JOTA manuscript No.
(will be inserted by the editor)

On the local convergence analysis of the Gradient Sampling method

Elias Salomão Helou · Sandra A. Santos · Lucas E. A. Simões

Received: date / Accepted: date

Elias Salomão Helou was partially supported by FAPESP grant 2013/07375-0 and by CNPq grant 311476/2014-7.

Sandra A. Santos was partially supported by CNPq grant 304032/2010-7, FAPESP grants 2013/05475-7 and 2013/07375-0 and PRONEX Optimization.

Lucas E. A. Simões was supported by FAPESP grant 2013/14615-7.

Elias Salomão Helou

Institute of Mathematical Sciences and Computation, University of São Paulo.

São Carlos - SP, Brazil.

E-mail: elias@icmc.usp.br

Sandra A. Santos

Department of Applied Mathematics, University of Campinas.

Campinas - SP, Brazil.

E-mail: sandra@ime.unicamp.br

Lucas E. A. Simões

Department of Applied Mathematics, University of Campinas.

Campinas - SP, Brazil.

E-mail: simoes.lea@gmail.com

Abstract The Gradient Sampling method is a recently developed tool for solving unconstrained nonsmooth optimization problems. Using just first order information about the objective function, it generalizes the steepest descent method, one of the most classical methods to minimize a smooth function. This manuscript aims at determining under which circumstances one can expect the same local convergence result of the Cauchy method for the Gradient Sampling algorithm. Additionally, at the end of this study, we show how to practically accomplish the required hypotheses during the execution of the algorithm.

Keywords nonsmooth nonconvex optimization · gradient sampling · local convergence · unconstrained minimization

Mathematics Subject Classification (2000) 90C30 · 65K05

1 Introduction

In the past fifteen years an algorithm known as *Gradient Sampling* (GS) has gained attention because of its good numerical results in solving nonsmooth optimization problems (specially for nonconvex objective functions) [1,2]. A strong appeal of the referred method is its intuitive functioning, which uses just first order information to find a descent direction and performs a line search procedure to find out the next iterate. Different from other optimization algorithms, it has a nondeterministic approach, since it randomly samples points around the current iterate in order to obtain a rich set of gradients to compute the search direction. Consequently, a good movement towards the solution is directly linked to a good set of sampled points.

In 2007, Kiwiel introduced a nonnormalized version of GS [3], which can be viewed as a generalization of the well known steepest descent method for smooth functions. Hence, it suggests that, in the best case scenario, the Gradient Sampling will have the same local convergence of the Cauchy method. Although this is reasonable to expect, for the best of our knowledge, there is no proof in the literature of local convergence rates for the GS method nor a clarification of hypotheses under which this can be established.

This theoretical paper has the goal to prove that, under special circumstances, one can achieve linear convergence of the GS method towards the optimal function value of the optimization problem. Moreover, we justify the hypotheses made along the manuscript with illustrative examples, which help us to understand when such a behavior cannot be expected.

The outline of this study is as follows. Section 2 presents the theoretical background and the GS algorithm. Section 3 shows an example that helps us to understand what kind of hypotheses are needed to obtain a good performance from the method. Section 4 is devoted to the theoretical results that prove linear convergence of the algorithm. Section 5 exhibits a practical implication of our study to the GS method with illustrative examples and comparative results. The final section is dedicated to the conclusions.

For clarity, we present some notations that appear along this manuscript:

- $\operatorname{co} \mathcal{X}$ is the convex hull of $\mathcal{X}$;
- $\mathcal{X}^{\#}$ is the cardinality of $\mathcal{X}$;

- $\mathcal{B}(x,r)$ and $\overline{\mathcal{B}}(x,r)$ are, respectively, the Euclidean open and closed balls with center at x and radius r;
- $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^n$;
- e is a vector of appropriate dimension with ones in all entries;
- $\mathcal{P}_{\mathcal{V}}(x)$ is the orthogonal projection of x into the vector space $\mathcal{V}$.

2 GS algorithm

The GS method has the aim to solve the following unconstrained optimization problem

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is a nonsmooth locally Lipschitz function, continuously differentiable in an open dense subset $\mathcal{D}\subset\mathbb{R}^n$. The function f can be either convex or nonconvex.

For objective functions that satisfy the properties required above, it is possible to define a generalization of the so called derivatives for smooth functions [4].

Definition 2.1 (Subdifferential set, subgradient, stationary point) The set given by

$$\overline{\partial} f(x) := \operatorname{co}\left\{\lim_{j\to\infty} \nabla f(x_j) \mid x_j\to x, x_j\in\mathcal{D}\right\}$$

is called the Clarke's subdifferential set for f at x and any $v\in\overline{\partial} f(x)$ is known as a subgradient of f at x. Moreover, if $0\in\overline{\partial} f(x)$, then we say that x is a stationary point for f.

Additionally, a set more aligned with the idea of sampling points around the current iterate can also be defined [5].

Definition 2.2 (ϵ-Subdifferential set, ϵ-subgradient, ϵ-stationary point) The ϵ-subdifferential set for f at x is given by

$$\overline{\partial}_\epsilon f(x) := \text{co}\ \overline{\partial} f(\mathcal{B}(x, \epsilon)).$$

Any $v \in \overline{\partial}_\epsilon f(x)$ is known as an ϵ-subgradient of f at x. Moreover, if $0 \in \overline{\partial}_\epsilon f(x)$, then we say that x is an ϵ-stationary point for f.

For completeness, we present the GS algorithm that will be treated along the subsequent sections. It is the same algorithm presented by the original authors but using the nonnormalized direction suggested by Kiwiel [3]. The procedures presented in Algorithm 1 show the importance of a good set of sampled points for the computation of the search direction. Therefore, it seems reasonable to accept that any local convergence result will be conditioned to a good set of sampled points.

The next section has the intent to help us to determine what is the mathematical meaning of the expression "good set of sampled points". However, in order to achieve such a goal, we need to have a more structured problem than the one presented in (1). From now on, we will be interested in the following nonsmooth optimization problem

$$\min_{x \in \mathbb{R}^n} \left(f(x) := \max_{1 \leq i \leq p} \{\phi_i(x)\} \right), \tag{2}$$

where the functions $\phi_i : \mathbb{R}^n \to \mathbb{R}$ are all of class C^2. It is important to stress that we only require this structure for the function f, we do not need to write

Algorithm 1: Nonnormalized version of the GS method.

Step 0. Given $x_0 \in \mathcal{D}$, $m \in \mathbb{N}$ with $m \geq n+1$, fixed real numbers ν_0, ϵ_0, $\nu_{\text{opt}}, \epsilon_{\text{opt}} > 0$ and $0 < \theta_\nu, \theta_\epsilon, \gamma, \beta < 1$, set $k = 0$.

