Raport Badawczy

Research Report

RB/38/2009

An improved convergence result for the discrete gradient and secant methods for nonsmooth optimization

K.C. Kiwiel

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences

POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6

01-447 Warszawa

tel.: (+48) (22) 3810100

fax: (+48) (22) 3810105

Kierownik Pracowni zgłaszający pracę:
Prof. dr hab. inż. Krzysztof C. Kiwiel

Warszawa 2009

Journal of Optimization Theory and Applications manuscript No.
(will be inserted by the editor)

An Improved Convergence Result for the Discrete Gradient and Secant Methods for Nonsmooth Optimization

K.C. Kiwiel

Received: March 18, 2009 / Accepted: date

Abstract We study a generalization of the non-derivative discrete gradient method of Bagirov et al. for minimizing a locally Lipschitz function f on $\mathbb{R}^n$. We strengthen the existing convergence result for this method by showing that it either drives the f-values to $-\infty$ or each of its cluster points is Clarke stationary for f, without requiring compactness of the level sets of f. Our generalization is an approximate bundle method, which also subsumes the secant method of Bagirov et al.

Keywords Nonsmooth optimization · Derivative-free optimization · Bundle methods · Discrete gradient

Mathematics Subject Classification (2000) 65K10 · 90C26

1 Introduction

We consider the recently proposed *discrete gradient* (DG) method [3] for minimizing a locally Lipschitzian function $f : \mathbb{R}^n \to \mathbb{R}$. In contrast with bundle methods (see, e.g., [11, 12] and the references in [3,5,7,14]) which require the computation of a single subgradient of f at each trial point, the DG method approximates subgradients by discrete gradients using f-values only. This is important for applications where subgradients are unavailable and derivative free methods are employed; see, e.g., [1, 2] and the references therein.

Our contributions can be summarized as follows. First, although the DG method [3, Alg. 7.1] has three nested loops, we show that it may be regarded as an instance of a bundle method with approximate subgradients and a single iteration loop; this simplifies its analysis. Second, we prove that this bundle method either drives the f-values to $-\infty$, or each of its cluster points is Clarke [8] stationary for f (see Thm. 3.1). This is significantly stronger than the result of [3, Thm. 7.1], which assumes

K.C. Kiwiel
Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01–447 Warsaw, Poland, E-mail: kiwiel@ibspan.waw.pl

additionally that f has bounded level sets and establishes stationarity only for cluster points of a subsequence generated in an outer loop (cf. Rem. 3.1(c)). Third, since the DG method needs unbounded storage (cf. Rems. 2.1(e) and 3.1(e)), we show how to use well-known bundle techniques [11, 12] to ensure bounded storage. Fourth, we note that our results extend easily to the secant method of [5] and the quasisecant method of [4].

We add that our proof technique is related to that employed in [15] for establishing global convergence of the gradient sampling algorithm [7].

The paper is organized as follows. Section 2 presents our bundle generalization of the DG and secant methods. Its convergence is analyzed in Section 3. Section 4 gives extensions for the quasisecant setting of [4] and the bundle setting of [14].

2 A bundle method with approximate subgradients

We assume that the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous. The Clarke *subdifferential* [8] of f at any point x is given by

$$\partial f(x) = \mathrm{co}\{\lim_j \nabla f(y^j) : y^j \to x \text{ and } \nabla f(y^j) \text{ exists for all } j\},$$

where co denotes the convex hull, and the Clarke *ε-subdifferential* [10] by

$$\partial_\varepsilon f(x) := \mathrm{co}\,\partial f(B(x,\varepsilon)), \tag{1}$$

where $B(x,\varepsilon) := \{y : |y-x| \le \varepsilon\}$ is the Euclidean ball centered at x with radius $\varepsilon \ge 0$. The mapping $\partial f(\cdot)$ closed. We say that a point $\bar{x}$ is *stationary* for f if $0 \in \partial f(\bar{x})$.

