SIAM J. OPTIM.
Vol. x, No. x, pp. xxx-xxx

MESH ADAPTIVE DIRECT SEARCH ALGORITHMS FOR CONSTRAINED OPTIMIZATION *

CHARLES AUDET † AND J.E. DENNIS JR. ‡

Abstract. This paper introduces the Mesh Adaptive Direct Search (MADS) class of algorithms for nonlinear optimization. MADS extends the Generalized Pattern Search (GPS) class by allowing local exploration, called *polling*, in a dense set of directions in the space of optimization variables. This means that under certain hypotheses, including a weak constraint qualification due to Rockafellar, MADS can treat constraints by the extreme *barrier* approach of setting the objective to infinity for infeasible points and treating the problem as unconstrained. The main GPS convergence result is to identify limit points where the Clarke generalized derivatives are nonnegative in a *finite* set of directions, called *refining directions*. Although in the unconstrained case, nonnegative combinations of these directions spans the whole space, the fact that there can only be finitely many GPS refining directions limits rigorous justification of the barrier approach to finitely many constraints for GPS. The MADS class of algorithms extend this result; the set of refining directions may even be dense in $\mathbb{R}^n$, although we give an example where it is not.

We present an implementable instance of MADS, and we illustrate and compare it with GPS on some test problems. We also illustrate the limitation of our results with examples.

Key words. Mesh adaptive direct search algorithms (MADS), convergence analysis, constrained optimization, nonsmooth analysis, Clarke derivatives, hypertangent, contingent cone.

1. Introduction.

We present and analyze a new *Mesh Adaptive Direct Search* (MADS) class of algorithms for minimizing a nonsmooth function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ under general constraints $x \in \Omega \neq \emptyset \subset \mathbb{R}^n$. For the form of the algorithm given here, the feasible region Ω may be defined through blackbox constraints given by an oracle, such as a computer code, that returns a yes or no indicating whether or not a specified trial point is feasible.

In the unconstrained case, where $\Omega = \mathbb{R}^n$, this new class of algorithms occupies a position somewhere between the Generalized Pattern Search (GPS) [22, 6] algorithms and the Coope and Price frame-based methods [10]. A key advantage of MADS over GPS is that local exploration of the space of variables is not restricted to a finite number of directions (called *poll directions*). This is the primary drawback of GPS algorithms in our opinion, and our main motivation in defining MADS was to overcome this restriction. MADS algorithms are frame-based methods. We propose a less general choice of frames than the choices allowed by Coope and Price, but they are specifically aimed at ensuring a dense set of polling directions, and they are effective and easy to implement. We illustrate with an example algorithm that we call LTMADS because it is based on a random lower triangular matrix.

The convergence analysis here is based on Clarke's calculus [8] for nonsmooth functions. It evolved from our previous work on GPS [3] where we give a hierarchy of convergence results for GPS that show the limitations inherent in the restriction to finitely many directions. Specifically, we show there that for unconstrained optimization, GPS produces a limit point at which the gradient is zero if the function at that

* Work of the first author was supported by FCAR grant NC72792 and NSERC grant 239436-01, and both authors were supported by AFOSR F49620-01-1-0013, The Boeing Company, Sandia LG-4253, ExxonMobil, and the LANL Computer Science Institute (LACSI) contract 03891-99-23.

†GERAD and Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal, C.P. 6079, Succ. Centre-ville, Montréal (Québec), H3C 3A7 Canada (Charles.Audet@gerad.ca, http://www.gerad.ca/Charles.Audet)

‡Computational and Applied Mathematics Department, Rice University - MS 134, 6100 Main Street, Houston, Texas, 77005-1892 (dennis@caam.rice.edu, http://www.caam.rice.edu/∼dennis)

point is strictly differentiable [16], but if the function f is only Lipschitz near such a limit point, then Clarke's generalized directional derivatives [8] are provably nonnegative only for a finite set of directions $\hat{D} \subset \mathbb{R}^n$ whose nonnegative linear combinations span the whole space:

$$f^\circ(\hat{x}; d) := \limsup_{y \to \hat{x},\ t \downarrow 0} \frac{f(y + td) - f(y)}{t} \ \geq\ 0 \quad \text{for all } d \in \hat{D}. \tag{1.1}$$

$\hat{D}$ is called the set of *refining directions*. This result (1.1) for GPS is not as strong as stating that the generalized derivative is nonnegative for every directions in $\mathbb{R}^n$, i.e., that the limit point is a Clarke stationary point, or equivalently that $0 \in \partial f(\hat{x})$, the generalized gradient of f at $\hat{x}$ defined by

$$f^\circ(\hat{x}; v) \geq 0 \ \text{ for all } v \in \mathbb{R}^n \Leftrightarrow 0 \in \partial f(\hat{x}) := \{s \in \mathbb{R}^n : f^\circ(\hat{x}; v) \geq v^T s \text{ for all } v \in \mathbb{R}^n\}. \tag{1.2}$$

Example F in [2] shows that indeed the GPS algorithm does not necessarily produce a Clarke stationary point for Lipschitz functions because of the restriction to finitely many poll directions. For the unconstrained case, this restriction can be overcome by assuming more smoothness for f, e.g., strict differentiability at $\hat{x}$ [3]. Strict differentiability is just the requirement that the generalized gradient is a singleton, i.e., that $\partial f(\hat{x}) = \{\nabla f(\hat{x})\}$ in addition to the requirement that f is Lipschitz near $\hat{x}$. However, the directional dependence of GPS in the presence of even bound constraints can not be overcome by any amount of smoothness, by using penalty functions, or by the use of the more flexible filter approach for handling constraints [4].

The MADS algorithms can generate a dense set of polling directions in $\mathbb{R}^n$. The set of directions for which we can show that the Clarke generalized derivatives are nonnegative, the refining directions, is a subset of this dense set. This does not necessarily ensure Clarke stationarity, but is stronger than if the poll directions belonged to a fixed finite set, as it it the case with GPS.

Besides the advantages for the unconstrained case of a dense set of polling directions, this also allows MADS to treat a wide class of nonlinear constraints by the "barrier" approach. By this we mean that the algorithm is not applied directly to f but to the barrier function f_Ω, defined to be equal to f on Ω and $+\infty$ outside Ω. This way of rejecting infeasible points was shown to be effective for GPS with linear constraints by Lewis and Torczon [17] if one included in the poll directions the tangent cone generators of the feasible region at boundary points near an iterate. For LTMADS, no special effort is needed for the barrier approach to be provably effective with probability 1 on constraints satisfying a reasonable constraint qualification due to Rockafellar [21] – that there exists a hypertangent vector at the limit point. A key advantage of the barrier approach is that one can avoid expensive function calls to f whenever a constraint is violated. Indeed, the question of feasibility of a trial point needs only a yes or no answer - the constraints do not need to be given by a known algebraic condition.

Thus the class of algorithms presented here differs significantly from previous GPS extensions [4, 18] to nonlinear constraints. Treating constraints as we do motivates us to use the generalization of the Clarke derivative presented in Jahn [14] where the evaluation of f is restricted to points in the domain Ω. Thus we use the following definition of the Clarke generalized derivative at $\hat{x} \in \Omega$ in the direction $v \in \mathbb{R}^n$:

$$f^\circ(\hat{x}; v) := \limsup_{\substack{y \to \hat{x},\ y \in \Omega \\ t \downarrow 0,\ y + tv \in \Omega}} \frac{f(y + tv) - f(y)}{t}. \tag{1.3}$$

Both definitions (1.1) and (1.3) coincide when $\Omega = \mathbb{R}^n$ or when $\hat{x} \in int(\Omega)$.

The main theoretical objective of the paper is to show that under appropriate assumptions, any MADS algorithm produces a constrained Clarke stationary point, *i.e.,* a limit point $\hat{x} \in \Omega$ satisfying the following necessary optimality condition

$$f^\circ(\hat{x}; v) \ \geq \ 0 \text{ for all } v \in T_\Omega^{Cl}(\hat{x}), \tag{1.4}$$

where $T_\Omega^{Cl}(\hat{x})$ is the Clarke tangent cone to Ω at $\hat{x}$ (see [8] or Definition 3.5).

The paper is organized as follows. Section 2 presents the theoretical MADS algorithm class and Section 3 contains its convergence analysis. We show that (1.4) holds, and discuss its consequences when the algorithm is applied to an unconstrained problem, or when the set Ω is regular in the sense of Definition 3.7 or [8]. We give some contraint qualification conditions ensuring that MADS produces a contingent KKT stationary point. Two implementable instances are proposed and analyzed in Section 4. Numerical experiments are conducted in Section 5 to compare MADS with standard GPS for an unconstrained, a bound constrained, a disk constrained, and a nasty exponentially constrained optimization problem. On an artificial example where GPS and Nelder-Mead are well known to stagnate, we show that MADS reaches the global optimum. We give a comparison on a parameter fitting problem in catalytic combustion kinetics on which we know that GPS performs well [13], and we give an example illustrating the power of being able to handle more constraints by the barrier approach. The final example shows the value of randomly generated polling directions for a problem with an ever narrower feasible region.

Notation. $\mathbb{R}, \mathbb{Z}$ and $\mathbb{N}$ respectively denote the sets of real numbers, integers and nonnegative integers. For $x \in \mathbb{R}^n$ and $\delta \in \mathbb{R}_+$, $B_\delta(x)$ denotes the open ball of radius δ centered at x. For a matrix D, the notation $d \in D$ indicates that d is a column of D.

2. Mesh Adaptive Direct Search algorithms. Given an initial iterate $x_0 \in \Omega$, a MADS algorithm attempts to locate a minimizer of the function f over Ω by evaluating f_Ω at some trial points. The algorithm does not require the use or the approximations of derivatives of f. This is useful when f is contaminated with noise, or when ∇f is unavailable, does not exists or cannot be accurately estimated, or when there are several local optima. MADS is an iterative algorithm where at each iteration (the iteration number is denoted by the index k) a finite number of trial points are generated and their objective function values are compared with the current incumbent value $f_\Omega(x_k)$, *i.e.,* the best feasible objective function value found so far. Each of these trial points lies on the *current mesh*, constructed from a finite fixed set of n_D directions $D \subset \mathbb{R}^n$ scaled by a *mesh size parameter* $\Delta_k^m \in \mathbb{R}_+$.

There are two restrictions on the set D. First, D must be a positive spanning set [11], *i.e.*, nonnegative linear combinations of its elements must span $\mathbb{R}^n$. Second, each direction $d_j \in D$ (for $j = 1, 2, \ldots, n_D$) must be the product $G\bar{z}_j$ of some fixed non-singular generating matrix $G \in \mathbb{R}^{n \times n}$ by an integer vector $\bar{z}_j \in \mathbb{Z}^n$. For convenience, the set D is also viewed as a real $n \times n_D$ matrix. Similarly, the matrix whose columns are $\bar{z}_j$, for $j = 1, 2, \ldots, n_D$ is denoted by $\bar{Z}$; we can therefore use matrix multiplication to write $D = G\bar{Z}$. This is all in common with GPS.

DEFINITION 2.1. *At iteration k, the current mesh is defined to be the following union:*

$$M_k = \bigcup_{x \in S_k} \left\{ x + \Delta_k^m D z : z \in \mathbb{N}^{n_D} \right\} \ ,$$

where S_k is the set of points where the objective function f had been evaluated by the start of iteration k.

In the definition above, the mesh is defined to be an union of sets over S_k. Defining the mesh this way ensures that all previously visited points lie on the mesh, and that new trial points can be selected around any of them. This definition of the mesh is identical to the one in [4] and generalizes the one in [3].