Step 1. Choose $\{x_{k,1}, \ldots, x_{k,m}\} \subset \mathcal{B}(x_k, \epsilon_k)$ with randomly, independently and uniformly sampled elements.

Step 2. Set $G_k = [\nabla f(x_k)\ \nabla f(x_{k,1})\ \ldots \nabla f(x_{k,m})]$ and find $g_k = G_k\lambda_k$, where λ_k solves

$$\min_{\lambda} \quad \frac{1}{2}\lambda^T G_k^T G_k \lambda$$
$$\text{s.t.} \quad e^T\lambda = 1,\ \lambda \geq 0.$$

Step 3. If $\nu_k < \nu_{\text{opt}}$ and $\epsilon_k < \epsilon_{\text{opt}}$, then STOP!
Otherwise, if $\|g_k\| < \nu_k$, then $\epsilon_{k+1} = \theta_\epsilon\epsilon_k$, $\nu_{k+1} = \theta_\nu\nu_k$, $x_{k+1} = x_k$ and go to Step 6.

Step 4. Do a backtracking line search and find the maximum $t_k \in \{1, \gamma, \gamma^2, \ldots\}$ such that

$$f(x_k + t_k d_k) < f(x_k) - \beta t_k \|g_k\|^2, \text{ where } d_k = -g_k.$$

Step 5. If $x_k + t_k d_k \in \mathcal{D}$, then set $x_{k+1} = x_k + t_k d_k$. Otherwise, find

$$x_{k+1} \in \mathcal{B}(x_k + t_k d_k, \min\{t_k, \epsilon_k\}\|d_k\|) \cap \mathcal{D},$$

such that $f(x_{k+1}) \leq f(x_k) - \beta t_k \|g_k\|^2$.

Step 6. Set $k \leftarrow k+1$ and go back to Step 1.

the function f in this explicit form, i.e., we do not need to know the functions ϕ_i, but only need to evaluate f, and its gradient, at the demanded points whenever the latter exists.

Finally, for functions satisfying (2), it is possible to define some sets that will be useful for us latter. The first one is called the active set of indices of f

at x and it is given by

$$\mathcal{I}(x) := \{i \in \{1, \ldots, p\} \mid f(x) = \phi_i(x)\},$$

whereas the other ones are defined below.

Definition 2.3 (U,V-spaces) Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is the continuous objective function of problem (2) and x is any point in $\mathbb{R}^n$. Then, we define

$$U(x) := \{s \in \mathbb{R}^n \mid [\nabla\phi_i(x) - \nabla\phi_j(x)]^T s = 0, \ \ \forall i, j \in \mathcal{I}(x), \ \ i \neq j\}$$

and $V(x) := U(x)^\perp$ as, respectively, the smooth and nonsmooth subspaces of f at x.

The vector spaces defined above are of great importance to us. Notice that they split the domain of the function in two subspaces: the one in which f behaves smoothly (U-space) and the other that captures all the nonsmoothness of the function (V-space).

3 Example

In this section we have the intent to determine what one needs to ask to be able to set a local convergence result for GS. In other words, we want to establish what are the minimal requirements to obtain a satisfactory movement towards the solution.

Given a function f of the kind presented in (2) and considering x_* as the limit point of the GS sequence, the first condition that is indispensable is that all of the functions ϕ_i, with $i \in \mathcal{I}(x_*)$, must strictly assume the maximum

at least at one of the sampled points or at the current iterate. In a more rigorous way, we are saying that, for a good movement towards the solution at some fixed iteration k, given any $i \in \mathcal{I}(x_*)$, there must be $x_{k,j}$, for some $j \in \{0, \ldots, m\}$ (here, and from now on, we define $x_{k,0} := x_k$), such that

$$\phi_i(x_{k,j}) > \phi_s(x_{k,j}), \text{ for any } s \in \{1, \ldots, p\} \setminus \{i\}. \tag{H_ϕ}$$

To highlight the plausibility of this hypothesis, we present an example where the absence of this assumption causes a bad GS behavior.

Let us consider a two-dimensional function $f : \mathbb{R}^2 \to \mathbb{R}$, with

$$f(x) = \max\{\phi_1(x), \phi_2(x), \phi_3(x)\},$$

where, for $x = [\xi_1 \quad \xi_2]^T$, we have

$$\phi_1(x) = \xi_1 + \xi_2,\ \phi_2(x) = -2\xi_1 + \xi_2 \text{ and } \phi_3(x) = \xi_1 - 2\xi_2.$$

Clearly, it is a convex function with $x_* = 0$ as its global minimizer. Furthermore, the lowest function value is given by $f(x_*) = 0$.

Suppose we want to start an iteration of the GS method with

$$x_0 = \begin{bmatrix} 0.5^l & 0.5^{2l} \end{bmatrix}^T, \text{ for any fixed } l \in \mathbb{N}.$$

Moreover, we assume that the method has sampled in such a way that

$$f(x_{0,i}) = \phi_2(x_{0,i}),\ \forall i \in \{1, 2, 3\} \text{ (assuming } m = 3).$$

Consequently, the function ϕ_3 does not assume the maximum at the sampled points nor at x_0. Step 2 returns us $g_0 = [0 \quad 1]^T$. Assuming that $\nu_0 = \epsilon_0 = 10^{-1}$

and $\epsilon_{\text{opt}} = \nu_{\text{opt}} = 10^{-6}$, the method does not stop neither jumps from Step 3 to Step 6.

Now, notice that, for all $t > 0$, we have

$$\phi_1(x_0 - tg_0) = 0.5^{2l} + 0.5^l - t;$$

$$\phi_2(x_0 - tg_0) = 0.5^{2l} - 2 \cdot 0.5^l - t;$$

$$\phi_3(x_0 - tg_0) = -2 \cdot 0.5^{2l} + 0.5^l + 2t.$$

Hence, for $t = O(1)$, we must have that $f(x_0 - tg_0) = \phi_3(x_0 - tg_0)$, while for $t = \rho \cdot 0.5^{2l}$, with $\rho \in (0, 1)$, we have $f(x_0 - tg_0) = \phi_1(x_0 - tg_0)$. Since $f(x_0) = -2 \cdot 0.5^{2l} + 0.5^l$, we see clearly that for any $t_0 \in \{1, \gamma, \gamma^2, \ldots\}$, we must have

$$f(x_0) < \phi_3(x_0 - t_0 g_0).$$

Consequently, in order to have a successful line search, we must be in a region of the domain where ϕ_3 does not assume the maximum, which is achieved by setting $t_0 = O(0.5^{2l})$.

Defining $x_1 = x_0 - t_0 g_0$, one can compute the reduction efficiency of the function value, which yields

$$\frac{f(x_1) - f(x_*)}{f(x_0) - f(x_*)} = 1 + O(0.5^l).$$

Hence, one can see that it is not possible to establish a linear convergence rate no matter how close we start from x_* ($l \to \infty$).

