Discrete gradient method: a derivative free method for nonsmooth optimization

Adil M. Bagirov[1], Bülent Karasözen[2], Meral Sezer[3]

Communicated by F. Giannessi

[1]Corresponding author, a.bagirov@ballarat.edu.au, Centre for Informatics and Applied Optimization, School of Information Technology and Mathematical Sciences, University of Ballarat, Victoria 3353, Australia

[2]Department of Mathematics & Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey

[3]Department of Mathematics, Middle East Technical University, Ankara, Turkey

Abstract

In this paper a new derivative-free method is developed for solving unconstrained nonsmooth optimization problems. This method is based on the notion of a discrete gradient. It is demonstrated that the discrete gradients can be used to approximate subgradients of a broad class of nonsmooth functions. It is also shown that the discrete gradients can be applied to find descent directions of nonsmooth functions. The preliminary results of numerical experiments with unconstrained nonsmooth optimization problems as well as the comparison of the proposed method with nonsmooth optimization solver DNLP from CONOPT-GAMS and derivative-free optimization solver CONDOR are presented.

Keywords: nonsmooth optimization, derivative-free optimization, subdifferential, discrete gradients.

Mathematical Subject Classification: 65K05, 90C25.

1 Introduction

Consider the following nonsmooth unconstrained minimization problem:

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad x \in \mathbb{R}^n \tag{1}$$

where the objective function f is assumed to be Lipschitz continuous.

Nonsmooth unconstrained optimization problems appear in many applications and in particular in data mining. Over more than four decades different methods have been developed to solve problem (1). We mention among them the bundle method and its different variations (see, for example, [1, 2, 3, 4, 5, 6, 25, 7]), algorithms based on smoothing techniques [8] and the random subgradient sampling algorithm [9].

In most of these algorithms at each iteration the computation of at least one subgradient or approximating gradient is required. However, there are many practical problems where the computation of even one subgradient is a difficult task. Therefore in such situations derivative free methods seem to be better choice since they do not use explicit computation of subgradients.

Among derivative free methods, the generalized pattern search methods are well-suited for nonsmooth optimization [10, 11]. However their convergence are proved under quite restrictive differentiability assumptions. It was shown in [11] that when the objective function f is continuously differentiable in $\mathbb{R}^n$ then the limit inferior of the norm of the gradient of the sequence of points generated by the generalized pattern search algorithm goes to zero. The paper [10] provides convergence analysis under less restrictive differentiability assumptions. It was shown that if f is strictly differentiable near the limit of any refining subsequence, the gradient at that point is zero. However, in many practically important problems this condition is not satisfied, because in such problems the objective functions are not differentiable at local minimizers.

In this paper we develop a new derivative free method based on the notion of a discrete gradient for solving unconstrained nonsmooth optimization problems. First, we describe an algorithm for the computation of the subgradients of a broad class

of non-regular functions. Then we prove that the discrete gradients can be used to approximate subdifferentials of such functions. We also describe an algorithm for the computation of the descent directions of nonsmooth functions using discrete gradients. It is shown that the proposed algorithm converges for a broad class of nonsmooth functions. Finally, we present the comparison of the proposed algorithm with one nonsmooth optimization solver, DNLP from GAMS and one derivative-free solver CONDOR using results of numerical experiments.

The structure of the paper is as follows. Section 2 provides some necessary definitions used in the following sections and Section 3 presents problems from data clustering where the objective functions are non-regular and nonsmooth. An algorithm for approximation of subgradients is described in Section 4. Discrete gradients and their application for the approximation of subdifferentials are given in Section 5. In Section 6 we develop an algorithm for the computation of a descent direction and Section 7 presents the discrete gradient method. Results of numerical experiments are given in Section 8. Section 9 concludes the paper.

2 Preliminaries

In this section we will give some definitions which are used in the subsequent sections.

2.1 The Clarke subdifferential

Let f be a function defined on $\mathbb{R}^n$. The function f is called locally Lipschitz continuous if for any bounded subset $X \subset \mathbb{R}^n$ there exists an $L > 0$ such that

$$|f(x) - f(y)| \leq L\|x - y\| \quad \forall x, y \in X.$$

We recall that a locally Lipschitz function f is differentiable almost everywhere and that we can define for it a Clarke subdifferential [12] by

$$\partial f(x) = \text{co}\left\{v \in \mathbb{R}^n : \exists (x^k \in D(f), x^k \to x, k \to +\infty) : v = \lim_{k \to +\infty} \nabla f(x^k)\right\},$$

here $D(f)$ denotes the set where f is differentiable, co denotes the convex hull of a set. It is shown in [12] that the mapping $\partial f(x)$ is upper semicontinuous and bounded on bounded sets.

The generalized directional derivative of f at x in the direction g is defined as

$$f^0(x,g) = \limsup_{y \to x, \alpha \to +0} \alpha^{-1}[f(y+\alpha g) - f(y)].$$

If the function f is locally Lipschitz continuous then the generalized directional derivative exists and

$$f^0(x,g) = \max\{\langle v, g\rangle : v \in \partial f(x)\}$$

where $\langle \cdot, \cdot \rangle$ stands for an inner product in $\mathbb{R}^n$. f is called a Clarke regular function on $\mathbb{R}^n$, if it is differentiable with respect to any direction $g \in \mathbb{R}^n$ and $f'(x,g) = f^0(x,g)$ for all $x, g \in \mathbb{R}^n$ where $f'(x,g)$ is a derivative of the function f at the point x with respect to the direction g:

$$f'(x,g) = \lim_{\alpha \to +0} \alpha^{-1}[f(x+\alpha g) - f(x)].$$

It is clear that directional derivative $f'(x,g)$ of the Clarke regular function f is upper semicontinuous with respect to x for all $g \in \mathbb{R}^n$.

