MINIMIZING NONCONVEX NONSMOOTH FUNCTIONS VIA CUTTING PLANES AND PROXIMITY CONTROL
A. FUDULI∗, M. GAUDIOSO§, AND G. GIALLOMBARDO§

Abstract. We describe an extension of the classical cutting plane algorithm to tackle the unconstrained minimization of a nonconvex, not necessarily differentiable function of several variables. The method is based on the construction of both a lower and an upper polyhedral approximation to the objective function and it is related to the use of the concept of proximal trajectory. Convergence to a stationary point is proved for locally Lipschitz functions.

Key words. Nonsmooth optimization, cutting planes, bundle methods, proximal trajectory.

AMS subject classifications. 90C26, 65K05

1. Introduction. Most of the numerical methods for solving nonsmooth optimization problems aim at minimizing convex functions of several variables, and convex analysis is in fact the background theory [7, 20]. Although generalized gradient theory [2] and codifferentiable functions theory [3] provide an interesting framework for dealing with nonsmooth nonconvex functions, apparently they have not yet been fully exploited from the numerical point of view.

Most of the existing algorithms for nonsmooth optimization fall in the class of the subgradient and space dilatation type algorithms [22], or of the bundle methods [5, 7, 15], or of the minmax type algorithms [18, 19] (convexity is not necessary in the latter).

In particular, the bundle methods family is based on the cutting plane method, firstly described in [1, 8], where the convexity of the objective function is the fundamental assumption. In fact the extension of the cutting plane method to the nonconvex case is not straightforward. A basic observation is that in general first order information do not provide any longer a lower approximation to the objective function, independently on the nonsmoothness assumption.

Thus, the optimization of the cutting plane approximation does not necessarily give an optimistic estimate of the obtainable reduction in the objective function. Moreover such model might even fail to interpolate the objective function at the points where its value is known.

On the other hand it is apparent that a number of ideas valid in the convex nonsmooth framework are valuable also in the treatment of the nonconvex case.

∗Dipartimento di Ingegneria dell’Innovazione, Università di Lecce, Via Monteroni, 73100 Lecce (LE), Italia. E-mail: antonio.fuduli@unile.it
§Dipartimento di Elettronica Informatica e Sistemistica, Università della Calabria, 87036 Rende (CS), Italia. E-mail: {gaudioso,giallo}@deis.unical.it
For example, search directions obtained as the opposite of a convex combination of gradients, relative to points close to each other, appear often to enjoy good descent properties for nonconvex functions too, especially when the contour lines have a narrow valley shape.

Thus it appears reasonable to claim that nonconvex nonsmooth minimization can benefit from the experience of convex optimization, but the approaches valid in the latter case cannot be trivially extended.

Most of the authors who have extended bundle methods to the nonconvex case have considered piecewise affine models embedding possible downward shifting of the affine pieces [12, 16, 21]. However the amount of the shifting appears somehow arbitrary.

In this paper we present an iterative algorithm which is still based on first order approximations to the objective function.

The main difference with other known methods is to make a distinction between affine pieces that exhibit a kind of convex or of nonconvex behavior relative to the current point in the iterative procedure. Furthermore the use of downward shifting is restricted to some particular cases.

The following notations are adopted throughout the paper. We denote by $\| \cdot \|$ the euclidean norm in $\mathbb{R}^n$, by $a^Tb$ the inner product of the vectors $a$ and $b$, and by $e$ a vector of ones of appropriate dimension. The generalized gradient of a Lipschitz function $f : \mathbb{R}^n \mapsto \mathbb{R}$ at any point $x$ is denoted by $\partial f(x)$.

2. The model. Consider the following unconstrained minimization problem:

$$\min_{x \in \mathbb{R}^n} f(x)$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is not necessarily differentiable.

We assume that $f$ is locally Lipschitz, i.e. it is Lipschitz on every bounded set. Since $f$ is locally Lipschitz, then it is differentiable almost everywhere. It is well known [2] that, under the above hypotheses, it is defined at each point $x$ the generalized gradient (or Clarke’s gradient or subdifferential)

$$\partial f(x) = \text{conv}\{g \mid g \in \mathbb{R}^n, \nabla f(x_k) \rightarrow g, \ x_k \rightarrow x, \ x_k \notin \Omega f\}$$

where $\Omega f$ is the set (of zero measure) where $f$ is not differentiable. An extension of the generalized gradient is the Goldstein $\epsilon$-subdifferential $\partial^\epsilon_G f(x)$ defined as

$$\partial^\epsilon_G f(x) = \text{conv}\{\partial f(y) \mid \|y - x\| \leq \epsilon\}.$$

We assume also that we are able to calculate at each point $x$ both the objective function value and a subgradient $g \in \partial f(x)$, i.e. an element of the generalized gradient.

