J Optim Theory Appl (2007) 134: 47–59
DOI 10.1007/s10957-007-9204-2

Pattern Search Method for Discrete L_1–Approximation

C. Bogani · M.G. Gasparo · A. Papini

Published online: 17 May 2007

Abstract We propose a pattern search method to solve a classical nonsmooth optimization problem. In a deep analogy with pattern search methods for linear constrained optimization, the set of search directions at each iteration is defined in such a way that it conforms to the local geometry of the set of points of nondifferentiability near the current iterate. This is crucial to ensure convergence. The approach presented here can be extended to wider classes of nonsmooth optimization problems. Numerical experiments seem to be encouraging.

Keywords Pattern search methods · Nonsmooth optimization · Linear L_1–estimation · Convex optimization

1 Introduction

Pattern search methods are a particular class of direct search methods first analyzed by Torczon for unconstrained optimization [1, 2]; they were successfully extended to bound-constrained and linearly-constrained optimization in [3, 4]. The research about

Communicated by G. Di Pillo.

This work was supported by M.U.R.S.T., Rome, Italy.

C. Bogani (✉)
Dipartimento di Matematica "Ulisse Dini", Università di Firenze, Firenze, Italy
e-mail: bogani@math.unifi.it

M.G. Gasparo · A. Papini
Dipartimento di Energetica "Sergio Stecco", Università di Firenze, Firenze, Italy

M.G. Gasparo
e-mail: mariagrazia.gasparo@unifi.it

A. Papini
e-mail: alessandra.papini@unifi.it

pattern search methods is still flourishing: several generalizations and extensions have been proposed recently (see e.g. [5–8] and the survey [9]).

Given an initial guess x^0 and a steplength $\Delta_0 > 0$, a pattern search method generates a sequence of iterates $\{x^k\}$ by conducting a series of exploratory moves around the current iterate x^k, along a finite set of search directions $\{d_1^k, d_2^k, \dots, d_{r_k}^k\}$. At the kth iteration the objective function f is evaluated at the trial points $x^k + \Delta_k d_1^k, x^k + \Delta_k d_2^k, \dots, x^k + \Delta_k d_{r_k}^k$: if $f(x^k + \Delta_k d_j^k) < f(x^k)$ for some j, the next iterate x^{k+1} is selected among the trial points in such a way that $f(x^{k+1}) < f(x^k)$ and the steplength is increased or left unaltered. Otherwise, we set $x^{k+1} := x^k$ and the steplength is reduced.

A key issue to design convergent pattern search methods is that the set of search directions must contain a descent direction for f at x^k (of course, if x^k is not a minimizer). In unconstrained strictly differentiable[1] optimization this can be easily ensured by using positive spanning sets of $\mathbb{R}^n$ [7–10]. For constrained problems this simple rule is no more valid when x^k lies on the boundary of the feasible region: in this case, the set of search directions must conform to the local geometry of the constraints [3, 4].

Convergence properties of pattern searches when applied to nonsmooth problems are studied in several recent papers. In [7], it is shown that, under mild assumptions, pattern search methods produce a (Clarke) stationary limit point if f is strictly differentiable at that point. But, if f is only Lipschitz near the limit point, all that we can say is that the (Clarke) generalized directional derivatives along the search directions are nonnegative. This result is tight because of the restriction to finitely many search directions [6]. On the other hand, in practice a pattern search algorithm may converge to a point where f is not strictly differentiable, whether or not it is a stationary point (see the examples in [6, 9]). In particular, quoting from [9], pattern searches "may not converge to a stationary point when applied to nonsmooth problems—especially when the loci of nonsmoothness are highly structured".

A way to cope with these potential pitfalls of pattern searches is described in [8], where it is shown that a (Clarke) stationary point can be found among the accumulation points, if the set of search directions is asymptotically dense in $\mathbb{R}^n$. Another possible approach consists in exploiting the structure of the set $\mathcal{S}$ where f is not strictly differentiable (when this structure is known), in order to choose adaptively suitable search directions. This is similar in principle to the idea developed in [4] for linearly constrained optimization.

In this paper we show the effectiveness of this approach, by considering the classical discrete L_1–approximation problem

$$\min_{x \in \mathbb{R}^n} f(x) := \|A^T x - b\|_1, \tag{1}$$

where $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^m$ are given with $m > n$. Problem (1) has been studied since the 50s and solved by linear programming techniques [11, 12]. Indeed, it can

[1] f is strictly differentiable at x if there exists a $D_s f(x) \in \mathbb{R}^n$ such that, for all $w \in \mathbb{R}^n$, $\lim_{y \to x, t \to 0^+} \frac{f(y+tw) - f(y)}{t} = D_s f(x)^T w$.

be restated as

$$\begin{aligned} &\min_{p,q,x} \sum_{j=1}^{m}(p_j+q_j), \\ &\text{s.t.} \quad A^T x - p + q = b, \\ &\qquad\; p \geq 0, \\ &\qquad\; q \geq 0 \end{aligned}$$

which is a classical linear program. More recently it has been solved by iterative methods based on affine scaling [13]. We remark that the objective function f is not differentiable on the union of m hyperplanes and (1) admits solutions which lie on some of these hyperplanes [11].

It is worthwhile to stress that our aim here is not simply to propose a new method for the L_1–approximation problem: some of the algorithms known in the literature are so specialized that pretending to outperform them is at best a waste of time. Instead we believe that the main contribution of this paper consists in proving that a suitable implementation of pattern searches can be successfully employed to solve (1), without making use of linear programming. This approach has the advantage of being readily extendible to more general classes of nonsmooth problems.

