Raport Badawczy

Research Report

RB/53/2006

Convergence of the gradient sampling algorithm for nonsmooth nonconvex 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) 8373578

fax: (+48) (22) 8372772

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

Warszawa 2006

CONVERGENCE OF THE GRADIENT SAMPLING ALGORITHM FOR NONSMOOTH NONCONVEX OPTIMIZATION

KRZYSZTOF C. KIWIEL*

Abstract. We study the gradient sampling algorithm of Burke, Lewis and Overton for minimizing a locally Lipschitz function f on $\mathbb{R}^n$ that is continuously differentiable on an open dense subset. We strengthen the existing convergence results for this algorithm, and introduce a slightly revised version for which stronger results are established without requiring compactness of the level sets of f. In particular, we show that with probability 1 the revised algorithm either drives the f-values to $-\infty$, or each of its cluster points is Clarke stationary for f. We also consider a simplified variant in which the differentiability check is skipped and the user can control the number of f-evaluations per iteration.

Key words. generalized gradient, nonsmooth optimization, subgradient, gradient sampling, nonconvex.

AMS subject classifications. 65K10, 90C26

1. Introduction.

In two recent papers [BLO02b, BLO05], Burke, Lewis and Overton introduced and established convergence of the *gradient sampling* (GS) algorithm for minimizing a locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ which is continuously differentiable on an open dense subset D of $\mathbb{R}^n$ and has bounded level sets.

At each iteration, the GS algorithm computes the gradient of f at the current iterate and at $m \geq n+1$ randomly generated nearby points. This bundle of gradients is used to find an approximate ϵ-steepest descent direction, where ϵ is the sampling radius, as the solution of a quadratic program. A standard Armijo line search along this direction produces a candidate for the next iterate, which is obtained by perturbing the candidate if necessary to stay in the set D where f is differentiable; the perturbation is random and small enough to maintain the Armijo sufficient descent property. The sampling radius ϵ may be fixed for all iterations or may be reduced dynamically. For ϵ fixed, the main convergence resnlt of [BLO05, Theorem 3.4] established that, with probability 1, the GS algorithm generates a sequence with a cluster point that is ϵ-stationary for f (as defined in section 2). For ϵ reduced dynamically, the result of [BLO05, Theorem 3.8] established that if the GS algorithm converges to a point, this limit point is stationary for f with probability 1.

The GS algorithm is not only very interesting in theory (especially due to its ingenious use of gradients instead of subgradients [BLO02a]), but also widely applicable and robust in practice [BHLO06, BLO04, BLO05, Lew05].

This paper provides stronger convergence results for the GS algorithm. For ϵ fixed, we show that with probability 1 *every* cluster point of the GS algorithm is ϵ-stationary for f (see Theorem 3.6). For ϵ reduced dynamically, we show that with probability 1 every cluster point of a well-defined subsequence is stationary for f (see Theorem 3.4), without assuming that the whole sequence converges. In both cases, we show that suitable stopping criteria ensure with probability 1 that the algorithm terminates with the required "optimality certificate" of [BLO05, p. 768]; this practical aspect was not analyzed in [BLO05, section 3].

We also introdnce a slight revision of the GS algorithm, in which the perturbation of the Armijo candidate is controlled by the current step size (instead of ϵ as in the

[*] Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01–447 Warsaw, Poland (kiwiel@ibspan.waw.pl).

original method; see (2.6)). This tiny modification enables us to derive much stronger convergence results; in particular, we can dispense with the assumption of [BLO05] that f has compact level sets. For ϵ fixed, we show that with probability 1 either the algorithm drives the f-values to $-\infty$, or every cluster point of a well-defined subsequence of its iterates is ϵ-stationary for f (see Theorem 3.5). For ϵ reduced dynamically, we show that with probability 1 the algorithm either drives the f-values to $-\infty$, or *each* of its cluster points is stationary for f (see Theorem 3.3); in a sense, this is the best result one can hope for. If $\inf f > -\infty$, in both cases suitable stopping criteria ensure with probability 1 that the algorithm terminates with the required "optimality certificate".

Our further modifications of the GS algorithm are intended to improve its performance in practice. Since the GS algorithm employs search directions of unit 2-norm, the number of f-evaluations per Armijo's line search can grow to infinity as the algorithm converges. To mitigate this drawback, we consider using an "unscaled" search direction, i.e., the negative of the convex combination of the gradients in the bundle whose norm is minimized. (This direction was used in [BLO02b, section 3] for a different line search.) The third alternative is to scale the direction so that its length equals ϵ and the Armijo liue search is made within the ball in which gradient sampling occurs.

Finally, we introduce a lower bound on step sizes tested by the Armijo search, accepting a null step size when this bound is reached. Here the idea is simple: when the search direction is good enough, a step size close to our lower bound should work, whereas if the search direction is poor, the Armijo search will produce a tiny step size auyway. In our limited Armijo line search (see Proc. 4.3) the number of f-evaluations can be controlled by the choice of an initial step size; in an extreme version, just one evaluation occurs. Further, for our limited line search there is no longer any need for keeping the iterates in the set D where f is differentiable. Skipping the differentiability check makes life easier for the user who provides gradient values and brings the simplified algorithm closer to the version implemented and tested in [BLO05, section 4].

Among other algorithms for minimizing locally Lipschitz functions, we should mention bundle methods (see the references in [BLO05, Kiw96]). Bundle methods require the computation of a single subgradient at each trial point in addition to the objective value. They generate search directions by solving quadratic programs based on accumulated subgradients, and employ line searches which either produce descent or find a new subgradient that modifies the next search direction. At first sight, they have little in common with the GS algorithm, which does not accumulate gradients. We believe, however, that deeper understanding of their similarities and differences should lead to new variants. The first step in this direction is made here: the proof technique of section 3 is borrowed from [Kiw96, section 3] and the limited line search of section 4.3 is inspired by null steps of bundle methods. We defer consideration of gradient sampling in bundle methods, as well as numerical comparisons, to future work.