Now we describe the basic idea of our method, focusing on the differences with respect to the methods tailored on the convex case. We denote by $x_j$ the current
estimate of the minimum in an iterative procedure and by \( g_j \) any subgradient of \( f \) at \( x_j \). The bundle of available information is the set of elements

\[
(x_i, f(x_i), g_i, \alpha_i, a_i) \quad i \in I
\]

where \( x_i, i \in I \), are the points touched in the procedure, \( g_i \) is a subgradient of \( f \) at \( x_i \), \( \alpha_i \) is the linearization error between the actual value of the objective function at \( x_j \) and the linear expansion generated at \( x_i \) and evaluated at \( x_j \), i.e.

\[
\alpha_i \triangleq f(x_j) - f(x_i) - g_i^T(x_j - x_i)
\]

and

\[
a_i \triangleq \|x_j - x_i\|.
\]

We recall that the classical cutting plane method \([1, 8]\) minimizes at each iteration the cutting plane function \( f_j(x) \) defined as

\[
f_j(x) = \max_{i \in I} \{ f(x_i) + g_i^T(x - x_i) \}.
\]

The minimization of \( f_j(x) \) can be put in a linear programming form as

\[
\begin{align*}
\min_{\eta, x} & \quad \eta \\
\text{subject to} & \quad \eta \geq f(x_i) + g_i^T(x - x_i) \quad i \in I,
\end{align*}
\]

which is equivalent to solve

\[
\begin{align*}
\min_{v, d} & \quad v \\
\text{subject to} & \quad v \geq g_i^T d - \alpha_i \quad i \in I,
\end{align*}
\]

where \( d \) is the “displacement” from \( x_j \), i.e. \( d \triangleq x - x_j \). In the sequel we will refer to the point \( x_j \) as to the “stability center”.

It is worth noting that in the nonconvex case \( \alpha_i \) may be negative, since the first order expansion at any point does not necessarily support from below the epigraph of the function.

Thus we partition the set \( I \) in two sets \( I_+ \) and \( I_- \) defined as follows

\[
I_+ \triangleq \{ i | \alpha_i \geq 0 \} \quad I_- \triangleq \{ i | \alpha_i < 0 \}.
\]

The bundles defined by the index sets \( I_+ \) and \( I_- \) are characterized by points that somehow exhibit respectively a “convex behavior” and a “concave behavior” relatively to \( x_j \). We observe that \( I_+ \) is never empty as at least the element \((x_j, f(x_j), g_j, 0, 0)\) belongs to the bundle.
The basic idea of our approach is to treat differently the two bundles in the construction of a piecewise affine model.

We define the following piecewise affine functions:

\[ \Delta^+(d) \overset{\Delta}{=} \max_{i \in I_+} \{ g_i^T d - \alpha_i \} \]

and

\[ \Delta^-(d) \overset{\Delta}{=} \min_{i \in I_-} \{ g_i^T d - \alpha_i \} \].

In fact \( \Delta^+(d) \) is intended as an approximation to the difference function

\[ h(d) \overset{\Delta}{=} f(x_j + d) - f(x_j) \]

which interpolates it at \( d = 0 \) (since the index \( j \) belongs to \( I_+ \)).

On the other hand \( \Delta^-(d) \) is a locally “pessimistic” approximation to the difference function \( h(d) \). Since at \( d = 0 \) we have \( \Delta^+(0) < \Delta^-(0) \), it appears reasonable to consider significant the approximation \( \Delta^+(d) \) as far as

\[ \Delta^+(d) \leq \Delta^-(d) \].

In other words we introduce a kind of trust region model \( S \) defined as:

\[ S = \{ d | \Delta^+(d) \leq \Delta^-(d) \} \].

In addition we introduce proximity control [10] into our approach by defining the “proximal trajectory” [4] of \( \Delta^+(d) \) as the optimal solution \( d_\gamma \) to the following convex quadratic program, parameterized in the nonnegative scalar parameter \( \gamma \), where the constraints ensure that \( d \in S \):

\[
\begin{align*}
QP(\gamma) \quad & \quad z_\gamma = \min_{v,d} \quad \gamma v + \frac{1}{2} \|d\|^2 \\
& \quad v \geq g_i^T d - \alpha_i \quad i \in I_+ \\
& \quad v \leq g_i^T d - \alpha_i \quad i \in I_- .
\end{align*}
\]

We observe that \( z_\gamma \leq 0 \), as the couple \((v, d) = (0, 0)\) is feasible; we have consequently that the optimal value of \( v \) cannot be positive.

An interesting insight can be gathered by considering the dual of the program \( QP(\gamma) \). It can be written in the form:

\[
\begin{align*}
DP(\gamma) \quad & \quad w_\gamma = \min_{\lambda \geq 0, \mu \geq 0} \quad \frac{1}{2} \| G_+ \lambda - G_- \mu \|^2 + \alpha_+^T \lambda - \alpha_-^T \mu \\
& \quad \quad e^T \lambda - e^T \mu = \gamma
\end{align*}
\]
where $G_+$ and $G_-$ are matrices whose columns are, respectively, the vectors $g_i, i \in I_+$, and $g_i, i \in I_-$. Analogously, the terms $\alpha_i, i \in I_+$, and $\alpha_i, i \in I_-$, are grouped in the vectors $\alpha_+$ and $\alpha_-$, respectively.

The optimal primal solution $(v_\gamma, d_\gamma)$ is related to the optimal dual solution $(\lambda_\gamma, \mu_\gamma)$ by the following formulae:

\begin{align}
d_\gamma &= -G_+ \lambda_\gamma + G_- \mu_\gamma \tag{2.4a} \\
v_\gamma &= -\frac{1}{\gamma} \left( \|d_\gamma\|^2 + \alpha_+^T \lambda_\gamma - \alpha_-^T \mu_\gamma \right) . \tag{2.4b}
\end{align}

We remark that the proximal trajectory emanates from the stability center $x_j$.

