L}}} (r_k^{i_k})^\top \left(\frac{\delta_k}{\alpha_k \|d_k\|} \right)
\end{aligned}
$$

and the proof is concluded. ■

3.2. The strictly differentiable case. To better understand Theorem 3 and the role of the lim sup term in (6), let us focus now on the case where $\mathcal{Z}_k$ is constant for all $k \in \mathcal{K}$ and f is strictly differentiable at x_*. As an example, let us look at the case of coordinate search, where $D_k = [I_n \ \ -I_n]$ for all k (and $I_n = [\, e_1 \ \cdots \ e_n \,]$ stands for the identity matrix of size n). Let us consider the calculation of overdetermined simplex gradients using all the poll points ($\mathcal{Z}_k = 2n$). It is easy to see that

$$
R_k \ = \ I_{2n} - S_k^\top (S_k S_k^\top)^{-1} S_k \ = \ 0.5 \begin{bmatrix} I_n & I_n \\ I_n & I_n \end{bmatrix}.
$$

Thus, what we get in this case from Theorem 3 are the following $2n$ inequalities

$$
\begin{aligned}
f'(x_*; e_i) &\geq \nabla_s f_*^\top e_i + 0.5 \left[f'(x_*; e_i) + f'(x_*; -e_i) \right], \quad i = 1, ..., n, \\
f'(x_*; -e_i) &\geq \nabla_s f_*^\top (-e_i) + 0.5 \left[f'(x_*; e_i) + f'(x_*; -e_i) \right], \quad i = 1, ..., n.
\end{aligned}
$$

Since f is strictly differentiable at x_*, we also get $f'(x_*; e_i) + f'(x_*; -e_i) = 0$ and, thus, the extra terms in the above inequalities (which come from the lim sup term in (6)) vanish. The following corollary summarizes a consequence of Theorem 3 in the strictly differentiable case.

Corollary 1. *Let the assumptions of Theorem 3 hold. Assume further that the function f is strictly differentiable at x_*, $\mathcal{Z}_k$ is constant for all $k \in \mathcal{K}$, and the normalized form of E_k given by $E_k/\|d_k\|$ converges to V_v in $\mathcal{K}$. Then, for the refining direction $v \in V_v$ given by (5),*

$$
f^\circ(x_*; v) \ = \ \left(\, f'(x_*; v) \ = \ \nabla f(x_*)^\top v \, \right) \ = \ \nabla_s f_*^\top v.
$$

Proof: First, we point out that

$$
R_k \ = \ I - (E_k/\|d_k\|)^\top \left((E_k/\|d_k\|)(E_k/\|d_k\|)^\top \right)^{-1} (E_k/\|d_k\|),
$$

and, as a result, $R_k \to R_* \equiv I - V_v^\top (V_v V_v^\top)^{-1} V_v$ in $\mathcal{K}$. The result stated in the corollary can then be obtained by replacing the last two inequalities of the proof of Theorem 3 by equalities. Note that the lim sup term in (6) is, in fact, always zero:

$$\left(I - V_v^\top (V_v V_v^\top)^{-1} V_v\right) f'(x_*; V_v) \;=\; \left(I - V_v^\top (V_v V_v^\top)^{-1} V_v\right) V_v^\top \nabla f(x_*) \;=\; 0,$$

where $f'(x_*; V_v)$ is the vector formed by the directional derivatives of f at x_* along the directions in V_v. ∎

Note that V_v depends on v since the normalization of the columns in E_k is done with respect to $\|d_k\|$, which, in turn, is associated with the refining direction v. Suppose now that Corollary 1 is applicable to a set of linearly independent refining directions $v \in V$ for which $V_v = V$ for all v. In this case, as a result of Corollary 1, applied for all $v \in V$, we would conclude that $\nabla_s f_* = \nabla f(x_*)$.

Our next theorem focuses exclusively on the case where f is strictly differentiable at the limit point x_* of a refining subsequence. The result of this theorem is only true for determined or overdetermined simplex gradients ($|\mathcal{Z}_k| \geq n$). However, it is true for any cardinal $|\mathcal{Z}_k| = |E_k| \geq n$ and it does not require any assumption on limits of normalized directions of E_k.

