SIAM J. Optim.
Vol. 14, No. 3, pp. 743–756

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 is related to the use of the concept of proximal trajectory.

Convergence to a stationary point is proved for weakly semismooth functions.

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

AMS subject classifications. 90C26, 65K05

DOI. 10.1137/S1052623402411459

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 [9, 22]. Although generalized gradient theory [2] and codifferentiable functions theory [4] 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 into the class of subgradient and space dilatation–type algorithms [24], bundle methods [7, 10, 18], or minmax-type algorithms [5, 21] (convexity is not necessary in the latter).

In particular, the bundle methods family is based on the cutting plane method, first described in [1, 11], 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 no longer provides a lower approximation to the objective function independently of 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 a 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.

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 de-

[*] Received by the editors July 17, 2002; accepted for publication (in revised form) July 30, 2003; published electronically January 30, 2004. This work was partially supported by the Italian Ministero dell'Istruzione, dell'Università e della Ricerca Scientifica, under FIRB project *Large Scale Nonlinear Optimization* (RBNE01WBBB).

http://www.siam.org/journals/siopt/14-3/41145.html

[†] Dipartimento di Ingegneria dell'Innovazione, Università di Lecce, Via Monteroni, 73100 Lecce (LE), Italia (antonio.fuduli@unile.it).

[‡] Dipartimento di Elettronica Informatica e Sistemistica, Università della Calabria, 87036 Rende (CS), Italia (gaudioso@deis.unical.it).

[§] Judge Institute of Management, University of Cambridge, Cambridge CB2 1AG, UK, and Dipartimento di Elettronica Informatica e Sistemistica, Università della Calabria, 87036 Rende (CS), Italia (g.giallombardo@jims.cam.ac.uk, giallo@deis.unical.it).

scent 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 [15, 19, 23]. 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 that our algorithm makes a distinction between affine pieces that exhibit a *convex* or a *concave* behavior relative to the current point in the iterative procedure. Furthermore, the use of downward shifting is restricted to some particular cases.

The following notation is 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, there is defined at each point x the generalized gradient (or Clarke's gradient or subdifferential)

$$\partial f(x) = \mathrm{conv}\{g \;|g \in \mathbb{R}^n, \nabla f(x_k) \to g,\; x_k \to x,\; x_k \notin \Omega_f\},$$

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

$$\partial^G_\epsilon f(x) = \mathrm{conv}\{\partial f(y) \;|\|y-x\| \le \epsilon\}.$$

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, α_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.,

$$\alpha_i \overset{\triangle}{=} f(x_j) - f(x_i) - g_i^T(x_j - x_i),$$

and

$$a_i \stackrel{\triangle}{=} \|x_j - x_i\| \ .$$

We recall that the classical cutting plane method [1, 11] minimizes at each iteration the cutting plane function $f_j(x)$ defined as

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

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

$$\left\{ \begin{array}{ll} \min\limits_{\eta, x} & \eta \\ & \eta \geq f(x_i) + g_i^T (x - x_i), \quad i \in I, \end{array} \right. \tag{2.1}$$

which is equivalent to solving

$$\left\{ \begin{array}{ll} \min\limits_{v, d} & v \\ & v \geq g_i^T d - \alpha_i, \quad i \in I, \end{array} \right. \tag{2.2}$$

where d is the "displacement" from x_j, i.e., $d \stackrel{\triangle}{=} x - x_j$. In what follows we will refer to the point x_j as the "stability center."