The paper is organized as follows. In Sect. 2 we recall some preliminary results on the objective function f and some elements of convex analysis which are basic to characterize the minimizers of convex nonsmooth functions. In Sect. 3 we consider the geometry of the set $\mathcal{S}$ and prove this nice property: if $x \in \mathcal{S}$ is not a minimizer, then a descent direction can be found among the generators of the polyhedral cones delimited by the hyperplanes of nondifferentiability passing through x. Section 4 is devoted to the definition of the pattern search algorithm suggested by the previous property and to the convergence analysis. In Sect. 5 we report the results of some numerical experiments. Some conclusions and perspectives are drawn in Sect. 6.

Throughout the paper, we adopt the following notations: a_j is the jth column of A; $\Pi_j := \{x \in \mathbb{R}^n : \rho_j(x) := a_j^T x - b_j = 0\}$ is the jth hyperplane of nondifferentiability; $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{Z}$ and $\mathbb{N}$ denote the sets of real, rational, integer and natural numbers respectively; $\|\cdot\|$ is the Euclidean norm; $\mathcal{K} := \langle u_1, u_2, \ldots, u_r\rangle^+ = \{\lambda_1 u_1 + \lambda_2 u_2 \ldots + \lambda_r u_r : \lambda_1, \lambda_2, \ldots, \lambda_r \geq 0\}$ is the polyhedral cone generated by $u_1, u_2, \ldots, u_r$; $\mathcal{K}^0 := \{v \in \mathbb{R}^n : u^T v \leq 0 \text{ for } u \in \mathcal{K}\}$ is the polar cone associated with $\mathcal{K}$.

2 Background

We consider problem (1) under the following assumptions:

Assumption 2.1 $\operatorname{rank}(A) = n$.

Assumption 2.2 $A \in \mathbb{Q}^{n\times m}$.

Assumption 2.1 ensures that the level sets of f are bounded (cf. Proposition 2.1 below). It is not required by simplex based methods, which attain convergence in a finite number of steps. On the other hand, it is required by methods converging to the solution only in the limit (see e.g. [13]). Assumption 2.2 is standard in the context of linearly constrained pattern search methods [4, 7, 9]. In our opinion this theoretical requirement is not a restriction in practice, since only rational numbers exist in floating point arithmetic.

Proposition 2.1 *The objective function f in (1) is convex. Moreover if Assumption 2.1 holds, then*

(i) *Given $\zeta \in \mathbb{R}^n$, the set $\mathcal{L}(\zeta) := \{x \in \mathbb{R}^n : f(x) \le f(\zeta)\}$ is compact and convex.*
(ii) *There exists $x^* \in \mathbb{R}^n$ s.t. $f(x^*) = \min_{x\in\mathbb{R}^n} f(x)$.*
(iii) *$X^* := \{x \in \mathbb{R}^n : f(x) = \min_{x\in\mathbb{R}^n} f(x)\}$ is compact and convex.*

Proof The closure of $\mathcal{L}(\zeta)$ follows from the continuity of f. To prove that $\mathcal{L}(\zeta)$ is bounded, it is sufficient to show that there exist $\beta, \hat{\gamma} > 0$ s.t. $f(x) \ge \hat{\gamma}\|x\| - \beta$ for $x \in \mathbb{R}^n$. We recall from [2] the uniform linear independence lemma: given a basis $\{d_1, d_2, \dots, d_n\}$ of $\mathbb{R}^n$, there exists $\gamma > 0$ s.t.

$$\max_{j=1,2,\dots,n} |d_j^T y| \ge \gamma\|y\| \quad \text{for } y \in \mathbb{R}^n. \tag{2}$$

Because of Assumption 2.1, the set $\{a_1, a_2, \dots, a_m\}$ contains $r \ge 1$ bases of $\mathbb{R}^n$. Let $\gamma_1, \gamma_2, \dots, \gamma_r$ be the corresponding constants in (2) and let $\hat{\gamma} := \min_j \gamma_j$. Let $\beta := \|b\|_\infty$ and, given $x \in \mathbb{R}^n$, let k s.t. $|a_k^T x| = \max_j |a_j^T x|$. Then $|a_k^T x| \ge \hat{\gamma}\|x\|$ and hence

$$\hat{\gamma}\|x\| - \beta \le |a_k^T x| - |b_k| \le |a_k^T x - b_k| \le f(x).$$

The convexity of $\mathcal{L}(\zeta)$ comes from the convexity of f. Items (ii) and (iii) follow trivially from item (i). □

Below, we recall some definitions and results of convex analysis useful to detect the local behavior of f and to define descent directions at points of nondifferentiability. Given $\varphi : \mathbb{R}^n \to \mathbb{R}$ convex and $x, v \in \mathbb{R}^n$, we denote by $D\varphi(x, v)$ (respectively $\varphi^0(x, v)$) the right directional derivative (respectively Clarke's generalized derivative) of φ at x along v.

Proposition 2.2 *Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be convex. Then:*

(i) *Given $x \in \mathbb{R}^n$, φ is regular at x, that is $\varphi^0(x, v) = D\varphi(x, v)$ for all $v \in \mathbb{R}^n$.*
(ii) *Given $x \in \mathbb{R}^n$, $D\varphi(x, \cdot) : \mathbb{R}^n \to \mathbb{R}$ is positively homogenous, sublinear and Lipschitz continuous.*

Proof See [14]. □

The following proposition emphasizes the role played by right directional derivatives in convex analysis.