Theorem 4. *Let $\{x_k\}_{k\in\mathcal{K}}$ be a refining subsequence converging to x_* such that $\{Y_k\}_{k\in\mathcal{K}}$ is Λ–poised and $|\mathcal{Z}_k| \geq n$ for all $k \in \mathcal{K}$. Let f be strictly differentiable at x_*. Then, there exists a subsequence of indices $\mathcal{L} \subseteq \mathcal{K}$ such that*

$$\lim_{\substack{k \to +\infty \\ k \in \mathcal{L}}} \nabla_s f(x_k) \;=\; \nabla f(x_*).$$

Proof: Since f is strictly differentiable at x_*, then it is Lipschitz continuous near x_* and we can apply Lemma 1. Let $\mathcal{L} \subseteq \mathcal{K}$ be the index set for which the corresponding subsequence of simplex gradients converges.

From the mean value theorem (3) for locally Lipschitz functions, we have, for all $i \in \mathcal{Z}_k$ and $k \in \mathcal{L}$ sufficiently large, that

$$f(x_k + \alpha_k d_k^i) - f(x_k) \;=\; \alpha_k \nabla f(z_k^i)^\top d_k^i,$$

where z_k^i is a point in the line segment $(x_k, x_k + \alpha_k d_k^i)$ and $\nabla f(z_k^i) \in \partial f(z_k^i)$. Now, because ∂f is locally bounded, the sequence $\{\nabla f(z_k^i)\}_{k\in\mathcal{L}}$ is bounded. But, since ∂f is a closed set-valued mapping and $z_k^i \to x_*$ for $k \in \mathcal{L}$, any limit point of $\{\nabla f(z_k^i)\}_{k\in\mathcal{L}}$ is necessarily in $\partial f(x_*)$. Thus, $\nabla f(z_k^i) \to \nabla f(x_*)$ for $k \in \mathcal{L}$.

Now we write, for all $i \in \mathcal{Z}_k$,

$$f(x_k + \alpha_k d_k^i) - f(x_k) \;=\; \alpha_k \nabla f(x_*)^\top d_k^i + \alpha_k[\nabla f(z_k^i) - \nabla f(x_*)]^\top d_k^i.$$

Let $\bar{r}_k$ denote the vector of dimension $|\mathcal{Z}_k|$ and components $[\nabla f(z_k^i) - \nabla f(x_*)]^\top d_k^i$. Then,

$$\delta_k \;=\; S_k^\top \nabla f(x_*) + \alpha_k \bar{r}_k$$

and

$$\nabla_s f(x_k) \;\equiv\; (S_k S_k^\top)^{-1} S_k \delta_k \;=\; \nabla f(x_*) + \alpha_k (S_k S_k^\top)^{-1} S_k \bar{r}_k.$$

Moreover, note that

$$\alpha_k (S_k S_k^\top)^{-1} S_k \bar{r}_k \;=\; \frac{\alpha_k}{\Delta_k} \left[(S_k/\Delta_k)(S_k/\Delta_k)^\top\right]^{-1} (S_k/\Delta_k)\bar{r}_k. \tag{7}$$

Now, let $\tilde{r}_k$ denote the vector of dimension $|\mathcal{Z}_k|$ and components $\|\nabla f(z_k^i) - \nabla f(x_*)\|$. One can easily prove that

$$\|\bar{r}_k\| \;\leq\; \max\{\|d_k^j\| : \; d_k^j \in E_k\} \|\tilde{r}_k\|.$$

Thus, from this bound, (7), and the Λ–poisedness of $\{Y_k\}_{k\in\mathcal{K}}$,

$$\left\|\alpha_k (S_k S_k^\top)^{-1} S_k \bar{r}_k\right\| \;\leq\; \frac{1}{\max\{\|d_k^j\| : \; d_k^j \in E_k\}} \|\Sigma_k^{-1}\| \|\bar{r}_k\| \;\leq\; \Lambda \, \|\tilde{r}_k\|.$$

The proof is thus concluded from the fact that $\tilde{r}_k \to 0$ for $k \in \mathcal{L}$. ■

The result of Theorem 4 cannot possibly be true for simplex gradients computed with less than $n+1$ points ($|\mathcal{Z}_k| < n$). Even in the smooth case such result would not be valid as one could infer from Theorem 1, where $\hat{V} \neq I$ when $q < n$. From the proof of Theorem 4, we have

$$\|\nabla f(x_*) - \nabla_s f(x_k)\| \;\leq\; \Lambda \, \|\tilde{r}_k\|, \quad \tilde{r}_k \;\to\; 0 \text{ (for } k \in \mathcal{L}),$$

which is a nonsmooth counterpart of Theorem 1.

