Journal of Computational and Applied Mathematics 14 (1986) 391–400
North-Holland

An aggregate subgradient method for nonsmooth and nonconvex minimization

Krzysztof C. KIWIEL
Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland

Received 6 August 1984

Abstract: This paper presents a readily implementable algorithm for minimizing a locally Lipschitz continuous function that is not necessarily convex or differentiable. This extension of the aggregate subgradient method differs from one developed by the author in the treatment of nonconvexity. Subgradient aggregation allows the user to control the number of constraints in search direction finding subproblems and, thus, trade-off subproblem solution effort for rate of convergence. All accumulation points of the algorithm are stationary. Moreover, the algorithm converges when the objective function happens to be convex.

Mathematics Subject Classification: Primary 65K05, Secondary 90C25.

Keywords: Nonsmooth optimization, nondifferentiable programming, locally Lipschitz functions, semismooth functions, descent methods.

1. Introduction

This paper presents a readily implementable algorithm for minimizing a locally Lipschitz continuous function $f: \mathbb{R}^N \to \mathbb{R}$ that is not necessarily convex or differentiable. We suppose that at each $x \in \mathbb{R}^N$ we can compute $f(x)$ and a certain subgradient $g_f(x) \in \partial f(x)$, i.e. an arbitrary element of the subdifferential $\partial f(x)$ of f at x. The method is an extension of one for the convex case given in [3]. To deal with nondifferentiability of f the method accumulates subgradient information collected at many trial points in the form of an aggregate subgradient. At each iteration a search direction is found by solving a quadratic programming problem with linear constraints generated by the aggregate subgradient and several past subgradients. Then a line search procedure finds the next approximation to a solution and the next trial point. The two-point line search is employed to detect discontinuities in the gradient of f. Subgradient aggregation allows the user to control the number of subproblem constraints and, thus, trade-off subproblem solution effort and storage for speed of convergence.

The method requires uniformly bounded storage, and hence may be regarded as an implementable version of the algorithms in [5,8]. It differs significantly from the previous extension [4] of the aggregate subgradient method [3] in the treatment of nonconvexity. More specifically, here we use subgradient locality measures of [8] to ensure that the local past subgradient information dominates the nonlocal one at search direction finding, while in [4] a resetting strategy is used for

dropping nonlocal past subgradients. We also note that our treatment of nonconvexity and subgradient aggregation rules are similar to those in [6], although the method of [6] uses entirely different search direction finding subproblems and line search rules.

We may add that one drawback of algorithms based on resetting strategies [3,6,7,10] is that they may require the solution of several subproblems at each iteration with a reset. On the other hand, efficient rules for selecting parameters of subgradient locality measures are still unknown (see [6]). Thus, much more theoretical and experimental work remains to be done before we fully recognize the relative merits of algorithms with subgradient locality measures and those based on resetting techniques.

The algorithm presented has stationary accumulation points, if any. Its line search procedure is implementable if f satisfies an additional mild semismoothness hypothesis (see Section 2). Slightly stronger hypotheses were used in [6,7,8,10]. Moreover, the algorithm converges if f is convex and attains its minimum. These convergence properties are the same as those in [4]. No such convergence results are known for other comparable algorithms. In effect, we give the first implementable and globally convergent method with subgradient locality measures.

The method is described in detail in Section 2. Its global convergence is established in Section 3. Finally, we have a conclusion section.

We shall use the following notation and terminology. $\mathbb{R}^N$ denotes the N-dimensional Euclidean space with the usual scalar product $\langle \cdot, \cdot \rangle$ and the associated norm $|\cdot|$. All vectors are row vectors.