Proposition 2.3 *Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be convex. Then $x \in \mathbb{R}^n$ is a minimizer if and only if $D\varphi(x, v) \geq 0$ for $v \in \mathbb{R}^n$.*

Proof Since φ is convex, given $t_1, t_2, t_3 \in \mathbb{R}$ s.t. $t_1 < t_2 < t_3$, the following inequality holds:

$$\frac{\varphi(x + t_2 v) - \varphi(x + t_1 v)}{t_2 - t_1} \leq \frac{\varphi(x + t_3 v) - \varphi(x + t_1 v)}{t_3 - t_1}.$$

From this, we deduce the inequality

$$\varphi(x + tv) \geq \varphi(x) + tD\varphi(x, v), \qquad \text{for } v \in \mathbb{R}^n \text{ and } t \geq 0,$$

which concludes the proof in one sense. The reverse implication is obvious. □

The definition of descent direction is now straightforward.

Definition 2.1 Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be convex and $x \in \mathbb{R}^n$. Then $v \in \mathbb{R}^n$ is said to be a descent direction at x if $D\varphi(x, v) < 0$.

Proposition 2.4 *Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be convex. Given $x \in \mathbb{R}^n$, the set $\mathcal{D}(x) := \{v \in \mathbb{R}^n : D\varphi(x, v) < 0\}$ is an open and convex cone, possibly empty.*

Proof It follows from Proposition 2.2(ii). □

3 Descent Directions

If $x \notin \mathcal{S}$, every positive spanning set of $\mathbb{R}^n$ is suitable to detect descent directions. If $x \in \mathcal{S}$, this is no more true. Indeed if $x \in \mathcal{S}$ is not a minimizer, then by Proposition 2.4 $\mathcal{D}(x)$ can be smaller than a halfspace. For example, let $f(x_1, x_2) := |x_1| + 5|x_2| + 2|x_1 + x_2 - 1|$. Clearly, f is not differentiable at the lines $\Pi_1 : x_1 = 0$, $\Pi_2 : x_2 = 0$ and $\Pi_3 : x_1 + x_2 = 1$. Let $x := (0, 0)^T \in \Pi_1 \cap \Pi_2$; it is easy to see that x is not a minimizer since

$$\mathcal{D}(x) = \left\{(v_1, v_2)^T \in \mathbb{R}^2 : -\frac{1}{7} < \frac{v_2}{v_1} < \frac{1}{3}, v_1 > 0\right\}.$$

Consider the positive spanning set $\mathcal{P} := \{(1, 1)^T, (-1, 1)^T, (-1, -1)^T, (1, -1)^T\}$; then, $\mathcal{P} \cap \mathcal{D}(x) = \emptyset$, i.e., $Df(x, d) \geq 0$, for $d \in \mathcal{P}$. If we consider in alternative $\mathcal{P}' := \{(1, 0)^T, (0, 1)^T, (-1, 0)^T, (0, -1)^T\}$, then $\mathcal{P}' \cap \mathcal{D}(x) = \{(1, 0)^T\}$. Therefore, $\mathcal{P}'$ is well designed to detect descent directions at x. Let us remark that $\mathcal{P}'$ contains the generators of the cones delimited by the lines of nondifferentiability Π_1 and Π_2 passing through x.

We introduce some basic notations. For $x \in \mathcal{S}$, we define

$$\mathcal{I}(x) = \{j : \rho_j(x) = 0\} := \{j_1, j_2, \ldots, j_{l(x)}\},$$

$$A(x) := [a_{j_1}\, a_{j_2}\, \ldots\, a_{j_{l(x)}}],$$

$$\mathcal{J}(x) := \{\sigma \in \mathbb{Z}^{l(x)} : \ |\sigma_j| = 1, \ j = 1, 2, \ldots, l(x)\},$$
$$\mathcal{K}_\sigma(x) := \langle \sigma_1 a_{j_1}, \sigma_2 a_{j_2}, \ldots, \sigma_{l(x)} a_{j_{l(x)}} \rangle^+, \quad \sigma \in \mathcal{J}(x).$$

From now on, we will assume for simplicity that $\mathcal{I}(x) = \{1, 2, \ldots, l(x)\}$.

The following lemma emphasizes the role played by the polar cones $\mathcal{K}_\sigma^0(x)$, when analyzing the directional derivatives of f at $x \in \mathcal{S}$.

Lemma 3.1 *Given $x \in \mathcal{S}$ and $\sigma \in \mathcal{J}(x)$, there exists $g_\sigma = g_\sigma(x) \in \mathbb{R}^n$ s.t. $Df(x,u) = g_\sigma^T u$ for $u \in \mathcal{K}_\sigma^0(x)$.*

Proof An easy calculation yields

$$Df(x, v) = \sum_{j=1}^{l(x)} |a_j^T v| + \sum_{j=l(x)+1}^{m} \operatorname{sign}(\rho_j(x)) a_j^T v, \tag{3}$$

for all $v \in \mathbb{R}^n$. Take $u \in \mathcal{K}_\sigma^0(x)$; then, we have $\sigma_j a_j^T u \le 0$, $j = 1, 2, \ldots, l(x)$, and from (3) it follows that $Df(x,u) = g_\sigma^T u$, where

$$g_\sigma := -\sum_{j=1}^{l(x)} \sigma_j a_j + \sum_{j=l(x)+1}^{m} \operatorname{sign}(\rho_j(x)) a_j.$$

□

Now, we are able to prove a key property of the descent directions at points of nondifferentiability.