The paper is organized as follows. A slightly revised version of the GS algorithm is presented in section 2. A convergence analysis of the original and revised versions is given in section 3. Various modifications are discussed in section 4.

2. The gradient sampling algorithm. As in [BLO05], we assume that the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous and continuously differentiable on an open dense subset D of $\mathbb{R}^n$. The Clarke *subdifferential* [Cla83] of f at any point x is given by

$$\bar{\partial} f(x) = \operatorname{co}\{\, \lim_j \nabla f(y^j) : y^j \to x, y^j \in D \,\},$$

where co denotes the convex hull, and the Clarke *ϵ-subdifferential* [Gol77] by

$$\bar{\partial}_\epsilon f(x) := \operatorname{co} \bar{\partial} f(B(x, \epsilon)), \tag{2.1}$$

where $B(x, \epsilon) := \{y : |y - x| \leq \epsilon\}$ is the ball centered at x with radius $\epsilon \geq 0$ and $|\cdot|$ is the 2-norm. The Clarke ϵ-subdifferential $\bar{\partial}_\epsilon f(x)$ is approximated by the set of [BLO05]

$$G_\epsilon(x) := \operatorname{cl} \operatorname{co} \nabla f(B(x, \epsilon) \cap D), \tag{2.2}$$

since $G_\epsilon(x) \subset \bar{\partial}_\epsilon f(x)$, and $\bar{\partial}_{\epsilon_1} f(x) \subset G_{\epsilon_2}(x)$ for $0 \leq \epsilon_1 < \epsilon_2$. We say that a point x is *stationary* for f if $0 \in \bar{\partial} f(x)$; x is called *ϵ-stationary* for f if $0 \in \bar{\partial}_\epsilon f(x)$.

We now state a slightly revised version of the GS algorithm [BLO05, section 2]. In particular, we don't require that the starting point $x^1 \in D$ is such that the level set $\{x : f(x) \leq f(x^1)\}$ is compact. For a closed convex set G, $\operatorname{Proj}(0 \,|\, G)$ is its minimum-norm element.

ALGORITHM 2.1 (*Revised GS algorithm*).

Step 0 (initialization). Select an initial point $x^1 \in D$, optimality tolerances $\nu_{\text{opt}}, \epsilon_{\text{opt}} \geq 0$, line search parameters β, γ in $(0, 1)$, reduction factors μ, θ in $(0, 1]$, a sampling radius $\epsilon_1 > 0$, a stationarity target $\nu_1 \geq 0$ and a sample size $m \geq n + 1$. Set $k := 1$.

Step 1 (approximate the Clarke ϵ-subdifferential by gradient sampling). Let $\{x^{ki}\}_{i=1}^m$ be sampled independently and uniformly from $B(x^k, \epsilon_k)$. If $\{x^{ki}\}_{i=1}^m \not\subset D$, then stop; otherwise, set

$$G_k := \operatorname{co}\{\, \nabla f(x^k), \nabla f(x^{k1}), \ldots, \nabla f(x^{km}) \,\}. \tag{2.3}$$

Step 2 (direction finding). Set $g^k := \operatorname{Proj}(0 \,|\, G_k)$.

Step 3 (stopping criterion). If $|g^k| \leq \nu_{\text{opt}}$ and $\epsilon_k \leq \epsilon_{\text{opt}}$, terminate.

Step 4 (sampling radius update). If $|g^k| \leq \nu_k$, set $\nu_{k+1} := \theta \nu_k$, $\epsilon_{k+1} := \mu \epsilon_k$, $t_k := 0$, $x^{k+1} := x^k$ and go to Step 7. Otherwise, set $\nu_{k+1} := \nu_k$, $\epsilon_{k+1} := \epsilon_k$ and

$$d^k := -g^k / |g^k|. \tag{2.4}$$

Step 5 (line search). Set the step size

$$t_k := \max\{\, t : f(x^k + t d^k) < f(x^k) - \beta t |g^k|, t \in \{1, \gamma, \gamma^2, \ldots\} \,\}. \tag{2.5}$$

Step 6 (updating). If $x^k + t_k d^k \in D$, set $x^{k+1} := x^k + t_k d^k$. Otherwise, let x^{k+1} be any point in D satisfying

$$f(x^{k+1}) < f(x^k) - \beta t_k |g^k|, \tag{2.6a}$$

$$|x^k + t_k d^k - x^{k+1}| \leq \min\{t_k, \epsilon_k\}. \tag{2.6b}$$

Step 7. Increase k by 1 and go to Step 1.

The algorithm keeps every iterate x^k in the set D. At Step 2, g^k is characterized by $g^k \in G_k$ and $\langle g, g^k \rangle \geq |g^k|^2$ for all $g \in G_k$; since $\nabla f(x^k) \in G_k$ by (2.3), (2.4) yields

$\langle \nabla f(x^k), d^k \rangle \leq -|g^k|$. Hence the Armijo line search (2.5) is well defined, because there is $\bar{t} > 0$ such that $f(x^k + td^k) < f(x^k) - \beta t|g^k|\ \forall t \in (0, \bar{t})$.