Let f be a locally Lipschitz continuous function defined on $\mathbb{R}^n$. For point x to be a minimum point of the function f on $\mathbb{R}^n$, it is necessary that

$$0 \in \partial f(x).$$

2.2 Semismooth functions and quasidifferentiability

The function $f : \mathbb{R}^n \to \mathbb{R}^1$ is called semismooth at $x \in \mathbb{R}^n$, if it is locally Lipschitz continuous at x and for every $g \in \mathbb{R}^n$, the limit

$$\lim_{g' \to g, \alpha \to +0} \langle v, g\rangle, \quad v \in \partial f(x + \alpha g')$$

exists. It should be noted that the class of semismooth functions is fairly wide and it contains convex, concave, max- and min-type functions [13]. The semismooth function f is directionally differentiable and

$$f'(x,g) = \lim_{g' \to g, \alpha \to +0} \langle v, g\rangle, \quad v \in \partial f(x + \alpha g').$$

Let f be a semismooth function defined on $\mathbb{R}^n$. Consider the following set at a point $x \in \mathbb{R}^n$ with respect to a given direction $g \in R^n,\ \|g\| = 1$:

$$R(x,g) = \text{co}\left\{v \in \mathbb{R}^n : \exists (v^k \in \partial f(x + \lambda_k g),\ \lambda_k \to +0,\ k \to +\infty) :\ v = \lim_{k \to +\infty} v^k\right\}.$$

It follows from the semismoothness of f that

$$f'(x,g) = \langle v, g\rangle \quad \forall v \in R(x,g)$$

and for any $\varepsilon > 0$ there exists $\lambda_0 > 0$ such that

$$\partial f(x + \lambda g) \subset R(x,g) + S_\varepsilon, \tag{2}$$

for all $\lambda \in (0, \lambda_0)$. Here

$$S_\varepsilon = \{v \in \mathbb{R}^n : \|v\| < \varepsilon\}.$$

A function f is called quasidifferentiable at a point x if it is locally Lipschitz continuous, directionally differentiable at this point and there exist convex, compact sets $\underline{\partial} f(x)$ and $\overline{\partial} f(x)$ such that:

$$f'(x,g) = \max_{u \in \underline{\partial} f(x)} \langle u, g\rangle + \min_{v \in \overline{\partial} f(x)} \langle v, g\rangle.$$

The set $\underline{\partial} f(x)$ is called a subdifferential, the set $\overline{\partial} f(x)$ is called a superdifferential and the pair of sets $[\underline{\partial} f(x), \overline{\partial} f(x)]$ is called a quasidifferential of the function f at a point x [14].

3 Data clustering as nonsmooth optimization problem

There are many applications when the objective and/or constraint functions are not regular functions. We will mention here only one of them, the cluster analysis problem, which is an important application area in data mining.

Clustering is also known as the unsupervised classification of patterns, deals with the problems of organization of a collection of patterns into clusters based

on similarity. There are many application of clustering in information retrieval, medicine etc.

In cluster analysis we assume that we have been given a finite set C of points in the n-dimensional space $\mathbb{R}^n$, that is

$$C = \{c^1, \ldots, c^m\}, \text{ where } c^i \in \mathbb{R}^n, \ i = 1, \ldots, m.$$

There are different types of clustering such as partition, packing, covering and hierarchical clustering. We consider here partition clustering, that is the distribution of the points of the set C into a given number q of disjoint subsets $C^i, \ i = 1, \ldots, q$ with respect to predefined criteria such that:

1) $C^i \neq \emptyset, \ i = 1, \ldots, q$;

2) $C^i \bigcap C^j = \emptyset, \ i, j = 1, \ldots, q, \ i \neq j$;

3) $C = \bigcup\limits_{i=1}^{q} C^i$.

The sets $C^i, \ i = 1, \ldots, q$ are called clusters. The strict application of these rules is called *hard clustering*, unlike *fuzzy clustering*, where the clusters are allowed to overlap. We assume that no constraints are imposed on the clusters $C^i, \ i = 1, \ldots, q$ that is we consider the hard unconstrained clustering problem.

We also assume that each cluster $C^i, \ i = 1, \ldots, q$ can be identified by its center (or centroid). There are different formulations of the clustering as an optimization problem. In [15, 16, 17, 18] the cluster analysis problem is reduced to the following nonsmooth optimization problem

$$\text{minimize} \ \ f(x^1, \ldots, x^q) \quad \text{subject to } (x^1, \ldots, x^q) \in \mathbb{R}^{n \times q}, \tag{3}$$

where

$$f(x^1, \ldots, x^q) = \frac{1}{m} \sum_{i=1}^{m} \min_{s=1,\ldots,q} \|x^s - c^i\|^2. \tag{4}$$

Here $\|\cdot\|$ is the Euclidean norm and $x^s \in \mathbb{R}^n$ stands for s-th cluster center. If $q > 1$, the objective function (4) in problem (3) is nonconvex and nonsmooth. Moreover, the function f is non-regular function and the computation of even one subgradient

of this function is quite difficult task. This function can be represented as the difference of two convex functions as follows

$$f(x) = f_1(x) - f_2(x)$$

where

$$f_1(x) = \frac{1}{m}\sum_{i=1}^{m}\sum_{s=1}^{q}\|x^s - c^i\|^2,$$

$$f_2(x) = \frac{1}{m}\sum_{i=1}^{m}\max_{s=1,\ldots,q}\sum_{k=1,k\neq s}^{q}\|x^k - c^i\|^2.$$

It is clear that the function f is quasidifferentiable and its subdifferential and superdifferential at any point are polytopes.