This example shows us that when H_ϕ does not hold, the convergence might be excessively slow. Therefore, it is reasonable to think that a local convergence result may assume H_ϕ. However, one might still ask if this condition is sufficient

for our goal. Unfortunately, H_ϕ is not enough for reaching our purposes, as the size of the sampling radius plays a key role as well (see Sections 4 and 5). Indeed, an additional condition must be taken into account: a restriction over the value of

$$\tau_k := \max_{1\leq i\leq m} \{\|x_{k,i} - x_k\|\}. \tag{3}$$

We state that assuming H_ϕ and $\tau_k \leq T\|x_k - x_*\|^2$, for some small enough $T > 0$, is sufficient to obtain a linear convergence rate of the function values.

Before we proceed with the expected proofs, we need to make an observation. The local convergence theory developed here is only applicable to functions that have local minimizers x_* that satisfy $\dim\{U(x_*)\} \neq 0$. We justify this statement by looking back at the example just exhibited. We have seen that the lack of fulfillment of H_ϕ leaves us with a bad behavior of the GS method. However, if H_ϕ were valid for this specific example, it would yield that $g_0 = 0$. So, by Step 3, we would have $x_1 = x_0$. Hence, with or without the condition H_ϕ, the theory we have developed here says nothing about the local convergence whenever $V(x^*) = \mathbb{R}^n$.

4 Convergence results

In this section we have established a local convergence result for the nonnormalized GS method under the more structured case defined in (2). In other words, we find $R \in (0, 1)$ such that, for some infinite index set $\mathcal{K} \subset \mathbb{N}$, we have

$$f(x_{k+1}) - f(x_*) \leq R\,[f(x_k) - f(x_*)]\,, \forall k \in \mathcal{K}.$$

For that goal, we need to assume a condition upon the derivatives of the functions ϕ_i [6].

Assumption 1 *For all $x \in \mathbb{R}^n$ with $|\mathcal{I}(x)| \geq 2$, the gradients $\{\nabla\phi_i(x)\}_{i\in\mathcal{I}(x)}$ compose an affinely independent set, that is,*

$$\sum_{i\in\mathcal{I}(x)} \alpha_i \nabla\phi_i(x) = 0 \quad \textit{and} \quad \sum_{i\in\mathcal{I}(x)} \alpha_i = 0 \iff \alpha_i = 0, \ \forall i \in \mathcal{I}(x).$$

It is not hard to show that if Assumption 1 holds, then $\mathcal{I}(x)^{\#} \leq n+1$. Additionally, this information together with $\dim\{U(x_*)\} \geq 1$ give us that $\mathcal{I}(x)^{\#} \leq n$.

Notice that, by the way the subdifferential set is defined, all functions with the representation given by (2) have the following property

$$\overline{\partial} f(x) := \operatorname{co}\{\nabla\phi_i(x) \mid i \in \mathcal{I}(x)\}.$$

Consequently, supposing that x_* is a stationary point for f, Assumption 1 tells us that there is only one convex combination of $\nabla\phi_i(x_*)$, with $i \in \mathcal{I}(x_*)$, such that it generates the null vector.

With the considerations made above, we are ready to present the results that will culminate in the main theorem. For an easier exposition of the statements, from now on we will assume that x_* is always a local minimizer for the function f (with $\dim\{U(x_*)\} \geq 1$). Moreover, we will consider that $\mathcal{I}(x_*)^{\#} \geq 2$, since otherwise the convergence would be to a point where f is smooth, which is not the case of interest. Additionally, without any loss of generality we assume that $\mathcal{I}(x_*) = \{2, \ldots, r\}$, with $r \in \{2, \ldots, p\}$. Finally, we

define $\lambda^* \in \mathbb{R}^r$ as the unique vector that satisfies

$$\lambda^* \geq 0, \quad \sum_{i=1}^{r} \lambda_i^* = 1 \quad \text{and} \quad \sum_{i=1}^{r} \lambda_i^* \nabla \phi_i(x_*) = 0. \tag{4}$$

Lemma 4.1 *Suppose f is given in the form of (2). Then, for any $d \in U(x_*)$, we must have*

$$d^T \left(\sum_{i=1}^{r} \lambda_i^* \nabla^2 \phi_i(x_*) \right) d \geq 0.$$

Proof Let us consider any vector $d \in U(x_*)$. Therefore, since $\phi_i \in C^2$ for all $i \in \{1, \ldots, p\}$, we can see by the Implicit Function Theorem [7, Appendix] that there are a sufficiently small $\delta > 0$ and a twice differentiable function $\gamma : (-\delta, \delta) \to \mathbb{R}^n$ such that $\gamma(0) = x_*$, $\gamma'(0) = d$ and

$$t \in (-\delta, \delta) \Rightarrow \phi_i(\gamma(t)) - \phi_r(\gamma(t)) = 0, \, \forall i \in \{1, \ldots, r-1\}.$$

Additionally, since x_* is a local minimizer of f, we must have that $t = 0$ is a local minimizer of the function $F(t) := \phi_r(\gamma(t))$. Consequently,

$$d^T \nabla^2 \phi_r(x_*) d + \nabla \phi_r(x_*)^T \gamma''(0) = F''(0) \geq 0. \tag{5}$$

Now, defining $\psi_i(x) := \phi_i(x) - \phi_r(x)$ for $i \in \{1, \ldots, r-1\}$, we must have that

$$\psi_i(\gamma(t)) = \psi_i(x_*) + t \nabla \psi_i(x_*)^T d + \frac{t^2}{2} \left(d^T \nabla^2 \psi_i(x_*) d + \nabla \psi_i(x_*)^T \gamma''(0) \right) + o(t^2).$$

Hence, since $\psi_i(\gamma(t)) = \psi_i(x_*) = 0$ for all $t \in (-\delta, \delta)$ and $\nabla \psi_i(x_*)^T d = 0$ for $i \in \{1, \ldots, r-1\}$, we see, by taking the limit $t \to 0$, that

$$d^T \nabla^2 \psi_i(x_*) d + \nabla \psi_i(x_*)^T \gamma''(0) = 0, \;\; \forall i \in \{1, \ldots, r-1\},$$

which yields

$$d^T \sum_{i=1}^{r-1} \lambda_i^* \left[\nabla^2 \phi_i(x_*) - \nabla^2 \phi_r(x_*) \right] d + \sum_{i=1}^{r-1} \lambda_i^* \left[\nabla \phi_i(x_*) - \nabla \phi_r(x_*) \right]^T \gamma''(0) = 0.$$

Finally, adding the last equation to (5) and recalling that $e^T\lambda^* = 1$, we have

$$d^T\left(\sum_{i=1}^{r}\lambda_i^*\nabla^2\phi_i(x_*)\right)d + \sum_{i=1}^{r}\lambda_i^*\nabla\phi_i(x_*)^T\gamma''(0) \geq 0,$$

which implies the desired result (because $\sum_{i=1}^{r}\lambda_i^*\nabla\phi_i(x_*) = 0$). □

The result presented above is a strong statement. It highlights that the matrix

$$\sum_{i=1}^{r}\lambda_i^*\nabla^2\phi_i(x_*)$$

will play the role of a generalized Hessian of the function f in the U-space. Therefore, from now on, we will assume that the above matrix will be positive definite in the subspace $U(x_*)$, i.e., there is $\mu > 0$ such that

$$d^T\left(\sum_{i=1}^{r}\lambda_i^*\nabla^2\phi_i(x_*)\right)d \geq \mu\|d\|^2, \ \ \forall d \in U(x_*). \tag{6}$$

In order to proceed with our goal, we present a technical lemma.