The only significant difference between Algorithm 2.1 and the original GS algorithm [BLO05, section 2] lies in the slightly stronger requirement (2.6). Namely, if $x^k + t_k d^k \notin D$, x^{k+1} can be found as follows. For $i = 1, 2, \ldots$, sample x^{k+1} from a uniform distribution on $B(x^k + t_k d^k, \min\{t_k, \epsilon_k\}/i)$ until $x^{k+1} \in D$ and (2.6a) holds. By (2.5) and the continuity of f, this procedure terminates with probability 1. In contrast, the original GS algorithm requires finding $\hat{x}^k$ in $B(x^k, \epsilon_k)$ such that $\hat{x}^k + t_k d^k \in D$ and (2.6a) holds for $x^{k+1} := \hat{x}^k + t_k d^k$; to this end, one can sample $\hat{x}^k$ from a uniform distribution on $B(x^k, \epsilon_k/i)$ until these requirements are met. Further, if (2.6) holds, then $\hat{x}^k := x^{k+1} - t_k d^k$ satisfies the requirements of the original GS algorithm. This is the only reason for including ϵ_k in (2.6b). On the other hand, the presence of t_k in (2.6b) yields $|x^{k+1} - x^k| \leq 2t_k$ (using $|d^k| = 1$ by (2.4)) and hence the highly useful consequence of (2.6a)

$$f(x^{k+1}) \leq f(x^k) - \beta\tfrac{1}{2}|x^{k+1} - x^k||g^k|. \tag{2.7}$$

Note that this key inequality (2.7) holds also when $x^{k+1} := x^k + t_k d^k$ at Step 6 (thanks to (2.5)), or when $x^{k+1} := x^k$ at Step 4.

The stopping criterion of Step 3 delivers the "optimality certificate" of [BLO05, p. 768]: the final values of $|g^k|$ and ϵ_k provide an estimate of nearness to Clarke stationarity.

3. Convergence analysis. We start with two techuical lemmas. The first lemma on approximate least-norm elemeuts is a simplified version of [BLO05, Lemma 3.1].

Lemma 3.1. *Let $\emptyset \neq C \subset \mathbb{R}^n$ be compact convex and $\beta \in (0, 1)$. If $0 \notin C$, there exists $\delta > 0$ such that $u, v \in C$ and $|u| \leq \operatorname{dist}(0 \,|\, C) + \delta$ imply $\langle v, u \rangle > \beta|u|^2$.*

Proof. If the assertion were false, we could pick two sequences $\{u^i\}, \{v^i\} \subset C$ satisfying $|u^i| \leq \operatorname{dist}(0 \,|\, C) + 1/i$ and $\langle v^i, u^i \rangle \leq \beta|u^i|^2$. By compactness, we may assume $u^i \to \bar{u} \in C$, $v^i \to \bar{v} \in C$; thus $\langle \bar{v}, \bar{u} \rangle \leq \beta|\bar{u}|^2$. However, $\bar{u} = \operatorname{Proj}(0 \,|\, C) \neq 0$ satisfies $\langle v, \bar{u} \rangle \geq |\bar{u}|^2$ for all $v \in C$, a contradiction. □

The next lemma recalls from [BLO05, Lemma 3.2] basic properties of the set of points close to a given point $\bar{x}$ that can be used to provide a δ-approximation to the least-norm element of $G_\epsilon(\bar{x})$; its second part summarizes some useful ideas from the proof of [BLO05, Theorem 3.4]. For $\epsilon, \delta > 0$ and $\bar{x}, x \in \mathbb{R}^n$, using the measure of proximity to ϵ-stationarity

$$\rho_\epsilon(\bar{x}) := \operatorname{dist}(0 \,|\, G_\epsilon(\bar{x})), \tag{3.1}$$

let

$$D_\epsilon^m(x) := \prod_1^m (B(x, \epsilon) \cap D) \subset \prod_1^m \mathbb{R}^n \tag{3.2}$$

and

$$V_\epsilon(\bar{x}, x, \delta) := \left\{ (y^1, \ldots, y^m) \in D_\epsilon^m(x) : \operatorname{dist}(0 \,|\, \operatorname{co}\{\nabla f(y^i)\}_{i=1}^m) \leq \rho_\epsilon(\bar{x}) + \delta \right\}. \tag{3.3}$$

Lemma 3.2. *Let $\epsilon > 0$ and $\bar{x} \in \mathbb{R}^n$.*

(i) *For any $\delta > 0$, there is $\tau > 0$ and a nonempty open set $\bar{V}$ satisfying $\bar{V} \subset V_\epsilon(\bar{x}, x, \delta)$ for all $x \in B(\bar{x}, \tau)$, and $\operatorname{dist}(0 \,|\, \operatorname{co}\{\nabla f(y^i)\}_{i=1}^m) \leq \rho_\epsilon(\bar{x}) + \delta$ for all $(y^1, \ldots, y^m) \in \bar{V}$.*

(ii) *Assuming* $0 \notin G_\epsilon(\bar{x})$, *pick* $\delta > 0$ *as in Lemma* 3.1 *for* $C := G_\epsilon(\bar{x})$, *and then* τ *and* $\bar{V}$ *as in statement* (i). *Suppose at iteration* k *of Algorithm* 2.1, *Step* 5 *is reached with* $x^k \in B(\bar{x}, \min\{\tau, \epsilon/3\})$, $\epsilon_k = \epsilon$ *and* $(x^{k1}, \dots, x^{km}) \in \bar{V}$. *Then* $t_k \geq \min\{1, \gamma\epsilon/3\}$.

(iii) *If* $\underline{\lim}_k \max\{|x^k - \bar{x}|, |g^k|, \epsilon_k\} = 0$ *with* $g^k \in \bar{\partial}_{\epsilon_k} f(x^k)$ *for all* k, *then* $0 \in \bar{\partial} f(\bar{x})$.

Proof. (i) Let $u \in \operatorname{co} \nabla f(B(\bar{x}, \epsilon) \cap D)$ be such that $|u| < \rho_\epsilon(\bar{x}) + \delta$. Then Carathéodory's theorem [Roc70] implies the existence of $(\bar{x}^1, \dots, \bar{x}^m) \in D_\epsilon^m(\bar{x})$ and $\bar{\lambda} \in \mathbb{R}_+^m$ with $\sum_{i=1}^m \bar{\lambda}_i = 1$ such that $u = \sum_{i=1}^m \bar{\lambda}_i \nabla f(\bar{x}^i)$. Since f is continuously differentiable on the open set D, there is $\bar{\epsilon} \in (0, \epsilon)$ such that the set $\bar{V} := \prod_{i=1}^m \operatorname{int} B(\bar{x}^i, \bar{\epsilon})$ lies in $D_{\epsilon - \bar{\epsilon}}^m(\bar{x})$ and $|\sum_{i=1}^m \bar{\lambda}_i \nabla f(y^i)| < \rho_\epsilon(\bar{x}) + \delta$ for all $(y^1, \dots, y^m) \in \bar{V}$. Hence for all $x \in B(\bar{x}, \tau)$ with $\tau := \bar{\epsilon}$, the fact that $B(\bar{x}, \epsilon - \bar{\epsilon}) \subset B(x, \epsilon)$ yields $\bar{V} \subset V_\epsilon(\bar{x}, x, \delta)$ by the definitions (3.2)–(3.3).