By Lebourg's mean value theorem (cf. [8, Theorem 2.3.7]), for each $d \in \mathbb{R}^n$,

$$f(x+\varepsilon d) - f(x) = \varepsilon\langle v(x,d,\varepsilon), d\rangle \quad \text{for some } v(x,d,\varepsilon) \in \partial f([x, x+\varepsilon d]). \tag{2}$$

Since $v(x,d,\varepsilon)$ may be hard to compute, for algorithmic purposes we assume that we have a simpler mapping $\gamma(x,d,\varepsilon)$ that satisfies the following. Let $\mathbb{S} := \{d : |d| = 1\}$.

Assumption 2.1 (1) *The mapping $\gamma(x,d,\varepsilon)$ has the mean value property*

$$f(x+\varepsilon d) - f(x) = \varepsilon\langle \gamma(x,d,\varepsilon), d\rangle \quad \textit{for all } x \in \mathbb{R}^n, d \in \mathbb{S}, \varepsilon > 0. \tag{3}$$

(2) *For each $x \in \mathbb{R}^n$ and $\varepsilon > 0$, there exists $L < \infty$ such that $\sup_{d\in\mathbb{S}} |\gamma(x,d,\varepsilon)| \le L$.*

(3) *The set $\Gamma_\varepsilon(x) := \mathrm{co}\,\gamma(x,\mathbb{S},\varepsilon)$ approximates $\partial f(x)$ in the following sense: for each $\bar{x} \in \mathbb{R}^n$ and $\rho > 0$, there exist $\tau > 0$ and $\bar{\varepsilon} > 0$ such that*

$$\Gamma_\varepsilon(B(\bar{x},\tau)) \subset \partial_\rho f(\bar{x}) + B(0,\rho) \quad \textit{for all } \varepsilon \in (0,\bar{\varepsilon}]. \tag{4}$$

Note that we may let $\Gamma_\varepsilon(x)$ be any set containing $\mathrm{co}\,\gamma(x,\mathbb{S},\varepsilon)$ for which (4) holds; e.g., we may replace $\Gamma_\varepsilon(x)$ by its closure. Thus our assumption holds in the DG framework of [3] and the secant framework of [5]. Incidentally, for $\gamma = v$ of (2) and $\Gamma_\varepsilon = \partial_\varepsilon f$, the inclusion in (4) holds when $\tau + \bar{\varepsilon} \le \rho$.

We now state a bundle method which generalizes the DG and secant methods. For a closed convex set C, $\mathrm{Proj}(0\,|\,C)$ is its minimum-norm element.

Algorithm 2.1 (bundle method with approximate subgradients)

Step 0 (*Initialization*). Select an initial point $x^1 \in \mathbb{R}^n$, optimality tolerances ν_{opt}, $\varepsilon_{\text{opt}} \geq 0$, a descent parameter $\kappa \in (0,1)$, reduction factors μ, θ in $(0,1)$, a locality radius $\varepsilon_1 > 0$, a stationarity target $\nu_1 > 0$ and an initial direction $d^0 \in \mathbb{S}$. Set $g^1 := \gamma(x^1, d^0, \varepsilon_1)$, $G_1 := \{g^1\}$ and $k := 1$.

Step 1 (*Direction finding*). Set $\hat{g}^k := \text{Proj}(0 | \operatorname{co} G_k)$.

Step 2 (*Stopping criterion*). If $|\hat{g}^k| \leq \nu_{\text{opt}}$ and $\varepsilon_k \leq \varepsilon_{\text{opt}}$, terminate.

Step 3 (*Locality radius update*). If $|\hat{g}^k| \leq \nu_k$, set $\nu_{k+1} := \theta \nu_k$, $\varepsilon_{k+1} := \mu \varepsilon_k$, $t_k := 0$, $d^k := d^{k-1}$ and go to Step 5. Otherwise, set $\nu_{k+1} := \nu_k$, $\varepsilon_{k+1} := \varepsilon_k$ and $d^k := -\hat{g}^k/|\hat{g}^k|$ (so that $d^k \in \mathbb{S}$).