It is worth noting that in the nonconvex case α_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 into two sets I_+ and I_-, defined as follows:

$$I_+ \stackrel{\triangle}{=} \{i | \alpha_i \geq 0\}, \qquad I_- \stackrel{\triangle}{=} \{i | \alpha_i < 0\}. \tag{2.3}$$

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" relative 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:

$$\Delta^+(d) \stackrel{\triangle}{=} \max_{i \in I_+} \left\{ g_i^T d - \alpha_i \right\}$$

and

$$\Delta^-(d) \stackrel{\triangle}{=} \min_{i \in I_-} \left\{ g_i^T d - \alpha_i \right\}.$$

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

$$h(d) \stackrel{\triangle}{=} f(x_j + d) - f(x_j),$$

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 $h(d)$. When $I_- \neq \emptyset$, since we have $\Delta^+(0) < \Delta^-(0)$, it appears reasonable to consider the approximation $\Delta^+(d)$ significant as far as

$$\Delta^+(d) \leq \Delta^-(d).$$

In other words we introduce a kind of trust region model $\mathcal{S}$ defined as

$$\mathcal{S} = \{d | \Delta^+(d) \le \Delta^-(d)\}.$$

In addition we introduce proximity control [13] into our approach by defining the "proximal trajectory" [6] of $\Delta^+(d)$ as the optimal solution d_γ to the following convex quadratic program, parameterized in the nonnegative scalar γ, where the constraints ensure that $d \in \mathcal{S}$:

$$QP(\gamma) \qquad \begin{cases} z_\gamma = \displaystyle\min_{v,d} \quad \gamma v + \frac{1}{2}\|d\|^2 \\ \qquad\qquad v \ge g_i^T d - \alpha_i, \quad i \in I_+, \\ \qquad\qquad v \le g_i^T d - \alpha_i, \quad i \in I_-. \end{cases}$$

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

The dual of the program $QP(\gamma)$ can be written in the form

$$DP(\gamma) \qquad \begin{cases} w_\gamma = \displaystyle\min_{\lambda \ge 0, \mu \ge 0} \quad \frac{1}{2}\|G_+\lambda - G_-\mu\|^2 + \alpha_+^T\lambda - \alpha_-^T\mu \\ \qquad\qquad e^T\lambda - e^T\mu = \gamma, \end{cases}$$

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 α_i, $i \in I_+$, and α_i, $i \in I_-$, are grouped in the vectors α_+ and α_-, respectively.

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

$$d_\gamma = -G_+\lambda_\gamma + G_-\mu_\gamma, \tag{2.4a}$$

$$v_\gamma = -\frac{1}{\gamma}\left(\|d_\gamma\|^2 + \alpha_+^T\lambda_\gamma - \alpha_-^T\mu_\gamma\right). \tag{2.4b}$$

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) $z_{\gamma_1} \le z_{\gamma_2}$*;*
(ii) $v_{\gamma_1} \le v_{\gamma_2}$*;*
(iii) $\|d_{\gamma_1}\| \ge \|d_{\gamma_2}\|$.

Proof. (i) From the definitions of z_γ, v_γ, and d_γ, and taking into account $\gamma_1 > \gamma_2 > 0$, it follows that

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

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

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

By adding and subtracting to the right-hand side

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

we would have

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

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

which is a contradiction, since, by the definitions, the right-hand side is the sum of two nonpositive quantities.

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

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

which contradicts the optimality of $(v_{\gamma_2}, d_{\gamma_2})$. □

LEMMA 2.2. *For any $\gamma > 0$ the following relations hold:*

(i) $\|d_\gamma\| \le 2\gamma\|g_j\|$;

(ii) $z_\gamma \ge -\frac{1}{2}\gamma^2\|g_j\|^2$;

(iii) $|v_\gamma| \ge \frac{1}{2\gamma}\|d_\gamma\|^2$.

Proof. (i) Since $z_\gamma \le 0$ we have

$$(v_\gamma, d_\gamma) \in \mathcal{D} \overset{\triangle}{=} \left\{(v,d) \mid \gamma v + \frac{1}{2}\|d\|^2 \le 0\right\}.$$

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

$$\gamma g_j^T d + \frac{1}{2}\|d\|^2. \tag{2.5}$$

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

(iii) The property follows as a consequence of $z_\gamma \le 0$. □

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 of 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 what follows we describe in detail the "main iteration" without indexing it for the sake of notational simplicity.