The mesh is conceptual in the sense that it is never actually constructed. In practice, one can easily make sure that the strategy for generating trial points is such that they all belong to the mesh. One simply has to verify in Definition 2.1 that x belongs to S_k and that z is an integer vector. The objective of the iteration is to find a trial mesh point with a lower objective function value than the current incumbent value $f_\Omega(x_k)$. Such a trial point is called an *improved mesh point*, and the iteration is called a *successful iteration*. There are no sufficient decrease requirements.

The evaluation of f_Ω at a trial point x is done as follows. First, the constraints defining Ω are tested to determine if x is feasible or not. Indeed, since some of the constraints defining Ω might be expensive or inconvenient to test, one would order the constraints to test the easiest ones first. If $x \notin \Omega$, then $f_\Omega(x)$ is set to $+\infty$ without evaluating $f(x)$. On the other hand, if $x \in \Omega$, then $f(x)$ is evaluated. This remark may seem obvious, but it saves computation, and it is needed in the proof of Theorem 3.12.

Each iteration is divided into two steps. The first, called the SEARCH step has the same flexibility as in GPS. It allows evaluation of f_Ω at any finite number of mesh points. Any strategy can be used in the SEARCH step to generate a finite number of trial mesh points. Restricting the SEARCH points to lie on the mesh is a way in which MADS is less general than the frame methods of Coope and Price [10]. The SEARCH is said to be empty when no trial points are considered. The drawback to the SEARCH flexibility is that it cannot be used in the convergence analysis – except to provide counterexamples as in [2]. More discussion of SEARCH steps is given in [1, 19, 6].

When an improved mesh point is generated, then the iteration may stop, or it may continue if the user hopes to find a better improved mesh point. In either case, the next iteration will be initiated with a new incumbent solution $x_{k+1} \in \Omega$ with $f_\Omega(x_{k+1}) < f_\Omega(x_k)$ and with a mesh size parameter Δ^m_{k+1} equal to or larger than Δ^m_k (the exact rules for updating this parameter are presented below, and they are the same as for GPS). Coarsening the mesh when improvements in f_Ω are obtained can speed convergence.

Whenever the SEARCH step fails to generate an improved mesh point, the second step, called the POLL, is invoked before terminating the iteration. The difference between the MADS and the GPS algorithms lies in this step. For this reason, our numerical comparisons in the sequel use empty, or very simple, SEARCH steps in order not to confound the point we wish to make about the value of the MADS POLL step.

When the iteration fails in generating an improved mesh point, then the next iteration is initiated from any point $x_{k+1} \in S_{k+1}$ with $f_\Omega(x_{k+1}) = f_\Omega(x_k)$; though there is usually a single such incumbent solution, and then x_{k+1} is set to x_k. The mesh size parameter Δ^m_{k+1} is reduced to increase the mesh resolution, and therefore to allow the evaluation of f at trial points closer to the incumbent solution.

More precisely, given a fixed rational number $\tau > 1$, and two integers $w^- \leq -1$

and $w^+ \geq 0$, the mesh size parameter is updated as follow :

$$\Delta_{k+1} = \tau^{w_k} \Delta_k$$
$$\text{for some } w_k \in \begin{cases} \{0, 1, \ldots, w^+\} & \text{if an improved mesh point is found} \\ \{w^-, w^- + 1, \ldots, -1\} & \text{otherwise.} \end{cases} \quad (2.1)$$

Everything up to this point in the section applies to both GPS and MADS. We now present the key difference between both classes of algorithms. For MADS, we introduce the *poll size parameter* $\Delta_k^p \in \mathbb{R}^+$ for iteration k. This new parameter dictates the magnitude of the distance from the trial points generated by the POLL step to the current incumbent solution x_k. In GPS, there is a single parameter to represent these quantities : $\Delta_k = \Delta_k^p = \Delta_k^m$. In MADS, the strategy for updating Δ_k^p must be such that $\Delta_k^m \leq \Delta_k^p$ for all k, and moreover, it must satisfy

$$\lim_{k \in K} \Delta_k^m = 0 \text{if and only if} \lim_{k \in K} \Delta_k^p = 0 \qquad \text{for any infinite subset of indices } K \quad (2.2)$$

An implementable updating strategy satisfying these requirements is presented in Section 4.

We now move away from the GPS terminology, and toward that of Coope and Price. The set of trial points considered during the POLL step is called a *frame* (the definition of Coope and Price is more general). The frame is constructed using a current incumbent solution x_k (called the *frame center*) and the poll and mesh size parameters Δ_k^p, Δ_k^m and a positive spanning matrix D_k.

DEFINITION 2.2. *At iteration k, the MADS frame is defined to be the set:*

$$P_k = \{x_k + \Delta_k^m d : d \in D_k\} \ \subset \ M_k \ ,$$

where D_k is a positive spanning set such that for each $d \in D_k$,

- *$d \neq 0$ can be written as a nonnegative integer combination of the directions in D :*
  $d = Du$ for some vector $u \in \mathbb{N}^{n_D}$ that may depend on the iteration number k
- *the distance from the frame center x_k to a poll point $x_k + \Delta_k^m d$ is bounded by a constant times the poll size parameter : $\Delta_k^m \|d\| \leq \Delta_k^p \max\{\|d'\| : d' \in D\}$*
- *limits (as defined in Coope and Price [9]) of the normalized sets D_k are positive spanning sets.*

If the POLL step fails to generate an improved mesh point then the frame is called a *minimal frame*, and the frame center x_k is said to be a *minimal frame center*. At each iteration, the columns of D_k are called the poll directions.

The algorithm is stated formally below. It is very similar to GPS, with differences in the POLL step, and in the new poll size parameter.

A GENERAL MADS ALGORITHM

- INITIALIZATION: Let $x_0 \in \Omega$, $\Delta_0^m \leq \Delta_0^p$, G, Z, τ, w^- and w^+ satisfy the requirements given above. Set the iteration counter k to 0.
- SEARCH AND POLL STEP: Perform the SEARCH and possibly the POLL steps (or only part of them) until an improved mesh point x_{k+1} is found on the mesh M_k (see Definition 2.1).
  - OPTIONAL SEARCH: Evaluate f_Ω on a finite subset of trial points on the mesh M_k.
  - LOCAL POLL: Evaluate f_Ω on the frame P_k (see Definition 2.2).
- PARAMETER UPDATE: Update Δ_{k+1}^m according to equation (2.1), and Δ_{k+1}^p according to (2.2). Increase $k \leftarrow k + 1$ and go back to the SEARCH and POLL step.

The crucial distinction with GPS is that if Δ_k^m goes to zero more rapidly than Δ_k^p, then the directions in D_k used to define the frame may be selected in a way so that they are not confined to a finite set. Note that in GPS both Δ_k^m and Δ_k^p are equal : a single parameter plays the role of the mesh and poll size parameters, and therefore, the number of positive spanning sets that can be formed by subsets of D is constant over all iterations. For example, suppose that in $\mathbb{R}^2$ the set D is composed of the eight directions $\{(d_1, d_2)^T \neq (0,0)^T : d_1, d_2 \in \{-1, 0, 1\}\}$. There are a total of 14 distinct positive bases that can be constructed from D. Figure 2.1 illustrates some possible frames in $\mathbb{R}^2$ for three values of $\Delta_k^m = \Delta_k^p$. Another figure in Section 4 contrasts with this one by illustrating how the new MADS algorithm may select the directions of D_k from a larger set.

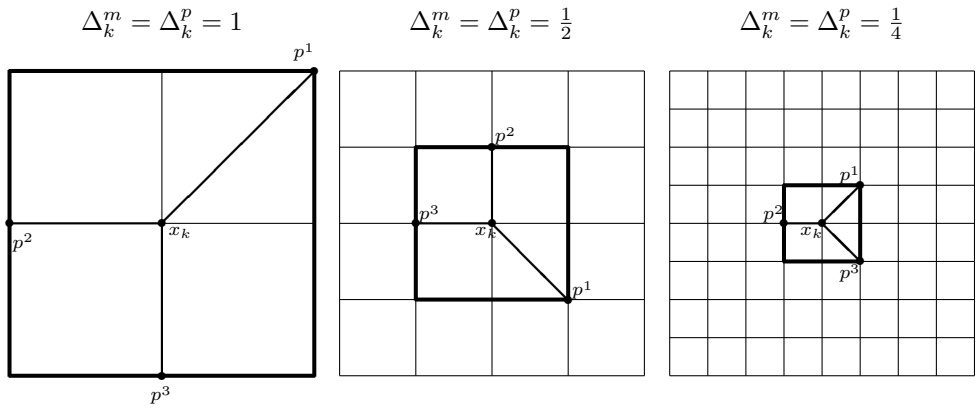

FIG. 2.1. *GPS (a special case of MADS) : Example of frames* $P_k = \{x_k + \Delta_k^m d : d \in D_k\} = \{p^1, p^2, p^3\}$ *for different values of* $\Delta_k^m = \Delta_k^p$. *In all three figures, the mesh* M_k *is the intersection of all lines.*

3. Convergence analysis of MADS. The convergence analysis below relies on the assumptions that $x_0 \in \Omega$, that $f(x_0)$ is finite, and that all iterates $\{x_k\}$ produced by the MADS algorithm lie in a compact set. Future work will relax the first assumption by incorporating the filter approach given in [4].

The section is divided in three subsections. The first recalls Torczon's [22] analysis of the behavior of the mesh size parameter and defines refining sequences as in [3]. It also defines the idea of a refining subsequence and a refining direction. The second subsection recalls the definitions of the hypertangent, Clarke, and contingent cones in addition to some results on generalized derivatives. The third contains a hierarchy of convergence results based on local properties of the feasible region Ω.

3.1. Preliminaries. Torczon [22] first showed the following result for unconstrained pattern search algorithms. Then Audet and Dennis [3] used the same technique for a description of GPS which is much closer to our description of MADS. The proof of this result for MADS is identical to that of GPS. The element necessary to the proof is that for any integer $N \geq 1$, the iterate x_N may be written as $x_N = x_0 + \sum_{k=0}^{N-1} \Delta_k D z_k$ for some vectors $z_k \in \mathbb{N}^{n_D}$. This is still true with our new way of defining the mesh and the frame (see Definitions 2.1 and 2.2).

PROPOSITION 3.1. *The poll and mesh size parameters produced by a MADS instance satisfy*

$$\liminf_{k\to+\infty} \Delta_k^p \;=\; \liminf_{k\to+\infty} \Delta_k^m \;=\; 0.$$

Since the mesh size parameter shrinks only at minimal frames, Proposition 3.1 guarantees that there are infinitely many minimal frame centers. The following definition specifies the subsequences of iterates and limit directions we use.

DEFINITION 3.2. *A subsequence of the MADS iterates consisting of minimal frame centers, $\{x_k\}_{k\in K}$ for some subset of indices K, is said to be a* refining subsequence *if $\{\Delta_k^p\}_{k\in K}$ converges to zero.*

If the limit $\lim_{k\in L} \frac{d_k}{\|d_k\|}$ exists for some subset $L \subseteq K$ with poll direction $d_k \in D_k$, and if $x_k + \Delta_k^m d_k \in \Omega$ for infinitely many $k \in L$, then this limit is said to be a refining direction for $\hat{x}$.

It is shown in [3], that there exists at least one convergent refining subsequence. We now present some definitions that will be used later to guarantee the existence of refining directions.

3.2. Cones and generalized derivatives. Three different types of tangent cones play a central role in our analysis. Their definition, and equivalent ones, may be found in [21, 8, 14]. After presenting them, we supply an example where the three cones differ to illustrate some of our results.