Step 4 (*Descent test*). If $f(x^k + \varepsilon_k d^k) - f(x^k) > -\kappa \varepsilon_k |\hat{g}^k|$, set $t_k := 0$; otherwise, choose a step size $t_k \geq \varepsilon_k$ such that $f(x^k + t_k d^k) \leq f(x^k) - \kappa t_k |\hat{g}^k|$.

Step 5 (*Updating*). Set $x^{k+1} := x^k + t_k d^k$.

Step 6 (*Bundle compression*). After a *null step* with $t_k = 0$ and $\varepsilon_{k+1} = \varepsilon_k$, choose a reduced bundle $\hat{G}_k \subset \{\hat{g}^k\} \cup G_k$ with $\hat{g}^k \in \operatorname{co} \hat{G}_k$; otherwise, set $\hat{G}_k := \emptyset$.

Step 7 (*Bundle addition*). Set $g^{k+1} := \gamma(x^{k+1}, d^k, \varepsilon_{k+1})$ and $G_{k+1} := \{g^{k+1}\} \cup \hat{G}_k$.

Step 8 (*Loop*). Increase k by 1 and go to Step 1.

A few comments on the method are in order.

Remark 2.1 (a) At Step 1, $\operatorname{co} G_k \subset \operatorname{co}\{g^j : \varepsilon_j = \varepsilon_k, j \leq k\}$ with $g^j = \gamma(x^k, \varepsilon_k, d^{j-1})$ give $\hat{g}^k \in \Gamma_{\varepsilon_k}(x^k)$ by Assumption 2.1(3); thus $\hat{g}^k$ is an *aggregate subgradient*. Step 1 may use the QP methods of [9, 13], which can solve efficiently related subproblems.

(b) The stopping criterion of Step 2 is motivated by the inclusion in (4) with $\bar{x} = x^k$; namely, if $\varepsilon_{\text{opt}} \leq \bar{\varepsilon}$ then $\hat{g}^k \in \Gamma_{\bar{\varepsilon}}(x^k)$ yields $\operatorname{dist}(0 | \partial_\rho f(x^k)) \leq |\hat{g}^k| + \rho$, so that the point x^k is approximately stationary if both $|\hat{g}^k|$ and ρ are small.

(c) At Step 3, the condition $|\hat{g}^k| \leq \nu_k$ detects progress in stationarity; then ν_k and ε_k are reduced and Step 6 drops past subgradients by setting $\hat{G}_k := \emptyset$.

(d) If Step 4 produces a null step with $t_k = 0$ and $\varepsilon_{k+1} = \varepsilon_k$, then by the mean value property (3), the next approximate subgradient g^{k+1} computed at Step 7 will satisfy $\langle g^{k+1}, d^k \rangle > -\kappa |\hat{g}^k|$; in other words, since $d^k := -\hat{g}^k/|\hat{g}^k|$, g^{k+1} will satisfy

$$\langle g^{k+1}, \hat{g}^k \rangle \leq \kappa |\hat{g}^k|^2. \tag{5}$$

In particular, since $\hat{g}^k$ is characterized by $\hat{g}^k \in \operatorname{co} G_k$ and $\langle g, \hat{g}^k \rangle \geq |\hat{g}^k|^2$ for all $g \in \operatorname{co} G_k$, we have $g^{k+1} \notin \operatorname{co} G_k$ by (5) with $\kappa < 1$. If a null step does not occur, we can try expansion, starting from $t := \varepsilon_k$ and setting $t := 2t$ until $f(x^k + t d^k) > f(x^k) - \kappa t |\hat{g}^k|$, in which case $t_k := t/2$ is accepted. At expansion we can replace κ by a parameter $\underline{\kappa} \in (0, \kappa]$. In practice expansion should stop when t is "too large" or $f(x^k + t d^k)$ is "too low"; otherwise it could drive $f(x^k + t d^k)$ to $-\infty$.