Theorem 3.1 *Let $x \in \mathcal{S}$ be given. If x is not a minimizer, then $Df(x,d) < 0$ for some $d \in \Gamma(x) := \cup_{\sigma \in \mathcal{J}(x)} \mathcal{G}_\sigma(x)$, where $\mathcal{G}_\sigma(x)$ is a set of generators for $\mathcal{K}_\sigma^0(x)$.*

Proof Since x is not a minimizer, by Proposition 2.3 it follows that a direction $u \in \mathbb{R}^n$ exists s.t. $Df(x,u) < 0$. For $j = 1, 2, \ldots, l(x)$, let us define $\sigma_j = 1$, if $a_j^T u \le 0$, and $\sigma_j = -1$, otherwise. Then, we have $\sigma_j a_j^T u \le 0$ for all j, so that $u \in \mathcal{K}_\sigma^0(x)$. Let $\mathcal{G}_\sigma(x) = \{z_1, \ldots, z_p\}$ be a set of generators for $\mathcal{K}_\sigma^0(x)$; then $u = \sum_{i=1}^p \alpha_i z_i$, with nonnegative coefficients $\alpha_1, \ldots, \alpha_p$. By applying Lemma 3.1, we get

$$Df(x,u) = g_\sigma^T \sum_{i=1}^p \alpha_i z_i = \sum_{i=1}^p \alpha_i g_\sigma^T z_i = \sum_{i=1}^p \alpha_i Df(x, z_i).$$

Since $Df(x,u) < 0$, it must be $Df(x, z_i) < 0$ for some i. □

Theorem 3.1 states that either $\Gamma(x)$ contains a descent direction for f at x or x is a minimizer. This emphasizes the advantages of knowing a set of generators for the polar cones $\mathcal{K}_\sigma^0(x)$, with $\sigma \in \mathcal{J}(x)$. If $\operatorname{rank}(A(x)) < l(x)$, the construction of $\mathcal{G}_\sigma(x)$ is nontrivial; we refer to [15] for a thorough discussion about this issue. Instead, a simple representation of $\mathcal{G}_\sigma(x)$ is available when $\operatorname{rank}(A(x)) = l(x)$: this recalls the nondegeneracy condition in [4] and implies $l(x) \le n$.

Proposition 3.1 *Let $x \in \mathcal{S}$ be given and assume that* $\operatorname{rank}(A(x)) = l(x)$. *Let $\mathcal{V}^+(x)$ be a positive basis for $N(A^T(x))$, the null space of $A^T(x)$, and denote by $\mathcal{W}(x) = \{w_1, w_2, \ldots, w_{l(x)}\}$ the set of columns of $-A(x)(A^T(x)A(x))^{-1}$. For $\sigma \in \mathcal{J}(x)$ define $\mathcal{W}_\sigma(x) := \{\sigma_1 w_1, \sigma_2 w_2, \ldots, \sigma_{l(x)} w_{l(x)}\}$. Then $\mathcal{W}_\sigma(x) \cup \mathcal{V}^+(x)$ is a set of generators for $\mathcal{K}^0_\sigma(x)$.*

Proof Let $A_\sigma(x) := [\sigma_1 a_1 \ \ldots \ \sigma_{l(x)} a_{l(x)}]$. Of course, $N(A^T_\sigma(x)) = N(A^T(x))$ and

$$-A_\sigma(x)(A^T_\sigma(x)A_\sigma(x))^{-1} = -A(x)(A^T(x)A(x))^{-1}\Sigma = [\sigma_1 w_1 \ \ldots \ \sigma_{l(x)} w_{l(x)}],$$

where $\Sigma := \operatorname{diag}(\sigma_1, \ldots, \sigma_{l(x)})$. Then the desired result follows from [4, Proposition 8.2]. □

Remark 3.1 Under the hypotheses of Proposition 3.1, we can choose $\Gamma(x) = \mathcal{W}(x) \cup -\mathcal{W}(x) \cup \mathcal{V}^+(x)$, so that $\Gamma(x)$ consists of at most $2n$ directions.

4 Algorithm and Convergence Results

Here we present and discuss our pattern search algorithm for solving (1). First of all we have to generalize some definitions and results of the previous section. So, given $x \in \mathbb{R}^n$ and $\eta > 0$, we denote by $\mathcal{I}(x, \eta)$ the set of the indices of the hyperplanes Π_j whose distance from x is less than or equal to η, i.e., $\mathcal{I}(x, \eta) := \{j : |\rho_j(x)|/\|a_j\| \le \eta\}$. Coherently, we define $l(x, \eta)$, $\mathcal{J}(x, \eta)$, $A(x, \eta)$, $\mathcal{W}(x, \eta)$, $\mathcal{V}^+(x, \eta)$ and $\mathcal{K}_\sigma(x, \eta)$ with $\sigma \in \mathcal{J}(x, \eta)$, instead of $l(x)$, $\mathcal{J}(x)$, $A(x)$, $\mathcal{W}(x)$, $\mathcal{V}^+(x)$ and $\mathcal{K}_\sigma(x)$ with $\sigma \in \mathcal{J}(x)$.

The following result is an extension of Theorem 3.1.

Theorem 4.1 *Let $x \in \mathcal{S}$ and $\eta > 0$ be given. If x is not a minimizer, then $Df(x, d) < 0$ for some $d \in \Gamma(x, \eta)$, where*

$$\Gamma(x, \eta) := \bigcup_{\sigma \in \mathcal{J}(x,\eta)} \mathcal{G}_\sigma(x, \eta) \tag{4}$$

and $\mathcal{G}_\sigma(x, \eta)$ is a set of generators for $\mathcal{K}^0_\sigma(x, \eta)$.