DEFINITION 3.3. *A vector $v \in \mathbb{R}^n$ is said to be a hypertangent vector to the set $\Omega \subset \mathbb{R}^n$ at the point $x \in \Omega$ if there exists a scalar $\epsilon > 0$ such that*

$$y + tw \in \Omega \qquad \text{for all } y \in \Omega \cap B_\epsilon(x), \quad w \in B_\epsilon(v) \quad \text{and} \quad 0 < t < \epsilon. \tag{3.1}$$

The set of hypertangent vectors to Ω at x is called the hypertangent cone to Ω at x *and is denoted by $T_\Omega^H(x)$.*

Since the definition of a hypertangent is rather technical and crucial to our results, we will pause for a short discussion. The reader could easily show that if Ω is a full dimensional polytope defined by linear constraints, then every direction from $\hat{x}$ into the interior of Ω is a hypertangent. That follows immediately from the following result relating hypertangents to the constraint qualification suggested by Gould and Tolle [12]. See also [5] for a discussion of the Gould and Tolle constraint qualification.

THEOREM 3.4. *Let $C : \mathbb{R}^n \to \mathbb{R}^m$ be continuously differentiable at a point $\hat{x} \in \Omega = \{x \in \mathbb{R}^n \;:\; C(x) \le 0\}$, and let $A = \{i \in \{1, 2, \ldots, m\} : c_i(\hat{x}) = 0\}$ be the active set at $\hat{x}$. Then $v \in \mathbb{R}^n$ is a hypertangent vector to Ω at $\hat{x}$ if and only if $\nabla c_i(\hat{x})^T v < 0$ for each $i \in A$ with $\nabla c_i(\hat{x}) \neq 0$.* *Proof.* Let v be a hypertangent vector to Ω at $\hat{x}$. Then, there exists an $\epsilon > 0$ such that $\hat{x} + tv \in \Omega$ for any $0 < t < \epsilon$. Let $i \in A$. Continuous differentiability of c_i at $\hat{x}$ implies that

$$\nabla c_i(\hat{x})^T v = \lim_{t\to 0} \frac{c_i(\hat{x} + tv) - c_i(\hat{x})}{t} \le 0 \; .$$

It only remains to show that $\nabla c_i(\hat{x})^T v \neq 0$ when $\nabla c_i(\hat{x}) \neq 0$. Suppose by way of contradiction that $\nabla c_i(\hat{x})^T v = 0$ and $\nabla c_i(\hat{x}) \neq 0$. Since the hypertangent cone is an open set [21], for any nonnegative $\delta \in \mathbb{R}$ sufficiently small, $v + \delta \nabla c_i(\hat{x})$ is a hypertangent vector to Ω at $\hat{x}$. It follows that

$$0 \;\ge\; \nabla c_i(\hat{x})^T (v + \delta \nabla c_i(\hat{x})) \;=\; \delta \|\nabla c_i(\hat{x})\|_2^2 \;>\; 0,$$

which is a contradiction. Thus, $\nabla c_i(\hat{x})^T v < 0$ when $\nabla c_i(\hat{x}) \neq 0$.

To prove the converse, let $i \in A$ be such that $\nabla c_i(\hat{x}) \neq 0$ and $v \in \mathbb{R}^n$ be such that $\|v\| = 1$ and $c_i(\hat{x})^T v < 0$. The product $c_i(y)w$ is a continuous function at $(y;\ w) = (\hat{x};\ v)$, and so there is some $\epsilon_1 > 0$ such that

$$c_i(y)w < 0 \text{ for all } y \in B_{\epsilon_1}(\hat{x}) \ \text{ and } \ w \in B_{\epsilon_1}(v). \tag{3.2}$$

Take $\epsilon = \min\{1, \frac{\epsilon_1}{3}\}$ and let y, w be in $B_\epsilon(\hat{x})$ and $B_\epsilon(v)$ respectively with $y \in \Omega$, and let $0 < t < \epsilon$. We will show that $y + tw \in \Omega$. Our construction ensures that $c_i(y) \leq 0$ and $\epsilon < \epsilon_1$, and so by the mean value theorem, we have

$$c_i(y + tw) \leq c_i(y + tw) - c_i(y) \ = \ \nabla c_i(y + \theta tw)(tw) \qquad \text{for some } \ \theta \in [0, 1] \tag{3.3}$$

But, $\|y + \theta tw - \hat{x}\| \leq \|y - \hat{x}\| + \theta t(\|w - v\| + \|v\|) < \epsilon + \epsilon(\epsilon + 1) \leq 3\epsilon \leq \epsilon_1$, thus $y + \theta tw \in B_{\epsilon_1}(\hat{x})$, and $w \in B_\epsilon(v) \subset B_{\epsilon_1}(v)$. It follows that equation (3.2) applies and therefore $\nabla c_i(y + \theta tw)^T w < 0$. Combining this with (3.3) and with the fact that $t > 0$ implies that $c_i(y + tw) \leq 0$. But c_i was any active component function, and so $C(y + tw) \leq 0$ and thus $y + tw \in \Omega$. ∎

Let us now present two other types of tangent cones.

DEFINITION 3.5. *A vector $v \in \mathbb{R}^n$ is said to be a Clarke tangent vector to the set $\Omega \subset \mathbb{R}^n$ at the point x in the closure of Ω if for every sequence $\{y_k\}$ of elements of Ω that converges to x and for every sequence of positive real numbers $\{t_k\}$ converging to zero, there exists a sequence of vectors $\{w_k\}$ converging to v such that $y_k + t_k w_k \in \Omega$. The set $T_\Omega^{Cl}(x)$ of all Clarke tangent vectors to Ω at x is called the Clarke tangent cone to Ω at x.*

DEFINITION 3.6. *A vector $v \in \mathbb{R}^n$ is said to be a tangent vector to the set $\Omega \subset \mathbb{R}^n$ at the point x in the closure of Ω if there exists a sequence $\{y_k\}$ of elements of Ω that converges to x and a sequence of positive real numbers $\{\lambda_k\}$ for which $v = \lim_k \lambda_k(y_k - x)$. The set $T_\Omega^{Co}(x)$ of all tangent vectors to Ω at x is called the contingent cone (or sequential Bouligand tangent cone) to Ω at x.*

DEFINITION 3.7. *The set Ω is said to be regular at x provided $T_\Omega^{Cl}(x) = T_\Omega^{Co}(x)$.*

Any convex set is regular at each of its points [8]. Both $T_\Omega^{Co}(x)$ and $T_\Omega^{Cl}(x)$ are closed cones, and both $T_\Omega^{Cl}(x)$ and $T_\Omega^{H}(x)$ are convex cones. Moreover, $T_\Omega^{H}(x) \subseteq T_\Omega^{Cl}(x) \subseteq T_\Omega^{Co}(x)$. Rockafellar [21] showed that $T_\Omega^{H}(x) = int(T_\Omega^{Cl}(x))$ whenever $T_\Omega^{H}(x)$ is nonempty.

Clarke [8] showed that the generalized derivative, as defined in equation (1.1), is Lipschitz continuous with respect to the direction $v \in \mathbb{R}^n$. We use Jahn's definition (1.3) of the Clarke generalized derivative, and need the following lemma to shows that it Lipschitz continuous with respect to v on the hypertangent cone.

LEMMA 3.8. *Let f be Lipschitz near $\hat{x} \in \Omega$ with Lipschitz constant λ. If u and v belong to $T_\Omega^{H}(\hat{x})$, then*

$$f^\circ(\hat{x}; u) \ \geq \ f^\circ(\hat{x}; v) - \lambda\|u - v\|.$$

Proof. Let f be Lipschitz near $\hat{x} \in \Omega$ with Lipschitz constant λ and let u and v belong to $T_\Omega^{H}(\hat{x})$. Let $\epsilon > 0$ be such that $y + tw \in \Omega$ whenever $y \in \Omega \cap B_\epsilon(\hat{x})$, $w \in B_\epsilon(u) \cup B_\epsilon(v)$ and $0 < t < \epsilon$. This can be done by taking ϵ to be the smaller of the values for u and v guaranteed by the definition of a hypertangent. In particular, if $y \in \Omega \cap B_\epsilon(\hat{x})$ and if $0 < t < \epsilon$, then both $y + tu$ and $y + tv$ belong to Ω. This allows

us to go from the first to the second line of the following chain of equalities:

$$\begin{aligned}
f^\circ(\hat{x};u) = & \limsup_{\substack{y\to\hat{x},\ y\in\Omega \\ t\downarrow 0,\ y+tu\in\Omega}} \frac{f(y+tu)-f(y)}{t} \\
= & \limsup_{\substack{y\to\hat{x},\ y\in\Omega \\ t\downarrow 0,\ y+tv\in\Omega}} \frac{f(y+tu)-f(y)}{t} \\
= & \limsup_{\substack{y\to\hat{x},\ y\in\Omega \\ t\downarrow 0,\ y+tv\in\Omega}} \frac{f(y+tv)-f(y)}{t} + \frac{f(y+tu)-f(y+tv)}{t} \\
= & f^\circ(\hat{x};v) + \limsup_{\substack{y\to\hat{x},\ y\in\Omega \\ t\downarrow 0,\ y+tv\in\Omega}} \frac{f(y+tu)-f(y+tv)}{t} \ \geq\ f^\circ(\hat{x};v) - \lambda\|u-v\|.
\end{aligned}$$

■

Based on the previous lemma, the next Proposition shows that the Clarke generalized derivative is continuous with respect to v on the Clarke tangent cone. The result is necessary to the proofs of Theorems 3.12 and 3.13.

PROPOSITION 3.9. *Let f be Lipschitz near $\hat{x}\in\Omega$. If $T_\Omega^H(\hat{x})\neq\emptyset$ and if $v\in T_\Omega^{Cl}(\hat{x})$ then*

$$f^\circ(\hat{x};v) \ =\ \lim_{\substack{w\to v, \\ w\in T_\Omega^H(\hat{x})}} f^\circ(\hat{x};w).$$

Proof. Let λ be a Lipschitz constant for f near $\hat{x}\in\Omega$ and let $\{w_k\}\subset T_\Omega^H(\hat{x})$ be a sequence of directions converging to a vector $v\in T_\Omega^{Cl}(\hat{x})$. By definition of the hypertangent cone, let $0<\epsilon_k<\frac{1}{k}$ be such that

$$y+tw\in\Omega \text{ whenever } y\in\Omega\cap B_{\epsilon_k}(\hat{x}), w\in B_{\epsilon_k}(w_k) \text{ and } 0<t<\epsilon_k. \tag{3.4}$$

We first show the inequality $f^\circ(\hat{x};v)\leq\lim_k f^\circ(\hat{x};w_k)$. Equation (3.4) implies that

$$\begin{aligned}
f^\circ(\hat{x};v) = & \limsup_{\substack{y\to\hat{x},\ y\in\Omega \\ t\downarrow 0,\ y+tv\in\Omega}} \frac{f(y+tv)-f(y)}{t} \\
= & \limsup_{\substack{y\to\hat{x},\ y\in\Omega \\ t\downarrow 0,\ y+tv\in\Omega \\ y+tw_k\in\Omega}} \frac{f(y+tv)-f(y)}{t} \\
\leq & \limsup_{\substack{y\to\hat{x},\ y\in\Omega \\ t\downarrow 0,\ y+tw_k\in\Omega}} \frac{f(y+tw_k)-f(y)}{t} - \frac{f(y+tw_k)-f(y+tv)}{t} \\
= & f^\circ(\hat{x};w_k) + \limsup_{\substack{y\to\hat{x},\ y\in\Omega \\ t\downarrow 0,\ y+tw_k\in\Omega}} \frac{f(y+tw_k)-f(y+tv)}{t}.
\end{aligned}$$