(e) After a null step, Step 6 may choose $\hat{G}_k$ by the well-known bundle strategies: *accumulation* with $\hat{G}_k := G_k$, *aggregation* with $\hat{G}_k := \{\hat{g}^k\} \cup G'_k$ for some $G'_k \subset G_k$, or *selection* with $\hat{G}_k \supset \{g^j : \lambda_j^k > 0\}$, where $\lambda_j^k \geq 0$ are multipliers such that $\hat{g}^k = \sum_j \lambda_j^k g^j$, $\sum_j \lambda_j^k = 1$. Of course, accumulation needs unbounded storage, whereas for any fixed $M_g \geq 1$, we can choose $|\hat{G}_k| \leq M_g$ via aggregation, or selection if $M_g \geq n+1$, since the QP methods of [9, 13] compute at most $n+1$ positive multipliers.

(f) If $t_k > 0$ or $\varepsilon_{k+1} < \varepsilon_k$, Step 7 may use an arbitrary $d^k \in \mathbb{S}$ for finding g^{k+1}.

3 Convergence analysis

Our main convergence result follows.

Theorem 3.1 *Let $\{x^k\}$ be a sequence generated by Algorithm 2.1 with $\varepsilon_{\text{opt}} = 0$ under Assumption 2.1. Then the algorithm does not stop and either $f(x^k) \downarrow -\infty$, or $v_k \downarrow 0$, $\varepsilon_k \downarrow 0$ and every cluster point of $\{x^k\}$ is stationary for f.*

Proof Since Step 2 is always reached with $\varepsilon_k > 0$, no termination occurs. If $f(x^k) \downarrow -\infty$, there is nothing to prove, and so assume $\inf_k f(x^k) > -\infty$. If $t_k > 0$, then Steps 4 and 5 with $|d^k| = 1$ yield $\kappa t_k |\hat{g}^k| = \kappa |x^{k+1} - x^k||\hat{g}^k| \le f(x^k) - f(x^{k+1})$. Summing this inequality (which holds also if $t_k = 0$ at Steps 3 or 4) gives

$$\sum_{k=1}^{\infty} t_k |\hat{g}^k| < \infty, \tag{6}$$

$$\sum_{k=1}^{\infty} |x^{k+1} - x^k||\hat{g}^k| < \infty. \tag{7}$$

Suppose there is k_1, $\bar{v} > 0$ and $\bar{\varepsilon} > 0$ such that $v_k = \bar{v}$ and $\varepsilon_k = \bar{\varepsilon}$ for all $k \ge k_1$. Using $|\hat{g}^k| \ge \bar{v}$ at Step 3 in (6) yields $t_k \to 0$. Pick $k_2 \ge k_1$ such that $t_k < \bar{\varepsilon}$ gives $t_k = 0$ at Step 4 for all $k \ge k_2$. Fix $k \ge k_2$. Since $\hat{g}^k \in \operatorname{co} \hat{G}_k$ at Step 6, at Step 7 we have $\hat{g}^k, g^{k+1} \in \operatorname{co} G_{k+1}$ and $|g^{k+1}| \le L$ for L given by Assumption 2.1(2) with $x = x^{k_2}$ and $\varepsilon = \bar{\varepsilon}$. On the next iteration, by Remark 2.1(d), $\langle \hat{g}^k, \hat{g}^{k+1} \rangle \ge |\hat{g}^{k+1}|^2$ and $\langle g^{k+1}, \hat{g}^{k+1} \rangle \ge |\hat{g}^{k+1}|^2$ at Step 1. The first inequality and expanding $|\hat{g}^{k+1} - \hat{g}^k|^2$ give

$$|\hat{g}^{k+1} - \hat{g}^k|^2 \le |\hat{g}^k|^2 - |\hat{g}^{k+1}|^2, \tag{8}$$

whereas the second one combined with (5) yields $\langle g^{k+1}, \hat{g}^{k+1} - \hat{g}^k \rangle \ge |\hat{g}^{k+1}|^2 - \kappa |\hat{g}^k|^2$, so using the Cauchy-Schwarz inequality and the bound $|g^{k+1}| \le L$, we obtain

$$L|\hat{g}^{k+1} - \hat{g}^k| \ge |\hat{g}^{k+1}|^2 - \kappa |\hat{g}^k|^2. \tag{9}$$

Now, by (8) with $|\hat{g}^k| \ge \bar{v}$, there is $\hat{v} \ge \bar{v}$ such that $|\hat{g}^k| \to \hat{v}$, and $\hat{g}^{k+1} - \hat{g}^k \to 0$, so in (9), the left-hand side converges to 0, whereas the right-hand side converges to $(1 - \kappa)\hat{v} > 0$ (since $\kappa < 1$). This contradiction implies that $v_k \downarrow 0$ and $\varepsilon_k \downarrow 0$.