Lemma 4.2 *Suppose f is given in the form of (2). Moreover, let us assume a sequence $\{s_k\} \subset \mathbb{R}^n$ such that $s_k \to 0$ and*

$$s_k = \mathcal{P}_{U(x_*)}(s_k) + o(\|s_k\|). \tag{7}$$

Then, for all sufficiently large k, the following must hold

i) $s_k^T\left(\sum_{i=1}^{r}\lambda_i^*\nabla^2\phi_i(x_*)\right)s_k \geq \frac{2}{3}\mu\|s_k\|^2$;

ii) $f(x_*) + \frac{\mu}{4}\|s_k\|^2 \leq f(x_* + s_k)$,

where μ is such that (6) holds.

Proof First, let us prove statement i). Since

$$\left\|\mathcal{P}_{U(x_*)}\left(s_k\right)\right\| \leq \|s_k\|,$$

notice that

$$\begin{aligned} s_k^T\left(\sum_{i=1}^r \lambda_i^* \nabla^2\phi_i(x_*)\right) s_k \geq\ & \mathcal{P}_{U(x_*)}\left(s_k\right)^T\left(\sum_{i=1}^r \lambda_i^* \nabla^2\phi_i(x_*)\right)\mathcal{P}_{U(x_*)}\left(s_k\right) \\ & - 2\left\|\mathcal{P}_{U(x_*)}\left(s_k\right)\right\| o(\|s_k\|) + o(\|s_k\|^2) \\ \geq\ & \mu\left\|\mathcal{P}_{U(x_*)}\left(s_k\right)\right\|^2 + o(\|s_k\|^2) \\ =\ & \mu\|s_k\|^2 + o(\|s_k\|^2) \\ & \text{(by relation (7)).} \end{aligned}$$

Hence, for a sufficiently large $k \in \mathbb{N}$, the first result is obtained.

Now, let us prove the second statement. Notice that, for a sufficiently large $k \in \mathbb{N}$, we must have

$$\begin{aligned} f(x_* + s_k) =\ & \max_{1\leq i\leq r} \phi_i(x_* + s_k) \\ \geq\ & \sum_{i=1}^r \lambda_i^* \phi_i(x_* + s_k) \\ =\ & \sum_{i=1}^r \lambda_i^* \left[\phi_i(x_*) + \nabla\phi_i(x_*)^T s_k + \frac{1}{2} s_k^T \nabla^2\phi_i(x_*) s_k\right] + o(\|s_k\|^2) \\ =\ & f(x_*) + \frac{1}{2} s_k^T\left(\sum_{i=1}^r \lambda_i^* \nabla^2\phi_i(x_*)\right) s_k + o(\|s_k\|^2). \end{aligned}$$

Consequently,

$$\frac{f(x_* + s_k) - f(x_*)}{\|s_k\|^2} \geq \frac{1}{2}\frac{s_k^T}{\|s_k\|}\left(\sum_{i=1}^r \lambda_i^* \nabla^2\phi_i(x_*)\right)\frac{s_k}{\|s_k\|} + \frac{o(\|s_k\|^2)}{\|s_k\|^2}.$$

Hence, for all $k \in \mathbb{R}^n$ sufficiently large, we must have

$$\frac{f(x_* + s_k) - f(x_*)}{\|s_k\|^2} \geq \frac{1}{4}\mu,$$

which is the desired result. □

Before we proceed, it is important to make an observation. By [3, Theorem 3.3], we know that, with probability one, $\epsilon_k, \nu_k \to 0$ in Algorithm 1. For such a fact to hold, there must exist an infinite index set $\mathcal{K} \subset \mathbb{N}$ such that

$$g_k \underset{k\in\mathcal{K}}{\to} 0. \tag{8}$$

The next lemma says that the statement "H_ϕ happens for all $k \in \mathcal{K}$ sufficiently large" is a sufficient condition for (8) to hold.

Lemma 4.3 *Suppose that $\mathcal{K} \subset \mathbb{N}$ is an infinite index set such that H_ϕ is satisfied for all $k \in \mathcal{K}$. Moreover, assume that the sequence $\{x_k\}$ was generated by Algorithm 1 and $x_k \to x_*$. Then, the following holds*

$$g_k \underset{k\in\mathcal{K}}{\to} 0.$$

Moreover,

$$\hat{\lambda}^k \underset{k\in\mathcal{K}}{\to} \lambda^*,$$

where $\hat{\lambda}^k \in \mathbb{R}^r$ and, for $i \in \{1, \ldots, r\}$,

$$\hat{\lambda}^k_i := \sum_{j\in\mathcal{J}_{k,i}} \lambda^k_j, \quad \text{with } \mathcal{J}_{k,i} := \{s \in \{0, \ldots, m\} \mid f(x_{k,s}) = \phi_i(x_{k,s})\}. \tag{9}$$

Proof We know that $\mathcal{I}(x_*) = \{1, \ldots, r\}$. Moreover, by [3, Theorem 3.3], we have that $\epsilon_k \to 0$. Consequently, for a sufficiently large $k \in \mathcal{K}$, we must have that only the functions ϕ_i, with $i \in \{1, \ldots, r\}$, assume the maximum at $x_k, x_{k,1}, \ldots, x_{k,m}$. Therefore, recalling the definition of τ_k in (3) and that g_k

solves the quadratic minimization problem of Step 2, we get

$$\begin{aligned}
\|g_k\| &= \left\| \sum_{i=1}^{r} \sum_{j \in \mathcal{J}_{k,i}} \lambda_j^k \nabla \phi_i(x_{k,j}) \right\| \\
&\leq \left\| \sum_{i=1}^{r} \sum_{j \in \mathcal{J}_{k,i}} \frac{\lambda_i^*}{\mathcal{J}_{k,i}^{\#}} \nabla \phi_i(x_{k,j}) \right\| \\
&\leq \left\| \sum_{i=1}^{r} \sum_{j \in \mathcal{J}_{k,i}} \frac{\lambda_i^*}{\mathcal{J}_{k,i}^{\#}} \nabla \phi_i(x_k) \right\| + O(\tau_k) \\
&= \left\| \sum_{i=1}^{r} \lambda_i^* \nabla \phi_i(x_k) \right\| + O(\tau_k). \qquad (10)
\end{aligned}$$