We say that a function $f\colon \mathbb{R}^N \to \mathbb{R}$ is locally Lipschitzian if for any bounded set $B \subset \mathbb{R}^N$ there exists a Lipschitz constant $L_f = L_f(B) < \infty$ such that

$$|f(x) - f(y)| \leq L_f |x - y| \quad \text{for all } x, y \in B.$$

The subdifferential (called generalized gradient in [1]) $\partial f(x)$ of f at x is the convex hull (conv) of all limits of sequences of the form $\{\nabla f(x^i)\colon x^i \to x$ and f is differentiable at each $x^i\}$, where $\nabla f(x^i)$ denotes the gradient of f at x^i. We say that a point $\bar{x} \in \mathbb{R}^N$ is stationary for f if $0 \in \partial f(\bar{x})$. Note that $0 \in \partial f(\hat{x})$ is a necessary condition for $\hat{x} \in \mathbb{R}^N$ to minimize f [1]. If f is convex, then for any $x \in \mathbb{R}^N$ and $\varepsilon \geq 0$

$$\partial_\varepsilon f(x) = \{g \in \mathbb{R}^N\colon f(y) \geq f(x) + \langle g, y - x \rangle - \varepsilon \quad \text{for all } y \in \mathbb{R}^N\}$$

is the ε-subdifferential of f at x.

2. The method

In this section we give motivation for the method and comment on its relations with other algorithms.

The algorithm to be described will generate a sequence of points $x^1, x^2, \ldots$ of $\mathbb{R}^N$, search directions $d^1, d^2, \ldots$ in $\mathbb{R}^N$ and nonnegative stepsizes $t_L^1, t_L^2, \ldots$, related by $x^{k+1} = x^k + t_L^k d^k$ for $k = 1, 2, \ldots$, where x^1 is a given starting point. The sequence $\{x^k\}$ is intended to converge to the required solution, and the algorithm is a descent method in the sense that $f(x^{k+1}) < f(x^k)$ if $x^{k+1} \neq x^k$. The algorithm will also calculate trial points $y^{k+1} = x^k + t_R^k d^k$ for $k = 1, 2, \ldots$, and subgradients $g^k = g_f(y^k)$ for all $k \geq 1$, where $y^1 = x^1$ and the trial stepsizes $t_R^k > 0$ satisfy $t_R^k = t_L^k$ if $t_L^k > 0$.

At the kth iteration, we associate with each past subgradient $g^j \in \partial f(y^j)$ the linearization of f

$$f_j(x) = f(y^j) + \langle g^j, x - y^j \rangle \quad \text{for all } x$$

and the following upper estimate of $|y^j - x^k|$

$$s_j^k = |y^j - x^j| + \sum_{i=j}^{k-1} |x^{i+1} - x^i|.$$

Since

$$f_j(x) = f_j^k + \langle g^j, x - x^k \rangle \quad \text{for all } x \tag{2.1}$$

with $f_j^k = f_j(x^k)$, we can compute $f_j(\cdot)$ and s_j^k recursively without storing the point y^j. Define the subgradient locality measure

$$\alpha_j^k = \max\left\{|f(x^k) - f_j^k|, \gamma\left(s_j^k\right)^2\right\} \tag{2.2}$$

where γ is a positive distance measure parameter that may be set equal to zero when f is convex. The value of $\alpha_j^k \geq 0$ indicates how much the subgradient $g^j \in \partial f(y^j)$ differs from being an element of $\partial f(x^k)$. In particular, $\alpha_j^k = 0$ implies $g^j \in \partial f(x^k)$, while in the convex case $g^j \in \partial_\varepsilon f(x^k)$ for $\varepsilon = \alpha_j^k$ [8]. Thus the triple (g^j, f_j^k, s_j^k) represents the subgradient information collected at the jth iteration. At the kth iteration we shall have a small subset J^k of $\{1, \ldots, k\}$ and the corresponding 'augmented' subgradients (g^j, f_j^k, s_j^k), $j \in j^k$. The rest of the past subgradient information will have been accumulated in the $(k-1)$st aggregate subgradient (p^{k-1}, f_p^k, s_p^k) satisfying

$$\left(p^{k-1}, f_p^k, s_p^k\right) \in \operatorname{conv}\left\{\left(g^j, f_j^k, s_j^k\right): j = 1, \ldots, k-1\right\}.$$