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

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

We remark that in general the "main iteration" maintains the (updated) bundle of information from previous iterations. Updating the bundle is necessary since the quantities α_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

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

1. *Construct the proximal trajectory d_γ for increasing values of γ and choose $\hat{\gamma}$ equal to the minimum value of $\gamma \in [\gamma_{min}, \gamma_{max}]$ such that*

$$f(y + d_\gamma) > f(y) + mv_\gamma$$

if such γ does exist. Otherwise set $\hat{\gamma} := \gamma_{max}$. If $\|d_{\hat{\gamma}}\| > \theta$, 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

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

If $\|g^*\| \leq \delta$, then STOP *(stationarity achieved).*
Else set $\gamma_{max} := \gamma_{max} - r(\gamma_{max} - \gamma_{min})$ and go to 1.

3. *Set $x_{\hat{\gamma}} := y + d_{\hat{\gamma}}$, calculate $g_{\hat{\gamma}} \in \partial f(x_{\hat{\gamma}})$, and set*

$$\alpha_{\hat{\gamma}} := f(y) - f(x_{\hat{\gamma}}) + g_{\hat{\gamma}}^T d_{\hat{\gamma}}.$$

4. (a) *If $\alpha_{\hat{\gamma}} < 0$ and $\|d_{\hat{\gamma}}\| > \epsilon$, then insert the element $(x_{\hat{\gamma}}, f(x_{\hat{\gamma}}), g_{\hat{\gamma}}, \alpha_{\hat{\gamma}}, \|d_{\hat{\gamma}}\|)$ 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 $g_{\hat{\gamma}}^T d_{\hat{\gamma}} \geq \rho v_{\hat{\gamma}}$, then insert the element $(x_{\hat{\gamma}}, f(x_{\hat{\gamma}}), g_{\hat{\gamma}}, \max(0, \alpha_{\hat{\gamma}}), \|d_{\hat{\gamma}}\|)$ into the bundle for an appropriate value of $i \in I_+$.*

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

5. *If* $\|d_{\hat\gamma}\| \le \theta$, *go to* 2. *If*

$$f(x_{\hat\gamma}) \le f(y) + mv_{\hat\gamma}, \tag{3.1}$$

set the new stability center $y := x_{\hat\gamma}$ *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* $(v_{\hat\gamma}, d_{\hat\gamma})$ *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 γ, or by adopting techniques of the type described in [6] (see also [14]).

The rationale of the test executed at step 2 is that the occurrence of a "small" (in norm) displacement d_γ corresponding to a "large" value of γ 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 $d_{\hat\gamma}$ (see Lemma 2.2(i)). On the other hand $\|d_{\hat\gamma}\| \le \theta$ is never a consequence of the choice of a too small $\hat\gamma$. In fact we note that if $\|g(y)\| > \delta$, it holds that

$$\|d_{\gamma_{min}}\| \le 2\gamma_{min}\|g(y)\| = \frac{2\|g(y)\|}{r\delta}\theta,$$

with the right-hand side strictly greater than θ.

We remark that the insertion of a bundle index into I_+ or I_- at step 4 is not simply based on the sign of α_i. In fact, in case $\alpha_i < 0$ and $a_i \le \epsilon$, the index i is inserted into I_+, and not into I_- as would be expected, and α_i is set equal to zero; that is, the related affine piece is shifted downward of a quantity equal to $|\alpha_i|$ (see also [23]). This is aimed at letting all elements of the Goldstein ϵ-subdifferential at y contribute to the construction of the polyhedral approximation $\Delta^+(d)$, and also guarantees 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 the same point solution $x_{\hat\gamma}$ being generated infinitely many times. To explain case (c) at step 4 we observe that the downward shifting of an affine piece, when $\hat\alpha < 0$, does not always cut out the point solution of $QP(\hat\gamma)$ generated at the previous iteration. A sufficient condition for such a cut to be effective is that $g_{\hat\gamma}^T d_{\hat\gamma} \ge \rho v_{\hat\gamma}$. If such a condition is not verified, we resort to a line search–type procedure which allows us to find a point $y + td_{\hat\gamma}$, with $t \in (0,1)$, satisfying $g(t)^T d_{\hat\gamma} \ge \rho v_{\hat\gamma}$, where $g(t) \in \partial f(y + td_{\hat\gamma})$ (see also [23]).