4. Applications in direct search methods

A point x_* at which f is locally Lipschitz is (Clarke) stationary if $f^\circ(x_*; d) \geq 0$, for all d in $\mathbb{R}^n$. If the function f is strictly differentiable at x_* then, for ensuring the stationarity of x_*, it suffices to show that $f^\circ(x_*; d) \geq 0, \forall d \in D$, where D is a positive spanning set for $\mathbb{R}^n$. In this context, the material of

Section 3 suggests a new stopping criterion for an algorithm that polls a positive basis at each iteration. In fact, if at an unsuccessful iteration

$$\nabla_s f(x_k)^\top (\alpha_k d) \;\geq\; -\epsilon_{tol}, \quad \forall d \in E_k,$$

for a given tolerance $\epsilon_{tol} > 0$, then it is probably safe to stop the algorithm. We should have $|\mathcal{Z}_k| \geq n+1$. A natural choice is $E_k = D_k$. Our numerical experience has shown, however, that the use of this stopping criterion has an effect similar to the use of a stopping criterion solely based on the size of α_k.

The simplex gradient can also be used to reorder the poll directions before sampling the poll points. This strategy was suggested by Custódio and Vicente [8], in the context of generalized pattern search, but it can be applied to any algorithm that polls using a positive spanning set. In fact, we can define a descent indicator by considering $-\nabla_s f(x_k)$ and order the poll vectors according to increasing magnitudes of the angles between this descent indicator and the poll directions. Based on a test set of smooth problems, and in the context of coordinate search, it has been observed that ordering the poll directions using simplex gradients can reduce the average number of function evaluations more than 50% [8]. Numerical results for the application of this strategy to nonsmooth problems will be reported in Section 6.

In the remaining of this section we are mainly interested in studying poisedness and Λ–poisedness of poll sets. The Λ–poisedness of the sequences of poll sets will be then analyzed in more detail in Section 5 for the context of particular algorithmic settings.

Positive bases for $\mathbb{R}^n$ must have between $n+1$ and $2n$ vectors (see [9]). Positive bases with $n+1$ ($2n$) elements are called minimal (maximal). The most used positive bases in practice probably are the ones of the form $[\, B \; -B \,]$ or $[\, B \; -\sum_{i=1}^n b_i \,]$, where B is a nonsingular matrix in $\mathbb{R}^{n\times n}$ (see [17]).

The question that arises is how to compute overdetermined simplex gradients from poll points defined by positive spanning sets, in other words how to identify poised poll sets. One possible approach is to use all the poll directions, in other words, all the vectors in each positive spanning set used for polling. It is easy to see that the corresponding overdetermined simplex gradients are well-defined in this circumstances (see Proposition 1). Furthermore, this proposition also tells us that if cosine measures of positive spanning sets are bounded away from zero then the corresponding poll sets are Λ–poised. It is known [15] that the cosine measure

$$\kappa(D) \;=\; \min_{v\in\mathbb{R}^n,\, v\neq 0} \max_{d\in D} \frac{v^\top d}{\|v\|\|d\|}$$

of a positive basis or positive spanning set D is always positive.

Proposition 1. *Let D be a positive spanning set for $\mathbb{R}^n$. Let $\|d\| \geq d_{min} > 0$ for all $d \in D$. Then D is full rank and*

$$\|\Sigma^{-1}\| \;\leq\; \frac{1}{d_{min}\kappa(D)},$$

where $D^\top = U\Sigma V^\top$.

Proof: Since $\|d\| \geq d_{min}, \forall d \in D$, we have

$$\begin{aligned}
\kappa(D) &= \min_{\|v\|=1} \max_{d\in D} \frac{v^\top d}{\|d\|} \\
&\leq \frac{1}{d_{min}} \min_{\|v\|=1} \max_{d\in D} |v^\top d| = \frac{1}{d_{min}} \min_{\|v\|=1} \|D^\top v\|_\infty \\
&\leq \frac{1}{d_{min}} \min_{\|v\|=1} \|D^\top v\|.
\end{aligned}$$

The Courant-Fischer inequalities applied to singular values (see, for example, [12]) allow us to conclude that

$$\kappa(D) \;\leq\; \frac{1}{d_{min}} \min_{\|v\|=1} \|D^\top v\| \;=\; \frac{1}{d_{min}\|\Sigma^{-1}\|}.$$

■

5. Two algorithmic contexts

This section is devoted to the validation of the conditions needed for the theorems stated in Section 3 in the context of two different direct search methods. These results where established for generalized pattern search (GPS) by Audet and Dennis [1], and for mesh adaptive direct search (MADS) by Audet and Dennis [3].