Proof Without loss of generality, we assume

$$\mathcal{I}(x, \eta) = \{1, 2, \ldots, l + r\} \supseteq \mathcal{I}(x) = \{1, 2, \ldots, l\}. \tag{5}$$

Given a descent direction $u \in \mathbb{R}^n$ for f at x, let $\sigma \in \mathcal{J}(x, \eta)$ be such that $\sigma_j = 1$, if $a_j^T u \le 0$, and $\sigma_j = -1$, otherwise, for $j = 1, 2, \ldots, l + r$. Then, u belongs to $\mathcal{K}^0_\sigma(x, \eta)$ and can be written as $u = \sum_{i=1}^q \alpha_i z_i$, where $\{z_1, \ldots, z_q\} = \mathcal{G}_\sigma(x, \eta)$ and the coefficients $\alpha_1, \ldots, \alpha_q$, are nonnegative. Since $\mathcal{K}^0_\sigma(x, \eta) \subseteq \mathcal{K}^0_{\sigma'}(x)$, with $\sigma' := (\sigma_1, \ldots, \sigma_l)^T$, we can still apply Lemma 3.1 as in Theorem 3.1 to conclude that

$$Df(x,u) = g_{\sigma'}^T \sum_{i=1}^{q} \alpha_i z_i = \sum_{i=1}^{q} \alpha_i g_{\sigma'}^T z_i = \sum_{i=1}^{q} \alpha_i Df(x,z_i) < 0.$$

Hence, $Df(x,z_i) < 0$ for some i. □

Remark 4.1 Let us consider more in general a set of indices $\bar{\mathcal{I}}(x)$ s.t. $\mathcal{I}(x) \subseteq \bar{\mathcal{I}}(x) \subseteq \mathcal{I}(x,\eta)$. Assuming (5), we can suppose $\bar{\mathcal{I}}(x) = \{1,\dots,l+s\}$ with $s \le r$. Further, let $\bar{\Gamma}(x)$ be a set of directions containing a set of generators for the polar cones $\bar{\mathcal{K}}_\sigma^0(x)$ associated with $\bar{\mathcal{K}}_\sigma(x) = \langle \sigma_1 a_1, \sigma_2 a_2, \dots, \sigma_{l+s} a_{l+s} \rangle^+$, for all $\sigma \in \mathbb{Z}^{l+s}$, $|\sigma_j| = 1$, $j = 1,\dots,l+s$. Then, it is seen easily that Theorem 4.1 still holds with $\bar{\Gamma}(x)$ in place of $\Gamma(x,\eta)$. In other words, if $Df(x,d) \ge 0$ for all $d \in \bar{\Gamma}(x)$ then x is a minimizer.

Proposition 3.1 and Remark 3.1 can be extended easily as well. Indeed, under the assumption $\operatorname{rank}(A(x)) = l(x)$, it can be proved by continuity arguments that there exists a sufficiently small $\bar{\eta} > 0$ such that $\operatorname{rank}(A(x,\eta)) = l(x,\eta)$ for $0 \le \eta \le \bar{\eta}$. In this case we can choose

$$\Gamma(x,\eta) = \mathcal{W}(x,\eta) \cup -\mathcal{W}(x,\eta) \cup \mathcal{V}^+(x,\eta). \tag{6}$$

Now we are in position to state the algorithm. Consider an iterate x^k which is not a minimizer. If $\mathcal{I}(x^k,\eta) = \emptyset$, then x^k is sufficiently far from $\mathcal{S}$ and every positive basis H of $\mathbb{R}^n$ is suitable to detect descent directions. Otherwise we define the set of search directions $\Gamma^k = \Gamma(x^k,\eta)$ according to (4). Further, we add to Γ^k a set of optional search directions Λ^k which is useful to explore more distant regions so as to improve the efficiency of the algorithm, but does not play any role in the convergence analysis. All pattern search methods are endowed with a similar tool: see e.g. the set L_k in [2] or the SEARCH step in [7, 8]. In order to fit our method into the pattern search framework the search directions must lie in a finite set and have the form Gz, where G is a nonsingular real *generating matrix* independent from k and z is an integer vector. This is ensured if all directions lie in a finite set of rational vectors. For this reason, we assume $H \subset \mathbb{Q}^n$ and $\Lambda^k \subset \Lambda \subset \mathbb{Q}^n$, with Λ finite. Assumption 2.2 guarantees the existence of rational generators for the polar cones, so that $\Gamma(x^k,\eta) \subset \mathbb{Q}^n$ either by (4) or by (6) in the nondegenerate case [4]. Moreover, we use the following simple rule to ensure that the collection $\{\Gamma^k, k \in \mathbb{N}\}$ is finite: if $\mathcal{I}(x^k,\eta) = \mathcal{I}(x^{k-p},\eta)$ for some $p > 0$, then we set $\Gamma^k = \Gamma^{k-p}$.

Algorithm 4.1 Data: $x^0 \in \mathbb{R}^n$; $\Delta_0, \eta > 0$; $\tau \in \mathbb{Q}$ and $\alpha_L, \alpha_U \in \mathbb{Z}$ s.t. $\tau > 1$ and $\alpha_L < 0 \le \alpha_U$; a positive rational basis H for $\mathbb{R}^n$; a finite set $\Lambda \subset \mathbb{Q}^n$.

```
Step 1 For $k = 0, 1, \dots,$ until convergence, execute the steps below:
  Step 1.1 Determine $\mathcal{I}(x^k,\eta)$
  Step 1.2 If $\mathcal{I}(x^k,\eta) = \emptyset$, then set $\Gamma^k := H$; else compute a rational set of
           generators $\mathcal{G}_\sigma(x^k,\eta)$ for $\mathcal{K}_\sigma^0(x^k,\eta)$, $\sigma \in \mathcal{J}(x^k,\eta)$, and set $\Gamma^k :=$
           $\cup_{\sigma \in \mathcal{J}(x^k,\eta)} \mathcal{G}_\sigma(x^k,\eta)$
  Step 1.3 Define an optional finite set of search directions $\Lambda^k \subset \Lambda$, possibly
           empty, and set $X^k := \{x^k + \Delta_k d : d \in \Gamma^k \cup \Lambda^k\}$
```

Step 1.4 If $f(x^k + \Delta_k d) < f(x^k)$ for some $d \in \Gamma^k$, then choose $x^{k+1} \in X^k$ s.t. $f(x^{k+1}) < f(x^k)$ and set $\Delta_{k+1} := \tau^{\alpha_k}\Delta_k$, with $\alpha_k \in \mathbb{Z}$ and $0 \le \alpha_k \le \alpha_U$; else set $x^{k+1} := x^k$ and $\Delta_{k+1} := \tau^{\alpha_k}\Delta_k$, with $\alpha_k \in \mathbb{Z}$ and $\alpha_L \le \alpha_k \le -1$.