Hence, since $x_k \to x_*$, $\epsilon_k \to 0$ and $\tau_k \in (0, \epsilon_k)$, we see that

$$g_k \underset{k \in \mathcal{K}}{\to} 0.$$

Moreover, it implies that

$$g_k = \sum_{i=1}^{r} \sum_{j \in \mathcal{J}_{k,i}} \lambda_j^k \nabla \phi_i(x_k) + O(\tau_k) = \sum_{i=1}^{r} \hat{\lambda}_i^k \nabla \phi_i(x_k) + O(\tau_k) \underset{k \in \mathcal{K}}{\to} 0.$$

Now, since $\lambda^* \in \mathbb{R}^r$ is the unique vector such that (4) holds, we must have that

$$\hat{\lambda}^k \underset{k \in \mathcal{K}}{\to} \lambda^*,$$

which ends the proof. □

The next technical lemma establishes sufficient conditions that can guarantee that the vector $x_k - x_*$ will be close enough to the subspace $U(x_*)$.

Lemma 4.4 *Suppose that the sequence $\{x_k\}$ was generated by Algorithm 1 and that $x_k \to x_*$. Moreover, assume that there is an infinite index set $\mathcal{K} \subset \mathbb{N}$ such that hypothesis H_ϕ is satisfied for all $k \in \mathcal{K}$. Then, there must exist*

$k' \in \mathcal{K}$, such that for all $k \in \mathcal{K}$ larger than k' and having $\tau_k \leq \alpha\|x_k - x_*\|^2$, for some $\alpha > 0$, the following must happen

i) For all $i \in \{1, \ldots, r-1\}$, we have that

$$|\phi_i(x_k) - \phi_r(x_k)| \leq 2\alpha L_{max}\|x_k - x_*\|^2,$$

where L_{max} is an upper bound for the Lipschitz constants of the functions ϕ_i around x_*;

ii) $x_k - x_* = \mathcal{P}_{U(x_*)}(x_k - x_*) + o(\|x_k - x_*\|)$.

Proof First, since H_ϕ holds, there are points $y_1, \ldots, y_r \in \overline{\mathcal{B}}(x_k, \tau_k)$ such that

$$\phi_r(y_r) > \phi_i(y_r) \ \text{ and } \ \phi_r(y_i) < \phi_i(y_i), \ \ i \in \{1, \ldots, r-1\}.$$

Therefore, defining $\psi_i := \phi_i - \phi_r$, we have, by the Intermediate Value Theorem, that there exists $z_i \in \overline{\mathcal{B}}(x_k, \tau_k)$ such that $\psi_i(z_i) = 0$, for all $i \in \{1, \ldots, r-1\}$. Consequently, considering a sufficiently large $k \in \mathcal{K}$ such that x_k is sufficiently close to x_* in order to have $L_{\max}$ as a valid upper bound for the Lipschitz constants of the functions ϕ_i, for $i \in \{1, \ldots, r\}$, the following holds

$$\begin{aligned}\phi_i(z_i) = \phi_r(z_i) &\Rightarrow |\phi_i(x_k) - \phi_r(x_k)| = |\phi_i(x_k) - \phi_i(z_i) + \phi_r(z_i) - \phi_r(x_k)| \\ &\Rightarrow |\phi_i(x_k) - \phi_r(x_k)| \leq 2L_{\max}\tau_k.\end{aligned}$$

Since we have supposed that $\tau_k \leq \alpha\|x_k - x_*\|^2$, the first result is obtained.

Now, let us consider the Taylor's expansion for the functions ϕ_i, with $i \in \{1, \ldots, r\}$. Then,

$$\phi_i(x_k) = \phi_i(x_*) + \nabla\phi_i(x_*)^T(x_k - x_*) + O(\|x_k - x_*\|^2).$$

So, for $i \in \{1, \ldots, r-1\}$,

$$\phi_i(x_k) - \phi_r(x_k) = \left[\nabla\phi_i(x_*) - \nabla\phi_r(x_*)\right]^T (x_k - x_*) + O(\|x_k - x_*\|^2),$$

which yields

$$\left[\nabla\phi_i(x_*) - \nabla\phi_r(x_*)\right]^T (x_k - x_*) = O(\|x_k - x_*\|^2).$$

Therefore, because of the definition of the subspace $U(x_*)$, we must have that

$$x_k - x_* = \mathcal{P}_{U(x_*)}(x_k - x_*) + o(\|x_k - x_*\|),$$

as desired. □

The following statement says that, under some hypothesis, the difference $f(x_k) - f(x_*)$ can be majored by a value proportional to $\|g_k\|^2$. This and the other subsequent results pursuit, for the nonsmooth case, equivalent statements of the well established local convergence result of the steepest descent method [8]. For this reason, the hypothesis about the value τ_k in Lemma 4.5 seems essential.

Lemma 4.5 *Suppose that the sequence $\{x_k\}$ was generated by Algorithm 1 and that $x_k \to x_*$. Additionally, assume that there exists an infinite index set $\mathcal{K} \subset \mathbb{N}$ such that H_ϕ holds for all $k \in \mathcal{K}$. Then, there must exist $k' \in \mathcal{K}$, such that*

$$k \geq k' \text{ and } \tau_k \leq \frac{\mu}{8L_{max}}\|x_k - x_*\|^2 \Rightarrow \frac{\mu}{4}\left[f(x_k) - f(x_*)\right] \leq \|g_k\|^2.$$

Proof First, let us consider a sufficiently large k such that only the functions ϕ_i with $i \in \mathcal{I}(x_*)$ assume the maximum at $\mathcal{B}(x_k, \epsilon_k)$. Moreover, suppose that

$$\tau_k \leq \frac{\mu}{8L_{\max}}\|x_k - x_*\|^2.$$

Now, using the definition of $\hat{\lambda}^k$ in (9), one can notice that

$$\begin{aligned}
f(x_*) &= \max_{1\leq i\leq r} \phi_i(x_*) \\
&= \max_{1\leq i\leq r} \left\{ \phi_i(x_k) + \nabla\phi_i(x_k)^T(x_* - x_k) + \frac{1}{2}(x_* - x_k)^T\nabla^2\phi_i(x_k)(x_* - x_k) \right\} \\
&\quad + o(\|x_k - x_*\|^2) \\
&\geq \sum_{i=1}^{r} \hat{\lambda}_i^k \left[\phi_i(x_k) + \nabla\phi_i(x_k)^T(x_* - x_k) + \frac{1}{2}(x_* - x_k)^T\nabla^2\phi_i(x_k)(x_* - x_k) \right] \\
&\quad + o(\|x_k - x_*\|^2).
\end{aligned}$$