This example demonstrates the importance of development of derivative-free methods for nonsmooth optimization.

4 Approximation of subgradients

We consider a locally Lipschitz continuous function f defined on $\mathbb{R}^n$ and assume that this function is quasidifferentiable. We also assume that both sets $\underline{\partial} f(x)$ and $\overline{\partial} f(x)$ at any $x \in \mathbb{R}^n$ are represented as a convex hull of a finite number of points that is at a point $x \in \mathbb{R}^n$ there exist sets

$$A = \{a^1, \ldots, a^m\}, \quad a^i \in \mathbb{R}^n, \quad i = 1, \ldots, m, m \geq 1$$

and

$$B = \{b^1, \ldots, b^p\}, \quad b^j \in \mathbb{R}^n, \quad j = 1, \ldots, p, p \geq 1$$

such that

$$\underline{\partial} f(x) = \operatorname{co} A, \quad \overline{\partial} f(x) = \operatorname{co} B.$$

In other words we assume that the subdifferential and the superdifferential of the function f are polytopes at any $x \in \mathbb{R}^n$. This assumption is true, for example, for functions represented as a maximum, minimum or max-min of a finite number of smooth functions.

We take a direction $g \in \mathbb{R}^n$ such that:

$$g = (g_1, \ldots, g_n), \quad |g_i| = 1, \ i = 1, \ldots, n$$

and consider the sequence of n vectors $e^j = e^j(\alpha), \ j = 1, \ldots, n$ with $\alpha \in (0, 1]$:

$$\begin{aligned} e^1 &= (\alpha g_1, 0, \ldots, 0), \\ e^2 &= (\alpha g_1, \alpha^2 g_2, 0, \ldots, 0), \\ \ldots &= \ldots\ldots\ldots \\ e^n &= (\alpha g_1, \alpha^2 g_2, \ldots, \alpha^n g_n). \end{aligned}$$

We introduce the following sets:

$$\underline{R}_0(g) \equiv \underline{R}_0 = A, \quad \overline{R}_0(g) \equiv \overline{R}_0 = B,$$

$$\underline{R}_j(g) = \left\{ v \in \underline{R}_{j-1}(g) : v_j g_j = \max\{w_j g_j : \ w \in \underline{R}_{j-1}(g)\} \right\},$$
$$\overline{R}_j(g) = \left\{ v \in \overline{R}_{j-1}(g) : v_j g_j = \min\{w_j g_j : \ w \in \overline{R}_{j-1}(g)\} \right\}.$$

It is clear that

$$\underline{R}_j(g) \neq \emptyset, \ \forall j \in \{0, \ldots, n\}, \quad \underline{R}_j(g) \subseteq \underline{R}_{j-1}(g), \quad \forall j \in \{1, \ldots, n\}$$

and

$$\overline{R}_j(g) \neq \emptyset, \ \forall j \in \{0, \ldots, n\}, \quad \overline{R}_j(g) \subseteq \overline{R}_{j-1}(g), \quad \forall j \in \{1, \ldots, n\}.$$

Moreover

$$v_r = w_r \quad \forall v, w \in \underline{R}_j(g), \ r = 1, \ldots, j \tag{5}$$

and

$$v_r = w_r \quad \forall v, w \in \overline{R}_j(g), \ r = 1, \ldots, j. \tag{6}$$

Proposition 1 *Assume that the function f is quasidifferentiable and its subdifferential and superdifferential are polytopes at a point x. Then the sets $\underline{R}_n(g)$ and $\overline{R}_n(g)$ are singletons.*

Proof: It follows from (5) that for any $v, w \in \underline{R}_n(g)$ and from (6) that for any $v, w \in \overline{R}_n(g)$

$$v_r = w_r, \quad r = 1, \ldots, n$$

that is $v = w$. $\triangle$

Consider the following two sets:

$$\underline{R}(x, e^j(\alpha)) = \left\{ v \in A : \langle v, e^j \rangle = \max_{u \in A} \langle u, e^j \rangle \right\},$$

$$\overline{R}(x, e^j(\alpha)) = \left\{ w \in B : \langle w, e^j \rangle = \min_{u \in B} \langle u, e^j \rangle \right\}.$$

We take any $a \in A$. If $a \notin \underline{R}_n(g)$ then there exists $r \in \{1, \ldots, n\}$ such that $a \in \underline{R}_t(g),\ t = 0, \ldots, r-1$ and $a \notin \underline{R}_r(g)$. It follows from $a \notin \underline{R}_r(g)$ that

$$v_r g_r > a_r g_r \quad \forall v \in \underline{R}_r(g).$$

For $a \in A,\ a \notin \underline{R}_n(g)$ we define

$$d(a) = v_r g_r - a_r g_r > 0$$

and then the following number

$$d_1 = \min_{a \in A \setminus \underline{R}_n(g)} d(a).$$

Since the set A is finite and $d(a) > 0$ for all $a \in A \setminus \underline{R}_n(g)$ it follows that $d_1 > 0$.