As k goes to infinity, one gets that $\left|\frac{f(y+tw_k)-f(y+tv)}{t}\right|\leq\lambda\|w_k-v\|$ goes to zero. Since $\{w_k\}$ was arbitrary in the hypertangent cone, it follows that

$$f^\circ(\hat{x};v) \ \leq\ \lim_{\substack{w\to v, \\ w\in T_\Omega^H(\hat{x})}} f^\circ(\hat{x};w).$$

Second, we show the reverse inequality: $f^\circ(\hat{x};v) \geq \lim_k f^\circ(\hat{x};w_k)$. Let us define $u_k = \frac{1}{k}w_k + (1-\frac{1}{k})v = w_k + (1-\frac{1}{k})(v-w_k)$. Observe that since the hypertangent cone is a convex set, and since v lies in the closure of the hypertangent cone, then it follows that $u_k \in T_\Omega^H(\hat{x})$ for every $k = 1, 2, \ldots$

We now consider the generalized directional derivative

$$f^\circ(\hat{x};u_k) = \limsup_{\substack{y \to \hat{x},\ y \in \Omega \\ t \downarrow 0,\ y+tu_k \in \Omega}} \frac{f(y+tu_k)-f(y)}{t}.$$

The fact that $u_k \in T_\Omega^H(\hat{x})$ implies that there exist $y_k \in \Omega \cap B_{\epsilon_k}(\hat{x})$ and $0 < \frac{t_k}{k} < \epsilon_k$ such that $y_k + t_k u_k \in \Omega$ and

$$\frac{f(y_k+t_ku_k)-f(y_k)}{t_k} \geq f^\circ(\hat{x};u_k) - \epsilon_k, \tag{3.5}$$

where ϵ_k is the constant from equation (3.4). We now define the sequence $z_k = y_k + \frac{t_k}{k}w_k \in \Omega$ converging to $\hat{x}$, and the sequence of scalars $h_k = (1-\frac{1}{k})t_k > 0$ converging to zero. Notice that

$$z_k + h_k v = y_k + t_k\left(\frac{1}{k}w_k + (1-\frac{1}{k})v\right) \ = \ y_k + t_k u_k \ \in \ \Omega,$$

and therefore

$$\begin{aligned}
f^\circ(\hat{x};v) &= \limsup_{\substack{z \to \hat{x},\ z \in \Omega \\ h \downarrow 0,\ z+hv \in \Omega}} \frac{f(z+hv)-f(z)}{t} \\
&\geq \lim_k \frac{f(z_k+h_kv)-f(z_k)}{h_k} \\
&= \lim_k \frac{f(y_k+t_ku_k)-f(y_k)}{(1-\frac{1}{k})t_k} + \frac{f(y_k)-f(y_k+\frac{t_k}{k}w_k)}{(1-\frac{1}{k})t_k} \\
\text{by equation (3.5)}: \quad &\geq \lim_k f^\circ(\hat{x};u_k) - \epsilon_k + \frac{f(y_k)-f(y_k+\frac{t_k}{k}w_k)}{(1-\frac{1}{k})t_k} \\
\text{by Lemma 3.8}: \quad &\geq \lim_k f^\circ(\hat{x};w_k) - \lambda\|u_k-w_k\| - \epsilon_k + \frac{\frac{1}{k}\|w_k\|}{(1-\frac{1}{k})} \\
&= \lim_k f^\circ(\hat{x};w_k) - \lambda\|v-w_k\| + \tfrac{1}{k}\|v\| \ = \ f^\circ(\hat{x};w_k).
\end{aligned}$$

■

Unfortunately, the above Proposition is not necessarily true when the hypertangent cone is empty : We cannot show in equation (3.4) that $y+tw_k$ belongs to Ω when $y \in \Omega$ is close to $\hat{x}$ and when $t > 0$ is small. The following example in $\mathbb{R}^2$ shows that in this case, the Clarke generalized derivative is not necessarily upper semi-continuous on the boundary of the Clarke tangent cone.

EXAMPLE 3.10. *Consider the continuous concave function in* $\mathbb{R}^2$ *:* $f(a,b) = -\max\{0,a\}$. *Moreover, define the feasible region* Ω *to be the union of*

$$\Omega_1 = \{(a,b)^T : a \geq 0,\ b \geq 0\} \quad \textit{with} \quad \Omega_2 = \{(a,b)^T : b = -a^2,\ a \leq 0\}.$$

One can verify that

$$T_\Omega^H(0) = \emptyset,\ T_\Omega^{Cl}(0) = \Omega_1, \quad \textit{and } T_\Omega^{Co}(0) = \Omega_1 \cup \{(a,0)^t : a \leq 0\},$$

and therefore Ω is not regular at the origin. We will show that $f^\circ(0;w)$ is positive for w in the interior of the Clarke tangent cone but $f^\circ(0;e_1) = -1$ with $e_1 = (1,0)^T$ on the boundary of the Clarke tangent cone.

Let $w = (w_1, w_2)^T$ be any direction in $int(T_\Omega^{Cl}(0))$. We will construct appropriate subsequences in order to compute a valid lower bound on $f^\circ(0;w)$. Define

$$y_k = \left(\frac{-w_1}{k},\ \frac{-w_1^2}{k^2}\right)^T \quad \text{and} \quad t_k = \frac{1}{k} \quad \text{for every positive integer } k.$$

One can easily check that $y_k \in \Omega_2 \subset \Omega$ and $y_k + t_k w = \left(0,\ \frac{1}{k}(v - \frac{w_1^2}{2})\right)^T \in \Omega_1 \subset \Omega$ for every $k > \frac{w_1^2}{w_2}$. It follows that

$$f^\circ(0;w) \geq \lim_{k\to\infty} \frac{f(y_k + t_k w) - f(y_k)}{t_k} = \lim_{k\to\infty} \frac{0-0}{\frac{1}{k}} = 0.$$

In particular, $f^\circ(0;(1,\epsilon)^T)$ is nonnegative for any $\epsilon > 0$. However, if one considers the direction $e_1 = (1,0)^T$ on the boundary of $T_\Omega^{Cl}(0)$ and computes the Clarke generalized derivative, then the origin cannot be approached by points $y_k = (a_k, b_k)^T \in \Omega$ with $b_k < 0$ and with $y_k + t_k e_1 \in \Omega$. A necessary condition for both sequences to be in Ω is that y_k belongs to Ω_1, where f is linear, and therefore, one gets that the Clarke generalized derivative is $f^\circ(0;e_1) = -1$.

This example shows that when the hypertangent cone is empty, but the interior of the Clarke tangent cone is nonempty, it is possible that $f^\circ(\hat{x};w)$ is nonnegative for every w in the interior of the Clarke tangent cone and jumps to a negative value on the boundary of the tangent cone: $f^\circ(\hat{x};e_1) < \limsup_{w\to e_1} f^\circ(\hat{x};w)$.

We now present two types of necessary optimality conditions based on the tangent cone definitions.

DEFINITION 3.11. *Let f be Lipschitz near $\hat{x} \in \Omega$. Then, $\hat{x}$ is said to be a Clarke, or contingent stationary point of f over Ω, if $f^\circ(\hat{x};v) \geq 0$ for every direction v in the Clarke, or contingent cone of Ω at $\hat{x}$, respectively.*

In addition, $\hat{x}$ is said to be a Clarke, or contingent KKT stationary point of f over Ω, if $-\nabla f(\hat{x})$ exists and belongs to the polar of the Clarke, or contingent cone of Ω at $\hat{x}$, respectively.

3.3. A hierarchy of convergence results for MADS. We now present our basic result on refining directions from which our hierarchy of results are derived.

THEOREM 3.12. *Let f be Lipschitz near a limit $\hat{x} \in \Omega$ of a refining subsequence, and $v \in T_\Omega^H(\hat{x})$ be a refining direction for $\hat{x}$. Then the generalized directional derivative of f at $\hat{x}$ in the direction v is nonnegative,* i.e., *$f^\circ(\hat{x};v) \geq 0$.* *Proof.* Let $\{x_k\}_{k\in K}$ be a refining subsequence converging to $\hat{x}$ and $v = \lim_{k\in L} \frac{d_k}{\|d_k\|} \in T_\Omega^H(\hat{x})$ be a refining direction for $\hat{x}$, with $d_k \in D_k$ for every $k \in L$. Since f is Lipschitz near $\hat{x}$, Proposition 3.9 ensures that $f^\circ(\hat{x};v) = \lim_{k\in L} f^\circ(\hat{x}; \frac{d_k}{\|d_k\|})$. But, for any $k \in L$, one can apply the definition of the Clarke generalized derivative with the roles of y and t played by x_k and $\Delta_k^m \|d_k\|$, respectively. Note that this last quantity indeed converges to zero since Definition 2.2 ensures that it is bounded above by $\Delta_k^p \max\{\|d'\| : d' \in D\}$, where D is a finite set of directions, and Equation (2.2) states that Δ_k^p goes to zero. Therefore

$$f^\circ(\hat{x};w_k) \geq \limsup_{k\in L} \frac{f(x_k + \Delta_k^m \|d_k\| \frac{d_k}{\|d_k\|}) - f(x_k)}{\Delta_k^m \|d_k\|} = \limsup_{k\in L} \frac{f(x_k + \Delta_k^m d_k) - f(x_k)}{\Delta_k^m \|d_k\|} \geq 0.$$

The last inequality follows from the fact that for each $k \in L$, $x_k + \Delta_k^m d_k \in \Omega$ and $f(x_k + \Delta_k^m d_k) = f_\Omega(x_k + \Delta_k^m d_k)$ was evaluated and compared by the algorithm to $f(x_k)$, but x_k is a minimal frame center. ∎

We now show that Clarke derivatives of f at the limit $\hat{x}$ of minimal frame centers, for meshes that get infinitely fine, are nonnegative for all directions in the hypertangent cone. Note that even though the algorithm is applied to f_Ω instead of f, the convergence results are linked to the local smoothness of f and not f_Ω, which is obviously discontinuous on the boundary of Ω. This is because we use (1.3) as the definition of the Clarke generalized derivative instead of (1.1). The constraint qualification used in these results is that the hypertangent cone is non-empty at the feasible limit point $\hat{x}$. Further discussion on non-empty hypertangent cones is found in Rockafellar [21].

THEOREM 3.13. *Let f be Lipschitz near a limit $\hat{x} \in \Omega$ of a refining subsequence, and assume that $T_\Omega^H(\hat{x}) \neq \emptyset$. If the set of refining directions for $\hat{x}$ is dense in $T_\Omega^H(\hat{x})$, then $\hat{x}$ is a Clarke stationary point of f on Ω. Proof.* The proof follows directly from Theorem 3.12 and Proposition 3.9.

A corollary to this result is that if f is strictly differentiable at $\hat{x}$, then it is a Clarke KKT point.

COROLLARY 3.14. *Let f be strictly differentiable at a limit $\hat{x} \in \Omega$ of a refining subsequence, and assume that $T_\Omega^H(\hat{x}) \neq \emptyset$. If the set of refining directions for $\hat{x}$ is dense in $T_\Omega^H(\hat{x})$, then $\hat{x}$ is a Clarke KKT stationary point of f over Ω. Proof.* Strict differentiability ensures that the gradient $\nabla f(\hat{x})$ exists and that $\nabla f(\hat{x})^T v = f^\circ(\hat{x}; v)$ for any directions. It follows directly from the previous proposition that $-\nabla f(\hat{x})^T v \leq 0$ for every direction v in $T_\Omega^{Cl}(\hat{x})$, thus $\hat{x}$ is a Clarke KKT stationary point. ∎

Our next two results are based on the definition of set regularity (see Definition 3.7).