In the convergence theorem below, we will invoke some results of Audet and Dennis [7], related to the so called *refining subsequences* $\{x_k\}_{k\in K}$, that is, subsequences of mesh local optimizers such that $\lim_{k\to+\infty,k\in K} \Delta_k = 0$. An iterate x^k is a *mesh local optimizer* when $f(x^k) \ge f(x^k + \Delta_k d)$ for all $d \in \Gamma^k$.

Theorem 4.2 *Under Assumptions* 2.1 *and* 2.2, *the sequence* $\{x^k\}$ *produced by the algorithm admits some limit points and each of them is a minimizer.*

Proof Since the iterates lie in the level set $\mathcal{L}(x^0)$, which is compact (cf. Proposition 2.1(i)), by [7, Theorem 3.6] there exists at least one convergent refining subsequence $\{x^k\}_{k\in K}$; let $\hat{x}$ denote its limit. We will show in a while that $\hat{x}$ is a minimizer; then, using the convexity of f and the monotonicity of $\{f(x^k)\}$, we can conclude, as in [1, Proposition 5.6], that each limit point of the iteration sequence $\{x^k\}$ is a minimizer.

Depending on the local smoothness of f, $\hat{x}$ satisfies different optimality conditions. Here we have to distinguish between two situations: either f is continuously differentiable, and hence strictly differentiable, near $\hat{x}$, or $\mathcal{I}(\hat{x}) \ne \emptyset$ (f regular, but not strictly differentiable, at $\hat{x}$). In the first case it is known that $\nabla f(\hat{x}) = 0$ (cf. [7, Theorem 3.9]): therefore, from Proposition 2.3 it follows that $\hat{x}$ is a minimizer.

In the second case, we need to show that there exist a subsequence $\{x^k\}_{k\in\bar{K}}$ with $\bar{K} \subseteq K$, a set of indices $\bar{\mathcal{I}} \subseteq \mathcal{I}(x^k, \eta)$ and a set of directions $\bar{\Gamma}$ such that $\mathcal{I}(x^k, \eta) = \bar{\mathcal{I}}$ and $\Gamma(x^k, \eta) = \bar{\Gamma}$ for all $k \in \bar{K}$. Indeed, if $\mathcal{I}(\hat{x}) \ne \emptyset$, there exists an $\epsilon > 0$ s.t. $\mathcal{I}(\hat{x}) \subseteq \mathcal{I}(x, \eta) \subseteq \mathcal{I}(\hat{x}, \eta)$ for $\|\hat{x} - x\| \le \epsilon$, and then $\mathcal{I}(\hat{x}) \subseteq \mathcal{I}(x^k, \eta) \subseteq \mathcal{I}(\hat{x}, \eta)$ for all but finitely many $k \in K$. Since each $\Gamma(x^k, \eta)$ is selected from a finite family of directions, there is an infinite subset of indices $\bar{K} \subseteq K$ such that $\mathcal{I}(x^k, \eta)$ and $\Gamma(x^k, \eta)$ are constant for $k \in \bar{K}$.

Then, we have that $Df(\hat{x}, d) \ge 0$ for all $d \in \bar{\Gamma}$ (cf. [7], Theorem 3.7 and bullet (v) at p. 900), and by Remark 4.1 $\hat{x}$ is a minimizer. □

5 Numerical Results

For simplicity we implemented the algorithm under the nondegeneracy assumption $\mathrm{rank}(A(x^k, \eta)) = l(x^k, \eta)$ whenever $\mathcal{I}(x^k, \eta) \ne \emptyset$. Of course, the choice of η is crucial. For a given value of η, it may happen that $\mathcal{I}(x^k, \eta) \ne \emptyset$ and the matrix $A(x^k, \eta)$ is column rank deficient for some k. In this case we adopt the strategy suggested in [4]: η is reduced by a factor γ until the cardinality of $\mathcal{I}(x^k, \eta)$ decreases and the columns of $A(x^k, \eta)$ are linearly independent. A safeguard minimal value $\eta_{\min} > 0$ is also given: if in the previous backtracking η becomes less than $\eta_{\min}$, failure is declared. Below we report some numerical results obtained with an initial value for η equal to 0.25, $\gamma = 0.5$ and $\eta_{\min} = 10^{-12}$.

If $\mathcal{I}(x^k, \eta) = \emptyset$, we take $\Gamma^k = H := \{\pm e_1, \pm e_2, \ldots, \pm e_n\}$. Otherwise, we use the SVD decomposition of $A(x^k, \eta)$ to compute $\mathcal{W}(x^k, \eta)$ and a basis $\mathcal{V}$ for $N(A(x^k, \eta)^T)$. We define $\mathcal{V}^+(x^k, \eta) := \mathcal{V} \cup -\mathcal{V}$, so that Γ^k consists of $2n$ directions. In general, such a procedure does not satisfy the theoretical requirement for rational sets of generators. However it works very well in practice, since in floating point arithmetic all numbers are rational. Moreover, the SVD decomposition is the best numerical tool to detect the rank of a matrix.