Before giving a formal description of the algorithm, we state some simple properties of problem $QP(\gamma)$.

**Lemma 2.1.** Let $\gamma_1 > \gamma_2 > 0$. Then the following relations hold:

(i) $z_{\gamma_1} \leq z_{\gamma_2}$;
(ii) $v_{\gamma_1} \leq v_{\gamma_2}$;
(iii) $\|d_{\gamma_1}\| \geq \|d_{\gamma_2}\|$.

**Proof.** (i) From the definitions of $z_\gamma$, $v_\gamma$ and $d_\gamma$, and taking into account $\gamma_1 > \gamma_2 > 0$, it follows

$$z_{\gamma_1} = \gamma_1 v_{\gamma_1} + \frac{1}{2} \|d_{\gamma_1}\|^2 \leq \gamma_1 v_{\gamma_2} + \frac{1}{2} \|d_{\gamma_2}\|^2 \leq \gamma_2 v_{\gamma_2} + \frac{1}{2} \|d_{\gamma_2}\|^2 = z_{\gamma_2}.$$ 

(ii) Assume $v_{\gamma_1} > v_{\gamma_2}$. Then, since $\gamma_1 > \gamma_2$, it holds

$$0 < (\gamma_1 - \gamma_2)(v_{\gamma_1} - v_{\gamma_2}) = \gamma_1 v_{\gamma_1} + \gamma_2 v_{\gamma_2} - (\gamma_1 v_{\gamma_2} + \gamma_2 v_{\gamma_1}).$$

By adding and subtracting to the r.h.s.

$$\frac{1}{2} \|d_{\gamma_1}\|^2 + \frac{1}{2} \|d_{\gamma_2}\|^2$$

we would have:

$$0 < \left( \gamma_1 v_{\gamma_1} + \frac{1}{2} \|d_{\gamma_1}\|^2 \right) - \left( \gamma_1 v_{\gamma_2} + \frac{1}{2} \|d_{\gamma_2}\|^2 \right) +$$

$$+ \left( \gamma_2 v_{\gamma_2} + \frac{1}{2} \|d_{\gamma_2}\|^2 \right) - \left( \gamma_2 v_{\gamma_1} + \frac{1}{2} \|d_{\gamma_1}\|^2 \right)$$

which is a contradiction, since, by the definitions, the r.h.s. is sum of two non-positive quantities.

(iii) Assume $\|d_{\gamma_1}\| < \|d_{\gamma_2}\|$. Then ii) implies

$$\gamma_2 v_{\gamma_1} + \frac{1}{2} \|d_{\gamma_1}\|^2 < \gamma_2 v_{\gamma_2} + \frac{1}{2} \|d_{\gamma_2}\|^2 = z_{\gamma_2}.$$
which contradicts the optimality of \((v, d)\). \qed

**Lemma 2.2.** For any \(\gamma > 0\) the following relations hold:

(i) \[\|d\| \leq 2\gamma \|g\|;\]

(ii) \[z \geq -\gamma^2 \|g\|^2;\]

(iii) \[|v| \geq \frac{1}{2\gamma \|d\|^2}.\]

**Proof.** (i) Since \(z \leq 0\) we have \((v, d) \in \mathcal{D} \triangleq \{(v, d) \mid \gamma v + \frac{1}{2}\|d\|^2 \leq 0\}\).

The property follows by noting that the objective function of \(QP(\gamma)\) is minorized by

\[\gamma g^T d + \frac{1}{2}\|d\|^2.\]

(ii) The property follows by noting that \(-\frac{1}{2}\gamma^2 \|g\|^2\) is the minimum value of the minorizing function (2.5).

(iii) Consequence of \(z \leq 0\). \qed

**3. The algorithm.** In this section we describe an algorithm based on repeatedly solving problem \(QP(\gamma)\), or equivalently \(DP(\gamma)\). The core of the algorithm is the “main iteration”, i.e. the set of steps where the stability center remains unchanged.

Two exits from the “main iteration” may occur:

(i) termination of the whole algorithm, due to the satisfaction of an approximate stationarity condition;

(ii) update of the stability center, due to the satisfaction of a sufficient decrease condition.

The initialization of the algorithm requires a starting point \(x_0 \in \mathbb{R}^n\). The initial stability center \(y\) is set equal to \(x_0\). The initial bundle is made up by just one element \((y, f(y), g(y), 0, 0)\), where \(g(y) \in \partial f(y)\), so that \(I_-\) is the empty set, while \(I_+\) is a singleton. The following global parameters are to be set:

- the stationarity tolerance \(\delta > 0\) and the proximity measure \(\epsilon > 0\);
- the descent parameter \(m \in (0, 1)\) and the cut parameter \(\rho \in (m, 1)\);
- the reduction parameter \(r \in (0, 1)\), and the increase parameter \(R > 1\).

A short description of the algorithm is the following:

**Algorithm Outline**

1. Initialization.
2. Execute the “main iteration”.
3. Update the bundle of information with respect to the new stability center and return to 2.
In the sequel we describe in details the “main iteration” without indexing it for sake of notation simplicity. Of course the stability center $y$ is to be intended as the current stability center.

The following local parameters are set each time the “main iteration” is entered:

- the proximity measure $\theta > 0$;
- the safeguard parameters $\gamma_{\text{min}}$ and $\gamma_{\text{max}}$, $0 < \gamma_{\text{min}} < \gamma_{\text{max}}$.