PROPOSITION 3.15. *Let f be Lipschitz near a limit $\hat{x} \in \Omega$ of a refining subsequence, and assume that $T_\Omega^H(\hat{x}) \neq \emptyset$. If the set of refining directions for $\hat{x}$ is dense in $T_\Omega^H(\hat{x})$, and if Ω is regular at $\hat{x}$, then $\hat{x}$ is a contingent stationary point of f over Ω. Proof.* The definition of regularity of a set ensures that $f^\circ(\hat{x}; w) \geq 0$ for all w in $T_\Omega^{Co}(\hat{x})$. ∎

The following result is the counterpart to Corollary 3.14 for contingent stationarity. The proof is omitted since it is essentially the same.

COROLLARY 3.16. *Let f be strictly differentiable at a limit $\hat{x} \in \Omega$ of a refining subsequence, and assume that $T_\Omega^H(\hat{x}) \neq \emptyset$. If the set of refining directions for $\hat{x}$ is dense in $T_\Omega^H(\hat{x})$, and if Ω is regular at $\hat{x}$, then $\hat{x}$ is a contingent KKT stationary point of f over Ω.*

Example F in [2] presents an instance of a GPS algorithm such that when applied to a given unconstrained optimization problem, generates a single limit point $\hat{x}$ which is not a Clarke stationary point. In fact, it is shown that f is differentiable but not strictly differentiable at $\hat{x}$ and $\nabla f(\hat{x})$ is nonzero. This unfortunate circumstance is due to the fact that GPS uses a finite number of poll directions. MADS can use infinitely many.

The following result shows that the algorithm ensures strong optimality conditions for unconstrained optimization, or when $\hat{x}$ is in the interior of Ω.

THEOREM 3.17. *Let f be Lipschitz near a limit $\hat{x}$ of a refining subsequence. If $\Omega = \mathbb{R}^n$, or if $\hat{x} \in int(\Omega)$, and if the set of refining directions for $\hat{x}$ is dense in $\mathbb{R}^n$, then $0 \in \partial f(\hat{x})$. Proof.* Let x be as in the statement of the result, then $T_\Omega^H(\hat{x}) = \mathbb{R}^n$. Combining the previous corollary with equation (1.2) yields the result. ∎

We have show that MADS, an algorithm that deals only with function values and does not evaluate or estimate derivatives, can produce a limit point $\hat{x}$ such that if the function f is Lipschitz near $\hat{x}$ and if the hypertangent cone to Ω at $\hat{x}$ is nonempty, then $\hat{x}$ is a Clarke stationary point. The main algorithmic condition in this result is that the set of refining directions is dense in $T_\Omega^H(\hat{x})$. In the general statement of the algorithm we did not present a strategy that would guarantee a dense set of refining directions in the hypertangent cone. We want to keep the algorithm framework as general as possible. There are different strategies that could be used to generate a dense set of poll directions. The selection of the set D_k could be done in a deterministic way or may use some randomness. We present, analyze, and test one strategy in the next section.

4. Practical implementation – LTMADS. We now present two examples of a stochastic implementation of the MADS algorithm. We call either variant LTMADS, because of the underlying lower triangular basis construction, and we show that with probability 1, the set of poll directions generated by the algorithm is dense in the whole space, and in particular in the hypertangent cone.

4.1. Implementable instances of a MADS algorithm. Let $G = I$, the identity matrix, and let $D = Z = [I \ -I], \tau = 4, w^- = -1$ and $w^+ = 1$ be the fixed algorithmic parameters. Choose $\Delta_0^m = 1, \Delta_0^p = 1$ to be the initial mesh and poll size parameters, and define the update rules as follows :

$$\Delta_{k+1}^m = \begin{cases} \frac{\Delta_k^m}{4} & \text{if } x_k \text{ is a minimal frame center} \\ 4\Delta_k^m & \text{if an improved mesh point is found, and if } \Delta_k^m \leq \frac{1}{4} \\ \Delta_k^m & \text{otherwise.} \end{cases}$$

A consequence of these rules is that the mesh size parameter is always a power of 4 and never exceeds 1. Thus, $\frac{1}{\sqrt{\Delta_k^m}} \geq 1$ is always a nonnegative power of 2 and hence integral. The positive spanning set D_k contains a positive basis of $n+1$ or $2n$ directions constructed as follows. It goes without saying that including more directions from D is allowed.

GENERATION OF THE POSITIVE BASIS D_k.

- BASIS CONSTRUCTION: Let B be a lower triangular matrix where each term on the diagonal is either plus or minus $\frac{1}{\sqrt{\Delta_k^m}}$, and the lower components are integers in the open interval $\left]\frac{-1}{\sqrt{\Delta_k^m}}, \frac{1}{\sqrt{\Delta_k^m}}\right[$ randomly chosen with equal probability.
- PERMUTATION OF LINES AND COLUMNS OF B: Let $\{i_1, i_2, \ldots, i_n\}$ and $\{j_1, j_2, \ldots, j_n\}$ be random permutations of the set $\{1, 2, \ldots, n\}$. Set $d_p^q = B_{i_p, j_q}$ for each p and q in $\{1, 2, \ldots, n\}$.
- COMPLETION TO A POSITIVE BASIS:
  - A minimal positive basis : $n+1$ directions.
    Set $d^{n+1} = -\sum_{i=1}^n d^i$ and let $D_k = \{d^1, d^2, \ldots, d^{n+1}\}$.
    Set the poll size parameter to $\Delta_k^p = n\sqrt{\Delta_k^m} \geq \Delta_k^m$.
  - A maximal positive basis : $2n$ directions.
    Set $d^{n+i} = -d^i$ for $i = 1, 2, \ldots, n$ and let $D_k = \{d^1, d^2, \ldots, d^{2n}\}$.
    Set the poll size parameter to $\Delta_k^p = \sqrt{\Delta_k^m} \geq \Delta_k^m$.

Some comments are in order, since MADS is allowed to be *opportunistic* and end a POLL step as soon as a better point is found, we want the order of the POLL directions we generate to be random as are the directions themselves. Thus, the purpose of the

second step is to permute the lines so that the line with $n-1$ zeroes is not always the first in D_k, and permute the columns so that the dense column is not always the first in D_k. The name LTMADS is based on the lower triangular matrix at the heart of the construction of the frames.

The following result shows that the frames generated by the LTMADS algorithm satisfy the conditions of Definition 2.2.

PROPOSITION 4.1. *At each iteration k, the procedure above yields a D_k and a MADS frame P_k such that:*

$$P_k = \{x_k + \Delta_k^m d : d \in D_k\} \ \subset \ M_k \ ,$$

where M_k is given by Definition 2.1 and D_k is a positive spanning set such that for each $d \in D_k$,

- *$d \neq 0$ can be written as a nonnegative integer combination of the directions in $D = [I \ -I]$: $d = Du$ for some vector $u \in \mathbb{N}^{n_D}$ that may depend on the iteration number k*
- *the distance from the frame center x_k to a poll point $x_k + \Delta_k^m d$ is bounded by a constant times the poll size parameter : $\Delta_k^m \|d\|_\infty \leq \Delta_k^p \max\{\|d'\|_\infty : d' \in D\}$.*
- *limits (as defined in Coope and Price [9]) of the normalized sets D_k are positive spanning sets.*

Proof. The first n columns of D_k form a basis of $\mathbb{R}^n$ because they are obtained by permuting rows and columns of the lower triangular matrix B, which is nonsingular because it has nonzero terms on the diagonal. Moreover, taking the last direction to be the negative of the sum of the others leads to a minimal positive basis, and combining the first n columns of D_k with their negatives gives a maximal positive basis [11].

Again by construction, D_k has all integral entries in the interval $[-\frac{1}{\sqrt{\Delta_k^m}}, \frac{1}{\sqrt{\Delta_k^m}}]$, and so clearly each column d of D_k can be written as a nonnegative integer combination of the columns of $D = [I, -I]$. Hence, the frame defined by D_k is on the mesh M_k.

Now the ℓ_∞ distance from the frame center to any poll point is $\|\Delta_k^m d\|_\infty = \Delta_k^m \|d\|_\infty$. There are two cases. If the maximal positive basis construction is used, then $\Delta_k^m \|d\|_\infty = \sqrt{\Delta_k^m} = \Delta_k^p$. If the minimal positive basis construction is used, then $\Delta_k^m \|d\|_\infty \leq n\sqrt{\Delta_k^m} = \Delta_m^p$. The proof of the second bullet follows by noticing that $\max\{\|d'\|_\infty : d' \in [I \ -I]\} = 1$.

The frame can be rewritten in the equivalent form $\{x_k + \sqrt{\Delta_k^m} v : v \in \mathcal{V}\}$ where $\mathcal{V}$ is a set whose columns are the same as those of B after permutation and multiplied by $\sqrt{\Delta_k^m}$. Coope and Price [9] show that a sufficient condition for the third bullet to hold is that each element of $|\mathcal{V}|$ is bounded above and that $|det(\mathcal{V})|$ is bounded below by positive constants that are independent of k. This is trivial to show with our construction. Indeed, each entry of $\mathcal{V}$ lies between -1 and 1 and every term on the diagonal is ± 1. B is a triangular matrix, and therefore $|det(\mathcal{V})| = 1$.

■

Figure 4.1 illustrates some possible poll sets in $\mathbb{R}^2$ for three values of Δ_k^m and Δ_k^p with minimal positive bases. With standard GPS, the frame would have to be chosen among the eight neighboring mesh points. With the new algorithm, the frame may be chosen among the mesh points lying inside the square with the dark contour. On can see that as Δ_k^m and Δ_k^p go to zero, the number of candidates for poll points increases rapidly. For the example illustrated in the figure, the sets of directions D_k are respectively $\{(-1,0)^T, (0,-1)^T, (1,1)^T\}, \{(-2,-1)^T, (1,2)^T, (2,-2)^T\}$

and $\{(4,1)^T,(-3,4)^T,(-4,-4)^T\}$. In the rightmost figure, there are a total of 104 distinct possible frames that MADS may choose from.