Notice that the search direction $d_{\hat\gamma}$ is calculated only at steps 1 and 6. This means that in passing through step 4(a), the reduction of $\hat\gamma$ does not cause an immediate change in $d_{\hat\gamma}$, and indeed the search direction used at step 5 is the one available right before such a reduction.

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

4. 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

$$\|g^*\| \leq \delta \quad \text{with } g^* \in \partial_\epsilon^G f(y).$$

Throughout the section we make the following assumptions:

(A1) f is weakly semismooth;

(A2) the set $\mathcal{F}_0 = \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\}$ is compact.

We recall that a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is weakly semismooth at x (see [16, 20, 23]) if it is Lipschitz around x and

$$\lim_{t \downarrow 0} g(t)^T d$$

exists for all $d \in \mathbb{R}^n$, where $g(t) \in \partial f(x+td)$. In particular, if f is weakly semismooth at x, the directional derivative $f'(x,d)$ of f along the direction d exists for all $d \in \mathbb{R}^n$ and

$$f'(x,d) = \lim_{t \downarrow 0} g(t)^T d.$$

Moreover, f is weakly semismooth on $\mathbb{R}^n$ if it is weakly semismooth at each $x \in \mathbb{R}^n$.

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

LEMMA 4.1. *Let $\{(v_{\hat{\gamma}}^{(k)}, d_{\hat{\gamma}}^{(k)})\}_{k \in \mathcal{K}}$ be a subsequence generated within a single "main iteration" such that*

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

and

$$f(y + d_{\hat{\gamma}}^{(k)}) - f(y) > m v_{\hat{\gamma}}^{(k)},$$

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

(i) *there exists an index $\bar{k}$ such that for each $k \geq \bar{k}$, $k \in \mathcal{K}$, every new bundle index is inserted into I_+ and $\hat{\gamma}$ remains unchanged;*

(ii) *step 4(c) of the algorithm is well posed; i.e., there exist two nonnegative scalars $t_1^{(k)}$ and $t_2^{(k)}$, $0 \leq t_1^{(k)} < t_2^{(k)} < 1$, such that for any $t \in [t_1^{(k)}, t_2^{(k)}]$ the condition*

$$g(t)^T d_{\hat{\gamma}}^{(k)} \geq \rho v_{\hat{\gamma}}^{(k)}$$

is satisfied for every $g(t) \in \partial f(y + t d_{\hat{\gamma}}^{(k)})$;

(iii) *whenever a new bundle index is inserted into I_+ the condition*

$$g_k^T d_{\hat{\gamma}}^{(k)} \geq \rho v_{\hat{\gamma}}^{(k)}$$

holds, where g_k is the subgradient corresponding to the new bundle element.

Proof. (i) We observe that an infinite sequence of bundle index insertions into I_- cannot take place, as a consequence of the reduction of $\hat{\gamma}$ at step 4(a) of the algorithm. In particular, no bundle index can be inserted into I_- as soon as $\hat{\gamma}$ falls below the threshold $\frac{\epsilon}{2\|g(y)\|}$.