We remark that in general the “main iteration” maintains the (updated) bundle of information from previous iterations. Updating of the bundle is necessary since the quantities $a_i$ and $a_i$ are dependent on the stability center.

Algorithm 3.1 (Main Iteration).

0. If $\|g(y)\| \leq \delta$ then STOP (stationarity achieved).

Set

$$\gamma_{\text{min}} := \frac{r\epsilon}{2\|g(y)\|}, \quad \gamma_{\text{max}} := R\gamma_{\text{min}}, \quad \theta := r\gamma_{\text{min}}\delta.$$  

1. Construct the proximal trajectory $d_{\gamma}$ for increasing values of $\gamma$ and choose $\hat{\gamma}$ equal to the minimum value of $\gamma \in (\gamma_{\text{min}}, \gamma_{\text{max}})$, such that:

$$f(y + d_{\gamma}) > f(y) + m\gamma$$

if such $\gamma$ does exist. Otherwise set $\hat{\gamma} := \gamma_{\text{max}}$. If $\|d_{\gamma}\| > \theta$ go to 3.

2. Set

$$I_{+} := I_{+} \setminus \{i \in I_{+} \mid a_i > \epsilon\}$$

and

$$I_{-} := I_{-} \setminus \{i \in I_{-} \mid a_i > \epsilon\}.$$  

Calculate

$$g^* = \min_{g \in \text{conv}\{g_i \mid i \in I_{+}\}} \|g\|.$$  

If $\|g^*\| \leq \delta$ then STOP (stationarity achieved).

Else set $\gamma_{\text{max}} := \gamma_{\text{max}} - r(\gamma_{\text{max}} - \gamma_{\text{min}})$ and go to 1.

3. Set $\hat{d} := d_{\hat{\gamma}}, \hat{v} := v_{\hat{\gamma}}, \hat{x} := y + \hat{d}$. Calculate $\hat{g} \in \partial f(\hat{x})$ and set

$$\hat{\alpha} := f(y) - f(\hat{x}) + \hat{g}^T \hat{d}.$$
4. (a) If $\hat{\alpha} < 0$ and $\|\hat{d}\| > \epsilon$ then insert the element $(\hat{x}, f(\hat{x}), \hat{y}, \hat{\alpha}, \|\hat{d}\|)$ into the bundle for an appropriate value of $i \in I_-$ and set $\hat{\gamma} := \hat{\gamma} - r(\hat{\gamma} - \gamma_{min})$.

(b) Else, if $\hat{g}^T \hat{d} \geq \rho \hat{v}$ then insert the element $(\hat{x}, f(\hat{x}), \hat{y}, \max(0, \hat{\alpha}), \|\hat{d}\|)$ into the bundle for an appropriate value of $i \in I_+$.

(c) Else find a scalar $t \in (0,1)$ such that $g(t) \in \partial f(y + td)$ satisfies the condition $g(t)^T \hat{d} \geq \rho \hat{v}$ and insert the element $(y + td, f(y + td), g(t), \max(0, \alpha_i), t\|d\|)$ into the bundle for an appropriate value of $i \in I_+$, where $\alpha_i = f(y) - f(y + td) + tg(t)^T \hat{d}$.

5. If $\|\hat{d}\| \leq \theta$ go to 2. If

\begin{equation}
  f(\hat{x}) \leq f(y) + \hat{m} \hat{v}
\end{equation}

set the new stability center $y := \hat{x}$ and EXIT from the main iteration.

6. Solve $QP(\hat{\gamma})$, or equivalently $DP(\hat{\gamma})$, obtain both the primal and the dual optimal solution $(\nu_{\hat{\gamma}}, d_{\hat{\gamma}})$ and $(\lambda_{\hat{\gamma}}, \mu_{\hat{\gamma}})$, and go to 3.

Some explanations are in order. The stationarity test at step 0 prevents the “main iteration” from being executed if enough information is already available to assess the stationarity of $y$.

The construction of the proximal trajectory at step 1 may be discretized by repeatedly solving $QP(\gamma)$ for increasing values of $\gamma$, or by adopting techniques of the type described in [4] (see also [11]).

The rationale of the test executed at step 2 is that the occurrence of a “small” (in norm) displacement $d_i$ corresponding to a “large” value of $\gamma$ denotes either that a stationary point has been reached or that the model is inconsistent. We discriminate between these two cases by considering the distance measures $a_i$ (bundle deletion at step 2). We observe that the choice of $\hat{\gamma}$ defines implicitly a constraint on the norm of $d_i$ (see Lemma 2.2-i). On the other hand $\|d_i\| \leq \theta$ is never consequence of the choice of a too small $\hat{\gamma}$. In fact we note that, if $\|g(y)\| > \delta$ it holds

$$\|d_{\gamma_{min}}\| \leq 2\gamma_{min} \|g(y)\| = \frac{2\|g(y)\|}{\rho \delta} g$$

with the right hand side strictly greater than $\theta$.