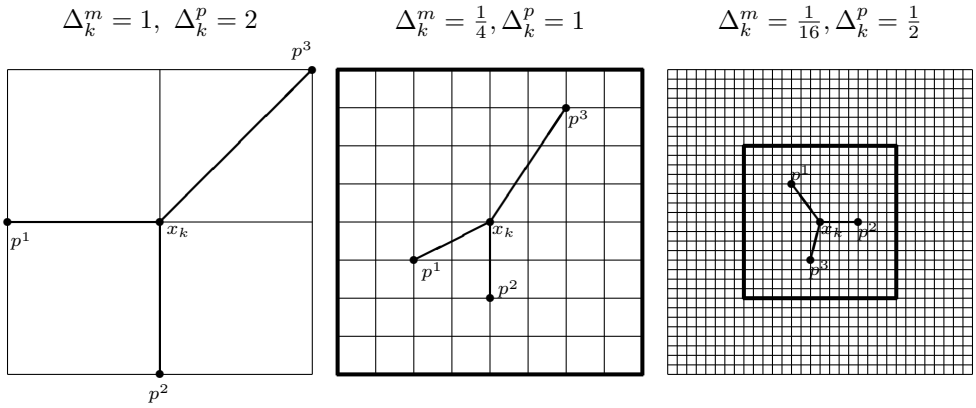

FIG. 4.1. *Example of frames* $P_k = \{x_k + \Delta_k^m d : d \in D_k\} = \{p^1, p^2, p^3\}$ *for different values of* Δ_k^m *and* Δ_k^p. *In all three figures, the mesh* M_k *is the intersection of all lines.*

In addition to an opportunistic strategy, *i.e.*, terminating a POLL step as soon as an improved mesh point is detected, a standard trick we use in GPS to improve the convergence speed consists in promoting a successful poll direction to the top of the list of directions for the next POLL step. We call this *dynamic ordering* of the polling directions. This strategy can not be directly implemented in MADS since at a successful iteration $k-1$, the poll size parameter is increased, and therefore a step of Δ_k in the successful direction will often be outside the mesh. The way we mimic GPS dynamic ordering in MADS is that when the previous iteration succeeded in finding an improved mesh point, we execute a simple one point *dynamic search* in the next iteration as follows. Suppose that $f_\Omega(x_k) < f_\Omega(x_{k-1})$ and that d is the direction for which $x_k = x_{k-1} + \Delta_{k-1}^m d$. Then, the trial point produced by the SEARCH step is $t = x_{k-1} + 4\Delta_{k-1}^m d$. Note that with this construction, if $\Delta_{k-1}^m < 1$, then $t = x_{k-1} + \Delta_k^m d$ and otherwise, $t = x_{k-1} + 4\Delta_k^m d$. In both cases t lies on the current mesh M_k. If this SEARCH finds a better point, then we go on to the next iteration, but if not, then we proceed to the POLL step. The reader will see in the numerical results below that this seems to be a good strategy.

4.2. Convergence analysis. The convergence results in Section 3.3 are based on the assumption that the set of refining directions for the limit of a refining sequence is dense in the hypertangent cone at that limit. The following result shows that the above instances of LTMADS generates a dense set of poll directions, and therefore, the convergence results based on the local smoothness of the objective function f and on the local topology of the feasible region Ω can be applied.

THEOREM 4.2. *Let* $\hat{x} \in \Omega$ *be the limit of a refining subsequence produced by either instance of LTMADS. Then the set of poll directions for the subsequence converging to* $\hat{x}$ *is dense in* $T_\Omega^H(\hat{x})$ *with probability one. Proof.* Let $\hat{x}$ be the limit of a refining subsequence $\{x_k\}_{k\in K}$ produced by one of the above instances of LTMADS (either with the minimal or maximal positive basis). Consider the sequence of positive bases $\{D_k\}_{k\in K}$. Each one of these bases is generated independently.

We use the notation $P[E]$ to denote the probability that E occurs. Let v be a direction in $\mathbb{R}^n$ with $\|v\|_\infty = 1$ such that $P[|v_j| = 1] \geq \frac{1}{n}$ and $P[v_j = 1 \mid |v_j| = 1] =$

$P[v_j = -1 \mid |v_j| = 1] = \frac{1}{2}$. We will find a lower bound on the probability that a normalized direction in D_k is arbitrarily close to the vector v.

Let k be an index of K. Recall that in the generation of the positive basis D_k, the column d^{i_1} is such that $|d^{i_1}_{j_1}| = \frac{1}{\sqrt{\Delta^m_k}}$, and the other components of d^{i_1} are random integers in the open interval $\left]\frac{-1}{\sqrt{\Delta^m_k}}, \frac{1}{\sqrt{\Delta^m_k}}\right[$. Set $u = \frac{d^{i_1}}{\|d^{i_1}\|_\infty}$. It follows by construction that $u = d^{i_1}\sqrt{\Delta^m_k}$ and $|u_{j_1}| = 1$. We will now show for any $0 < \epsilon < 1$, that the probability that $\|u - v\|_\infty < \epsilon$ is bounded below by some non-negative number independent of k, as $k \in K$ goes to infinity. Let us estimate the probability that $|u_j - v_j| < \epsilon$ for each j. For $j = j_1$ we have

$$\begin{aligned} P[|u_{j_1} - v_{j_1}| < \epsilon] &\geq P[u_{j_1} = v_{j_1} = 1] + P[u_{j_1} = v_{j_1} = -1] \\ &= P[u_{j_1} = 1] \times P[v_{j_1} = 1] + P[u_{j_1} = -1] \times P[v_{j_1} = -1] \\ &\geq \frac{1}{2} \times \frac{1}{2n} + \frac{1}{2} \times \frac{1}{2n} \;=\; \frac{1}{2n}. \end{aligned}$$

For $j \in \{j_2, j_3, \ldots, j_n\}$ we have

$$P[|u_j - v_j| < \epsilon] = P[v_j - \epsilon < u_j < v_j + \epsilon] \;=\; P\left[\frac{v_j - \epsilon}{\sqrt{\Delta^m_k}} < d^{i_1}_j < \frac{v_j + \epsilon}{\sqrt{\Delta^m_k}}\right].$$

We will use the fact that the number of integers in the interval $\left[\frac{v_j - \epsilon}{\sqrt{\Delta^m_k}}, \frac{v_j + \epsilon}{\sqrt{\Delta^m_k}}\right] \cap \left]\frac{-1}{\sqrt{\Delta^m_k}}, \frac{1}{\sqrt{\Delta^m_k}}\right[$ is bounded below by the value $\frac{\epsilon}{\sqrt{\Delta^m_k}} - 1$. Now, since the bases D_k are independently generated, and since $d^{i_1}_j$ is an integer randomly chosen with equal probability among the $\frac{2}{\sqrt{\Delta^m_k}} - 1$ integers in the interval $\left]\frac{-1}{\sqrt{\Delta^m_k}}, \frac{1}{\sqrt{\Delta^m_k}}\right[$, then it follows that

$$P[|u_j - v_j| < \epsilon] \geq \frac{\frac{\epsilon}{\sqrt{\Delta^m_k}} - 1}{\frac{2}{\sqrt{\Delta^m_k}} - 1} \;>\; \frac{\epsilon - \sqrt{\Delta^m_k}}{2}.$$

Recall that $\hat{x}$ is the limit of a refining subsequence, and so, there exists an integer α such that $\sqrt{\Delta^m_k} \leq \frac{\epsilon}{2}$ whenever $\alpha \leq k \in K$, and so

$$P[|u_j - v_j| < \epsilon] \geq \frac{\epsilon - \sqrt{\Delta^m_k}}{2} \;\geq\; \frac{\epsilon}{4} \qquad \text{for any } k \in K \text{ with } k \geq \alpha.$$

It follows that

$$P[\|u - v\|_\infty < \epsilon] \;=\; \prod_{j=1}^{n} P[|u_j - v_j| < \epsilon] \geq \frac{\left(\frac{\epsilon}{4}\right)^{n-1}}{2n} \qquad \text{for any } k \in K \text{ with } k \geq \alpha.$$

We have shown when k is sufficiently large, that $P[\|u - v\|_\infty < \epsilon]$ is larger than a strictly positive constant, and is independent of Δ^m_k. Thus, there will be a poll direction in D_k for some $k \in K$ arbitrarily close to any direction $v \in \mathbb{R}^n$, and in particular to any direction $v \in T^H_\Omega(\hat{x})$. ∎

The proof of the previous result shows that the set of directions consisting of the d^{i_1} directions over all iterations is dense in $\mathbb{R}^n$. Nevertheless, we require the algorithm

to use a positive spanning set at each iteration instead of a single poll direction. This ensures that any limit of a refining subsequence is the limit of minimal frame centers on meshes that get infinitely fine. At this limit point, the set of refining directions is generated from the set of poll directions which is dense in LTMADS and finite in GPS. Therefore with both MADS and GPS, the set of directions for which the Clarke generalized derivatives are nonnegative positively span the whole space. However, only LTMADS allow the possibility for the set of refining directions to be dense.

5. Numerical results. In this section, we look at 4 test problems. The first problem is unconstrained, but GPS is well-known to stagnate if it is given an unsuitable set of directions. MADS has no problem converging quickly to a global optimizer. The second case is a bound constrained chemical engineering problem where GPS is known to behave well enough to justify publication of the results [13]. Still, on the whole, MADS does better. The third case is a simple nonlinearly constrained problem where GPS and our filter version of GPS are both known to converge short of an optimizer. As the theory given here predicts, MADS has no difficulty.

The last example is such that the feasible region gets narrow very quickly. This is meant to be a test for any derivative-free feasible point algorithm - like GPS or MADS with the extreme barrier approach to constraints. MADS does better than GPS with the filter or the barrier, both of which stagnate due to the limitation of finitely many POLL directions. MADS stops making progress when the mesh size gets smaller than the precision of the arithmetic. This is not to say that in exact arithmetic MADS would have converged to $-\infty$, the optimizer. Most likely, a SEARCH procedure would allow even more progress in all the algorithms, but that is not the point. The point is to test the POLL procedure.

Of course, even when one tries to choose carefully, 4 examples are not conclusive evidence. However, we believe that these numerical results coupled with the more powerful theory for MADS make a good case for MADS versus GPS.

5.1. An unconstrained problem where GPS does poorly. Consider the unconstrained optimization problem in $\mathbb{R}^2$ presented in [15] where GPS algorithms are known to converge to non-stationary points :

$$f(x) \;=\; \left(1 - \exp(-\|x\|^2\right) \times \max\{\|x-c\|^2, \|x-d\|^2\},$$

where $c = -d = (30, 80)^T$. Figure 5.1 shows level sets of this function. It can be shown that f is locally Lipschitz and and strictly differentiable at its global minimizer $(0, 0)^T$.

The GPS and MADS runs are initiated at $x_0 = (-3.3, 1.2)^T$. The gradient of f exists and is non-zero at that point, and therefore both GPS and MADS will move away from it. Since there is some randomness involved in the MADS instances described in Section 4.1, we ran it a total of 5 times, to see how it compares to our standard NOMAD implementation of GPS. Figure 5.2 shows a log plot of the progress of the objective function value for each sets of runs. The runs were stopped when a minimal frame with poll size parameter less than 10^{-10} was detected. For GPS, the maximal $2n$ positive basis refers to the set of positive and negative coordinate directions, and the two minimal $n+1$ positive bases are $\{(1,0)^T, (0,1)^T, (-1,-1)^T\}$ and $\{(1,0)^T, (-0.5, 0.866025)^T, (-0.5, -0.866025)^T\}$.