5.1. Generalized pattern search. GPS allows the use of different positive spanning sets D_k at each iteration, but all D_k must be chosen from a positive spanning set D. As a result, the number of distinct positive spanning sets D_k that is possible to consider is finite and, thus, it is also finite the number of different direction sets $E_k \subseteq D_k$ used in the computation of simplex gradients. As a result, all refining subsequences $\{Y_k\}_{k \in \mathcal{K}}$ of poised poll sets are Λ–poised, for some $\Lambda > 0$ only dependent on D. The computation of poised poll sets Y_k, for overdetermined simplex gradients, can adopt, for instance, the choice $E_k = D_k$.

The existence of a convergent refining subsequence for a sequence of iterates generated by GPS is proved in [1, Theorem 3.6]). From the finiteness of D, we trivially guarantee $\alpha_k \|d_k\| \to 0$ and the existence of refining directions.

5.2. Mesh adaptive direct search. The poll set or frame in MADS is of the form $\{x_k + \Delta_k^m d : \ d \in D_k\}$, where $\Delta_k^m > 0$ represents a mesh size parameter and D_k is a positive spanning set not necessarily extracted from a single positive spanning set D. One can have, in MADS, an infinite number of distinct positive spanning sets D_k, but each d in D_k must be a nonnegative integer combination of directions in a fixed positive basis D. MADS considers also a poll size parameter $\Delta_k^p > 0$, but we omit that part of the description of the algorithm since it plays no role in our discussion. In the context of our paper we have $\alpha_k = \Delta_k^m$.

The existence of a convergent refining subsequence for a sequence of iterates generated by MADS is proved in [3]. From the relationship between Δ_k^m and Δ_k^p, it is known that $\alpha_k \|d_k\| \to 0$ for all refining subsequences. Refining directions are guaranteed to exist in the unconstrained case.

Audet and Dennis [3, Proposition 4.2] suggested a practical implementation of MADS, called LTMADS, that generates a dense set of poll directions in $\mathbb{R}^n$ with probability one, satisfying all MADS requirements. The positive spanning sets D_k in LTMADS are of the form $[\, B_k \ {-B_k} \,]$ or $[\, B_k \ -\sum_{i=1}^n (b_k)_i \,]$.

Let us start by looking at the maximal case $[\, B_k \ {-B_k} \,]$. If we are interested in overdetermined simplex gradients one can set $E_k = D_k = [\, B_k \ - B_k \,]$. In this case, $S_k = \alpha_k [\, B_k \ - B_k \,]$ and $\Delta_k = \alpha_k \max\{\|(b_k)_i\| : \ (b_k)_i \in B_k\}$.

Now let us look at the minimal case $[\, B_k \ -\sum_{i=1}^n (b_k)_i \,]$. The use of overdetermined simplex gradients is also straightforward. We can set $E_k = D_k = [\, B_k \ \ -\sum_{i=1}^n (b_k)_i \,]$. In this case, $S_k = \alpha_k [\, B_k \ -\sum_{i=1}^n (b_k)_i \,]$ and $\Delta_k = \alpha_k \max\{\| -\sum_{j=1}^n (b_k)_j \|, \|(b_k)_i\| : \ (b_k)_i \in B_k\}$.

From the fact that the smallest singular value of a matrix does not decrease when rows or columns are added, we can infer, for both cases, that the corresponding sequences of sample points $\{Y_k\}_{k\in\mathcal{K}}$ are Λ–poised if the inverse of the matrix $\alpha_k B_k/\Delta_k$ is uniformly bounded in $\mathcal{K}$. Let us see that that is the case for the maximal case. The definition of Δ_k is slightly different in the minimal case, but the proof is similar.