We remark that the insertion of a bundle index into $I_+$ or $I_-$ at step 4 is not simply based on the sign of $\alpha_i$. In fact, in case $\alpha_i < 0$ and $a_i \leq \epsilon$, the index $i$ is inserted into $I_+$ and not into $I_-$ as it would be expected, and $\alpha_i$ is set equal to zero, that is the related affine piece is shifted downward of a quantity equal to $|\alpha_i|$ (see also [21]). This is aimed at letting all the elements of the Goldstein $\epsilon$-subdifferential at $y$ to contribute to the construction of the polyhedral approximation $\Delta^+(d)$, and to guarantee that the model interpolates the objective function at $y$. Furthermore the
reduction of $\hat{\gamma}$, whenever a bundle index is inserted into $I_-$, is aimed at avoiding that the same point solution $\hat{x}$ is generated infinitely many times. To explain the case (c) at step 4 we need to observe that not always the downward shifting of an affine piece, when $\hat{\alpha} < 0$, cuts out the point solution of $QP(\hat{\gamma})$ generated at the previous iteration. A sufficient condition for such cut is $\hat{g}^T \hat{d} \geq \rho \hat{v}$. If such condition is not verified we resort to a line search type procedure which allows us to find a point $y + t\hat{d}$, with $t \in (0, 1)$, satisfying $g(t)^T \hat{d} \geq \rho \hat{v}$, where $g(t) \in \partial f(y + td)$ (see also [21]).

Finally we observe that every time the stability center is updated, the parameters $\alpha_i$ and $a_i$ are to be updated for each element of the bundle as well, which may result in changing the attribution of the corresponding index $i$ from $I_+$ to $I_-$ and viceversa.

4. The convergence. In this section we prove the termination of the algorithm at a point satisfying an approximate stationarity condition. In particular we prove that, for any given $\epsilon > 0$ and $\delta > 0$, it is possible to set the input parameters such that, after a finite number of “main iteration” executions, the algorithm stops at a point $y$ satisfying the condition $\|g^*\| \leq \delta$, with $g^* \in \partial f(y)$.

Throughout the section we make the following assumptions:

**A1** $f$ is locally Lipschitz;
**A2** $f$ is weakly semismooth (see [17, 21]);
**A3** the set $F_0 = \{ x \in \mathbb{R}^n \mid f(x) \leq f(x_0) \}$ is compact.

Before proving finite termination of the “main iteration” we introduce the following lemma.

**Lemma 4.1.** Let $\{(v^{(k)}_{\hat{\gamma}}, d^{(k)}_{\hat{\gamma}})\}_{k \in \mathcal{K}}$ be a subsequence generated within a single “main iteration” such that

$$\|d^{(k)}_{\hat{\gamma}}\| > \theta$$

and

$$f(y + d^{(k)}_{\hat{\gamma}}) - f(y) > m v^{(k)}_{\hat{\gamma}}$$

with the algorithm looping from step 3 to step 6. Then the following hold:

(a) There exists an index $\hat{k}$ such that for each $k \geq \hat{k}$, $k \in \mathcal{K}$, every new bundle index is inserted into $I_+$ and $\hat{\gamma}$ remains unchanged with $\hat{\gamma} < \frac{\epsilon}{2\|g(y)\|}$.

(b) If $\rho v^{(k)}_{\hat{\gamma}} > g^{(k)^T} d^{(k)}_{\hat{\gamma}}$ then there exist two nonnegative scalars $t_1$ and $t_2$, $0 \leq t_1 < t_2 < 1$, such that

$$g(t)^T d^{(k)}_{\hat{\gamma}} \geq \rho v^{(k)}_{\hat{\gamma}} \quad \forall t \in [t_1, t_2].$$
Whenever a new bundle index is inserted into $I_+$ the condition
\[ g^{(k)}^T d^{(k)}_\gamma \geq \rho v^{(k)}_\gamma \]
holds.

Proof. (a) We observe that every time a bundle index is inserted into $I_-$, $\gamma$ is reduced. Thus $k$ is the first index such that $\gamma$ reduces below the threshold $\frac{\epsilon}{2\|g(y)\|}$. In fact for all $k \geq \bar{k}$, $\|d^{(k)}_\gamma\| \leq \epsilon$ which implies both insertion into $I_+$ and no reduction of $\gamma$.

(b) First we observe that if $\rho v^{(k)}_\gamma > g^{(k)}^T d^{(k)}_\gamma$, taking into account that by hypothesis the sufficient decrease condition is not satisfied, from the mean value theorem [2] it follows that there exists a scalar $\bar{t} \in (0, 1)$ such that
\[ \rho v^{(k)}_\gamma < m v^{(k)}_\gamma < f(y + d^{(k)}_\gamma) - f(y) = g(y + \bar{t}d^{(k)}_\gamma)^T d^{(k)}_\gamma \]
where $g(y + \bar{t}d^{(k)}_\gamma) \in \partial f(y + \bar{t}d^{(k)}_\gamma)$. In addition, the existence of an entire interval $[t_1, t_2]$ such that the condition
\[ g(t)^T d^{(k)}_\gamma \geq \rho v^{(k)}_\gamma \quad \forall t \in [t_1, t_2] \]
is satisfied, follows from the weakly semismoothness assumption (see also [13, 17, 21]).

(c) We observe that the condition $g^{(k)}^T d^{(k)}_\gamma \geq \rho v^{(k)}_\gamma$ is either ensured by construction or by the fact that it holds
\[ g^{(k)}^T d^{(k)}_\gamma \geq g^{(k)}^T d^{(k)}_\gamma - \bar{\alpha} = f(y + d^{(k)}_\gamma) - f(y) > m v^{(k)}_\gamma > \rho v^{(k)}_\gamma \]
whenever $\bar{\alpha} \geq 0$. \hfill \Box

Now we can prove finite termination of the “main iteration”.