However, assuming, without loss of generality, that

$$\max_{1\leq i\leq r} \phi_i(x_k) = \phi_r(x_k)$$

and because of the implication i) of Lemma 4.4, we have that

$$\begin{aligned}
\sum_{i=1}^{r} \hat{\lambda}_i^k \phi_i(x_k) &\geq \max_{1\leq i\leq r} \phi_i(x_k) - \frac{\mu}{8L_{\max}} 2L_{\max}\|x_k - x_*\|^2 \\
&= f(x_k) - \frac{\mu}{4}\|x_k - x_*\|^2.
\end{aligned}$$

Additionally, since the derivatives of ϕ_i are all Lipschitz continuous, we must have that

$$\begin{aligned}
\sum_{i=1}^{r} \hat{\lambda}_i^k \nabla\phi_i(x_k)^T(x_* - x_k) &= \sum_{i=1}^{r} \sum_{j\in\mathcal{J}_{k,i}} \lambda_j^k \nabla\phi_i(x_k)^T(x_* - x_k) \\
&= \sum_{i=1}^{r} \sum_{j\in\mathcal{J}_{k,i}} \lambda_j^k \nabla\phi_i(x_{k,j})^T(x_* - x_k) + o(\|x_k - x_*\|^2) \\
&= g_k^T(x_* - x_k) + o(\|x_k - x_*\|^2).
\end{aligned}$$

Still, because of the Hessian's continuity of the functions ϕ_i and by the Lemmas 4.2, 4.3 and 4.4, we see, for a sufficiently large k, that

$$\begin{aligned}(x_* - x_k)^T \sum_{i=1}^{r} \hat{\lambda}_i^k \nabla^2 \phi_i(x_k)(x_* - x_k) &= (x_* - x_k)^T \sum_{i=1}^{r} \lambda_i^* \nabla^2 \phi_i(x_*)(x_* - x_k) \\ &\quad + o(\|x_k - x_*\|^2) \\ &\geq \frac{2}{3}\mu\|x_k - x_*\|^2 + o(\|x_k - x_*\|^2).\end{aligned}$$

Therefore, the following must hold

$$\begin{aligned}f(x_*) &\geq f(x_k) + g_k^T(x_* - x_k) + \frac{2}{3}\mu\|x_k - x_*\|^2 \\ &\quad - \frac{1}{4}\mu\|x_k - x_*\|^2 + o(\|x_k - x_*\|^2) \\ &= f(x_k) + g_k^T(x_* - x_k) + \frac{5}{12}\mu\|x_k - x_*\|^2 + o(\|x_k - x_*\|^2).\end{aligned}$$

Consequently, for k sufficiently large, the following holds

$$f(x_k) - f(x_*) \leq g_k^T(x_k - x_*) \leq \|g_k\|\|x_k - x_*\|. \tag{11}$$

Then, looking at Lemma 4.4, we see that for a sufficiently large k, the hypothesis of Lemma 4.2 will hold for $s_k = x_k - x_*$. So, by implication ii) of Lemma 4.2, we get

$$\|x_k - x_*\| \leq 2\sqrt{\frac{f(x_k) - f(x_*)}{\mu}}.$$

Finally, putting together the last relation with (11) we obtain the desired result. □

The result presented below guarantees a sufficient decrease for the function value. This lemma will be of great importance in our main theorem.

Lemma 4.6 *Suppose that the sequence* $\{x_k\}$ *was generated by Algorithm 1 and that* $x_k \to x_*$. *Moreover, assume that there is an infinite index set* $\mathcal{K} \subset \mathbb{N}$ *such that hypothesis* H_ϕ *is satisfied for all* $k \in \mathcal{K}$. *Then, there must exist* $k' \in \mathcal{K}$, *such that for all* $k \in \mathcal{K}$ *larger than* k', *the following must happen*

$$t \leq \frac{1-\beta}{M} \Rightarrow f(x_k - tg_k) < f(x_k) - \beta t \|g_k\|^2,$$

where M *is a positive real number such that, for a sufficiently small neighborhood* $\mathcal{V}$ *of* x_*, *we have*

$$\max_{1\leq i\leq r} \left\{\|\nabla^2 \phi_i(x)\|\right\} \leq M, \ \ \forall x \in \mathcal{V}. \tag{12}$$

Proof Let us consider an index $k \in \mathcal{K}$ sufficiently large such that the only functions that do assume the maximum at the points $x_k, x_{k,1}, \ldots, x_{k,m}$ are those with indices $i \in \{1, \ldots, r\}$. Then, considering a fixed $t \leq 1/(2M)$, we have

$$\begin{aligned} f(x_k - tg_k) = & \max_{1\leq i\leq r} \left\{\phi_i(x_k) - t\nabla\phi_i(x_k)^T g_k + \frac{t^2}{2} g_k^T \nabla^2 \phi_i(x_k) g_k\right\} \\ & + o(\|g_k\|^2) \\ \leq & f(x_k) + \max_{1\leq i\leq r} \left\{-t\nabla\phi_i(x_k)^T g_k\right\} \\ & + \frac{t^2}{2} \max_{1\leq i\leq r} \left\{g_k^T \nabla^2 \phi_i(x_k) g_k\right\} + o(\|g_k\|^2). \end{aligned}$$

Additionally, since $\tau_k = O(\|g_k\|^2)$, notice that

$$\max_{1\leq i\leq r} \left\{-t\nabla\phi_i(x_k)^T g_k\right\} = \max_{0\leq i\leq m} \left\{-t\nabla f(x_{k,i})^T g_k\right\} + o(\|g_k\|^2).$$

Moreover, from convex analysis, we know that

$$g_k \ \text{ solves } \min_{g\in \mathrm{co}\{\nabla f(x_{k,i})\}_{i=0}^m} \|g\| \Leftrightarrow \langle g - g_k, -g_k\rangle \leq 0, \ \forall g \in \mathrm{co}\{\nabla f(x_{k,i})\}_{i=0}^m,$$

which yields

$$\max_{0\leq i\leq m}\left\{-t\nabla f(x_{k,i})^T g_k\right\} \leq -t\|g_k\|^2.$$

Hence, it implies that

$$f(x_k - tg_k) \leq f(x_k) - t\|g_k\|^2 + \frac{t^2}{2}\max_{1\leq i\leq r}\left\{g_k^T\nabla^2\phi_i(x_k)g_k\right\} + o(\|g_k\|^2).$$

Moreover, since $x_k \to x_*$, we must have that

$$\max_{1\leq i\leq r}\left\{g_k^T\nabla^2\phi_i(x_k)g_k\right\} \leq M\|g_k\|^2, \text{ for all } x_k \text{ close enough to } x_*.$$

Therefore, since g_k tends to the null vector for indices in $\mathcal{K}$ (by Lemma 4.3), there must exist a sufficiently large $k' \in \mathcal{K}$ such that for all $k \in \mathcal{K}$ larger than k', we have

$$\begin{aligned} f(x_k - tg_k) &< f(x_k) - t\|g_k\|^2 + t^2M\|g_k\|^2 \\ &= f(x_k) - t\|g_k\|^2(1 - Mt) \\ &\quad (\text{since } t \leq (1-\beta)/M) \\ &\leq f(x_k) - \beta t\|g_k\|^2, \end{aligned}$$

which completes the proof. □

Finally, we are able to prove the main result of this manuscript. It establishes, under special conditions, that the GS method, in fact, shares the linear convergence of the steepest descent method.