Now we take any $b \in B$. If $b \notin \overline{R}_n(g)$ then there exists $r \in \{1, \ldots, n\}$ such that $b \in \overline{R}_t(g),\ t = 0, \ldots, r-1$ and $b \notin \overline{R}_r(g)$. It follows from $b \notin \overline{R}_r(g)$ that

$$v_r g_r < b_r g_r \quad \forall v \in \overline{R}_r(g).$$

For $b \in B,\ b \notin \overline{R}_n(g)$ we define

$$d(b) = b_r g_r - v_r g_r > 0$$

and then the following number

$$d_2 = \min_{b \in B \setminus \overline{R}_n(g)} d(b).$$

Since the set B is finite and $d(b) > 0$ for all $b \in B \setminus \overline{R}_n(g)$ it follows that $d_2 > 0$. Let

$$\bar{d} = \min\{d_1,\ d_2\}.$$

Since the subdifferential $\underline{\partial} f(x)$ and the superdifferential $\overline{\partial} f(x)$ are bounded on any bounded subset $X \subset \mathbb{R}^n$, there exists $D > 0$ such that $\|v\| \leq D$ and $\|w\| \leq D$ for all $v \in \underline{\partial} f(y),\ w \in \overline{\partial} f(y)$ and $y \in X$.

Take any $r, j \in \{1, \ldots, n\},\ r < j$. Then for all $v, w \in \underline{\partial} f(x),\ x \in X$ and $\alpha \in (0, 1]$ we have

$$\left| \sum_{t=r+1}^{j} (v_t - w_t) \alpha^{t-r} g_t \right| < 2D\alpha n.$$

Let $\alpha_0 = \min\{1, \bar{d}/(4Dn)\}$. Then for any $\alpha \in (0, \alpha_0]$

$$\left| \sum_{t=r+1}^{j} (v_t - w_t) \alpha^{t-r} g_t \right| < \frac{\bar{d}}{2}. \tag{7}$$

In a similar way we can show that for all $v, w \in \overline{\partial} f(x),\ x \in X$ and $\alpha \in (0, \alpha_0]$

$$\left| \sum_{t=r+1}^{j} (v_t - w_t) \alpha^{t-r} g_t \right| < \frac{\bar{d}}{2}. \tag{8}$$

Proposition 2 *Assume that the function f is quasidifferentiable and its subdifferential and superdifferential are polytopes at a point x. Then there exists $\alpha_0 > 0$ such that*

$$\underline{R}(x, e^j(\alpha)) \subset \underline{R}_j(g), \quad \overline{R}(x, e^j(\alpha)) \subset \overline{R}_j(g), \quad j = 1, \ldots, n$$

for all $\alpha \in (0, \alpha_0)$.

Proof: We will prove that $\underline{R}(x, e^j(\alpha)) \subset \underline{R}_j(g)$. The second inclusion can be proved in a similar way. Assume the contrary. Then there exists $y \in \underline{R}(x, e^j(\alpha))$ such that $y \notin \underline{R}_j(g)$. Consequently there exists $r \in \{1, \ldots, n\},\ r \leq j$ such that $y \notin \underline{R}_r(g)$ and $y \in \underline{R}_t(g)$ for any $t = 0, \ldots, r-1$. We take any $v \in \underline{R}_j(g)$. From (5) we have

$$v_t g_t = y_t g_t, \quad t = 1, \ldots, r-1, \quad v_r g_r \geq y_r g_r + \bar{d}.$$

It follows from (7) that for

$$\begin{aligned} \langle v, e^j \rangle - \langle y, e^j \rangle &= \sum_{t=1}^{j} (v_t - y_t) \alpha^t g_t \\ &= \alpha^r \left[v_r g_r - y_r g_r + \sum_{t=r+1}^{j} (v_t - y_t) \alpha^{t-r} g_t \right] \\ &> \alpha^r \bar{d}/2 > 0. \end{aligned}$$

Since $\langle y, e^j\rangle = \max\{\langle u, e^j\rangle : \ u \in \underline{\partial} f(x)\}$ and $v \in \underline{\partial} f(x)$ we get the contradiction

$$\langle y, e^j\rangle \geq \langle v, e^j\rangle > \langle y, e^j\rangle + \alpha^r \bar{d}/2.$$

$\triangle$

Corollary 1 *Assume that the function f is quasidifferentiable and its subdifferential and superdifferential are polytopes at a point x. Then there exits $\alpha_0 > 0$ such that*

$$f'(x, e^j(\alpha)) = f'(x, e^{j-1}(\alpha)) + v_j \alpha^j g_j + w_j \alpha^j g_j, \quad \forall v \in \underline{R}_j(g), \ \ w \in \overline{R}_j(g), \ \ j = 1, \ldots, n$$

for all $\alpha \in (0, \alpha_0]$.

Proof: Proposition 2 implies that $\underline{R}(x, e^j(\alpha)) \subset \underline{R}_j(g)$ and $\overline{R}(x, e^j(\alpha)) \subset \overline{R}_j(g), \ \ j = 1, \ldots, n$. Then there exist $v \in \underline{R}_j(g), \ w \in \overline{R}_j(g), \ v^0 \in \underline{R}_{j-1}(g), \ w^0 \in \overline{R}_{j-1}(g)$ such that

$$f'(x, e^j(\alpha)) - f'(x, e^{j-1}(\alpha)) = \langle v + w, e^j\rangle - \langle v^0 + w^0, e^{j-1}\rangle.$$

It follows from (5) and (6) that

$$\begin{aligned} f'(x, e^j(\alpha)) - f'(x, e^{j-1}(\alpha)) &= (v_j + w_j)\alpha^j g_j + \sum_{t=1}^{j-1} [(v_t + w_t) - (v_t^0 + w_t^0)]\alpha^t g_t \\ &= v_j \alpha^j g_j + w^j \alpha^j g_j. \end{aligned}$$