**Lemma 4.2.** The “main iteration” terminates after a finite number of steps provided that
\[ \gamma_{min} < \frac{\epsilon}{2\|g(y)\|}, \quad \theta < \gamma_{min} \delta, \quad \gamma_{min} < \gamma_{max}. \]

Proof. To prove finiteness of the “main iteration” it is necessary to demonstrate that in a finite number of steps either the stop at step 2 or the exit at step 5 is achieved.

We start by proving that the algorithm cannot pass infinitely many times through step 2. Assume by contradiction that such case occurs and let us index by $k \in \mathcal{K}$ all the quantities referred to the $k$-th passage. We have
\[ \|d^{(k)}_\gamma\| \leq \theta \]
Minimizing nonconvex nonsmooth functions via cutting planes and proximity control

and

\[ \|g^{(k)}\| > \delta. \]

Observe that \( \hat{\gamma} < \gamma_{\text{max}} \) and that by construction \( \gamma_{\text{max}} \) reduces in a finite number of steps below the threshold \( \frac{\epsilon}{\|g(y)\|} \). Thus it follows that asymptotically \( \|d^{(k)}_{\hat{\gamma}}\| \leq \epsilon \), which in turn implies that the indices of the new bundle elements are asymptotically inserted into \( I_+ \) and they are never removed.

Moreover the bundle insertion rules at step 4 allow to insert an index into \( I_- \) only if \( \|d^{(k)}_{\hat{\gamma}}\| > \epsilon \), and this implies that whenever a passage at step 2 occurs, all the elements with index \( i \in I_- \) are removed.

From the above considerations, taking into account (2.4a) and the constraint \( e^T \lambda - e^T \mu = \hat{\gamma} \) in the dual problem \( DP(\hat{\gamma}) \), we note that there exists an index \( k \) such that for all \( k \geq \bar{k} \) the direction \( d^{(k)}_{\hat{\gamma}} \) can be expressed in the form

\[ d^{(k)}_{\hat{\gamma}} = -\hat{\gamma}g^{(k)}, \]

with \( g^{(k)} \in \text{conv}\{g_i \mid i \in I^{(k)}_+\} \). But we are assuming \( \|d^{(k)}_{\hat{\gamma}}\| \leq \theta \) and \( \|g^{(k)}\| > \delta \), so we would have

\[ \theta \geq \|d^{(k)}_{\hat{\gamma}}\| = \hat{\gamma}\|g^{(k)}\| > \gamma_{\min}\|g^{*(k)}\| > \frac{\theta}{\delta} \delta = \theta \]

that is impossible.

So far we have proved that an infinite number of passages through step 2 cannot occur. To complete the proof of termination we need to show that it is impossible to have infinitely many times \( \|d_{\hat{\gamma}}\| > \theta \) and the descent condition (3.1) not satisfied, with the algorithm looping between steps 3 and 6.

Indexing again by \( k \in \mathcal{K} \) the \( k \)-th passage through such a loop, we observe that, from part (a) of lemma 4.1, there exists an index \( k \) such that for every \( k \geq \bar{k} \) the index of each new bundle element is put in \( I_+ \) with \( \hat{\gamma} \) remaining unchanged. Under such condition, for \( k \geq \bar{k} \) the sequence \( \{\gamma^{(k)}_{\hat{\gamma}}\} \) is monotonically nondecreasing, bounded and hence it is convergent. Moreover, since the sequence \( \{d^{(k)}_{\hat{\gamma}}\} \) is bounded in norm, it admits a convergent subsequence, say \( \{d^{(k)}_{\hat{\gamma}}\}_{k \in K' \subseteq K} \).

The above considerations imply that also the sequence \( \{v^{(k)}_{\hat{\gamma}}\}_{k \in K' \subseteq K} \) is convergent to a nonpositive limit, say \( \bar{v} \). Now assume that \( \bar{v} < 0 \), let \( s \) and \( t \) be two successive indices in \( K' \) and \( \beta^{(s)} = \max\{0, \alpha^{(s)}\} \). We have

\[ v^{(t)}_{\hat{\gamma}} \geq g^{(s)^T} d^{(t)}_{\hat{\gamma}} - \beta^{(s)}, \]

\[ f(y + d^{(s)}_{\hat{\gamma}}) - f(y) > m v^{(s)}_{\hat{\gamma}} \]
and

\[ g(s)^T d(s) \gamma \geq \rho v(s) \hat{\gamma} . \]

We note that

\[ g(s)^T d(s) \gamma - \beta(s) \geq \rho v(s) \hat{\gamma} . \]  

(4.2)

In fact two cases may occur. If \( \beta(s) = \alpha(s) \triangleq f(y) - f(y + d(s)) + g(s)^T d(s) \), taking into account that \( \rho > m \), it holds

\[ g(s)^T d(s) - \beta(s) = f(y + d(s)) - f(y) > m v(s) > \rho v(s) . \]

On the other hand if \( \beta(s) = 0 \), it holds

\[ g(s)^T d(s) - \beta(s) \geq \rho v(s) \hat{\gamma} . \]

Combining (4.1) and (4.2) we obtain

\[ v(t) \hat{\gamma} - \rho v(s) \hat{\gamma} \geq g(s)^T (d(t) - d(s)) \]

and passing to the limit

\[ (1 - \rho) \hat{v} \geq 0 \]

which contradicts \( \hat{v} < 0 \). Hence we conclude that \( \hat{v} = 0 \), but this in turn contradicts the fact that \( \|d(k)\| > \theta \) for all \( k \in K \).