The matrix B_k in LTMADS results from row and column permutations of a lower triangular matrix L_k, where each diagonal element is given by $\pm 1/\sqrt{\alpha_k}$ and the lower diagonal elements are integers in the open interval $\left(-1/\sqrt{\alpha_k}, 1/\sqrt{\alpha_k}\right)$. Thus, since the 2-norm of a matrix is invariant under row and column permutations and from the property of singular values mentioned above,

$$\|\Sigma_k^{-1}\| \;\leq\; \|(\alpha_k B_k/\Delta_k)^{-1}\| \;=\; \|(\alpha_k L_k/\Delta_k)^{-1}\|. \tag{8}$$

One can see that $\alpha_k L_k$ is a lower triangular matrix with diagonal elements $\pm\sqrt{\alpha_k}$ and lower diagonal elements in the interval $(-\sqrt{\alpha_k}, \sqrt{\alpha_k})$. So, the norms of the columns of $\alpha_k L_k$ are in $[\sqrt{\alpha_k}, \sqrt{n\alpha_k})$ and one can observe that $\alpha_k L_k/\Delta_k$ is a lower triangular matrix with diagonal elements in $(1/\sqrt{n}, 1]$ in absolute value.

The 1-norm of the inverse of a nonsingular lower triangular matrix L of dimension n can be bounded by

$$\|L^{-1}\|_1 \;\leq\; \frac{(\beta_1+1)^{n-1}}{\beta_2},$$

where $\beta_1 = \max_{i>j} |\ell_{ij}|/|\ell_{ii}|$ and $\beta_2 = \min_i |\ell_{ii}|$ ([16]; see also [11]). Thus, we obtain (with $\beta_1 = 1$ and $\beta_2 = 1/\sqrt{n}$):

$$\|(\alpha_k L_k/\Delta_k)^{-1}\| \;\leq\; \sqrt{n}\|(\alpha_k L_k/\Delta_k)^{-1}\|_1 \;\leq\; n\,2^{n-1}. \tag{9}$$

Finally, from (8) and (9), we conclude that $\{Y_k\}_{k\in\mathcal{K}}$ is Λ–poised with $\Lambda = n2^{n-1}$.

6. Numerical experiments

We collected a set of nonsmooth functions from the nonsmooth optimization literature. As far as we could verify all the functions are continuous. Several types of nondifferentiability are represented. The list of problems is given in Table 1.

In Table 2 we report the results of two (generalized) pattern search methods on this test set. The `basic` version corresponds to a simple implementation

problem	**source**	**dimension**
activefaces	[10]	20
elattar	[18]	6
EVD61	[18]	6
filter	[18]	9
goffin	[18]	50
HS78	[18]	5
L1HILB	[18]	50
MXHILB	[18]	50
osborne2	[18]	11
PBC1	[18]	5
polak2	[18]	10
shor	[18]	5
wong1	[18]	7
wong2	[18]	10

TABLE 1. Test set of nonsmooth functions.

of coordinate search with opportunistic pooling, where the positive basis used for polling is $[\, I \ -I \,]$. No search step is considered. The mesh size parameter is halved in unsuccessful iterations and kept constant in successful iterations. The other version `order` differs from the basic one only in the fact that the polling vectors are ordered according to increasing angles with a descent indicator (the negative simplex gradient). All previously sample points are candidates for the simplex gradient calculations (`store-all` mode in [8]). Before polling one attempts to build a simplex gradient from a set of Λ-poised points (Λ was set to 100) with a number of points as large as possible but between $(n+1)/2$ and $2n+1$.

The results show clearly that the ordering strategy based on simplex gradients for nonsmooth functions leads to better performance. The average reduction in function evaluations was around 27%. In some cases the reduction is significant and when an increase occurs it is relatively small. The average reduction of function evaluations reported in [8] for similar simplex derivatives based strategies was around 50% for continuously differentiable problems. The application of direct search methods to nonsmooth functions is however less well understood in practice and the sources for different numerical behavior are greater. In this paper we analyze some properties of

		fevals		**fvalue**	
problem	**fbest**	`basic`	`order`	`basic`	`order`
activefaces	0.00e+00	913	713	2.30e+00	2.30e+00
elattar	5.60e-01	1635	569	6.66e+00	6.91e-01
EVD61	3.49e-02	538	335	3.16e-01	9.07e-02
filter	6.19e-03	370	333	9.50e-03	9.50e-03
goffin	0.00e+00	22526	17038	0.00e+00	0.00e+00
HS78	-2.92e+00	329	212	-1.52e+00	2.07e-04
L1HILB	0.00e+00	3473240	7660	2.33e+00	2.20e-01
MXHILB	0.00e+00	26824	3164	1.24e+00	1.24e+00
osborne2	4.80e-02	727	761	2.80e-01	1.01e-01
PBC1	2.23e-02	287	264	4.39e-01	4.34e-01
polak2	5.46e+01	2179	1739	5.46e+01	5.46e+01
shor	2.26e+01	215	257	2.43e+01	2.34e+01
wong1	6.81e+02	343	366	6.85e+02	6.85e+02
wong2	2.43e+01	819	763	3.97e+01	2.58e+01

TABLE 2. Ordering poll vectors using simplex gradients on a set of nonsmooth problems. **fbest** is the best function value reported in the source reference, **fevals** is the number of functions evaluations taken, and **fvalue** is the final function value computed.

simplex gradients which tend to support the improvement observed in the numerical results.