Theorem 4.1 *Suppose that the sequence $\{x_k\}$ was generated by Algorithm 1 and that $x_k \to x_*$. Additionally, assume that there exists an infinite index set*

$\mathcal{K} \subset \mathbb{N}$ such that H_ϕ holds for all $k \in \mathcal{K}$. Then, there must exist $k' \in \mathcal{K}$, such that

$$k \geq k',\ \tau_k \leq \frac{\mu}{8L_{max}}\|x_k - x_*\|^2 \text{ and } x_{k+1} = x_k - t_k g_k, \tag{13}$$

implies

$$f(x_{k+1}) - f(x_*) \leq \left(1 - \mu\gamma\frac{\beta(1-\beta)}{4M}\right)[f(x_k) - f(x_*)].$$

Proof First, let us suppose that we have $k' \in \mathcal{K}$ large enough so that Lemmas 4.5 and 4.6 hold. Then, assuming $k \geq k'$ and (13), one can notice that since t_k is obtained using Step 4 of Algorithm 1, we must have, by Lemma 4.6, that

$$t_k \geq \gamma\frac{1-\beta}{M}.$$

Therefore,

$$f(x_{k+1}) \leq f(x_k) - \beta t_k\|g_k\|^2 \leq f(x_k) - \gamma\frac{\beta(1-\beta)}{M}\|g_k\|^2.$$

Consequently, by Lemma 4.5, we see that

$$f(x_{k+1}) - f(x_k) \leq -\gamma\frac{\beta(1-\beta)}{M}\frac{\mu}{4}[f(x_k) - f(x_*)],$$

which yields

$$f(x_{k+1}) - f(x_*) \leq \left(1 - \mu\gamma\frac{\beta(1-\beta)}{4M}\right)[f(x_k) - f(x_*)],$$

as desired. □

5 Practical implications

By the results from the last section, we see that an essential hypothesis to obtain those statements is

$$\tau_k \leq \frac{\mu}{8L_{\max}}\|x_k - x_*\|^2. \tag{14}$$

By Step 2, the probability of such a condition to hold in any fixed $k \in \mathbb{N}$ is directly linked to the value of ϵ_k. If ϵ_k is significantly larger than the upper bound required for τ_k, then the probability of (14) to happen is low. On the other hand, if ϵ_k is small enough, such a condition has a high probability to hold.

At least theoretically, we have a strong argument to request that

$$\epsilon_k \approx \frac{\mu}{8L_{\max}}\|x_k - x_*\|^2.$$

Unfortunately, the knowledge of μ and $L_{\max}$ is not a reality for most of the problems. Moreover, x_* is the ultimate goal of GS, which implies that $\|x_k - x_*\|$ can not be directly computed. Therefore, it seems difficult to guarantee this approximation. Luckily, such a requirement is not impossible to be satisfied in practice.

Indeed, let us consider the infinite index set $\mathcal{K} \subset \mathbb{N}$ such that, for all $k \in \mathcal{K}$, we have that H_ϕ holds. Then, by (10), we see that

$$\|g_k\| \leq \left\| \sum_{i=1}^{r} \lambda_i^* \nabla \phi_i(x_k) \right\| + O(\tau_k).$$

For now, let us assume that, for all $k \in \mathbb{N}$, we know how to force $\tau_k = O(\|g_k\|^{2+\rho})$, for some fixed $\rho > 0$. Then,

$$\begin{aligned}\|g_k\| \left[1 + O(\|g_k\|^{1+\rho})\right] &\leq \left\|\sum_{i=1}^{r} \lambda_i^* \nabla \phi_i(x_k)\right\| \\ &= \left\|\sum_{i=1}^{r} \lambda_i^* \left[\nabla \phi_i(x_*) + \nabla^2 \phi_i(x_*)(x_k - x_*)\right]\right\| \\ &\quad + O(\|x_k - x_*\|^2).\end{aligned}$$

Consequently, since $\sum_{i=1}^{r} \lambda_i^* \nabla \phi_i(x_*) = 0$, we obtain $\|g_k\| = O(\|x_k - x_*\|)$. So, since we are assuming that $\tau_k = O(\|g_k\|^{2+\rho})$ and due to the limitation presented in (12), it yields

$$\tau_k = O(\|x_k - x_*\|^{2+\rho}).$$

Therefore, for any sufficiently large $k \in \mathcal{K}$, we must have that (14) is satisfied, which is our desired hypothesis.

The only gap that we have left here is how to ensure $\tau_k = O(\|g_k\|^{2+\rho})$. For this aim, we just need to set the following relation in Step 0:

$$\theta_\epsilon = (\theta_\nu)^{2+\rho}, \text{ for any desired value of } \rho > 0. \tag{15}$$

Indeed, defining l_k as the number of times the algorithm has reduced the sampling radius until the iteration k, and assuming that $x_{k+1} = x_k - t_k g_k$, we have

$$\tau_k \leq \epsilon_k = (\theta_\epsilon)^{l_k} \epsilon_0 = \left[(\theta_\nu)^{l_k}\right]^{2+\rho} \epsilon_0 = \left(\frac{\nu_k}{\nu_0}\right)^{2+\rho} \epsilon_0.$$

Hence, since $x_{k+1} = x_k - t_k g_k$, it yields that $\|g_k\| \geq \nu_k$, which guarantees $\tau_k = O(\|g_k\|^{2+\rho})$.

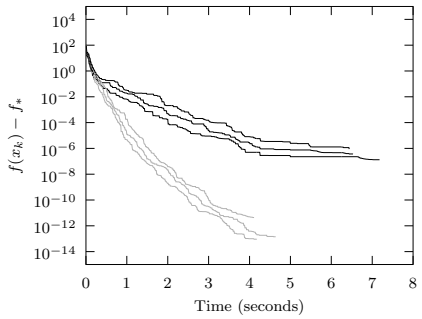
As a result, (15) gives a practical implication to the GS method. In fact, what one really needs to ask is that $\epsilon_k \leq \nu_k^{2+\rho}$, for all sufficiently large k. The equality (15) is just a way to ensure this relation between ϵ_k and ν_k, for all $k \geq 1$. For the best of our knowledge, there is no previous study that uses theoretical arguments to help a potential user to set the parameter values of θ_ν and θ_ϵ.