Next, suppose $\{x^k\}$ has a cluster point $\bar{x}$. We claim that

$$\underline{\lim}_k \max\{|x^k - \bar{x}|, |\hat{g}^k|\} = 0. \tag{10}$$

This is clear if $x^k \to \bar{x}$, since $v_k \downarrow 0$ and $|\hat{g}^k| \le v_k$ when $v_{k+1} < v_k$, so assume $x^k \not\to \bar{x}$. Arguing by contradiction, suppose (10) is false. Then there exist $\bar{v} > 0$, $\bar{k}$ and an infinite set $K := \{k : k \ge \bar{k}, |x^k - \bar{x}| \le \bar{v}\}$ such that $|\hat{g}^k| > \bar{v}$ for all $k \in K$, so (7) gives $\sum_{k \in K} |x^{k+1} - x^k| < \infty$. Since $x^k \not\to \bar{x}$, there is $\varepsilon > 0$ such that for each $k \in K$ with $|x^k - \bar{x}| \le \bar{v}/2$ there exists $k' > k$ satisfying $|x^{k'} - x^k| > \varepsilon$ and $|x^i - \bar{x}| \le \bar{v}$ for all $k \le i < k'$. Therefore, by the triangle inequality, we have $\varepsilon < |x^{k'} - x^k| \le \sum_{i=k}^{k'-1} |x^{i+1} - x^i|$ with the right side being less than ε for large $k \in K$ from $\sum_{k \in K} |x^{k+1} - x^k| < \infty$, a contradiction. Therefore, (10) must hold also when $x^k \not\to \bar{x}$.

By (10), there exists an infinite set K such that $x^k \xrightarrow{K} \bar{x}$ and $\hat{g}^k \xrightarrow{K} 0$. According to Assumption 2.1(3), for any $\rho > 0$, there exist $\tau > 0$ and $\tilde{\varepsilon} > 0$ such that (4) holds.

Since $\hat{g}^k \in \Gamma_{\varepsilon_k}(x^k)$ by Remark 2.1(a), $x^k \underset{K}{\to} \bar{x}$ and $\varepsilon_k \to 0$, (4) with $x^k \in B(\bar{x}, \tau)$ and $\varepsilon_k \le \bar{\varepsilon}$ for $k \in K$ large enough yields $\hat{g}^k \in \partial_\rho f(\bar{x}) + B(0, \rho)$. Since $\hat{g}^k \underset{K}{\to} 0$, $\rho > 0$ is arbitrary and the mapping $\partial f(\bar{x})$ is closed, $0 \in \partial f(\bar{x})$. □

Remark 3.1 (a) Suppose Step 4 is modified so that if $t_k \ge \varepsilon_k$, then actually $t_k := \arg\max\{t \ge 0 : f(x^k + td^k) \le f(x^k) - \underline{\kappa} t|\hat{g}^k|\}$, where $\underline{\kappa} \in (0, \kappa]$ (cf. Remark 2.1(d)); such t_k exists if $\inf f > -\infty$. Then Theorem 3.1 remains true (by its proof).

(b) For k in $J := \{k : |\hat{g}^k| \le \nu_k\}$, Step 3 may choose $\nu_{k+1}, \varepsilon_{k+1} > 0$ in other ways that ensure $\nu_k, \varepsilon_k \to 0$ when the set J is infinite (see, e.g., [3, Alg. 7.1]). Theorem 3.1 remains true. Indeed, in its proof, suppose J is finite, let $k_1 := 1 + \max_{k \in J} k$ and use $\bar{\nu} := \nu_{k_1}$, $\bar{\varepsilon} := \varepsilon_{k_1}$ to obtain a contradiction from (8), (9) as before. Therefore, J must be infinite, $\nu_k, \varepsilon_k \to 0$ and $\hat{g}^k \underset{J}{\to} 0$, so that $\underline{\lim}_k |\hat{g}^k| = 0$ in the proof of (10).