. $\triangle$

4.1 Computation of subgradients

Let $g \in \mathbb{R}^n, \ \ |g_i| = 1, \ i = 1, \ldots, n$ be a given vector and $\lambda > 0, \ \ \alpha > 0$ be given numbers. We define the following points

$$x^0 = x, \ \ x^j = x^0 + \lambda e^j(\alpha), \ \ j = 1, \ldots, n.$$

It is clear that

$$x^j = x^{j-1} + (0, \ldots, 0, \lambda \alpha^j g_j, 0, \ldots, 0), \ \ j = 1, \ldots, n.$$

Let $v = v(\alpha, \lambda) \in \mathbb{R}^n$ be a vector with the following coordinates:

$$v_j = (\lambda \alpha^j g_j)^{-1} \left[f(x^j) - f(x^{j-1}) \right], \ \ j = 1, \ldots, n. \tag{9}$$

For any fixed $g \in \mathbb{R}^n$, $|g_i| = 1,\ i = 1, \ldots, n$ and $\alpha > 0$ we introduce the following set:

$$V(g, \alpha) = \left\{ w \in \mathbb{R}^n : \ \exists (\lambda_k \to +0, \ \ k \to +\infty), \ \ w = \lim_{k \to +\infty} v(\alpha, \lambda_k) \right\}.$$

Proposition 3 *Assume that f is a quasidifferentiable function and its subdifferential and superdifferential are polytopes at x. Then there exists $\alpha_0 > 0$ such that*

$$V(g, \alpha) \subset \partial f(x)$$

for any $\alpha \in (0, \alpha_0]$.

Proof: It follows from the definition of vectors $v = v(g, \alpha)$ that

$$\begin{aligned} v_j &= (\lambda \alpha^j g_j)^{-1} \left[f(x^j) - f(x^{j-1}) \right] \\ &= (\lambda \alpha^j g_j)^{-1} \left[f(x^j) - f(x) - (f(x^{j-1}) - f(x)) \right] \\ &= (\lambda \alpha^j g_j)^{-1} \left[\lambda f'(x, e^j) - \lambda f'(x, e^{j-1}) + o(\lambda, e^j) - o(\lambda, e^{j-1}) \right] \end{aligned}$$

where

$$\lambda^{-1} o(\lambda, e^i) \to 0, \quad \lambda \to +0, \ \ i = j-1, j.$$

We take $w \in \underline{R}_n(g)$ and $y \in \overline{R}_n(g)$. By Proposition 1 w and y are unique. Since $\underline{R}_n(g) = \underline{R}(x, e^n)$ and $\overline{R}_n(g) = \overline{R}(x, e^n)$ it follows from Proposition 4.2 [14] (p. 146) that $w + y \in \partial f(x)$. The inclusions $w \in \underline{R}_n(g)$ and $y \in \overline{R}_n(g)$ imply that $w \in \underline{R}_j(g)$ and $y \in \overline{R}_j(g)$ for all $j \in \{1, \ldots, n\}$. Then it follows from Corollary 1 that there exists $\alpha_0 > 0$ such that

$$\begin{aligned} v_j(\alpha, \lambda) &= (\lambda \alpha^j g_j)^{-1} \left[\lambda \alpha^j g_j (w_j + y_j) + o(\lambda, e^j) - o(\lambda, e^{j-1}) \right] \\ &= w_j + y_j + (\lambda \alpha^j g_j)^{-1} \left[o(\lambda, e^j) - o(\lambda, e^{j-1}) \right] \end{aligned}$$

for all $\alpha \in (0, \alpha_0]$. Then for any fixed $\alpha \in (0, \alpha_0]$ and $g \in \mathbb{R}^n$ we have

$$\lim_{\lambda \to +0} |v_j(\alpha, \lambda) - (w_j + y_j)| = 0.$$

Consequently

$$\lim_{\lambda \to +0} v(\alpha, \lambda) = w + y \in \partial f(x).$$

$\triangle$

Remark 1 It follows from Proposition 3 that in order to approximate subgradients of quasidifferentiable functions one can choose a vector $g \in \mathbb{R}^n$ such that $|g_i| = 1,\ i = 1, \ldots, n$, sufficiently small $\alpha > 0,\ \lambda > 0$ and apply (9) to compute a vector $v(\alpha, \lambda)$. This vector is an approximation to a certain subgradient.

Remark 2 A class of quasidifferentiable functions presents a broad class of nonsmooth functions, including many interesting non-regular functions. Thus the scheme proposed in this section allows one to approximate subgradients of a broad class of nonsmooth functions.

5 Computation of subdifferentials and discrete gradients

In previous section we have demonstrated an algorithm for the computation of subgradients. In this section we consider an algorithm for the computation of subdifferentials. This algorithm is based on the notion of a discrete gradient. We start with the definition of the discrete gradient, which was introduced in [19] (for more details, see also [20, 21]).

Let f be a locally Lipschitz continuous function defined on $\mathbb{R}^n$. Let

$$S_1 = \{g \in \mathbb{R}^n : \|g\| = 1\}, \quad G = \{e \in \mathbb{R}^n : e = (e_1, \ldots, e_n), |e_j| = 1,\ j = 1, \ldots, n\},$$

$$P = \{z(\lambda) : z(\lambda) \in \mathbb{R}^1,\ z(\lambda) > 0,\ \lambda > 0,\ \lambda^{-1} z(\lambda) \to 0,\ \lambda \to 0\}.$$

Here S_1 is the unit sphere, G is the set of vertices of the unit hypercube in $\mathbb{R}^n$ and P is the set of univariate positive infinitesimal functions.