Remark. Since \( \gamma_{\text{min}} = \frac{r \epsilon}{2\|g(y)\|} \) and \( \theta = r \gamma_{\text{min}} \delta \) it follows that

\[ \theta \geq \frac{r^2 \epsilon \delta}{2L_0} \]

(4.3)

where \( L_0 \) is the Lipschitz constant of \( f \) on the set \( F_0 \).

Now we are ready to prove the overall finiteness of the algorithm.

THEOREM 4.4. For any \( \epsilon > 0 \) and \( \delta > 0 \), the algorithm stops in a finite number of “main iterations” at a point satisfying the approximate stationarity condition

\[ \|g^*\| \leq \delta \quad \text{with} \quad g^* \in \partial^G f(y) . \]  

(4.4)
Proof. The approximate stationarity condition (4.4) is exactly the stopping condition tested at step 2 of the “main iteration”. Now suppose that it is not verified for an infinite number of “main iteration” executions. From Lemma 4.2 it follows that infinitely many times the descent condition is satisfied. Let $y^{(k)}$ be the stability center at $k$-th passage through “main iteration”; then $\|d^{(k)}_\gamma\| > \theta^{(k)}$,

$$f(y^{(k+1)}) \leq f(y^{(k)}) + mv^{(k)}_\gamma$$

and

$$f(y^{(k+1)}) - f(y^{(0)}) \leq m \sum_{i=0}^{k} v^{(i)}_\gamma .$$

Passing to the limit, taking into account that from (4.3) $\|d^{(i)}_\gamma\|$ is bounded away from zero, and then from Lemma 2.2 $v^{(i)}_\gamma$ is bounded away from zero, we would have

$$\lim_{k \to \infty} f(y^{(k+1)}) - f(y^{(0)}) \leq -\infty$$

which is impossible, since $f$ is bounded from below as consequence of assumptions A1 and A3. \qed

5. Practical implementation and numerical results. The algorithm described in section 3 cannot be immediately implemented as it may require unbounded storage. In fact it does not encompass any mechanism to control the growth of the bundle size. Also the convergence properties described in section 4 are derived under the hypothesis that the bundle size can grow indefinitely.

Thus, before passing to the implementation issues, it is necessary to consider explicitly that the bundle has finite size and to show that convergence is retained under such hypothesis. A possible way to tackle the problem is to introduce an aggregation technique scheme of the type devised by Kiwiel [9] and widely used in bundle methods [7]. In particular let $\hat{x}$ be the point generated at step 3 of the “main iteration”, obtained by solving $QP(\hat{\gamma})$ or $DP(\hat{\gamma})$. If we define the aggregate quantities

$$g_+ = \frac{G_+ \lambda_\gamma}{e^T \lambda_\gamma} , \quad \alpha_+ = \frac{\alpha^T \lambda_\gamma}{e^T \lambda_\gamma}$$

and, in case $\mu_\gamma \neq 0$

$$g_- = \frac{G_- \mu_\gamma}{e^T \mu_\gamma} , \quad \alpha_- = \frac{\alpha^T \mu_\gamma}{e^T \mu_\gamma}$$
it is easy to verify that the aggregate problem $QP^a(\hat{\gamma})$

$$
\begin{align*}
\min_{v,d} & \quad \hat{\gamma}v + \frac{1}{2}\|d\|^2 \\
& v \geq g^T_d - \alpha^+ \\
& v \geq g^T_i d - \alpha_i \quad i \in \tilde{I}_+ \\
& v \leq g^T_d - \alpha^- \\
& v \leq g^T_i d - \alpha_i \quad i \in \tilde{I}_-
\end{align*}
$$

has the same optimal solution $(v_\hat{\gamma}, d_\hat{\gamma})$ as $QP(\hat{\gamma})$, where $\tilde{I}_+$ and $\tilde{I}_-$ are arbitrary subsets of $I_+$ and $I_-$ respectively.

Of course, in case $\mu_\hat{\gamma} = 0$, the formulation of the aggregate problem does not contain the constraint $v \leq g^T d - \alpha^-$ and $(v_\hat{\gamma}, d_\hat{\gamma})$ is still optimal.

The introduction of the aggregation allows in fact to delete (part of) the bundle and consequently to keep the bundle size bounded. On the other hand the insertion of a new bundle element at step 4 can be implemented without any modification. The convergence of the algorithm is not affected by the aggregation mechanism. The key argument is in fact that the monotonicity of the sequence $\{z^{(k)}_\hat{\gamma}\}$, necessary in the proof of Lemma 4.2, is still guaranteed.

The algorithm, encompassing the aggregation scheme, has been implemented in double precision Fortan-77 under a Windows Me system. The code, called NCVX, has been tested on a set [14] of 25 problems available on the web at the URL http://www.cs.cas.cz/~luksan/test.html. All test problems, except problem 1, are nonsmooth.