The set of optional search directions Λ^k is defined as follows. We set $\Lambda^k := \emptyset$ if x^k is a mesh local optimizer. Otherwise let $d^k \in \Gamma^k$ be a direction such that $f(x^k + \Delta_k d^k) = \min\{f(x^k + \Delta_k d),\ d \in \Gamma^k\}$. Since further improvement could possibly be made along d^k, we define $\Lambda^k := \{2d^k, 3d^k, \ldots, Md^k\}$ for some M and we try to find the minimum of f over the trial points $x^k + j\Delta_k d^k$, with $j = 1, 2, \ldots, M$.

In order to save computational effort, actually we approximate only such minimum value: we determine first the values of the real parameter $\lambda \in [1, M]$ for which $x^k + \lambda\Delta_k d^k$ belongs to a hyperplane of nondifferentiability and denote these values by λ_q, with $q = 1, 2, \ldots, t$; we define $\lambda_0 = 1$, $\lambda_{t+1} = M$ and assume $\lambda_0 < \lambda_1 < \cdots < \lambda_t < \lambda_{t+1}$. Then, we search for an index $\bar{q} \in \{1, 2, \ldots, t+1\}$ s.t. $f(x^k + \lfloor\lambda_{\bar{q}}\rfloor\Delta_k d^k) > f(x^k + \lfloor\lambda_{\bar{q}-1}\rfloor\Delta_k d^k)$. If such a $\bar{q}$ exists, we set $x^{k+1} := x^k + \lfloor\lambda_{\bar{q}-1}\rfloor\Delta_k d^k$; otherwise $x^{k+1} := x^k + M\Delta_k d^k$. This procedure has the flavor of the linesearch strategies described in [12, 13]. In the experiments, we used $M = 2^{10}$.

As regards the steplength update, we used $\tau = 2$, $\alpha_L = -10$ and $\alpha_U = 1$. More precisely, the steplength is divided by 1024 when $x_{k+1} = x_k$ and is doubled when $x_{k+1} \neq x_k$. Moreover, the steplength is kept unaltered on successful iterations in two cases: when the previous iteration was unsuccessful and when the set of search directions is the same over three consecutive iterations.

We implemented our algorithm in MATLAB 7.0.1 on a processor Pentium 4 (2.80 GHz) and compared its behavior with the algorithms by Barrodale and Roberts [11] and Coleman and Li [13], ad hoc tailored to problem (1). The first one is a modified version of the standard simplex method, the second is a quadratically convergent algorithm, based on the interior point approach. As another term of comparison, we used also the classical simplex method and the interior point implemented in the general purpose linear programming solver LINPROG of the MATLAB OPTIMIZATION TOOLBOX. We compared the methods taking into account the number of iterations required to solve a problem within a given accuracy; moreover, since the five methods may spend considerably different amount of time in each iteration, we considered also the CPU time.

We solved several L_1–approximation problems where a real function $g(t)$, evaluated over a uniform mesh of m points in an interval I, is approximated by a polynomial of degree $n-1$. The results that we report in Tables 1 and 2, which are referred to $g(t) = \exp(-t)$ and $g(t) = \sin(t)$, well represent the behavior of the five methods. In the tables we show the number of performed iterations and the CPU time in centiseconds; the methods are denoted by: PS (our algorithm), BR (Barrodale and Roberts' algorithm), CL (Coleman and Li's algorithm), SX (LINPROG: simplex method), IP (LINPROG: interior-point method). These results were obtained with x^0 equal to the least squares solution of the overdetermined linear system $A^T x = b$ as suggested by Coleman and Li; LINPROG uses built–in starting points, as well as the Barrodale

Table 1 Iterations/CPU times (cs) for $g(t) = \exp(-t)$, $t \in [-1, 1]$

n	PS	BR	CL	SX	IP
$m = 500$					
2	14/2.3	6/1.7	3/3.0	498/217.0	10/350.2
3	21/4.1	6/2.1	8/7.6	749/292.2	12/418.3
4	22/5.1	10/3.2	8/7.9	1000/377.2	12/423.1
5	27/7.1	13/3.5	13/13.1	1250/459.1	14/486.9
$m = 1000$					
2	12/4.3	6/5.0	7/22.9	998/812.7	12/66.9
3	15/6.4	6/6.2	11/38.1	1499/1096.5	13/76.3
4	19/8.9	11/9.3	10/36.6	2000/1382.8	14/86.1
5	21/11.6	12/9.7	13/49.5	2500/1765.6	14/95.5

Table 2 Iterations/CPU times (cs) for $g(t) = \sin(t)$, $t \in [0, 1]$

m	PS	BR	CL	SX	IP
$n = 4$					
100	21/1.5	11/0.4	7/0.6	233/38.3	10/14.1
200	17/1.9	10/0.7	8/1.7	467/96.9	12/43.3
300	21/3.3	11/1.1	9/3.6	706/199.2	12/107.0
400	19/3.7	12/1.9	7/4.5	941/332.3	13/241.4
500	18/4.0	13/2.6	10/9.9	1175/511.1	12/413.6
1000	20/10.2	12/6.7	11/39.8	2355/1758.2	13/77.2
$n = 5$					
100	26/2.2	12/0.5	8/0.7	313/41.4	26/27.5
200	19/2.3	9/1.0	11/2.5	636/124.8	16/55.9
300	22/3.8	9/1.7	10/4.1	959/260.6	fail
400	25/5.2	15/2.3	12/7.6	1282/439.5	fail
500	29/7.7	16/3.4	10/10.3	1599/662.7	fail
1000	27/14.3	17/9.0	12/45.5	3200/2298.8	26/149.4

and Roberts' algorithm. The initial steplength is $\Delta_0 = 10^{-6}\|x^0\|_\infty$. The stopping criterion is $\Delta_k < \mathrm{Tol}(1 + \|x^k\|_\infty)$, with $\mathrm{Tol} = 10^{-12}$.