We take any $g \in S_1$ and define $|g_i| = \max\{|g_k|, \quad k = 1, \ldots, n\}$. We also take any $e = (e_1, \ldots, e_n) \in G$, a positive number $\alpha \in (0, 1]$ and define the sequence of n vectors $e^j(\alpha),\ j = 1, \ldots, n$ as in Section 4. Then for given $x \in \mathbb{R}^n$ and $z \in P$ we

define a sequence of $n+1$ points as follows:

$$\begin{aligned} x^0 &= x+ \quad \lambda g, \\ x^1 &= x^0+ \quad z(\lambda)e^1(\alpha), \\ x^2 &= x^0+ \quad z(\lambda)e^2(\alpha), \\ \ldots &= \ldots \quad \ldots \\ x^n &= x^0+ \quad z(\lambda)e^n(\alpha). \end{aligned}$$

Definition 1 *The discrete gradient of the function f at the point $x \in \mathbb{R}^n$ is the vector $\Gamma^i(x,g,e,z,\lambda,\alpha) = (\Gamma^i_1,\ldots,\Gamma^i_n) \in \mathbb{R}^n, g \in S_1$ with the following coordinates:*

$$\Gamma^i_j = [z(\lambda)\alpha^j e_j)]^{-1}\left[f(x^j) - f(x^{j-1})\right], \quad j = 1,\ldots,n, \quad j \neq i,$$

$$\Gamma^i_i = (\lambda g_i)^{-1}\left[f(x+\lambda g) - f(x) - \lambda \sum_{j=1, j\neq i}^{n} \Gamma^i_j g_j\right].$$

It follows from the definition that

$$f(x+\lambda g) - f(x) = \lambda\langle \Gamma^i(x,g,e,z,\lambda,\alpha), g\rangle \tag{10}$$

for all $g \in S_1,\ e \in G,\ z \in P,\ \lambda > 0,\ \alpha > 0$.

Remark 3 Definition 1 slightly differs from the definition of discrete gradients in [19, 20, 21] and this difference is in the definition of i-th coordinate of the discrete gradient.

Remark 4 One can see that the discrete gradient is defined with respect to a given direction $g \in S_1$ and in order to compute the discrete gradient $\Gamma^i(x,g,e,z,\lambda,\alpha)$ first we define a sequence of points $x^0,\ldots,x^n$ and compute the values of the function f at these points that is we compute $n+2$ values of this function including the point x. $n-1$ coordinates of the discrete gradient are defined similar to those of the vector $v(\alpha,\lambda)$ from Section 4 and i-th coordinate is defined so that to satisfy the equality (10) which can be considered as some version of the mean value theorem.

Proposition 4 *Let f be a locally Lipschitz continuous function defined on $\mathbb{R}^n$ and $L > 0$ is its Lipschitz constant. Then for any $x \in \mathbb{R}^n,\ g \in S_1,\ e \in G,\ \lambda > 0,\ z \in P,\ \alpha > 0$*

$$\|\Gamma^i\| \leq C(n)L$$

where

$$C(n) = (n^2 + 2n^{3/2} - 2n^{1/2})^{1/2}.$$

Proof: It follows from the definition of the discrete gradients that

$$|\Gamma_j^i| \leq L \ \ \forall\ j = 1, \ldots, n,\ j \neq i.$$

Then for $j = i$ we get

$$\begin{aligned} |\Gamma_i^i| &\leq \frac{\lambda L\|g\| + L\lambda \sum_{j=1, j\neq i}^n |g_j|}{\lambda |g_i|} \\ &= L\left(\frac{\|g\|}{|g_i|} + \sum_{j=1, j\neq i}^n \frac{|g_j|}{|g_i|}\right). \end{aligned}$$

Since $|g_i| = \max\{|g_j|,\ j = 1, \ldots, n\}$ we have

$$\frac{|g_j|}{|g_i|} \leq 1,\ j = 1, \ldots, n \ \text{ and } \ \frac{\|g\|}{|g_i|} \leq n^{1/2}.$$

Consequently

$$|\Gamma_i^i| \leq L(n + n^{1/2} - 1).$$

Thus

$$\|\Gamma^i\| \leq C(n)L$$

where

$$C(n) = (n^2 + 2n^{3/2} - 2n^{1/2})^{1/2}.$$

$\triangle$

For a given $\alpha > 0$ we define the following set:

$$B(x, \alpha) = \{v \in \mathbb{R}^n : \ \exists (g \in S_1,\ e \in G,\ z_k \in P,\ z_k \to +0,\ \lambda_k \to +0,\ k \to +\infty),$$

$$v = \lim_{k \to +\infty} \Gamma^i(x, g, e, z_k, \lambda_k, \alpha)\}. \tag{11}$$

Proposition 5 *Assume that f is semismooth, quasidifferentiable function and its subdifferential and superdifferential are polytopes at a point x. Then there exists $\alpha_0 > 0$ such that*

$$co\, B(x, \alpha) \subset \partial f(x)$$

for all $\alpha \in (0, \alpha_0]$.

Proof: Since the function f is semismooth it follows from (2) that for any $\varepsilon > 0$ there exists $\lambda_0 > 0$ such that

$$v \in R(x, g) + S_\varepsilon \tag{12}$$

for all $v \in \partial f(x + \lambda g)$ and $\lambda \in (0, \lambda_0)$. We take any $\lambda \in (0, \lambda_0)$. It follows from Proposition 3 and the definition of the discrete gradient that there exist $\alpha_0 > 0$ and $z_0(\lambda) \in P$ such that for any $\alpha \in (0, \alpha_0],\ z \in P, z(\lambda) < z_0(\lambda)$ can be found $v \in \partial f(x + \lambda g)$ so that

$$|\Gamma_j^i - v_j| < \varepsilon, \quad j = 1, \ldots, n,\ j \neq i.$$

(12) implies that for v can be found $w \in R(x, g)$ such that

$$\|v - w\| < \varepsilon.$$

Then

$$|\Gamma_j^i - w_j| < 2\varepsilon, \quad j = 1, \ldots, n,\ j \neq i. \tag{13}$$