Similarly to (2.1) and (2.2), let us define the $(k-1)$st aggregate linearization

$$\tilde{f}^{k-1}(x) = f_p^k + \langle p^{k-1}, x - x^k \rangle \quad \text{for all } x$$

and the corresponding measure

$$\alpha_p^k = \max\left\{|f(x^k) - f_p^k|, \gamma\left(s_p^k\right)^2\right\} \tag{2.3}$$

which indicates how far p^{k-1} deviates from being a member of $\partial f(x^k)$. For instance, in the convex case $p^{k-1} \in \partial_\varepsilon f(x^k)$ for $\varepsilon = \alpha_p^k$ [2]. The available subgradients define the following piecewise linear polyhedral approximation to f

$$\hat{f}^k(x) = f(x^k) + \max\left\{-\alpha_j^k + \langle g^j, x - x^k \rangle: j \in J^k; -\alpha_p^k + \langle p^{k-1}, x - x^k \rangle\right\}.$$

To justify the above construction, we note that in the convex case $f_j(\cdot)$ and $\tilde{f}^{k-1}(\cdot)$ approximate $f(\cdot)$ from below, so that for $\gamma = 0$ we have

$$\alpha_j^k = f(x^k) - f_j^k \geq 0, \quad \alpha_p^k = f(x^k) - f_p^k \geq 0,$$

$$\hat{f}^k(x) = \max\left\{f_j(x): j \in J^k; \tilde{f}^{k-1}(x)\right\} \leq f(x) \quad \text{for all } x,$$

$$\hat{f}^k(y^j) = f(y^j) \quad \text{for all } j \in J^k,$$

see [3,4]. Since we want fo find a descent direction for f at x^k, we shall compute d^k to

$$\text{minimize } \hat{f}^k(x^k + d) + \tfrac{1}{2}|d|^2 \quad \text{over all } d \in \mathbb{R}^N,$$

where the regularizing term $\frac{1}{2}|d|^2$ will tend to keep $x^k + d^k$ in the region where $\hat{f}^k(\cdot)$ is a close approximation to $f(\cdot)$.

For convenience, we shall now state the method in detail, and then comment on its rules in what follows.

2.1. Algorithm

Step 0: Initialization. Select the starting point $x^1 \in \mathbb{R}^N$ and a final accuracy tolerance $\varepsilon_s \geq 0$. Choose fixed positive line search parameters m_L, m_R, m_α and $\bar{t}$ satisfying $m_L + m_\alpha < m_R < 1$ and $\bar{t} \leq 1$, a distance measure parameter $\gamma > 0$ ($\gamma = 0$ if f is convex), and a distance reset parameter $\bar{a} > 0$. Set $J^1 = \{1\}$, $y^1 = x^1$, $g^1 = p^0 = g_f(y^1)$, $f_1^1 = f_p^1 = f(y^1)$ and $s_1^1 = s_p^1 = 0$. Set the locality radius $a^1 = 0$ and the reset indicator $r_a^1 = 1$. Set the counters $k = 1$, $l = 0$ and $k(0) = 1$.

Step 1: Direction finding. Find the solution $(d^k, \tilde{v}^k)$ to the following kth quadratic programming problem

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}|d|^2 + \tilde{v} \text{ over all } (d, \tilde{v}) \in R^{N+1} \\ &\text{satisfying } -\alpha_j^k + \langle g^j, d\rangle \leq \tilde{v}, \quad j \in J^k, \\ &\qquad\qquad -\alpha_p^k + \langle p^{k-1}, d\rangle \leq \tilde{v} \quad \text{if } r_a^k = 0, \end{aligned} \tag{2.4}$$

where α_j^k and α_p^k are given by (2.2) and (2.3). Find Lagrange multipliers λ_j^k, $j \in J^k$, and λ_p^k of (2.4), setting $\lambda_p^k = 0$ if $r_a^k = 1$. Set

$$\left(p^k, \tilde{f}_p^k, \tilde{s}_p^k\right) = \sum_{j \in J^k} \lambda_j^k \left(g^j, f_j^k, s_j^k\right) + \lambda_p^k \left(p^{k-1}, f_p^k, s_p^k\right), \tag{2.5}$$

$$\tilde{\alpha}_p^k = \max\left\{|f(x^k) - \tilde{f}_p^k|, \gamma\left(\tilde{s}_p^k\right)^2\right\}, \tag{2.6}$$

$$v^k = -\left\{|p^k|^2 + \tilde{\alpha}_p^k\right\}, \tag{2.7}$$

$$w^k = \tfrac{1}{2}|p^k|^2 + \tilde{\alpha}_p^k. \tag{2.8}$$

Step 2: Stopping criterion. If $w^k \leq \varepsilon_s$, terminate; otherwise, continue.