Finally, we present three illustrative examples in order to stress the importance of the relation (14). We have compared the number of iterations and time (in seconds) versus the distance of the current function value to the minimum function value f_* reached along twenty runs. For each example, we exhibit the median with the first and third quartiles of those runs. The curves in black stand for the GS method with the parameters suggested by the original authors [1], whereas the grey curves represent the same GS method but now using:

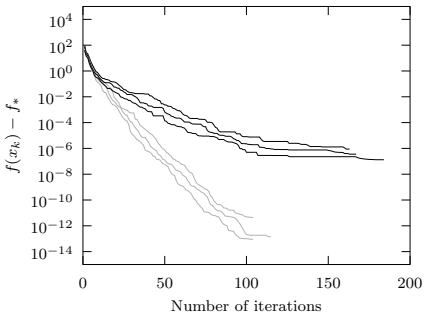
$$\nu_k = 10^{-(k+1)}, \forall k \geq 0; \ \epsilon_0 = \nu_0, \ \epsilon_1 = \nu_1^{1.5} \ \text{ and } \ \epsilon_k = \nu_k^{2.25}, \forall k \geq 2.$$

All the results were obtained using `Matlab` and its function `quadprog` for addressing the GS subproblem of Step 2.

In Figures 1 and 2, it is possible to see that relation (14) allows the method to keep behaving with linear rate of convergence until the final iterations, a characteristic that is not preserved for the usual GS method. Lastly, Figure 3 illustrates the necessity of Assumption 1, since it does not hold for the function `MAXQ`.

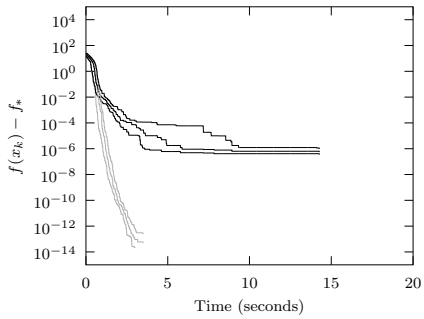


(a) $n = 10$

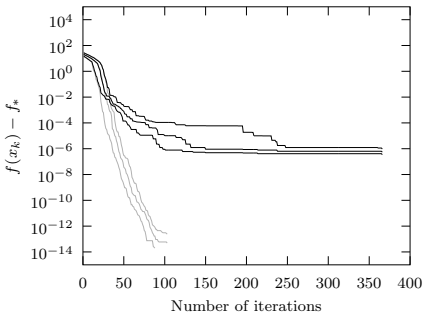


(b) $n = 10$

Fig. 1 Results for the nonsmooth convex function `Chained CB3 II` [9]. It satisfies $\dim\{U(x_*)\} \geq 1$.



(a) $n = 10$

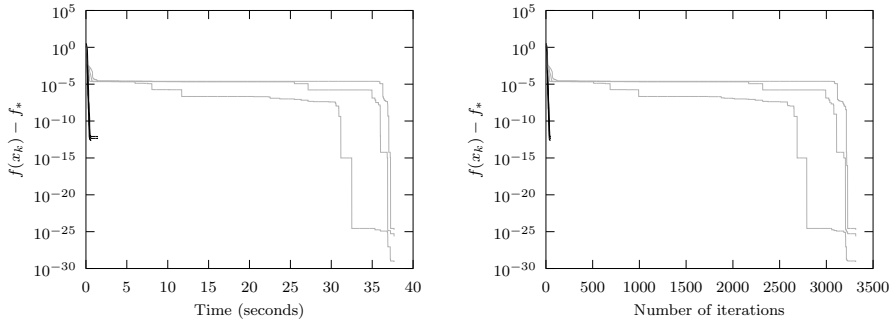


(b) $n = 10$

Fig. 2 Results for the nonsmooth nonconvex function `Chained Crescent I` [9]. It satisfies $\dim\{U(x_*)\} \geq 1$.

6 Conclusions

In this manuscript, we have established a linear local convergence result for the function value sequence generated by the nonnormalized version of the GS method. Our analysis does not provide any kind of local convergence result for functions such that $V(x_*) = \mathbb{R}^n$. Moreover, as it is reasonable to expect, for nonsmooth functions for which $\dim\{U(x_*)\} \geq 1$, a good decrease of the



(a) $n = 5$ (b) $n = 5$

Fig. 3 Results for the nonsmooth convex function `MAXQ` [9]. It does not satisfy Assumption 1.

function values is strongly dependent on a good set of sampled points. This set needs to cover all the functions ϕ_i such that $i \in \mathcal{I}(x_*)$. More than that, a restriction over the size of τ_k is also a crucial hypothesis.

Although the assumption over τ_k seems impracticable to be verified, we have shown that such a requirement can be satisfied by tunning properly the values of the parameters θ_ν and θ_ϵ. We believe this is an important implication, since as far we are concerned, there is no previous theoretical argumentation that corroborates any particular choices of such parameters.

In conclusion, this study reinforces what was already a belief in the nonsmooth field, by giving a theoretical proof and establishing in which circumstances one can expect linear local convergence of GS.

References

1. Burke, J.V., Lewis, A.S., Overton, M.L.: A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. SIAM Journal on Optimization **15**(3), 751–779 (2005)

2. Burke, J.V., Henrion, D., Lewis, A.S., Overton, M.L.: Stabilization via nonsmooth, nonconvex optimization. IEEE Transactions on Automatic Control **51**(11), 1760–1769 (2006)
3. Kiwiel, K.C.: Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. SIAM Journal on Optimization **18**(2), 379–388 (2007)
4. Clarke, F.H.: Optimization and nonsmooth analysis, vol. 5. SIAM, Montreal, Canada (1990)
5. Goldstein, A.A.: Optimization of Lipschitz continuous functions. Mathematical Programming **13**(1), 14–22 (1977)
6. Mifflin, R., Sagastizábal, C.: VU-decomposition derivatives for convex max-functions. In: M. Théra, R. Tichatschke (eds.) Ill-posed Variational Problems and Regularization Techniques, *Lecture Notes in Economics and Mathematical Systems*, vol. 477, pp. 167–186. Springer Berlin Heidelberg (1999)
7. Nocedal, J., Wright, S.: Numerical Optimization. Springer-Verlag, New York (2006)
8. Bonnans, J.F., Gilbert, J.C., Lemaréchal, C., Sagastizábal, C.A.: Numerical optimization: theoretical and practical aspects, 2nd edn. Springer-Verlag Berlin Heidelberg (2006)
9. Skajaa, A.: Limited memory BFGS for nonsmooth optimization. Master's thesis, Courant Institute of Mathematical Science, New York University (2010)