(ii) Let $\hat{G}_k := \operatorname{co}\{\nabla f(x^{ki})\}_{i=1}^m$. Since $(x^{k1}, \dots, x^{km}) \in \bar{V} \subset V_\epsilon(\bar{x}, \bar{x}, \delta)$ in statement (i), we get $\operatorname{dist}(0 \,|\, \hat{G}_k) \leq \rho_\epsilon(\bar{x}) + \delta$ and $\hat{G}_k \subset G_\epsilon(\bar{x})$ from (3.3), (3.2) and (2.2). We also have $\nabla f(x^k) \in G_\epsilon(\bar{x})$ from $x^k \in B(\bar{x}, \epsilon/3) \cap D$. Thus, by (2.3) and the construction of g^k at Step 2, $g^k \in G_\epsilon(\bar{x})$ and $|g^k| \leq \rho_\epsilon(\bar{x}) + \delta$. Hence by (3.1) and the choice of δ in Lemma 3.1,

$$\langle v, g^k \rangle > \beta |g^k|^2 \quad \text{for all } v \in G_\epsilon(\bar{x}). \tag{3.4}$$

Suppose for contradiction that $t_k < \min\{1, \gamma\epsilon/3\}$. Then by construction (cf. (2.5))

$$-\beta\gamma^{-1} t_k |g^k| \leq f(x^k + \gamma^{-1} t_k d^k) - f(x^k),$$

whereas Lebourg's mean value theorem (cf. [Cla83, Theorem 2.3.7]) yields the existence of $\tilde{x}^k \in [x^k + \gamma^{-1} t_k d^k, x^k]$ and $v^k \in \bar{\partial} f(\tilde{x}^k)$ such that

$$f(x^k + \gamma^{-1} t_k d^k) - f(x^k) = \gamma^{-1} t_k \langle v^k, d^k \rangle.$$

Hence using $d^k := -g^k/|g^k|$ gives $\langle v^k, g^k \rangle \leq \beta |g^k|^2$, so $v^k \notin G_\epsilon(\bar{x})$ by (3.4). But $\gamma^{-1} t_k |d^k| < \epsilon/3$ and $|x^k - \bar{x}| \leq \epsilon/3$ imply $\tilde{x}^k \in B(\bar{x}, 2\epsilon/3)$ and thus $v^k \in G_\epsilon(\bar{x})$, a contradiction.

(iii) Note that $g^k \in \bar{\partial}_{\epsilon_k} f(x^k)$ at Step 2 by (2.1), whereas $\bar{\partial}_\cdot f(\cdot)$ is closed. □

As discussed in section 2, Algorithm 2.1 is a special case of the GS algorithm, which in turn corresponds to removing t_k in the right-hand side of (2.6b) and requiring that the level set $\{x : f(x) \leq f(x^1)\}$ be bounded. Therefore, we give convergence results separately for Algorithm 2.1 and the original GS algorithm. We start with the case where ϵ_k and ν_k are allowed to decrease.

THEOREM 3.3. *Let* $\{x^k\}$ *be a sequence generated by Algorithm* 2.1 *with* $\nu_1 > \nu_{\mathrm{opt}} = \epsilon_{\mathrm{opt}} = 0$ *and* $\mu, \theta < 1$. *With probability* 1 *the algorithm doesn't stop and either* $f(x^k) \downarrow -\infty$, *or* $\nu_k \downarrow 0$, $\epsilon_k \downarrow 0$ *and every cluster point of* $\{x^k\}$ *is stationary for* f.

Proof. (i) Since termination in Step 1 has zero probability, we may assume it doesn't occur. Similarly, if $f(x^k) \downarrow -\infty$, there is nothing to prove, so assume $\inf_k f(x^k) > -\infty$. Then summing $\beta t_k |g^k| \leq f(x^k) - f(x^{k+1})$ (cf. (2.6a)) and relation (2.7) gives

$$\sum_{k=1}^{\infty} t_k |g^k| < \infty, \tag{3.5}$$

$$\sum_{k=1}^{\infty} |x^{k+1} - x^k| |g^k| < \infty. \tag{3.6}$$

(ii) Suppose there is k_1, $\bar\nu > 0$ and $\bar\epsilon > 0$ such that $\nu_k = \bar\nu$ and $\epsilon_k = \bar\epsilon$ for all $k \ge k_1$. Using $|g^k| \ge \bar\nu$ in (3.5)–(3.6) yields $t_k \to 0$, $\sum_k |x^{k+1} - x^k| < \infty$, and hence the existence of a point $\bar x$ such that $x^k \to \bar x$. Let $\epsilon := \bar\epsilon$. First, suppose $0 \notin G_\epsilon(\bar x)$. For δ, τ and $\bar V$ chosen as in Lemma 3.2(ii), we can pick $k_2 \ge k_1$ such that $x^k \in B(\bar x, \min\{\tau, \epsilon/3\})$ and $t_k < \min\{1, \gamma\epsilon/3\}$ yield $(x^{k1}, \ldots, x^{km}) \notin \bar V$ for all $k \ge k_2$. This event has probability 0, since for each $k \ge k_2$, $(x^{k1}, \ldots, x^{km})$ is sampled independently and uniformly from $D_\epsilon^m(x^k)$, which contains the open set $\bar V \ne \emptyset$. Second, suppose $0 \in G_\epsilon(\bar x)$. For $\delta := \bar\nu/2$ and τ, $\bar V$ chosen as in Lemma 3.2(i), we can pick $k_3 \ge k_1$ such that $x^k \in B(\bar x, \tau)$, $\bar\nu \le |g^k| \le \operatorname{dist}(0 \,|\, \operatorname{co}\{\nabla f(x^{ki})\}_{i=1}^m)$ and $\rho_\epsilon(\bar x) = 0$ imply $(x^{k1}, \ldots, x^{km}) \notin \bar V$ for all $k \ge k_3$. This event has probability 0 as well.