As it was to be expected, the ad hoc taylored methods BR and CL require the lowest number of iterations, while the general purpose methods reveal some difficulties: a very large number of iterations in the case of SX and some lack of robustness for IP. This is revealed by the fail entries of Table 2: in these cases, IP reaches the maximum number of iterations (85) without meeting the default stopping criteria. The behavior of PS can be placed in the middle: it is as much robust as best algorithms are, while requiring a reasonably greater—by a factor about 2 or 3—number of iterations.

However, looking at the CPU time, we see that in almost all the experiments our pattern search implementation compares favorably with all the other algorithms, but BR. We remark in particular that the CPU time for CL increases rapidly with m, due to the fact that each iteration requires the solution of a linear least squares problem of dimension $m \times n$. Finally, it is worth to note that the apparent incongruity in the results of `IP` depends on the way the linear algebra is handled in the code (see [16, pp. 2–75]).

6 Concluding Remarks

We have proposed a pattern search algorithm for the classical nonsmooth optimization problem (1). The results of the numerical experiments seem encouraging in two respects. First, they validate our idea that by exploiting the knowledge of the set of nondifferentiability it is possible to devise convergent pattern search methods for problem (1) with, all things considered, good practical behavior. Second, we look at problem (1) as at a particular instance of the piecewise differentiable problem and solve it without using linear programming. This approach has the advantage of being applicable to more general nonsmooth problems.

For example, it is straightforward to see that our method can be applied to the problem

$$\min_{x\in\mathbb{R}^n} f(x) := g^T x + \|A^T x - b\|_1, \tag{7}$$

where $g \in \mathbb{R}^n$ is given, as long as the level sets of f are bounded. Several problems, such as the linear feasibility problem (i.e., finding feasible points of linearly constrained programs) can be recast as (7). It will be interesting to study the possibility of extending the method to the corresponding box-constrained problem

$$\min_{x} f(x) := g^T x + \|A^T x - b\|_1 \quad \text{s.t.} \quad \|x\|_\infty \le \delta$$

which arises e.g. when solving constrained nonlinear programs via sequential linear programming [17] or systems of nonlinear equations via linear programming [18]. General linear constraints can be considered also, as well as the presence of nonlinear terms in the objective function. We anticipate more work along these directions.

Acknowledgements We thank the referees for careful reading of an earlier version of the paper and for many insightful comments and criticisms. Their questions and suggestions helped us to greatly improve the presentation.

References

1. Torczon, V.: On the convergence of the multidirectional search algorithm. SIAM J. Optim. **1**, 123–145 (1991)
2. Torczon, V.: On the convergence of pattern search algorithms. SIAM J. Optim. **7**, 1–25 (1997)
3. Lewis, R.M., Torczon, V.: Pattern search algorithms for bound constrained minimization. SIAM J. Optim. **9**, 1082–1099 (1999)
4. Lewis, R.M., Torczon, V.: Pattern search methods for linearly constrained minimization. SIAM J. Optim. **10**, 917–941 (2000)
5. Abramson, M.A., Audet, C., Dennis, J.E. Jr.: Generalized pattern searches with derivative information. Math. Program. **100B**, 3–25 (2004)
6. Audet, C.: Convergence results for generalized pattern search algorithms are tight. Optim. Eng. **5**, 101–122 (2004)

7. Audet, C., Dennis, J.E. Jr.: Analysis of generalized pattern searches. SIAM J. Optim. **13**, 889–903 (2003)
8. Audet, C., Dennis, J.E. Jr.: Mesh adaptive direct search algorithms for constrained optimization. Technical Report G-2004-04, Les Cahiers du GERAD, Montréal, Quebec, Canada (2004)
9. Kolda, T.G., Lewis, R.M., Torczon, V.: Optimization by direct search: new perspectives on some classical and modern methods. SIAM Rev. **45**, 385–482 (2003)
10. Lewis, R.M., Torczon, V.: Rank ordering and positive basis in pattern search algorithms. Technical Report TR-96-71, ICASE, NASA Langley Research Center (1996)
11. Barrodale, I., Roberts, F.D.K.: An improved algorithm for discrete L_1–approximation. SIAM J. Numer. Anal. **10**, 839–848 (1973)
12. Bartels, R.H., Conn, A.R., Sinclair, J.W.: Minimization techniques for piecewise differentiable functions: The L_1 solution to an overdetermined linear system. SIAM J. Numer. Anal. **15**, 224–241 (1978)
13. Coleman, T.F., Li, Y.: A globally and quadratically convergent affine scaling method for linear L_1 problems. Math. Program. **56**, 189–222 (1992)
14. Clarke, F.H.: Optimization and Nonsmooth Analysis. Wiley, New York (1983)
15. Abramson, M.A., Brezhneva, O.A., Dennis, J.E. Jr.: Pattern search methods in the presence of degeneracy. Technical Report TR03-09, Rice University, Department of Computational and Applied Mathematics (2005)
16. Optimization Toolbox User's guide, Version 3. The Mathworks, Natick (2004)
17. Byrd, R.H., Gould, N.I.M., Nocedal, J., Waltz, R.A.: An algorithm for nonlinear optimization using linear programming and equality constrained subproblems. Math. Program. **100B**, 27–48 (2004)
18. Duff, I.S., Nocedal, J., Reid, J.K.: The use of linear programming for the solution of sparse sets of nonlinear equations. SIAM J. Sci. Stat. Comput. **8**, 99–108 (1987)