(c) For the DG method, our Theorem 3.1 is significantly stronger than the result of [3, Thm. 7.1], which assumes additionally that the set $\{x : f(x) \le f(x^1)\}$ is bounded and the modifications of (a), (b) above are employed. This result says only that each accumulation point of the subsequence $\{x^k\}_{k \in J}$ is stationary for f (with $J := \{k : |\hat{g}^k| \le \nu_k\}$). Note that the DG method [3, Alg. 7.1] has three nested loops: an outer loop concerns iterations in J, a middle loop starts/stops whenever descent or stationarity progress occurs, and an inner loop works for consecutive null steps.

(d) As in (c) above, for the secant method [5, Alg. 3] our Theorem 3.1 is significantly stronger than the result of [5, Thm. 2], which assumes additionally that the set $\{x : f(x) \le f(x^1)\}$ is bounded and either Step 4 is modified as in (a) above or it employs the step size expansion of Remark 2.1(d). Here the three nested loops of (c) above correspond to Algorithms 3, 2 and 1 in [5], respectively.

(e) Both DG and secant methods use accumulation (cf. Remark 2.1(e)), for which storage cannot be bounded a priori. In contrast, Algorithm 2.1 may employ aggregation or selection to ensure bounded storage and work per iteration.

4 Extensions

Inspection of the proof of Theorem 3.1 reveals that Theorem 3.1 remains true if Assumption 2.1 on the algorithmic mapping $\gamma(x, d, \varepsilon)$ is replaced by the following.

Assumption 4.1 (1) *If* $d = -\hat{g}/|\hat{g}|$ *and* $f(x + \varepsilon d) - f(x) > -\kappa\varepsilon|\hat{g}|$ *for some* $x, \hat{g} \in \mathbb{R}^n$, $\hat{g} \ne 0$ *and* $\varepsilon > 0$, *then* $\langle \gamma(x, d, \varepsilon), d\rangle \ge -\bar{\kappa}|\hat{g}|$, *where* $\bar{\kappa} \in (\kappa, 1)$ *is fixed.*

(2) *For each* $x \in \mathbb{R}^n$ *and* $\varepsilon > 0$, *there exists* $L < \infty$ *such that* $\sup_{d \in \mathbb{S}} |\gamma(x, d, \varepsilon)| \le L$.

(3) *For each* $\bar{x} \in \mathbb{R}^n$ *and* $\rho > 0$, *there exist* $\tau > 0$ *and* $\bar{\varepsilon} > 0$ *such that the approximation property* (4) *holds, where* $\Gamma_\varepsilon(x) := \operatorname{co} \gamma(x, \mathbb{S}, \varepsilon)$.

Indeed, under Assumption 4.1(1), we may replace κ by $\bar{\kappa}$ in (5), (9) and below.

First, we note that Assumption 4.1 holds in the quasisecant setting of [4], where the secant property (3) is replaced by the following *quasisecant property*

$$f(x + \varepsilon d) - f(x) \le \varepsilon\langle \gamma(x, d, \varepsilon), d\rangle \quad \text{for all } x \in \mathbb{R}^n, d \in \mathbb{S}, \varepsilon > 0, \tag{11}$$

and $\partial f(\bar{x})$ replaces $\partial_\rho f(\bar{x})$ in (4). In effect, Remark 3.1(d) applies to the three algorithms of [4] as well as to those in [5].