(iii) Consider the event where $\nu_k \downarrow 0$, $\epsilon_k \downarrow 0$ and $\{x^k\}$ has a cluster point $\bar x$. If $x^k \to \bar x$, $0 \in \bar\partial f(\bar x)$ by Lemma 3.2(iii). If $x^k \not\to \bar x$, we claim that $\underline{\lim}_k \max\{|x^k - \bar x|, |g^k|\} = 0$. Otherwise, there exist $\bar\nu > 0$, $\bar k$ and an infinite set $K := \{k : k \ge \bar k, |x^k - \bar x| \le \bar\nu\}$ such that $|g^k| > \bar\nu$ for all $k \in K$, so (3.6) gives $\sum_{k\in K} |x^{k+1} - x^k| < \infty$. Since $x^k \not\to \bar x$, there is $\epsilon > 0$ such that for each $k \in K$ with $|x^k - \bar x| \le \bar\nu/2$ there exists $k' > k$ satisfying $|x^{k'} - x^k| > \epsilon$ and $|x^i - \bar x| \le \bar\nu$ for all $k \le i < k'$. Therefore, by the triangle inequality, we have $\epsilon < |x^{k'} - x^k| \le \sum_{i=k}^{k'-1} |x^{i+1} - x^i|$ 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, $\underline{\lim}_k \max\{|x^k - \bar x|, |g^k|\} = 0$ yields $0 \in \bar\partial f(\bar x)$ by Lemma 3.2(iii). □

THEOREM 3.4. *Let $\{x^k\}$ be a sequence generated by the original GS algorithm with $\nu_1 > \nu_{\rm opt} = \epsilon_{\rm opt} = 0$ and $\mu, \theta < 1$. Suppose the level set $\{x : f(x) \le f(x^1)\}$ is bounded. Then with probability 1 the algorithm doesn't stop, $\nu_k \downarrow 0$, $\epsilon_k \downarrow 0$, there is a subsequence $K \subset \{1, 2, \ldots\}$ such that $g^k \xrightarrow{K} 0$ and every cluster point of $\{x^k\}_{k\in K}$ is stationary for f.*

Proof. It suffices to reconsider part (ii) of the proof of Theorem 3.3 (since for $\nu_k \downarrow 0$, we can take $K := \{k : \nu_{k+1} < \nu_k\}$).

Thus suppose there is k_1, $\bar\nu > 0$ and $\bar\epsilon > 0$ such that $\nu_k = \bar\nu$ and $\epsilon_k = \bar\epsilon$ for all $k \ge k_1$. Using $|g^k| \ge \bar\nu$ in (3.5) yields $t_k \to 0$. Since $\{f(x^k)\}$ is decreasing and the set $\{x : f(x) \le f(x^1)\}$ is compact, there are a set $J \subset \{1, 2, \ldots\}$ and a point $\bar x$ such that $x^k \xrightarrow{J} \bar x$. Since $t_k \xrightarrow{J} 0$ as well, arguing as in part (ii) of the proof of Theorem 3.3 we deduce the existence of k_4 and an open set $\bar V \ne \emptyset$ such that $(x^{k1}, \ldots, x^{km}) \notin \bar V \subset D_\epsilon^m(x^k)$ for all $k \ge k_4$, $k \in J$, and again conclude that this event has probability 0. □

Our convergence results for fixed sampling radius follow.

THEOREM 3.5. *Let $\{x^k\}$ be a sequence generated by Algorithm 2.1 with $\nu_1 = \nu_{\rm opt} = 0$, $\epsilon_1 = \epsilon_{\rm opt} = \epsilon > 0$ and $\mu = 1$. With probability 1 either the algorithm terminates at some iteration k with $0 \in G_\epsilon(x^k)$, or $f(x^k) \downarrow -\infty$, or there is a subsequence $K \subset \{1, 2, \ldots\}$ such that $g^k \xrightarrow{K} 0$ and every cluster point $\bar x$ of $\{x^k\}_{k\in K}$ satisfies $0 \in \bar\partial_\epsilon f(\bar x)$.*

Proof. If the algorithm terminates at iteration k, then with probability 1 it does so at Step 3 with $0 = g^k \in G_\epsilon(x^k)$. Hence we may assume that no termination occurs and $\inf_k f(x^k) > -\infty$.

By the proof of Theorem 3.3, the event $\bar\nu := \inf_k |g^k| > 0$ has probability 0. In the remaining case of $\inf_k |g^k| = 0$, the conclusion follows from the closedness of $\bar\partial_\epsilon f(\cdot)$. □

THEOREM 3.6. *Let $\{x^k\}$ be a sequence generated by the original GS algorithm with $\nu_1 = \nu_{\rm opt} = 0$, $\epsilon_1 = \epsilon_{\rm opt} = \epsilon > 0$ and $\mu = 1$. Suppose the set $\{x : f(x) \le f(x^1)\}$*

is bounded. With probability 1 *either the algorithm terminates at some iteration* k *with* $0 \in G_\epsilon(x^k)$, *or* $g^k \to 0$ *and every cluster point* $\bar{x}$ *of* $\{x^k\}$ *satisfies* $0 \in \bar{\partial}_\epsilon f(\bar{x})$.