Since the function f is semismooth and $w \in R(x, g)$ we get that $f'(x, g) = \langle w, g \rangle$. Consequently

$$f(x + \lambda g) - f(x) = \lambda \langle w, g \rangle + o(\lambda, g) \tag{14}$$

where $\lambda^{-1} o(\lambda, g) \to 0$ as $\lambda \to +0$. It follows from (10) that

$$f(x + \lambda g) - f(x) = \lambda \langle \Gamma^i(x, g, e, z, \lambda, \alpha), g \rangle.$$

The latter together with (14) implies

$$\Gamma_i^i - w_i = \sum_{j=1, j \neq i}^{n} (w_j - \Gamma_j^i) \frac{g_j}{g_i} + (\lambda g_i)^{-1} o(\lambda, g).$$

Taking into account (13) we get

$$|\Gamma_i^i - w_i| \leq 2(n-1)\varepsilon + n^{1/2}\lambda^{-1}|o(\lambda, g)|. \tag{15}$$

Since $\varepsilon > 0$ is arbitrary it follows from (13) and (15) that

$$\lim_{k \to +\infty} \Gamma^i(x, g, e, z_k, \lambda_k, \alpha) = w \in \partial f(x).$$

$\triangle$

Remark 5 Proposition 5 implies that discrete gradients can be applied to approximate subdifferentials of a broad class of semismooth, quasidifferentiable functions.

Remark 6 One can see that the discrete gradient contains three parameters: $\lambda > 0,\ z \in P$ and $\alpha > 0$. $z \in P$ is used to exploit semismoothness of the function f and it can be chosen sufficiently small. If f is semismooth quasidifferentiable function and its subdifferential and superdifferential are polytopes at any $x \in \mathbb{R}^n$ then for any $\delta > 0$ there exists $\alpha_0 > 0$ such that $\alpha \in (0, \alpha_0]$ for all $y \in S_\delta(x)$. The most important parameter is $\lambda > 0$. In the sequel we assume that $z \in P$ and $\alpha > 0$ are sufficiently small.

Consider the following set at a point $x \in \mathbb{R}^n$:

$$D_0(x, \lambda) = \operatorname{cl}\operatorname{co}\left\{v \in \mathbb{R}^n : \ \exists (g \in S_1, e \in G, z \in P) : \quad v = \Gamma^i(x, g, e, \lambda, z, \alpha)\right\}.$$

Proposition 4 implies that the set $D_0(x, \lambda)$ is compact and it is also convex for any $x \in \mathbb{R}^n$.

Corollary 2 *Let f be a quasidifferentiable semismooth function. Assume that in the equality*

$$f(x + \lambda g) - f(x) = \lambda f'(x, g) + o(\lambda, g),\ g \in S_1$$

$\lambda^{-1} o(\lambda, g) \to 0$ as $\lambda \to +0$ uniformly with respect to $g \in S_1$. Then for any $\varepsilon > 0$ there exists $\lambda_0 > 0$ such that

$$D_0(x, \lambda) \subset \partial f(x) + S_\varepsilon$$

for all $\lambda \in (0, \lambda_0)$.

Proof: We take $\varepsilon > 0$ and set

$$\bar{\varepsilon} = \varepsilon / \bar{Q}$$

where

$$\bar{Q} = \left(4n^2 + 4n\sqrt{n} - 6n - 4\sqrt{n} + 3\right)^{1/2}.$$

It follows from the proof of Proposition 5 and upper semicontinuity of the subdifferential $\partial f(x)$ that for $\bar{\varepsilon} > 0$ there exists $\lambda_1 > 0$ such that

$$\min_{v \in \partial f(x)} \sum_{j=1, j \neq i}^{n} \left(\Gamma_j^i(x, g, e, \lambda, z, \alpha) - v_j\right)^2 < \bar{\varepsilon}, \quad j = 1, \ldots, n, \ j \neq i \tag{16}$$

for all $\lambda \in (0, \lambda_1)$. Let

$$A_0 = \operatorname{Argmin}_{v \in \partial f(x)} \sum_{j=1, j \neq i}^{n} \left(\Gamma_j^i(x, g, e, \lambda, z, \alpha) - v_j\right)^2.$$

It follows from (15) and the assumption of the proposition that for $\bar{\varepsilon} > 0$ there exists $\lambda_2 > 0$ such that

$$\min_{v \in A_0} \left|\Gamma_i^i(x, g, e, \lambda, z, \alpha) - v_i\right| \leq \left(2(n-1) + n^{1/2}\right) \bar{\varepsilon} \tag{17}$$

for all $g \in S_1$ and $\lambda \in (0, \lambda_2)$. Let $\lambda_0 = \min(\lambda_1, \lambda_2)$. Then (16) and (17) imply that

$$\min_{v \in \partial f(x)} \|\Gamma^i(x, g, e, \lambda, z, \alpha) - v_i\| \leq \varepsilon$$

for all $g \in S_1$ and $\lambda \in (0, \lambda_0)$. $\triangle$

Corollary 2 shows that the set $D_0(x, \lambda)$ is an approximation to the subdifferential $\partial f(x)$ for sufficiently small $\lambda > 0$. However it is true at a given point. In order to get convergence results for a minimization algorithm based on discrete gradients we need some relationship between the set $D_0(x, \lambda)$ and $\partial f(x)$ in some neighborhood of a given point x. We will consider functions satisfying the following assumption.