Step 3: Line search. By a line search procedure as given below, find two stepsizes t_L^k and t_R^k such that $0 \leq t_L^k \leq t_R^k \leq 1$ and such that the two corresponding points $x^{k+1} = x^k + t_L^k d^k$ and $y^{k+1} = x^k + t_R^k d^k$ satisfy

$$f(x^{k+1}) \leq f(x^k) + m_L t_L^k v^k, \tag{2.9a}$$

and either a serious step is taken: $t_L^k = t_R^k > 0$ and either

$$t_L^k \geq \bar{t} \quad \text{or} \quad \alpha(x^k, x^{k+1}) > m_\alpha |v^k|, \tag{2.9b}$$

or a null step occurs: $t_L^k = 0$ and

$$t_R^k \leq \bar{t}, \tag{2.9c}$$

$$-\alpha(x^{k+1}, y^{k+1}) + \langle g_f(y^{k+1}), d^k\rangle \geq m_R v^k, \tag{2.9d}$$

where

$$\alpha(x, y) = \max\{|f(x) - \tilde{f}(x; y)|, \gamma|x - y|^2\},$$
$$\tilde{f}(x; y) = f(x) - f(y) - \langle g_f(y), x - y\rangle. \tag{2.10}$$

If $t_L^k > 0$, set $k(l+1) = k+1$ and increase the counter of serious steps l by 1.

Step 4: Augmented subgradient updating. Select a set J^{k+1} satisfying

$$\{k+1, k(l)\} \subset J^{k+1} \subset J^k \cup \{k+1\},$$

set $g^{k+1} = g_f(y^{k+1})$ and

$$f_{k+1}^{k+1} = f(y^{k+1}) + \langle g^{k+1}, x^{k+1} - y^{k+1}\rangle,$$
$$f_j^{k+1} = f_j^k + \langle g^j, x^{k+1} - x^k\rangle, \quad j \in J^{k+1}\setminus\{k+1\},$$
$$f_p^{k+1} = \tilde{f}_p^k + \langle p^k, x^{k+1} - x^k\rangle,$$
$$s_{k+1}^{k+1} = |y^{k+1} - x^{k+1}|,$$
$$s_j^{k+1} = s_j^k + |x^{k+1} - x^k|, \quad j \in J^{k+1}\setminus\{k+1\},$$
$$s_p^{k+1} = \tilde{s}_p^k + |x^{k+1} - x^k|,$$
$$a^{k+1} = \max\{a^k + |x^{k+1} - x^k|, s_{k+1}^{k+1}\}.$$

Step 5: Distance resetting. If $t_L^k = 0$ or $a^{k+1} \leqslant \bar{a}$, set $r_a^{k+1} = 0$ and go to Step 6. Otherwise, set $r_a^{k+1} = 1$ and delete from J^{k+1} some indices j with the largest values of $s_j^{k+1} > 0$ so that the reset value a^{k+1} satisfies

$$a^{k+1} = \max\{s_j^{k+1}: j \in J^{k+1}\} \leqslant \bar{a}.$$

Step 6. Increase k by 1 and go to Step 1.

A few remarks on the algorithm are in order.

Our rules for aggregating and reducing the past subgradient information stem from the observation that we always have (see [3,4,8]).

$$\lambda_j^k \geqslant 0, \quad j \in J^k, \quad \lambda_p^k \geqslant 0, \quad \sum_{j \in J^k} \lambda_j^k + \lambda_p^k = 1, \qquad \lambda_p^k = 0 \quad \text{if } r_a^k = 1. \tag{2.11}$$

For any $k \geqslant 1$, let

$$k_a(k) = \max\{j: j \leqslant k \text{ and } r_a^j = 1\},$$
$$J_a^k = J^{k_a(k)} \cup \{j: k_a(k) < j \leqslant k\}.$$