(ii) Since the directional derivative $f'(y + t^{(k)} d_{\hat{\gamma}}^{(k)}, d_{\hat{\gamma}}^{(k)})$ exists for any $t^{(k)} \geq 0$, from the mean value theorem (see [3], Chap. 3, Prop. 3.1) it follows that

$$f(y + d_{\hat{\gamma}}^{(k)}) - f(y) = c \tag{4.1}$$

for some $c \in [f'_{\rm inf}, f'_{\rm sup}]$, where

$$f'_{\rm inf} \triangleq \inf_{0 \le t^{(k)} \le 1} f'(y + t^{(k)} d_{\hat\gamma}^{(k)}, d_{\hat\gamma}^{(k)}) \quad \text{and} \quad f'_{\rm sup} \triangleq \sup_{0 \le t^{(k)} \le 1} f'(y + t^{(k)} d_{\hat\gamma}^{(k)}, d_{\hat\gamma}^{(k)}).$$

Moreover, taking into account that the sufficient decrease condition is not satisfied, i.e.,

$$\rho v_{\hat\gamma}^{(k)} < m v_{\hat\gamma}^{(k)} < f(y + d_{\hat\gamma}^{(k)}) - f(y),$$

by (4.1) and the definition of $f'_{\rm sup}$ there exists a scalar $\bar t^{(k)} \in (0,1)$ such that

$$\rho v_{\hat\gamma}^{(k)} < f'(y + \bar t^{(k)} d_{\hat\gamma}^{(k)}, d_{\hat\gamma}^{(k)}).$$

Thus the thesis follows as a consequence of the weakly semismoothness assumption.

(iii) We observe that the condition $g_k^T d_{\hat\gamma}^{(k)} \ge \rho v_{\hat\gamma}^{(k)}$ is ensured either by construction or by the fact that

$$g_k^T d_{\hat\gamma}^{(k)} \ge g_k^T d_{\hat\gamma}^{(k)} - \alpha_{\hat\gamma}^{(k)} = f(y + d_{\hat\gamma}^{(k)}) - f(y) > m v_{\hat\gamma}^{(k)} > \rho v_{\hat\gamma}^{(k)}$$

whenever $\alpha_{\hat\gamma}^{(k)} \ge 0$. □

Now we can prove finite termination of the "main iteration."

LEMMA 4.2. *The* "main iteration" *terminates after a finite number of steps.*

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 a case occurs, and let us index by $k \in \mathcal{K}$ all the quantities referred to in the kth passage. We have

$$\|d_{\hat\gamma}^{(k)}\| \le \theta$$

and

$$\|{g^*}^{(k)}\| > \delta.$$

Observe that $\hat\gamma \le \gamma_{max}$ and that by construction γ_{max} falls in a finite number of steps below the threshold $\frac{\epsilon}{2\|g(y)\|}$. Thus, from Lemma 2.2(i), it follows that asymptotically $\|d_{\hat\gamma}^{(k)}\| \le \epsilon$, which in turn implies that the indices of the new bundle elements are asymptotically inserted into I_+ and are never removed.

Moreover, the bundle insertion rules at step 4 allow us to insert an index into I_- only if $\|d_{\hat\gamma}\| > \epsilon$, 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 $DP(\hat\gamma)$, it follows that there exists an index $\bar k \in \mathcal{K}$ such that for all $k \ge \bar k$ the direction $d_{\hat\gamma}^{(k)}$ can be expressed in the form

$$d_{\hat\gamma}^{(k)} = -\hat\gamma g^{(k)},$$

with $g^{(k)} \in \text{conv}\{g_i \mid i \in I_+^{(k)}\}$. But since $\|d_{\hat\gamma}^{(k)}\| \le \theta$ and $\|{g^*}^{(k)}\| > \delta$, we have

$$\theta \ge \|d_{\hat\gamma}^{(k)}\| = \hat\gamma \|g^{(k)}\| \ge \gamma_{min} \|{g^*}^{(k)}\| > \frac{\theta}{\delta}\delta = \theta,$$

reaching a contradiction.

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 kth passage through such a loop, we observe that, by Lemma 4.1(i), there exists an index $\bar{k}$ 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 a condition, for $k \geq \bar{k}$ the sequence $\{z_{\hat{\gamma}}^{(k)}\}$ is monotonically nondecreasing, bounded, and hence convergent. Moreover, since the sequence $\{d_{\hat{\gamma}}^{(k)}\}$ is bounded in norm, it admits a convergent subsequence, say $\{d_{\hat{\gamma}}^{(k)}\}_{k \in \mathcal{K}' \subseteq \mathcal{K}}$.