The input parameters have been set as follows: $\epsilon = 0.1, \delta = 10^{-4}, m = 0.2, \rho = 0.5, r = 0.5, R = 10^3$. In table 5.1 we report the computational results in terms of number $N_f$ of function evaluations.

At each iteration we solve the dual program $DP(\gamma)$, by using the subroutine DQPROG provided by the IMSL library and based on M.J.D. Powell’s implementation of the Goldfarb and Idnani [6] dual quadratic programming algorithm.

In testing the algorithm, we have always adopted the same set of input parameters, with no tuning based on specific test problem, aiming at checking algorithm robustness more than efficiency.

Only for problem 23 (marked by “*” in table 5.1) we have set $m = 0.8$, as with standard $m = 0.2$ we have incurred in failure of the quadratic subprogram solver due to accumulation of rounding errors.
Minimizing nonconvex nonsmooth functions via cutting planes and proximity control

<table>
<thead>
<tr>
<th>#</th>
<th>Problem</th>
<th>n</th>
<th>$f^*$</th>
<th>$N_f$</th>
<th>$f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rosenbrock</td>
<td>2</td>
<td>0</td>
<td>70</td>
<td>5.009E-07</td>
</tr>
<tr>
<td>2</td>
<td>Crescent</td>
<td>2</td>
<td>0</td>
<td>22</td>
<td>8.022E-06</td>
</tr>
<tr>
<td>3</td>
<td>CB2</td>
<td>2</td>
<td>1.9522245</td>
<td>18</td>
<td>1.9522245</td>
</tr>
<tr>
<td>4</td>
<td>CB3</td>
<td>2</td>
<td>2</td>
<td>15</td>
<td>2.0000001</td>
</tr>
<tr>
<td>5</td>
<td>DEM</td>
<td>2</td>
<td>-3</td>
<td>21</td>
<td>-2.9999999</td>
</tr>
<tr>
<td>6</td>
<td>QL</td>
<td>2</td>
<td>7.2</td>
<td>28</td>
<td>7.2000005</td>
</tr>
<tr>
<td>7</td>
<td>LQ</td>
<td>2</td>
<td>-1.4142136</td>
<td>9</td>
<td>-1.4142135</td>
</tr>
<tr>
<td>8</td>
<td>Mifflin1</td>
<td>2</td>
<td>-1</td>
<td>127</td>
<td>-0.9999977</td>
</tr>
<tr>
<td>9</td>
<td>Mifflin2</td>
<td>2</td>
<td>-1</td>
<td>13</td>
<td>-1.0000000</td>
</tr>
<tr>
<td>10</td>
<td>Rosen-Suzuki</td>
<td>4</td>
<td>-44</td>
<td>29</td>
<td>-44.000000</td>
</tr>
<tr>
<td>11</td>
<td>Shor</td>
<td>5</td>
<td>22.600162</td>
<td>44</td>
<td>22.600162</td>
</tr>
<tr>
<td>12</td>
<td>Maxquad</td>
<td>10</td>
<td>-0.8414083</td>
<td>56</td>
<td>-0.8414078</td>
</tr>
<tr>
<td>13</td>
<td>Maxq</td>
<td>20</td>
<td>0</td>
<td>293</td>
<td>1.660E-07</td>
</tr>
<tr>
<td>14</td>
<td>Maxl</td>
<td>20</td>
<td>0</td>
<td>44</td>
<td>1.110E-15</td>
</tr>
<tr>
<td>15</td>
<td>Goffin</td>
<td>50</td>
<td>0</td>
<td>148</td>
<td>1.142E-13</td>
</tr>
<tr>
<td>16</td>
<td>El-Attar</td>
<td>6</td>
<td>0.5598131</td>
<td>152</td>
<td>0.5598163</td>
</tr>
<tr>
<td>17</td>
<td>Wolfe</td>
<td>2</td>
<td>-8</td>
<td>21</td>
<td>-7.9999998</td>
</tr>
<tr>
<td>18</td>
<td>MXHILB</td>
<td>50</td>
<td>0</td>
<td>33</td>
<td>1.768E-05</td>
</tr>
<tr>
<td>19</td>
<td>L1HILB</td>
<td>50</td>
<td>0</td>
<td>104</td>
<td>6.978E-07</td>
</tr>
<tr>
<td>20</td>
<td>Colville1</td>
<td>5</td>
<td>-32.348679</td>
<td>47</td>
<td>-32.348679</td>
</tr>
<tr>
<td>21</td>
<td>Gill</td>
<td>10</td>
<td>9.7857721</td>
<td>164</td>
<td>9.7857746</td>
</tr>
<tr>
<td>22</td>
<td>HS78</td>
<td>5</td>
<td>-2.9197004</td>
<td>159</td>
<td>-2.9196589</td>
</tr>
<tr>
<td>23*</td>
<td>TR48</td>
<td>48</td>
<td>-638565</td>
<td>353</td>
<td>-638565.00</td>
</tr>
<tr>
<td>24</td>
<td>Shell Dual</td>
<td>15</td>
<td>32.348679</td>
<td>1497</td>
<td>32.349404</td>
</tr>
<tr>
<td>25</td>
<td>Steiner2</td>
<td>12</td>
<td>16.703838</td>
<td>196</td>
<td>16.703838</td>
</tr>
</tbody>
</table>

Table 5.1
NCVX: computational results

REFERENCES