Proof. Arguing by contradiction, it suffices to consider the case where there are a set $J \subset \{1, 2, \ldots\}$ and $\bar{\nu} > 0$ such that $\inf_{k \in J} |g^k| \geq \bar{\nu}$. Since $\{f(x^k)\}$ is decreasing and the set $\{x : f(x) \leq f(x^1)\}$ is compact, we may assume with no loss of generality that there is a point $\bar{x}$ such that $x^k \xrightarrow{J} \bar{x}$. Since (3.5) gives $t_k \xrightarrow{J} 0$, arguing as in part (ii) of the proof of Theorem 3.3 we deduce the existence of k_5 and an open set $\bar{V} \neq \emptyset$ such that $(x^{k1}, \ldots, x^{km}) \notin \bar{V} \subset D_\epsilon^m(x^k)$ for all $k \geq k_5$, $k \in J$. This event has probability 0, since for each k, $(x^{k1}, \ldots, x^{km})$ is sampled independently and uniformly from $D_\epsilon^m(x^k)$. □

A few comments and comparisons with the results of [BLO05, section 3] are in order.

Remark 3.7.

(i) Since the framework of [BLO05, section 3] requires compactness of the level set $\{x : f(x) \leq f(x^1)\}$, it has no results comparable to our Theorems 3.3 and 3.5.

(ii) Theorem 3.3 is essentially the best one can hope for. In particular, it implies that for positive optimality tolerances ν_{opt} and ϵ_{opt}, with probability 1 either $f(x^k) \downarrow -\infty$ or the algorithm terminates with the required "optimality certificate" of [BLO05, p. 768].

(iii) Theorem 3.3 is stronger than Theorem 3.4. Of course, Theorem 3.3 relies on our inclusion of t_k in the right-hand side of (2.6b), but this should be cheap in practice. With this fairly mild qualification, Theorem 3.3 gives a positive answer to the final open question of [BLO05, section 3] on whether all cluster points of the algorithm are stationary.

(iv) Theorem 3.4 subsumes [BLO05, Theorem 3.8], which assumes that $\{x^k\}$ converges.

(v) Theorem 3.6 subsumes [BLO05, Theorem 3.4], which only asserts the existence of a subsequence $K \subset \{1, 2, \ldots\}$ such that $\rho_\epsilon(x^k) \xrightarrow{K} 0$ and every cluster point $\bar{x}$ of $\{x^k\}_{k \in K}$ satisfies $0 \in \bar{\partial}_\epsilon f(\bar{x})$, without showing that $\inf_k |g^k| = 0$. In contrast, Theorem 3.6 implies that for a positive optimality tolerance ν_{opt}, with probability 1 the algorithm terminates when the required "optimality certificate" is reached (similarly for Theorem 3.5 if $\inf f > -\infty$). A result similar to Theorem 3.6 is given in [BLO05, Cor. 3.5.1] only for the case where the objective f is continuously differentiable everywhere. Finally, Theorem 3.6 disproves the conjecture raised in the open question number 2 at the end of [BLO05, section 3] that a counterexample with $\overline{\lim}_{k \in K} |g^k| > 0$ should exist.

4. Modifications.

Although our revision of Step 6 yields stronger theoretical results, it makes no difference to its implementation in practice when, as explained in [BLO05, section 4], it is not possible or practical to check whether the iterates lie in the set D where f is differentiable. Further, the implementation of [BLO05, section 4] obtained best results for the Armijo parameter $\beta = 0$ (although $\beta > 0$ is required in theory). Thus there is still the need for further study of line searches. In this section we propose several themes, supported by theory, that might prove useful in improving the practical performance of the method.

4.1. Non-normalized search directions.

Since the GS algorithm employs search directions $d^k := -g^k/|g^k|$ of unit norm, the number of f-evaluations per Armijo's line search (cf. (2.5)) can grow to infinity. This will happen in the generic case where $x^{k+1} = x^k + t_k d^k$ for almost all k and $t_k = |x^{k+1} - x^k| \to 0$ (e.g., $\{x^k\}$ converges). To mitigate this drawback, let's consider using $d^k := -g^k$ as in the steepest descent method with $d^k = -\nabla f(x^k)$ in the smooth case.

Formally, suppose relations (2.4)–(2.6) in Algorithm 2.1 are replaced by

$$d^k := -g^k, \tag{4.1}$$

$$t_k := \max\{\, t : f(x^k + td^k) < f(x^k) - \beta t|g^k|^2, t \in \{1, \gamma, \gamma^2, \ldots\}\,\}, \tag{4.2}$$

$$f(x^{k+1}) < f(x^k) - \beta t_k|g^k|^2, \tag{4.3a}$$

$$|x^k + t_k d^k - x^{k+1}| \le \min\{t_k, \epsilon_k\}|d^k|. \tag{4.3b}$$

Then (2.7) still holds, since $|x^{k+1} - x^k| \le 2t_k|d^k| = 2t_k|g^k|$. Lemma 3.2(ii) is replaced by

Lemma 4.1. *Let $\epsilon > 0$ and $\bar{x} \in \mathbb{R}^n$. Assuming $0 \notin G_\epsilon(\bar{x})$, pick $\delta > 0$ as in Lemma 3.1 for $C := G_\epsilon(\bar{x})$, and then τ and $\bar{V}$ as in Lemma 3.2(i). Suppose $x^k \in B(\bar{x}, \min\{\tau, \epsilon/3\})$, $\epsilon_k = \epsilon$ and $(x^{k1}, \ldots, x^{km}) \in \bar{V}$. Then $t_k \ge \min\{1, \gamma\epsilon/3\bar{\kappa}\}$, where $\bar{\kappa}$ is the Lipschitz constant of f on $B(\bar{x}, 2\epsilon)$.*

Proof. In the proof of Lemma 3.2(ii), assuming $t_k < \min\{1, \gamma\epsilon/3\bar{\kappa}\}$, use $d^k := -g^k$ to get $\langle v^k, g^k\rangle \le \beta|g^k|^2$ as before. Since $\gamma^{-1}t_k|d^k| < \epsilon/3$ yields $v^k \in G_\epsilon(\bar{x})$ as before, note that $|d^k| = |g^k| \le \bar{\kappa}$, since $|x^k - \bar{x}| \le \epsilon/3$ implies $g^k \in G_\epsilon(x^k) \subset G_{1.5\epsilon}(\bar{x})$ and hence $|g^k| \le \bar{\kappa}$. □

With the above replacements, the proofs of section 3 are modified in obvious ways. For instance, in the proof of Theorem 3.3, using (4.3a), we can replace (3.5) by

$$\sum_{k=1}^{\infty} t_k|g^k|^2 < \infty, \tag{4.4}$$

and in its part (ii) we can consider $t_k < \min\{1, \gamma\epsilon/3\bar{\kappa}\}$. In effect, Theorems 3.3–3.6 hold for this variant as well.