The above considerations imply also that the sequence $\{v_{\hat{\gamma}}^{(k)}\}_{k \in \mathcal{K}' \subseteq \mathcal{K}}$ is convergent to a nonpositive limit, say $\bar{v}$. Now assume that $\bar{v} < 0$, let i and j be two successive indices in $\mathcal{K}'$, and let $\beta_i = \max\{0, \alpha_i\}$, with $\alpha_i = f(y) - f(y + d_{\hat{\gamma}}^{(i)}) + g_i^T d_{\hat{\gamma}}^{(i)}$ and $g_i \in \partial f(y + d_{\hat{\gamma}}^{(i)})$. We have

$$v_{\hat{\gamma}}^{(j)} \geq g_i^T d_{\hat{\gamma}}^{(j)} - \beta_i, \tag{4.2}$$

$$f(y + d_{\hat{\gamma}}^{(i)}) - f(y) > m v_{\hat{\gamma}}^{(i)},$$

and

$$g_i^T d_{\hat{\gamma}}^{(i)} \geq \rho v_{\hat{\gamma}}^{(i)}.$$

We note that

$$g_i^T d_{\hat{\gamma}}^{(i)} - \beta_i \geq \rho v_{\hat{\gamma}}^{(i)}. \tag{4.3}$$

This inequality is trivial for $\beta_i = 0$. If, on the other hand, $\beta_i = \alpha_i$, then taking into account that $\rho > m$, it holds that

$$g_i^T d_{\hat{\gamma}}^{(i)} - \beta_i = f(y + d_{\hat{\gamma}}^{(i)}) - f(y) > m v_{\hat{\gamma}}^{(i)} > \rho v_{\hat{\gamma}}^{(i)}.$$

Combining (4.2) and (4.3) we obtain

$$v_{\hat{\gamma}}^{(j)} - \rho v_{\hat{\gamma}}^{(i)} \geq g_i^T (d_{\hat{\gamma}}^{(j)} - d_{\hat{\gamma}}^{(i)}),$$

and passing to the limit

$$(1 - \rho)\bar{v} \geq 0,$$

which contradicts $\bar{v} < 0$. Hence we conclude that $\bar{v} = 0$, which, by Lemma 2.2(iii), contradicts the fact that $\|d_{\hat{\gamma}}^{(k)}\| > \theta$ for all $k \in \mathcal{K}$. □

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

$$\theta \geq \frac{r^2 \epsilon \delta}{2L_0}, \tag{4.4}$$

where L_0 is the Lipschitz constant of f on the set $\mathcal{F}_0$.

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

THEOREM 4.3. *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*

$$\|g^*\| \leq \delta \qquad \text{with } g^* \in \partial_\epsilon^G f(y). \tag{4.5}$$

Proof. The approximate stationarity condition (4.5) 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 the kth passage through "main iteration"; then $\|d_{\hat{\gamma}}^{(k)}\| > \theta^{(k)}$,

$$f(y^{(k+1)}) \leq f(y^{(k)}) + mv_{\hat{\gamma}}^{(k)},$$

and

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

Now consider that by (4.4) $\|d_{\hat{\gamma}}^{(i)}\|$ is bounded away from zero. Then from Lemma 2.2(iii) it follows that $v_{\hat{\gamma}}^{(i)}$ is bounded away from zero as well. Therefore, by passing to the limit we obtain

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

which is a contradiction, since f is bounded from below as a consequence of assumptions (A1) and (A2). □