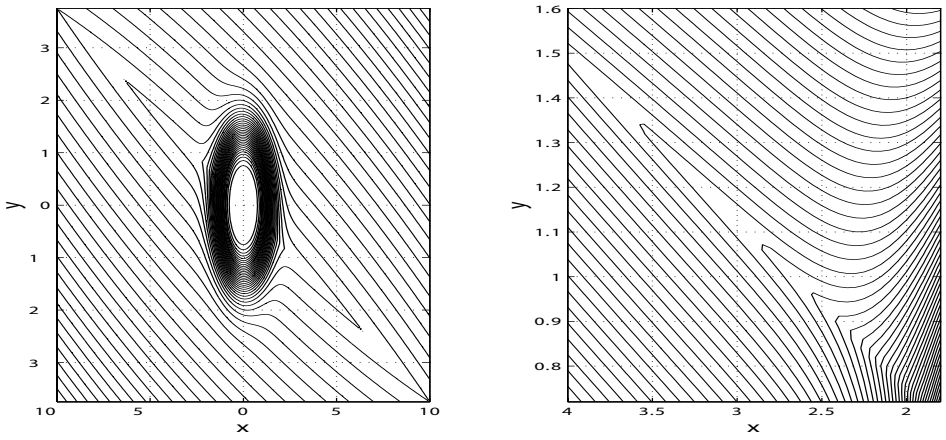

FIG. 5.1. *Level sets of f.*

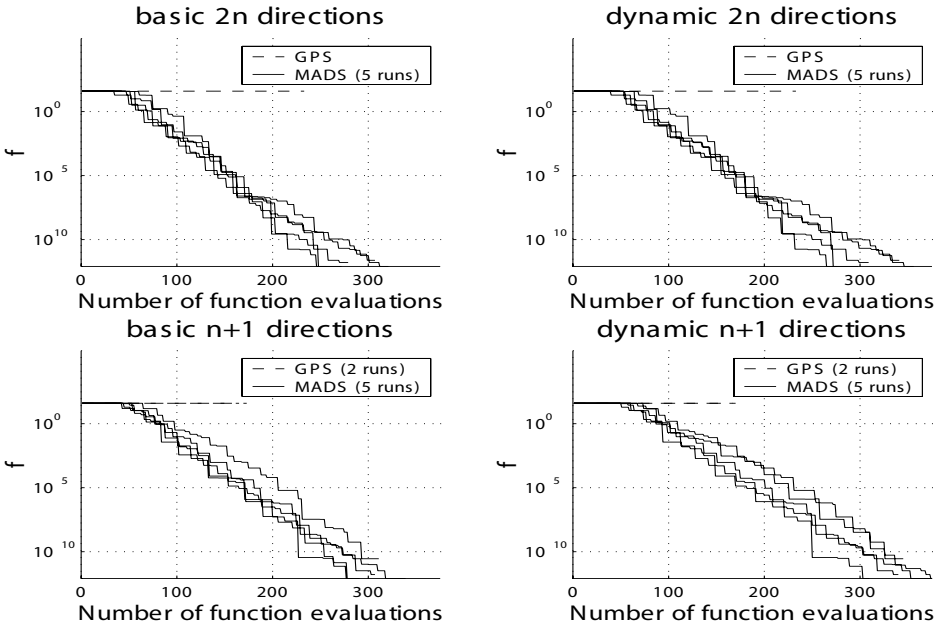

FIG. 5.2. *Progression of the objective function value vs the number of evaluations.*

Every GPS run converged to the nearby point $(-3.2, 1.2)^T$, where f is not differentiable. As proved in [3], the Clarke generalized directional derivatives are nonnegative for D at that point, but, it is not an optimizer. One can see by looking at the level sets of f that no descent directions at $(-3.2, 1.2)^T$ can be generated by the GPS algorithm. However, all MADS runs eventually generated good directions and converged to the origin, the global optimal solution. Figure 5.1 suggests that even if randomness appears in these instances of MADS, the behavior of the algorithm is very stable in converging quickly to the origin.

5.2. A test problem where GPS does well. The above was one of our motivational examples for MADS, and so we next tried a test case where GPS was known to behave so well that the results merited a publication. Hayes *et al.* [13] describe a method for evaluating the kinetic constants in a rate expression for catalytic combustion applications using experimental light-off curves. The method uses a transient one-dimensional single channel monolith finite element reactor model to simulate reactor performance. The objective is to find the values of four parameters in a way such that the model estimates as closely as possible (in a weighted least square sense) an experimental conversion rate. This is a bound constrained nonsmooth optimization problem in $\mathbb{R}^4_+$, where the objective function measures the error between experimental data and values predicted by the model.

For the three sets of experimental data analyzed in [13], we ran the instances of GPS and MADS discussed above. The algorithms terminate whenever a a minimal frame center with poll size parameter equal to 2^{-6} is detected, or whenever 500 functions evaluations are performed, whichever comes first. Figures 5.3, 5.4, and 5.5 show the progression of the objective function value versus the number of evaluations for each data set.

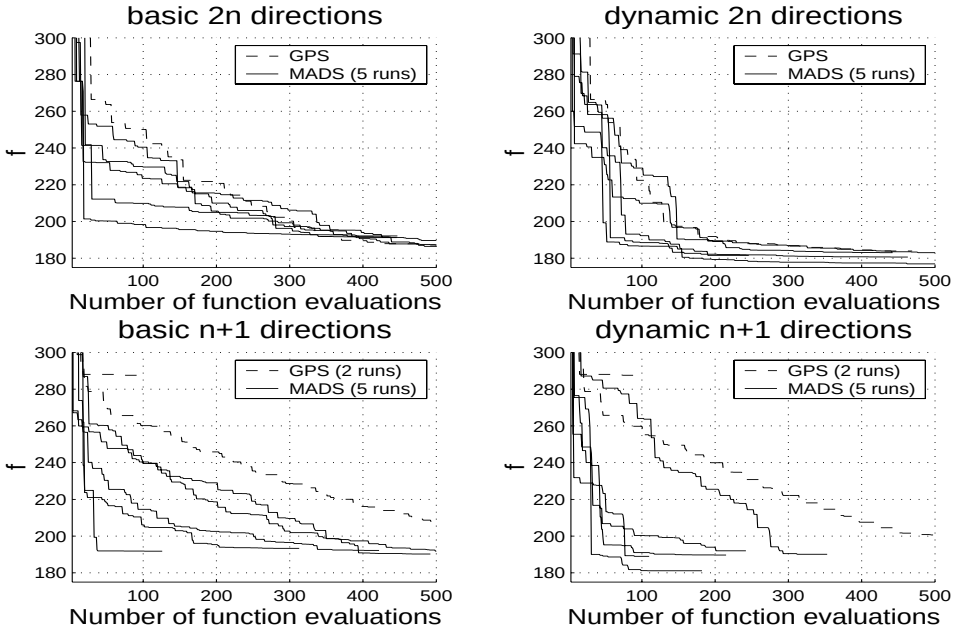

FIG. 5.3. *Data set 1 – Progression of the objective function value vs the number of evaluations.*

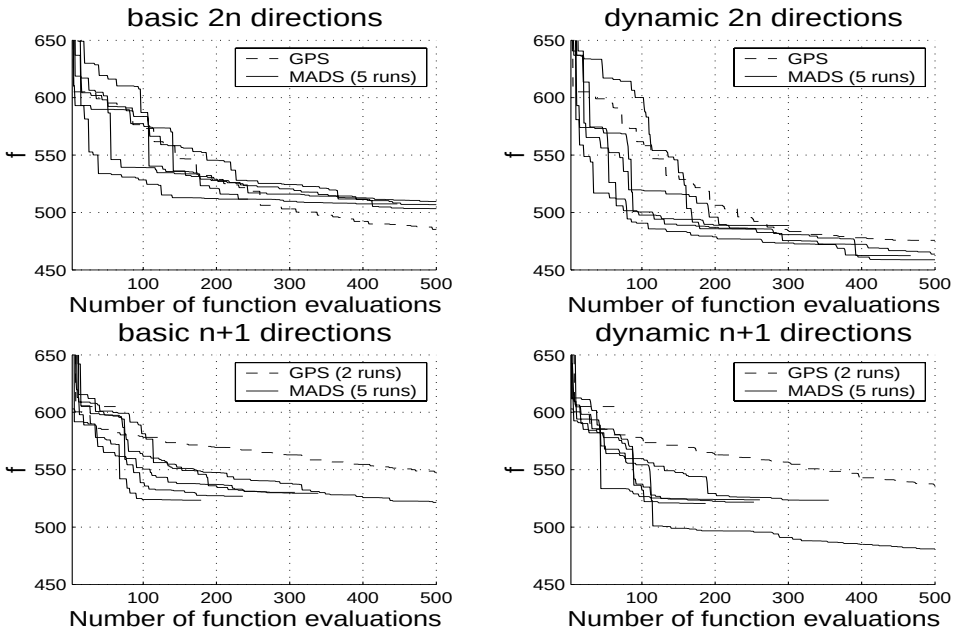

FIG. 5.4. *Data set 2 – Progression of the objective function value vs the number of evaluations.*

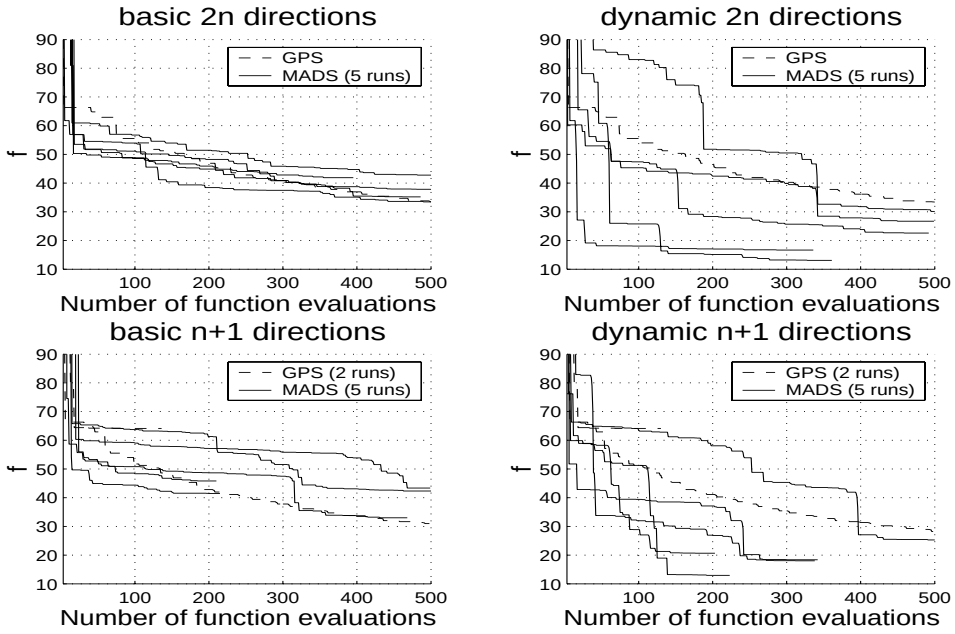

FIG. 5.5. *Data set 3 – Progression of the objective function value vs the number of evaluations.*

The plots suggest that the objective function value without the dynamic search procedure decreases more steadily with GPS than with MADS. This is because GPS uses a fixed set of poll directions that we know to be excellent for this problem. By allowing more directions, MADS eventually generates a steep descent direction, and the dynamic runs capitalizes on this by evaluating f further in that direction thus sharply reducing the objective function value in a few evaluations. In general, if the number of function evaluations is limited to a fixed number, then it appears that MADS gives better result than GPS. For all three data sets, the dynamic runs are preferable to the basic runs. It also appears that MADS runs with maximal $2n$ basis perform better than the minimal $n+1$ runs. In each of the three data sets, the best overall solution was always produced by MADS with the dynamic 2n directions.

The quality of the best solutions produced by GPS and MADS can be visualized in Figure 5.6 where the difference between the experimental and predicted conversions are plotted versus time. A perfect model with perfectly tuned parameters would have had a difference of zero everywhere. The superiority of the solution produced by MADS versus GPS is mostly visible for the third data set.