We now show how to satisfy Assumption 4.1 in the traditional bundle framework (cf. [12, 14]) where a *subgradient mapping* $g(\cdot) \in \partial f(\cdot)$ is available and the objective f is *upper semidifferentiable* [6], i.e., as noted in [16, Eq. (5)],

$$\overline{\lim_{t \downarrow 0}} \{[f(x+td) - f(x)]/t - \langle g(x+td), d\rangle\} \le 0 \quad \text{for every } x \text{ and } d. \tag{12}$$

Then $\gamma(x,d,\varepsilon)$ may be constructed by the following procedure (cf. [12, p. 103]).

Procedure 4.1 (line search at x along the direction $d = -\hat{g}/|\hat{g}|$ with $\hat{g} \neq 0$)

(a) Set $t_L := 0$ and $t := t_U := \varepsilon$.
(b) If $f(x+td) \le f(x) - \kappa t|\hat{g}|$ set $t_L := t$, otherwise set $t_U := t$.
(c) If $t_L = \varepsilon$ or $\langle g(x+td), d\rangle \ge -\bar{\kappa}|\hat{g}|$, return $\gamma(x,d,\varepsilon) := g(x+td)$.
(d) Choose $t \in [t_L + 0.1(t_U - t_L), t_U - 0.1(t_U - t_L)]$ and go to (b).

Thus, by construction $\gamma(x,d,\varepsilon) := g(x+td)$ satisfies Assumption 4.1(1) with $t \le \varepsilon$. Hence, for Assumption 4.1(2) we may take L as the Lipschitz constant of f on $B(x, 2\varepsilon)$, and the inclusion in (4) holds when $\tau + \bar{\varepsilon} \le \rho$.

References

1. Audet, C., Béchard, V., Le Digabel, S.: Nonsmooth optimization through mesh adaptive direct search and variable neighborhood search. J. Global Optim. **41**, 299–318 (2008)
2. Audet, C., Dennis Jr., J.E.: Mesh adaptive direct search algorithms for constrained optimization. SIAM J. Optim. **17**, 188–217 (2006)
3. Bagirov, A.M., Karazözen, B., Sezer, M.: Discrete gradient method: Derivative-free method for nonsmooth optimization. J. Optim. Theory Appl. **137**, 317–334 (2008)
4. Bagirov, A.M., Nazari Ganjehlou, A.: A quasisecant method for minimizing nonsmooth functions. Tech. rep., Centre for Informatics and Applied Optimization, University of Ballarat, Victoria, Australia (2009). Available at the URL http://www.optimization-online.org/DB_FILE/2009/03/2251.pdf.
5. Bagirov, A.M., Nazari Ganjehlou, A.: A secant method for nonsmooth optimization. Tech. rep., Centre for Informatics and Applied Optimization, University of Ballarat, Victoria, Australia (2009)
6. Bihain, A.: Optimization of upper semidifferentiable functions. J. Optim. Theory Appl. **44**, 545–568 (1984)
7. Burke, J.V., Lewis, A.S., Overton, M.L.: A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. SIAM J. Optim. **15**, 751–779 (2005)
8. Clarke, F.H.: Optimization and Nonsmooth Analysis. Wiley, New York (1983)
9. Frangioni, A.: Solving semidefinite quadratic problems within nonsmooth optimization algorithms. Comput. Oper. Res. **23**, 1099–1118 (1996)
10. Goldstein, A.A.: Optimization of Lipschitz continuous functions. Math. Programming **13**, 14–22 (1977)
11. Hiriart-Urruty, J.B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms. Springer, Berlin (1993)
12. Kiwiel, K.C.: Methods of Descent for Nondifferentiable Optimization. Lecture Notes in Mathematics 1133. Springer-Verlag, Berlin (1985)
13. Kiwiel, K.C.: A Cholesky dual method for proximal piecewise linear programming. Numer. Math. **68**, 325–340 (1994)
14. Kiwiel, K.C.: Restricted step and Levenberg-Marquardt techniques in proximal bundle methods for nonconvex nondifferentiable optimization. SIAM J. Optim. **6**, 227–249 (1996)
15. Kiwiel, K.C.: Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. SIAM J. Optim. **18**, 379–388 (2007)
16. Lemaréchal, C.: An introduction to the theory of nonsmooth optimization. Optimization **17**, 827–858 (1986)