5. Practical implementation and numerical results. The algorithm described in section 3 cannot be immediately implemented, since 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 take into account explicitly that the bundle has finite size and to show that convergence is retained under such a hypothesis. A possible way to tackle the problem is to introduce an aggregation technique scheme of the type devised by Kiwiel [12] and widely used in bundle methods [10]. 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

$$g_+ \triangleq \frac{G_+\lambda_{\hat{\gamma}}}{e^T\lambda_{\hat{\gamma}}}, \qquad \alpha^+ \triangleq \frac{\alpha_+^T\lambda_{\hat{\gamma}}}{e^T\lambda_{\hat{\gamma}}}$$

and, in case $\mu_{\hat{\gamma}} \neq 0$,

$$g_- \triangleq \frac{G_-\mu_{\hat{\gamma}}}{e^T\mu_{\hat{\gamma}}}, \qquad \alpha^- \triangleq \frac{\alpha_-^T\mu_{\hat{\gamma}}}{e^T\mu_{\hat{\gamma}}},$$

it is easy to verify that the aggregate problem

$$QP^a(\hat{\gamma}) \qquad \begin{cases} \displaystyle\min_{v,d} \quad \hat{\gamma} v + \frac{1}{2}\|d\|^2 \\ \qquad v \geq g_+^T d - \alpha^+, \\ \qquad v \geq g_i^T d - \alpha_i, \quad i \in \bar{I}_+, \\ \qquad v \leq g_-^T d - \alpha^-, \\ \qquad v \leq g_i^T d - \alpha_i, \quad i \in \bar{I}_-, \end{cases}$$

has the same optimal solution $(v_{\hat{\gamma}}, d_{\hat{\gamma}})$ as $QP(\hat{\gamma})$, where $\bar{I}_+$ and $\bar{I}_-$ are arbitrary subsets of I_+ and I_-, respectively. Of course, in case $I_- = \emptyset$ or $\mu_{\hat{\gamma}} = 0$, 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.

On the basis of the above observations it is possible to embed an aggregation scheme into the algorithm. Suppose that at a certain execution of the "main iteration," the quadratic program $QP(\hat{\gamma})$ (or $DP(\hat{\gamma})$) is solved, and the corresponding optimal dual vector $(\lambda_{\hat{\gamma}}, \mu_{\hat{\gamma}})$ is calculated. Then, once the quantities g_+, α^+, g_-, α^- have been calculated as well, it is possible to construct the aggregate problem $QP^a(\hat{\gamma})$ by inserting the aggregated constraints into $QP(\hat{\gamma})$ and deleting part of its bundle elements. Thus, next time the quadratic program must be solved, it can be obtained by inserting the new constraint, corresponding to the new bundle element calculated at step 3 of the "main iteration," into the aggregated problem $QP^a(\hat{\gamma})$. Of course, such an aggregation task will only be carried out each time a given maximal bundle dimension is reached.

The convergence of the algorithm is not affected by the aggregation mechanism. Indeed the key argument is that the monotonicity of the sequence $\{z_{\hat{\gamma}}^{(k)}\}$, necessary in the proof of Lemma 4.2, is still guaranteed.

The algorithm, encompassing the aggregation scheme, has been implemented in double precision Fortran-77 under a Windows ME system. The code, called NCVX, has been tested on a set [17] of 25 problems available on the web at the URL http://www.cs.cas.cz/~luksan/test.html. All test problems, except the Rosenbrock problem, are nonsmooth.

We have not implemented the construction of the proximal trajectory at step 1 of the "main iteration," and we have always set $\hat{\gamma} = 10\gamma_{min}$. Each test has returned the same number of function evaluations as the number of subgradient evaluations. In fact the condition at step 1 of the algorithm has always been satisfied by the initial choice of $\hat{\gamma}$ and step 4(c) has never been entered.

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 N_f function evaluations. By f^* and f we indicate the minimum value of the objective function and the function value reached by the algorithm when the stopping criterion is met, respectively.

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 [8] dual quadratic programming algorithm.

In testing the algorithm, we have always adopted the same set of input parameters,

TABLE 5.1
NCVX: Computational results.

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

with no tuning based on any specific test problem, aiming at checking algorithm robustness more than efficiency. For problem 23 (marked by "$*$" in Table 5.1) we have set $m = 0.8$, as with standard $m = 0.2$ the quadratic subprogram solver failed due to the accumulation of rounding errors.