Although $d^k := -g^k$ may be better than $d^k := -g^k/|g^k|$ asymptotically, it can be worse initially when $|g^k|$ is still "large" (this, of course, depends on problem scaling). In general, we may wish to scale d^k so that the first trial point $x^k + d^k$ is at a "reasonable" distance from x^k; using the sampling radius ϵ_k as this distance gives the variant analyzed below.

4.2. Searching within the trust region. To restrict the Armijo line search to the sampled trust region $B(x^k, \epsilon_k)$, suppose relations (2.4)–(2.6) in Algorithm 2.1 are replaced by

$$d^k := -\epsilon_k g^k/|g^k|, \tag{4.5}$$

$$t_k := \max\{\, t : f(x^k + td^k) < f(x^k) - \beta t\epsilon_k|g^k|, t \in \{1, \gamma, \gamma^2, \ldots\}\,\}, \tag{4.6}$$

$$f(x^{k+1}) < f(x^k) - \beta t_k\epsilon_k|g^k|, \tag{4.7a}$$

$$|x^k + t_k d^k - x^{k+1}| \le \min\{t_k, \epsilon_k\}|d^k|. \tag{4.7b}$$

Then (2.7) still holds, since $|x^{k+1} - x^k| \le 2t_k|d^k| = 2t_k\epsilon_k$. Lemma 3.2(ii) is replaced by

LEMMA 4.2. *Let $\epsilon > 0$ and $\bar{x} \in \mathbb{R}^n$. Assuming $0 \notin G_\epsilon(\bar{x})$, pick $\delta > 0$ as in Lemma 3.1 for $C := G_\epsilon(\bar{x})$, and then τ and $\bar{V}$ as in Lemma 3.2(i). Suppose $x^k \in B(\bar{x}, \min\{\tau, \epsilon/3\})$, $\epsilon_k = \epsilon$ and $(x^{k1}, \ldots, x^{km}) \in \bar{V}$. Then $t_k \ge \gamma/3$.*

Proof. In the proof of Lemma 3.2(ii), for $t_k < \gamma/3$, use $d^k := -\epsilon_k g^k/|g^k|$ to get $\langle v^k, g^k\rangle \le \beta|g^k|^2$ as before, and then $v^k \in G_\epsilon(\bar{x})$ from $\gamma^{-1}t_k|d^k| < \epsilon/3$ with $|d^k| = \epsilon_k = \epsilon$. □

As in section 4.1, we deduce that Theorems 3.3–3.6 hold for this variant as well, since in the proof of Theorem 3.3, using (4.7a), we can replace (3.5) by

$$\sum_{k=1}^{\infty} t_k\epsilon_k|g^k| < \infty. \tag{4.8}$$

4.3. Limiting the line search. Note that relations (2.4)–(2.6), (4.1)–(4.3) and (4.5)–(4.7) have the form

$$d^k := -\alpha_k g^k \quad \text{with} \quad \alpha_k > 0, \tag{4.9}$$

$$t_k := \max\{\, t : f(x^k + td^k) < f(x^k) - \beta t|d^k||g^k|, t \in \{1, \gamma, \gamma^2, \ldots\}\,\}, \tag{4.10}$$

$$f(x^{k+1}) < f(x^k) - \beta t_k|d^k||g^k|, \tag{4.11a}$$

$$|x^k + t_k d^k - x^{k+1}| \le \min\{t_k, \epsilon_k\}|d^k|, \tag{4.11b}$$

where $\alpha_k := 1/|g^k|$ in sections 2–3, $\alpha_k := 1$ in section 4.1, and $\alpha_k := \epsilon_k/|g^k|$ in section 4.2. The corresponding lower bounds on t_k produced by Lemmas 3.2(ii), 4.1 and 4.2 have the form $t_k \ge \min\{1, \gamma\epsilon/3|d^k|\}$. Procedure 4.3 below tests only step sizes that satisfy this bound. It finds $t_k \ge \min\{1, \gamma\epsilon/3|d^k|\}$ when the search direction is good enough (see Lemma 4.4 below). Otherwise, a *null step* with $t_k := 0$ occurs; then Step 1 resamples (most of) the gradient bundle G_k, so that eventually the search direction is improved sufficiently (unless x^k is already stationary). It will be seen that this null step/resampling mechanism obviates the need for the iterates to be in the set D where f is differentiable.

PROCEDURE 4.3 (*limited Armijo line search*).

(i) Choose an initial step size $t \ge \min\{1, \gamma\epsilon_k/3|d^k|\}$.

(ii) If $f(x^k + td^k) < f(x^k) - \beta t|d^k||g^k|$, return $t_k := t$.

(iii) If $t \le \min\{1/\gamma, \epsilon_k/3|d^k|\}$, return $t_k := 0$.

(iv) Set $t := \gamma t$ and go to (ii).