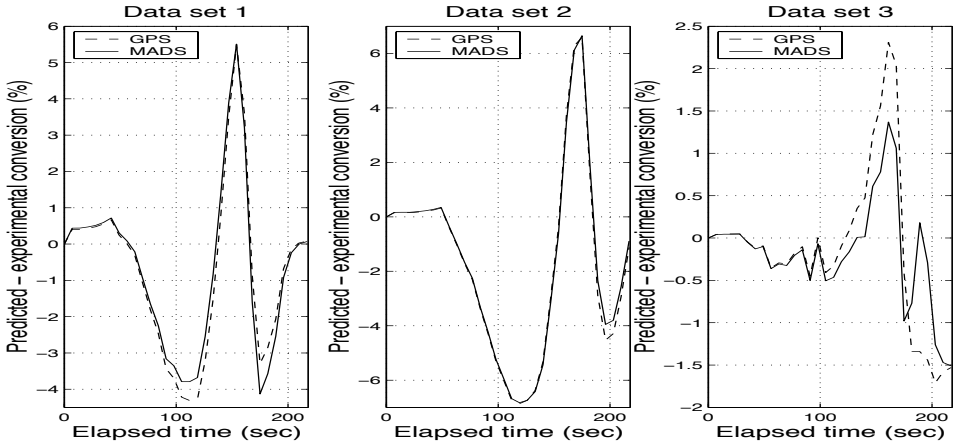

FIG. 5.6. *Conversion rate error versus time.*

5.3. A quadratically constrained problem. The third example shows again the difficulty caused by being restricted to a finite number of polling directions. This is a problem with a linear objective and a disk constraint, surely the simplest nonlinearly constrained problem imaginable.

$$\begin{aligned} \min_{x=(a,b)^T} \quad & a+b \\ \text{s.t.} \quad & a^2+b^2 \le 6. \end{aligned}$$

The starting point is the origin, and the same stopping criteria as for the first example was used, $\Delta^p < 10^{-10}$, with the following results. For GPS we always used $D_k = D = [I, -I]$ with dynamic ordering. The GPS filter method of [4] reached the point $x_* = (-1.42, -1.99)^T$ with $f(x_*) = -3.417$. The GPS using the barrier approach to the constraint got to $x_* = (-1.41, -2)^T$ with $f(x_*) = -3.414$. All 5 runs of the MADS method of the previous section ended with $f(x_*) = -3.464$, which is the global optimum, though the arguments varied between -1.715 and -1.749 for all the runs. The progression of the algorithms is illustrated in Figure 5.7.

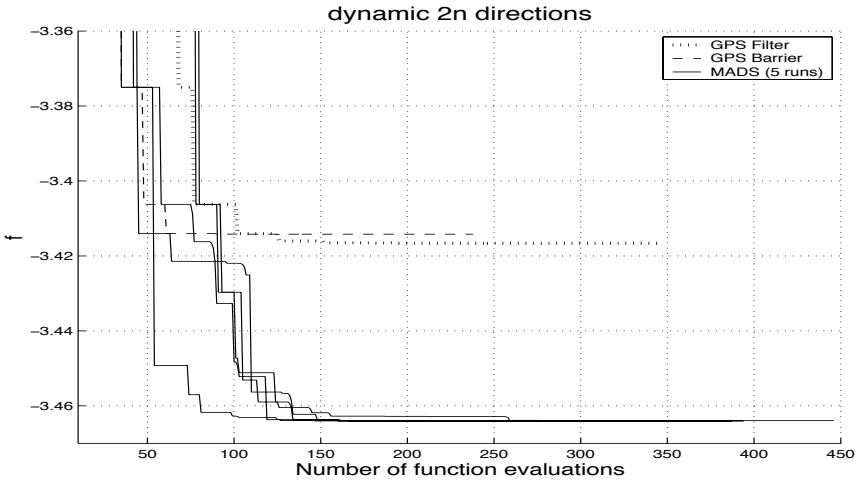

FIG. 5.7. *Progression of the objective function value vs the number of evaluations on an easy nonlinear problem*

Thus the GPS filter method did slightly better than the GPS barrier, and MADS solved the problem easily, as the theory given here predicts.

5.4. A nasty constrained problem for both GPS and MADS. This last example does not satisfy the hypotheses of any GPS or MADS theorems because the optimizer is at $-\infty$. But, it is intended to see how well the various algorithms track a feasible region that gets narrow quickly. For this reason, the objective is not meant to be an issue, it is linear.

$$\begin{aligned} \min_{x=(a,b)^T} \quad & a \\ \text{s.t.} \quad & e^a \le b \le 2e^a. \end{aligned}$$

The starting point is $(0, 1)^T$, and the stopping criteria is activated when $\Delta_k^m < 10^{-323}$ *i.e.*, when the mesh size parameter drops below the smallest positive representable

number. We admit that this is excessive, but we wanted to run the algorithms to their limits. The same strategies as in Section 5.3 are used.

The progression of the algorithms is illustrated in Figure 5.8. Both GPS with the barrier and filter approaches to constraints converged quickly to points where the standard 2n basis does not contain a feasible descent direction. The filter GPS approach to constraints did better than the GPS barrier approach, because it is allowed to become infeasible.

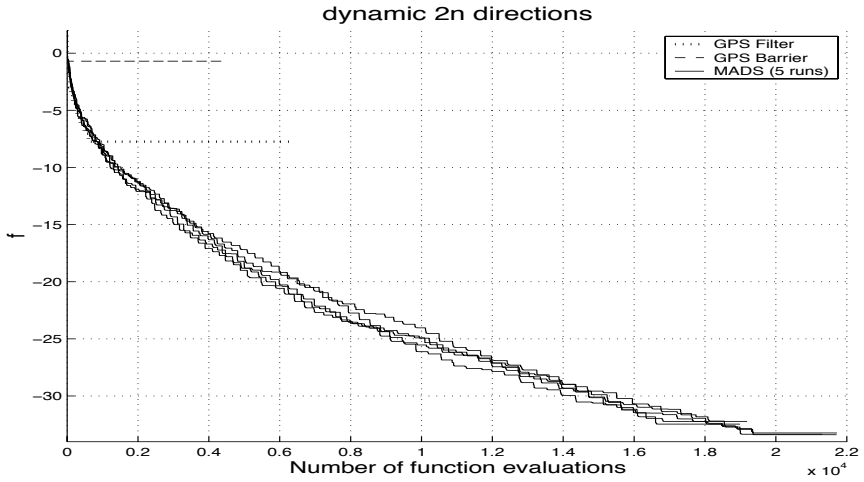

FIG. 5.8. *Progression of the objective function value vs the number of evaluations on a difficult nonlinear problem*

All 5 runs of the LTMADS method of the previous section ended with roughly the same solution, a point where $a \pm \Delta_k^p = a$, which all one can ask. The fact that LTMADS uses a dense set of poll directions explains why it does better.

The feasible region is very narrow, and therefore it gets quite improbable that the MADS poll directions generate a feasible point, and when such a feasible point is generated it is always very close to the frame center since the mesh and poll parameters are very small.

Even if the algorithm instances failed to solve this problem to optimality and converged to points that are not Clarke stationary points, the GPS and MADS convergence theory is not violated – yet. In all cases, there is a set of directions that positively span $\mathbb{R}^2$ such that for each directions either the Clarke generalized derivative is non-negative or is an infeasible directions.

6. Discussion. GPS is a valuable algorithm, but the application of nonsmooth analysis techniques in [3] showed its limitations due to the finite choice of directions in [3]. MADS removes the restriction of GPS to finitely many poll directions. We have long felt that this was the major impediment to stronger proofs of optimality for GPS limit points (and better behavior), and in this paper we do find more satisfying optimality conditions for MADS in addition to opening new possibilities for handling nonlinear constraints.

We gave a stochastic version of MADS, LTMADS, which performed well, especially for a first implementation. We expect others will find more, and perhaps better, implementations. Our priority is to explore MADS for constraints.

We think that the work here is readily applied to choosing templates for implicit filtering [7], another very successful algorithm for nasty nonlinear problems.

Finally, we wish to thank Gilles Couture for coding NOMAD, the C++ implementation of MADS and GPS, and to acknowledge useful discussions with Andrew Booker and Mark Abramson.

REFERENCES

[1] Abramson M. A. , Audet C. and Dennis J.E.Jr. (2003), "Generalized pattern searches with derivative information," Technical Report *TR02-10*, Department of Computational and Applied Mathematics, Rice University, Houston Texas, to appear in *Mathematical Programming Series B*.

[2] Audet C. (2002), "Convergence results for pattern search algorithms are tight," *Les Cahiers du GERAD* G-2002-56, Montréal. To appear in *Optimization and Engineering*.

[3] Audet C. and Dennis J.E.Jr. (2003), "Analysis of generalized pattern searches," *SIAM Journal on Optimization* 13, 889-903.

[4] Audet C. and Dennis J.E.Jr. (2000): "A pattern search filter method for nonlinear programming without derivatives," Technical Report *TR00-09*, Department of Computational and Applied Mathematics, Rice University, Houston Texas, to appear in *SIAM Journal on Optimization*.

[5] Avriel M. (1976) "Nonlinear Programming Analysis and Methods," Prentice-Hall, Englewood Cliffs, NJ.

[6] Booker A.J., Dennis J.E.Jr, Frank P.D., Serafini D.B., Torczon V. and Trosset M.W. (1999), "A rigorous framework for optimization of expensive functions by surrogates," *Structural Optimization* Vol.17 No.1, 1-13.

[7] Choi T.D. and Kelley C.T. (1999), "Superlinear convergence and implicit filtering," *SIAM Journal on Optimization* Vol.10 No.4, 1149–1162.

[8] Clarke, F.H. (1990) "Optimization and Nonsmooth Analysis," SIAM Classics in Applied Mathematics Vol.5, Philadelphia.

[9] Coope I.D. and Price C.J. (2000), "Frame-based methods for unconstrained optimization," *Journal of Optimization Theory and Application* Vol. 107, 261–274.

[10] Coope I.D. and Price C.J. (2001), "On the convergence of grid-based methods for unconstrained optimization," *SIAM Journal on Optimization* Vol.11, 859–869.

[11] Davis C. (1954): Theory of positive linear dependence, *Amer. J. Math.*, **76**, 733–746.

[12] Gould F.J. and Tolle J.W. (1972), "Geometry of optimality conditions and constraint qualifications," *Mathematical Programming*, **2**, 1–18.

[13] Hayes R.E., Bertrand F.H., Audet C. and Kolaczkowski S.T. (2002), "Catalytic Combustion Kinetics: Using a Direct Search Algorithm to Evaluate Kinetic Parameters from Light-Off Curves," *Les Cahiers du GERAD* G-2002-20, Montréal. To appear in *The Canadian journal of Chemical Engineering*.

[14] Jahn J. (1994), "Introduction to the Theory of Nonlinear Optimization," Springer, Berlin.

[15] Kolda T.G., Lewis R.M. and Torczon V. (2003), "Optimization by direct search : New perspectives on some classical and modern methods," *SIAM Review*, Vol.46, 385–482.

[16] Leach E.B. (1961) "A note on inverse function theorem," Proc. AMS Vol.12, 694-697.

[17] Lewis R.M. and Torczon V. (1996), "Rank ordering and positive basis in pattern search algorithms," *ICASE* TR 96-71.

[18] Lewis R.M. and Torczon V. (2002), "A globally convergent augmented Lagrangian pattern search algorithm for optimization with general constraints and simple bounds," *SIAM Journal on Optimization*, Vol.12 No.4, 1075-1089.

[19] McKay M.D., Conover W.J. and Beckman R.J. (1979): A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics*, Vol.21, 239–245.

[20] Rockafellar R.T. (1970) "Convex analysis," Princeton University Press.

[21] Rockafellar R.T. (1980) "Generalized directional derivatives and subgradients of nonconvex functions," *Canadian Journal of Mathematics* Vol.32, 157–180.

[22] Torczon V. (1997), "On the Convergence of Pattern Search Algorithms," *SIAM Journal on Optimization* Vol.7 No.1, 1–25.