REFERENCES

[1] E. W. CHENEY AND A. A. GOLDSTEIN, *Newton's method for convex programming and Tchebycheff approximation*, Numer. Math., 1 (1959), pp. 253–268.

[2] F. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.

[3] V. F. DEMYANOV AND A. RUBINOV, *Quasidifferential Calculus*, Optimization Software Inc., New York, 1986.

[4] V. F. DEMYANOV AND A. RUBINOV, *Constructive Nonsmooth Analysis*, Peter Lang, Frankfurt am Main, Germany, 1995.

[5] G. DI PILLO, L. GRIPPO, AND S. LUCIDI, *A smooth method for the finite minimax problem*, Math. Program., 60 (1993), pp. 187–214.

[6] A. FUDULI AND M. GAUDIOSO, *The Proximal Trajectory Algorithm for Convex Minimization*, Tech. Report 7/98, Laboratorio di Logistica, Dipartimento di Elettronica Informatica e Sistemistica, Università della Calabria, Italy, 1998.

[7] M. GAUDIOSO, *Nonsmooth optimization*, in Handbook of Applied Optimization, M. G. C. Resende and P. Pardalos, eds., Oxford University Press, New York, 2002, pp. 299–310.

[8] D. GOLDFARB AND A. IDNANI, *A numerically stable dual method for solving strictly convex quadratic program*, Math. Program., 27 (1983), pp. 1–33.

[9] J.-B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms. Vol.* I, Springer-Verlag, Berlin, 1993.

[10] J.-B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms. Vol.* II, Springer-Verlag, Berlin, 1993.

[11] J. E. KELLEY, JR., *The cutting-plane method for solving convex programs*, J. Soc. Indust. Appl.

Math., 8 (1960), pp. 703–712.

[12] K. C. Kiwiel, *An aggregate subgradient method for nonsmooth convex minimization*, Math. Program., 27 (1983), pp. 320–341.

[13] K. C. Kiwiel, *Proximity control in bundle methods for convex nondifferentiable minimization*, Math. Program., 46 (1990), pp. 105–122.

[14] K. C. Kiwiel, *Finding normal solutions in piecewise linear programming*, Appl. Math. Optim., 32 (1995), pp. 235–254.

[15] K. C. Kiwiel, *Restricted step and Levenberg–Marquardt techniques in proximal bundle methods for nonconvex nondifferentiable optimization*, SIAM J. Optim., 6 (1996), pp. 227–249.

[16] C. Lemaréchal, *A view of line-searches*, in Optimization and Optimal Control, Lecture Notes in Control and Inform. Sci. 30, A. Auslender, W. Oettli, and J. Stoer, eds., Springer-Verlag, Berlin, New York, 1981, pp. 59–78.

[17] L. Lukšan and J. Vlček, *Test Problems for Nonsmooth Unconstrained and Linearly Constrained Optimization*, Tech. Report 798, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, 2000.

[18] M. Mäkelä, *Survey of bundle methods for nonsmooth optimization*, Optim. Methods Softw., 17 (2002), pp. 1–29.

[19] M. Mäkelä and P. Neittaanmäki, *Nonsmooth Optimization*, World Scientific, River Edge, NJ, 1992.

[20] R. Mifflin, *An algorithm for constrained optimization with semismooth functions*, Math. Oper. Res., 2 (1977), pp. 191–207.

[21] E. Polak, D. Mayne, and J. Higgins, *A superlinear convergent algorithm for min-max problems*, in Proceedings of the 28th IEEE Conference on Decision and Control, Tampa, FL, 1989, pp. 894–898.

[22] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[23] H. Schramm and J. Zowe, *A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results*, SIAM J. Optim., 2 (1992), pp. 121–152.

[24] N. Shor, *Minimizations Methods for Nondifferentiable Functions*, Springer-Verlag, Berlin, 1985.