An inductive argument based on (2.5), (2.11) and the algorithm's rules yields (see [3,4]) that we always have

$$(p^k, \tilde{f}_p^k, \tilde{s}_p^k) \in \operatorname{conv}\{(g^j, f_j^k, s_j^k): j \in J_a^k\}, \tag{2.12}$$

$$a^k = \max\{s_j^k: j \in J_a^k\}, \tag{2.13}$$

and hence, since $s_j^k \geq |x^k - y^j|$,

$$\max\{|y^j - x^k|: j \in J_a^k\} \leq a^k. \tag{2.14}$$

The above relations say that $(p^k, \tilde{f}_p^k, \tilde{s}_p^k)$ aggregates the past subgradient information collected from the ball around x^k of radius a^k. We use distance resets only to ensure that the locality radius a^k stays locally uniformly bounded, since otherwise the convergence analysis of Section 3 would require, as in [8], and additional assumption on the boundedness of the entire sequence $\{y^k\}$.

The stopping criterion admits of the following interpretation. A small value of w^k indicates both that $|p^k|$ is small and that p^k is close to $\partial f(x^k)$, because the value of the subgradient locality measure $\tilde{\alpha}_p^k$ is small. Thus the null vector is close to $\partial f(x^k)$, i.e. x^k is approximately stationary. Also if f is convex then (see [4])

$$p^k \in \partial_\varepsilon f(x^k) \quad \text{for } \varepsilon = \tilde{\alpha}_p^k. \tag{2.15}$$

Our line search requirements (2.9) are modifications of those in [4,8]. They ensure that each serious step decreases significantly the objective value, while each null step results in a significant modification of the next search direction finding subproblem. The following procedure (taken from [4]) may be used for implementing Step 3.

2.2. Line Search Procedure

(a) Set $t_L = 0$, $t = t_U = 1$ and $m = (m_R - m_\alpha + m_L)/2$.
(b) If $f(x^k + td^k) \leq f(x^k) + mtv^k$ set $t_L = t$; otherwise set $t_U = t$.
(c) If $f(x^k + td^k) \leq f(x^k) + m_L tv^k$ and either $t \geq \bar{t}$ or $\alpha(x^k, x^k + td^k) > m_\alpha |v^k|$ set $t_L^k = t_R^k = t$ and return.
(d) If $t < \bar{t}$ and $-\alpha(x^k, x^k + td^k) + \langle g_f(x^k + td), d^k \rangle \geq m_R v^k$ set $t_R^k = t$, $t_L^k = 0$ and return.
(e) Choose $t \in [t_L + 0.1(t_U - t_L), t_U - 0.1(t_U - t_L)]$ by some interpolation procedure and go to (b).

Step (e) of the above procedure may use various interpolation formulae; see, for instance, [9]. Suppose that f satisfies the following 'semismoothes' hypothesis:

$$\begin{aligned} &\text{for any } x,\ d \in \mathbb{R}^N \text{ and sequences } \{\bar{g}^i\} \subset \mathbb{R}^N \text{ and } \{t^i\} \subset \mathbb{R}_+ \text{ satisfying } t^i \downarrow 0 \\ &\text{and } \bar{g}^i \in \partial f(x + t^i d) \text{ one has } \limsup_{i \to \infty} \langle \bar{g}^i, d \rangle \geq \liminf_{i \to \infty} [f(x + t^i d) - f(x)]/t^i. \end{aligned} \tag{2.16}$$

Then convergence of our line search procedure can be established as in [7] and [8] since examination of the proof of Theorem 4.1 in [7] reveals that it remains valid if one uses (2.16) instead of the definition of weak upper semismoothness in [7] (which is obtained by interchanging limsup and liminf in (2.16)).

The user can control storage and work per iteration by choosing the number of elements of J^k. The algorithm converges even if $J^k = \{k, k(l)\}$ for all k. Of course, one may expect faster convergence if more subgradients are used for search direction finding.

Suitable values for the line search parameters are $m_L = 0.1$, $m_R = 0.3$ and $m_\alpha = 0.1$. The role of $\bar{a}$ is secondary, hence any large value, say $\bar{a} = 10^5$, should suffice. The choice of the distance measure parameter γ is more delicate in the nonconvex case (see [6]), and should be based on an experiment.