References

[1] C. AUDET AND J. E. DENNIS JR., *Analysis of generalized pattern searches*, SIAM J. Optim., 13 (2003), pp. 889–903.

[2] ——, *A pattern search filter method for nonlinear programming without derivatives*, SIAM J. Optim., 14 (2004), pp. 980–1010.

[3] ——, *Mesh adaptive direct search algorithms for constrained optimization*, SIAM J. Optim., 17 (2006), pp. 188–217.

[4] D. M. BORTZ AND C. T. KELLEY, *The simplex gradient and noisy optimization problems*, in Computational Methods in Optimal Design and Control, Progress in Systems and Control Theory, edited by J. T. Borggaard, J. Burns, E. Cliff, and S. Schreck, vol. 24, Birkhäuser, Boston, 1998, pp. 77–90.

[5] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983. Reissued by SIAM, Philadelphia, 1990.

[6] A. R. CONN, K. SCHEINBERG, AND L. N. VICENTE, *Geometry of sample sets in derivative free optimization. Part II: polynomial regression and underdetermined interpolation*, Tech. Report 05-15, Departamento de Matemática, Universidade de Coimbra, Portugal, 2005.

[7] ——, *Geometry of interpolation sets in derivative free optimization*, Math. Program., (2006, to appear).

[8] A. L. Custódio and L. N. Vicente, *Using sampling and simplex derivatives in pattern search methods*, Tech. Report 04-35, Departamento de Matemática, Universidade de Coimbra, Portugal, 2004. Revised November 2005, June 2006.
[9] C. Davis, *Theory of positive linear dependence*, Amer. J. Math., 76 (1954), pp. 733–746.
[10] M. Haarala, *Large-scale nonsmooth optimization*, PhD thesis, University of Jyväskylä, Finland, 2004.
[11] N. Higham, *A survey of condition number estimation for triangular matrices*, SIAM Rev., 29 (1987), pp. 575–596.
[12] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1999.
[13] C. T. Kelley, *Detection and remediation of stagnation in the Nelder-Mead algorithm using a sufficient decrease condition*, SIAM J. Optim., 10 (1999), pp. 43–55.
[14] ——, *Iterative Methods for Optimization*, SIAM, Philadelphia, 1999.
[15] T. G. Kolda, R. M. Lewis, and V. Torczon, *Optimization by direct search: New perspectives on some classical and modern methods*, SIAM Rev., 45 (2003), pp. 385–482.
[16] F. Lemeire, *Bounds for condition numbers of triangular and trapezoid matrices*, BIT, 15 (1975), pp. 58–64.
[17] R. M. Lewis and V. Torczon, *Rank ordering and positive bases in pattern search algorithms*, Tech. Report 96-71, Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, 1996.
[18] L. Lukšan and J. Vlček, *Test problems for nonsmooth unconstrained and lineraly constrained optimization*, Tech. Report 798, Institute of Computer Science, Academy of Sciences of the Czech Republic, 2000.
[19] R. Mifflin, *A superlinearly convergent algorithm for minimization without evaluating derivatives*, Math. Program., 9 (1975), pp. 100–117.
[20] M. J. D. Powell, *Direct search algorithms for optimization calculations*, Acta Numer., 7 (1998), pp. 287–336.

A. L. Custódio
Departamento de Matemática, FCT-UNL, Quinta da Torre 2829-516 Caparica, Portugal (alcustodio@fct.unl.pt).

J. E. Dennis Jr.
Department of Computational and Applied Mathematics, Rice University – MS 134, 6100 South Main Street, Houston, Texas, 77005-1892, USA (dennis@rice.edu).

L. N. Vicente
Departamento de Matemática, Universidade de Coimbra, 3001-454 Coimbra, Portugal (lnv@mat.uc.pt).