LEMMA 4.4. *Let $\epsilon > 0$ and $\bar{x} \in \mathbb{R}^n$. Assuming $0 \notin G_\epsilon(\bar{x})$, pick $\delta > 0$ as in Lemma 3.1 for $C := G_\epsilon(\bar{x})$, and then τ and $\bar{V}$ as in Lemma 3.2(i). Suppose $x^k \in B(\bar{x}, \min\{\tau, \epsilon/3\})$, $\epsilon_k = \epsilon$, $(x^{k1}, \ldots, x^{km}) \in \bar{V}$ and $d^k := -\alpha_k g^k$ with $\alpha_k > 0$. Then Procedure 4.4 finds a step size $t_k \ge \min\{1, \gamma\epsilon/3|d^k|\}$, and the conclusions of Lemmas 3.2(ii), 4.1 and 4.2 hold for $\alpha_k = 1/|g^k|$, 1 and $\epsilon_k/|g^k|$, respectively.*

Proof. As in the proof of Lemma 3.2(ii), using relation (3.4) and the form of $d^k := -\alpha_k g^k$, we obtain $\langle v, d^k\rangle < -\beta|d^k||g^k|$ for all $v \in G_\epsilon(\bar{x})$. Let $t \in (0, \epsilon/3|d^k|]$.

By Lebourg's mean value theorem, $f(x^k+td^k)-f(x^k)=t\langle v,d^k\rangle$ for some $v\in\bar{\partial}f(x)$ with $x\in[x^k+td^k,x^k]$. Then $t|d^k|\le\epsilon/3$ and $|x^k-\bar{x}|\le\epsilon/3$ imply $x\in B(\bar{x},2\epsilon/3)$ and hence $v\in G_\epsilon(\bar{x})$. Therefore, $f(x^k+td^k)<f(x^k)-\beta t|d^k||g^k|\ \forall t\in(0,\epsilon/3|d^k|]$, and the conclusion follows from the rules of Procedure 4.3. □

Remark 4.5.

(i) We conclude from Lemma 4.4 that Theorems 3.3–3.6 remain valid for step sizes t_k produced by Procedure 4.3 instead of the standard Armijo searches (2.5), (4.2) and (4.6). This follows easily from the proofs of section 3 and the remarks in sections 4.1–4.2.

(ii) The number of f-evaluations made by Procedure 4.3 can be controlled via the choice of the initial step size t at step (i). For instance, if $t:=\min\{1,\epsilon_k/3|d^k|\}$, then only one evaluation occurs, and the procedure returns either $t_k:=t$ or $t_k:=0$. If the initial step size t looks "too small", e.g., $f(x^k+td^k)<f(x^k)-0.5t|d^k||g^k|$, we can try expansion by setting $t:=t/\gamma$ until $f(x^k+td^k)\ge f(x^k)-\beta t|d^k||g^k|$, in which case $t_k:=\gamma t$ is returned. Further, step (iii) of Procedure 4.3 can use a smaller threshold $0<\underline{t}<\min\{1/\gamma,\epsilon_k/3|d^k|\}$, returning $t_k:=0$ if $t\le\underline{t}$. Alternatively, the stopping criterion of step (iii) can be ignored until a given number of f-evaluations is reached. Such variations do not impair Lemma 4.4. Note that the theoretical guidelines above leave much freedom for implementations. For instance, choosing a smaller $\underline{t}$ involves the trade-off between the cost of additional f-evaluations during the line search versus the cost of evaluating m gradients at Step 1. On the other hand, using a positive $\underline{t}$ is essential in practice because otherwise an infinite loop might occur due to rounding errors.

(iii) Once Procedure 4.3 replaces the standard Armijo searches (2.5), (4.2) and (4.6), there is no longer any need for keeping x^k in D and including $\nabla f(x^k)$ in G_k at Step 1. This leads to the following simplified variant of Algorithm 2.1. At Step 0, select any $x^1\in\mathbb{R}^n$. At Step 1, set $G_k:=\text{co}\{\nabla f(x^{ki})\}_{i=1}^m$. At Step 5, find t_k via Procedure 4.3. Finally, at Step 6, set $x^{k+1}:=x^k+t_kd^k$. Then the requirements of (4.11) are met if $t_k>0$, whereas the key inequality (2.7) holds always. In effect, Theorems 3.3–3.6 remain valid for this variant. Further, Theorems 3.3–3.6 still hold if the differentiability check of Step 1 is skipped, since $\{x^{ki}\}_{i=1}^m\subset D$ with probability 1. In other words, we may skip the differentiability check of Steps 1 and 6 as in the implementation of [BLO05, section 4], assuming that the user provides reasonable replacements for the gradient at points where it is not defined or that such points are not encountered. In this setting, we may also include $\nabla f(x^k)$ in G_k: although we can't show that the event $x^k\notin D$ has probability zero, it is unlikely to occur in practice.

Acknowledgments. I would like to thank the Associate Editor and the two anonymous referees for their help in improving the exposition of the main results of this paper.

REFERENCES

[BHLO06] J. V. Burke, D. Henrion, A. S. Lewis, and M. L. Overton, *Stabilization via nonsmooth, nonconvex optimization*, IEEE Trans. Automat. Control, ? (2006), to appear.

[BLO02a] J. V. Burke, A. S. Lewis, and M. L. Overton, *Approximating subdifferentials by random sampling of gradients*, Math. Oper. Res., 27 (2002), pp. 567–584.

[BLO02b] ———, *Two numerical methods for optimizing matrix stability*, Linear Algebra Appl., 351-352 (2002), pp. 147–184.

[BLO04] ———, *Pseudospectral components and the distance to uncontollability*, SIAM J. Optim., 26 (2004), pp. 350–361.

[BLO05] ———, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM J. Optim., 15 (2005), pp. 751–779.

[Cla83] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[Gol77] A. A. Goldstein, *Optimization of Lipschitz continuous functions*, Math. Programming, 13 (1977), pp. 14–22.

[Kiw96] 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.

[Lew05] A. S. Lewis, *Local structure and algorithms in nonsmooth optimization*, in Optimization and Applications, F. Jarre, C. Lemaréchal, and J. Zowe, eds., Mathematisches Forschungsinstitut Oberwolfach, Oberwolfach, Germany, 2005, to appear.

[Roc70] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.