We may add that for $\gamma = 0$ subproblem (2.4) reduces to the one used in [4]. On the other hand, subproblem (2.4) is used in [8] with $J^k = \{1, \ldots, k\}$, $r_a^k = 1$ and $\alpha_j^k = \alpha_M(x^k, y^j)$, where

$$\alpha_M(x, y) = \max\{f(x) - \bar{f}(x; y), \gamma|x - y|^2\}.$$

Our definition of $\alpha(\cdot, \cdot)$ (see (2.10)) will allow for choosing a small value of γ in the nonconvex case. As far as the theory is concerned, one may use $\alpha_M(\cdot, \cdot)$ instead of $\alpha(\cdot, \cdot)$ in Algorithm 2.1, deleting the absolute value sign in (2.2), (2.3) and (2.6).

3. Convergence

In this section we shall establish global convergence of the method. In the absence of convexity, we will content ourselves with finding stationary points for f. We suppose that each execution of Line Search Procedure (section 2.2) is finite, e.g. that f has the additional semismoothness property (2.16), and that the final accuracy tolerance ε_s is set to zero.

First, we consider the case of finite termination.

Lemma 3.1. *If Algorithm 2.1 terminates at the kth iteration, then x^k is stationary for f.*

Proof. If $w^k = \frac{1}{2}|p^k|^2 + \tilde{\alpha}_p^k \leq \varepsilon_s = 0$, then $p^k = 0$, $\tilde{\alpha}_p^k = 0$ and $\gamma \tilde{s}_p^k = 0$. Thus if f is convex then $0 \in \partial f(x^k)$ from (2.15), while in the nonconvex case ($\gamma > 0$) we have $(p^k, \tilde{s}_p^k) = (0, 0)$ in (2.12), so $0 = p^k$ can be expressed as a convex combination of g^j with $|y^j - x^k| \leq s_j^k = 0$, i.e. $g^j \in \partial f(x^k)$, and hence $0 \in \partial f(x^k)$ from the convexity of $\partial f(x^k)$. □

From now on we suppose that the method computes an infinite sequence $\{x^k\}$. Note that, by construction,

$$x^k = x^{k(l)} \quad \text{if } k(l) \leq k < k(l+1), \tag{3.1}$$

where we set $k(l+1) = +\infty$ if the algorithm generates only a finite number l of serious steps. We shall need the following result, which follows directly from the proof of Lemma 3.3 in [4] and the algorithm's rules.

Lemma 3.2. *Let $\bar{x} \in \mathbb{R}^N$, $\varepsilon > 0$ and $B = \{x \in \mathbb{R}^N : |x - \bar{x}| \leq \varepsilon\}$. Then there exists a constant $C < \infty$ such that if $x^{k(l)} \in B$ and $k(l) \leq k < k(l+1)$ then $s_{k+1}^{k+1} \leq C$, $a^k \leq \max\{\bar{a}, C\}$, and $|x^{k+1} - x^k| \leq Ct_L^k$. Moreover, if $x^k \in B$ and $t_L^k = 0$ then*

$$0 \leq w^{k+1} \leq w^k - (1 - m_R)^2 (w^k)^2 / 8C^2. \tag{3.2}$$

We conclude from the above lemma that the stationarity measure w^k of the current point x^k decreases significantly after each null step ($m_R < 1$). A crucial asymptotic property of $\{w^k\}$ is given in

Lemma 3.3. *Suppose that there exist $\bar{x} \in \mathbb{R}^N$ and an infinite set $K \subset \{1, 2, \ldots\}$ such that $x^k \overset{K}{\to} \bar{x}$ and $w^k \overset{K}{\to} 0$. Then $0 \in \partial f(\bar{x})$.*

Proof. Suppose that $x^k \overset{K}{\to} \bar{x}$ and $w^k = \frac{1}{2}|p^k|^2 + \tilde{\alpha}_p^k \overset{K}{\to} 0$, so that $p^k \overset{K}{\to} 0$ and $\tilde{\alpha}_p^k \overset{K}{\to} 0$. If f is convex then let $k \in K$ tend to infinity in (2.15) and use the definition of ε-subdifferential to deduce that $0 \in \partial f(\bar{x})$. Next, suppose that $\gamma > 0$, so that $\tilde{s}_p^k \overset{K}{\to} 0$, because $\tilde{\alpha}_p^k \overset{K}{\to} 0$. Since $x^k \overset{K}{\to} \bar{x}$, we deduce from (3.1) and Lemma 3.2 the boundeness of $\{a^k\}_{k \in K}$. Thus we have $(p^k, \tilde{s}^k) \overset{K}{\to} (0, 0)$ and bounded $\{a^k\}_{k \in K}$ in (2.12)–(2.14), so we may use the proof of Lemma 6 in [8] to obtain $0 \in \partial f(\bar{x})$.

We may now consider the case of a finite number of serious steps.

Lemma 3.4. *Suppose that $x^k = x^{k(l)} = \bar{x}$ for some fixed l and all $k \geq k(l)$. Then $0 \in \partial f(\bar{x})$.*

Proof. If $t_L^k = 0$ for all large k, then Lemma 3.2 yields, by (3.2), that $w^k \downarrow 0$, so the desired conclusion follows from Lemma 3.3.

Let us now consider the remaining case of infinitely many serious steps.

Lemma 3.5. *Suppose that there exist $\bar{x} \in \mathbb{R}^N$ and an infinite set $L \subset \{1, 2, \ldots\}$ such that $\{x^{k(l)}\}_{l \in L} \to \bar{x}$. Then $0 \in \partial f(\bar{x})$.*

Proof. Let $K = \{k(l+1) - 1: l \in L\}$. In view of (3.1) and Lemma 3.3, we need only show that $w^k \overset{K}{\to} 0$. To obtain a contradiction, suppose that $w^k \geq \bar{w} > 0$ for some $\bar{w}$ and all large $k \in K$. Since $x^k \overset{K}{\to} \bar{x}$ and $\{f(x^k)\}$ is nonincreasing, we have $f(x^k) \downarrow f(\bar{x})$ by the continuity of f, so (2.9a) yields $t_L^k v^k \to 0$. But $|v^k| \geq w^k \geq \bar{w}$ for all large $k \in K$ from (2.7)–(2.8), so $t_L^k \overset{K}{\to} 0$, and we obtain $|x^{k+1} - x^k| \overset{K}{\to} 0$ from Lemma 3.2. Thus both $\{x^k\}_{k \in K}$ and $\{x^{k+1}\}_{k \in K}$ converge to $\bar{x}$, and the properties of $\alpha(\cdot, \cdot)$ (see [9]) imply that $\alpha(x^k, x^{k+1}) \overset{K}{\to} 0$. Hence we have $t_L^k < \bar{t}$ and $\alpha(x^k, x^{k+1}) < m_\alpha |v^k|$ for all large $k \in K$, since $t_L^k \to 0 < \bar{t}$ and $|v^k| \geq \bar{w} > 0$ for large $k \in K$, and we obtain a contradiction with (2.9b) and the definition of K. Therefore, we must have $w^k \overset{K}{\to} 0$, as desired. □

Combining (3.1) with Lemmas 3.4 and 3.5, we deduce our main result.

Theorem 3.6. *Every accumulation point of an infinite sequence $\{x^k\}$ generated by Algorithm 2.1 is stationary for f.*

As in [4], the above result may be strengthened in the convex case as follows.

Theorem 3.7. *If f is convex, then the sequence $\{x^k\}$ calculated by Algorithm 2.1 is minimizing, i.e. $f(x^k) \downarrow \inf\{f(x): x \in \mathbb{R}^N\}$. If additionally f attains its minimum value, then $\{x^k\}$ convergers to a minimum point of f.*

Proof. As in [4], one can easily extend the analysis of [3] to Algorithm 2.1.

We may add that the proofs of Lemmas 3.4 and 3.5 imply the following result. If the level set $\{x \in \mathbb{R}^N: f(x) \leq f(x^1)\}$ is bounded and the final accuracy tolerance ε_s is positive, then the algorithm will terminate in a finite number of iterations with $w^k \leq \varepsilon_s$, i.e. with an approximately stationary point x^k.

4. Conclusions

We have given a readily implementable algorithm for minimizing nonsmooth functions. It uses subgradient locality measures of [8] to deal with nonconvexity. This makes it different from another extension [4] of [3], which employs resets for dropping obsolete subgradients. The algorithm presented and the one in [4] have stationary accumulation points, if any, and converge whenever the objective function happens to be convex. No such results are known for other comparable methods [6,7,8,10].

Our computational experience [2] indicates that the method presented and the one in [4] may perform differently on certain classes of optimization problems. This experience will be reported elsewhere.

Acknowledgment

I would like to thank Andrzej Ruszczyński for the numerous discussions we have had on this work.

Appendix

A preliminary version of the algorithm has been implemented in standard FORTRAN on an ODRA 1325 computer (single precision of eleven digits). The performance of the method is illustrated by the following two highly nonconvex examples. We used the parameters $m_1 = 0.1$, $m_R = 0.3$, $m_\alpha = 0.1$, $\tilde{t} = 0.01$, $a = 10^3$ and $\gamma = 1$.

The first example is given by

$$f(x) = \max\{|10x_1^2 - 10x_2|, |x_1 - 1|\}, \quad x \in \mathbb{R}^2,$$

which has a unique minimizer $\hat{x} = (1, 1)$ with $f(\hat{x}) = 0$. For $x^1 = (-1.2, 1)$ and $\varepsilon_s = 10^{-8}$, the method stopped with $x^{20} = (1, 1 - 10^{-10})$ and $f(x^{20}) = 10^{-10}$ after 20 iterations and 48 function and subgradient evaluations.

The objective of the second problem

$$f(x) = \max\{x_1^2 + (x_2 - 1)^2 + x_2 - 1, -x_1^2 - (x_2 - 1)^2 + x_2 + 1\}$$

has a unique minimizer $\hat{x} = (0, 0)$ with $f(\hat{x}) = 0$. This function has narrow crescent-shaped level sets which force the algorithm to make very short steps. Accordingly, we demanded low accuracy by using $\varepsilon_s = 10^{-5}$. Starting from $x^1 = (-1.5, 2)$, the algorithm terminated with $x_1^{24} = -5.10^{-4}$, $x_2^{24} = -8.10^{-8}$ and $f(x^{24}) = 3.10^{-7}$ after 33 function and subgradient evaluations.

For both problems the sets J^k were chosen as in [3] so that they had at most three elements.

References

[1] F.H. Clarke, A new approach to Lagrange multipliers, *Math. Oper. Res.* **1** (1976) 165–174.

[2] K.C. Kiwiel, Efficient algorithms for nonsmooth optimization and their applications, Ph.D. dissertation, Dep. of Electronics, Technical University of Warsaw, Poland, 1982.

[3] K.C. Kiwiel, An aggregate subgradient method for nonsmooth convex minimization, *Math. Programming*, **27** (1983) 320–341.
[4] K.C. Kiwiel, A linearization algorithm for nonsmooth minimization, *Math. Oper. Res.*, to appear.
[5] C. Lemarechal, Nonsmooth optimization and descent methods, RR78-4, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1978.
[6] C. Lemarechal, J.-J. Strodiot and A. Bihain, On a bundle algorithm for nonsmooth minimization, in: O.L. Mangasarian, R.R. Meyer and S.M. Robinson, eds. *Nonlinear Programming 4* (Academic Press, New York, 1981) 245–282.
[7] R. Mifflin, An algorithm for constrained optimization with semismooth functions, *Math. Oper. Res.* **2** (1977) 959–972.
[8] R. Mifflin, A modification and an extension of Lemarechal's algorithm for nonsmooth minimization, in: D.C. Sorensen and R.J.B. Wets, Eds., *Nondifferential and Variational Techniques in Optimization*, Mathematical Programming Study **17** (North-Holland, Amsterdam, 1982) pp. 77–90.
[9] R. Mifflin, Stationarity and superlinear convergence of an algorithm for univariate locally Lipschitz constrained minimization, TR-82-2, Dept. of Pure and Applied Mathematics, Washington State University, Pulman, Washington, 1982.
[10] E. Polak, D.Q. Mayne and Y. Wardi, On the extension of constrained optimization algorithms from differentiable to nondifferentiable problems, *SIAM J. Control Optim.* **21** (1983) 179–203.