Assumption 1 *Let $x \in \mathbb{R}^n$ be a given point. For any $\varepsilon > 0$ there exist $\delta > 0$ and $\lambda_0 > 0$ such that*

$$D_0(y, \lambda) \subset \partial f(x + \bar{S}_\varepsilon) + S_\varepsilon \tag{18}$$

for all $y \in S_\delta(x)$ and $\lambda \in (0, \lambda_0)$. Here

$$\partial f(x + \bar{S}_\varepsilon) = \bigcup_{y \in \bar{S}_\varepsilon(x)} \partial f(y), \quad \bar{S}_\varepsilon(x) = \{y \in \mathbb{R}^n : \ \|x - y\| \leq \varepsilon\}.$$

Corollary 2 shows that the class of functions satisfying Assumption 1 is fairly broad.

5.1 A necessary condition for a minimum

Consider problem (1) where $f : \mathbb{R}^n \to \mathbb{R}^1$ is arbitrary function.

Proposition 6 *Let $x^* \in \mathbb{R}^n$ be a local minimizer of the function f. Then there exists $\lambda_0 > 0$ such that*

$$0 \in D_0(x, \lambda)$$

for all $\lambda \in (0, \lambda_0)$.

Proof: Since $x^* \in \mathbb{R}^n$ is a local minimizer of the function f on $\mathbb{R}^n$ then there exists $\lambda_0 > 0$ such that

$$f(x^*) \leq f(x^* + \lambda g), \quad \forall\ g \in S_1,\ \lambda \in (0, \lambda_0).$$

On the other hand

$$0 \leq f(x^* + \lambda g) - f(x^*) = \lambda \langle \Gamma^i(x^*, g, e, z, \lambda), g \rangle \leq \max\{\langle v, g \rangle : \quad v \in D_0(x^*, \lambda)\}.$$

Since $D_0(x^*, \lambda)$ is compact and convex set it follows that

$$\max\{\langle v, g \rangle :\ v \in D_0(x^*, \lambda)\} \geq 0$$

for all $g \in S_1$ and therefore

$$\max\{\langle v, g \rangle :\ v \in D_0(x, \lambda)\} \geq 0$$

for all $g \in \mathbb{R}^n$. The latter means that $0 \in D_0(x, \lambda)$.

$\triangle$

Proposition 7 *Let $0 \notin D_0(x, \lambda)$ for a given $\lambda > 0$ and $v^0 \in \mathbb{R}^n$ be a solution to the following problem:*

$$\text{minimize} \quad \|v\|^2 \quad \text{subject to} \quad v \in D_0(x, \lambda).$$

Then the direction $g^0 = -\|v^0\|^{-1} v^0$ is a descent direction.

Proof: It is clear that $\|v^0\| > 0$ and it follows from the necessary condition for a minimum that

$$\langle v^0, v - v^0 \rangle \geq 0 \ \ \forall \ v \in D_0(x, \lambda).$$

The latter implies that

$$\max \left\{ \langle v, g^0 \rangle : \ v \in D_0(x, \lambda) \right\} = -\|v^0\|.$$

We have from (10)

$$\begin{aligned} f(x + \lambda g^0) - f(x) &= \lambda \langle \Gamma^i(x, g^0, e, \lambda, z, \alpha), g \rangle \\ &\leq \lambda \max \left\{ \langle v, g^0 \rangle : \ v \in D_0(x, \lambda) \right\} \\ &= -\lambda \|v^0\|. \end{aligned}$$

$\triangle$

Proposition 7 shows how the set $D_0(x, \lambda)$ can be used to compute descent directions. However, in many cases the computation of the set $D_0(x, \lambda)$ is not possible. In the next section we propose an algorithm for the computation of descent directions using a few discrete gradients from $D_0(x, \lambda)$.

6 Computation of descent directions

In this section we describe an algorithm for the computation of descent directions of the objective function f of Problem (1).

Let $z \in P, \lambda > 0, \alpha \in (0, 1]$, the number $c \in (0, 1)$ and a tolerance $\delta > 0$ be given.

Algorithm 1 An algorithm for the computation of the descent direction.

Step 1. Choose any $g^1 \in S_1, e \in G$, compute $i = \text{argmax} \{|g_j|, \ j = 1, \ldots, n\}$ and a discrete gradient $v^1 = \Gamma^i(x, g^1, e, z, \lambda, \alpha)$. Set $\overline{D}_1(x) = \{v^1\}$ and $k = 1$.

Step 2. Compute the vector $\|w^k\|^2 = \min\{\|w\|^2 : w \in \overline{D}_k(x)\}$. If

$$\|w^k\| \leq \delta, \tag{19}$$

then stop. Otherwise go to Step 3.

Step 3. Compute the search direction by $g^{k+1} = -\|w^k\|^{-1} w^k$.

Step 4. If

$$f(x + \lambda g^{k+1}) - f(x) \leq -c\lambda \|w^k\|, \tag{20}$$

then stop. Otherwise go to Step 5.

Step 5. Compute $i = \text{argmax}\,\{|g_j^{k+1}| : \ j = 1, \ldots, n\}$ and a discrete gradient

$$v^{k+1} = \Gamma^i(x, g^{k+1}, e, z, \lambda, \alpha),$$

construct the set $\overline{D}_{k+1}(x) = \text{co}\,\{\overline{D}_k(x) \bigcup \{v^{k+1}\}\}$, set $k = k + 1$ and go to Step 2.

We give some explanations to Algorithm 1. In Step 1 we compute the discrete gradient with respect to an initial direction $g^1 \in \mathbb{R}^n$. The distance between the convex hull $\overline{D}_k(x)$ of all computed discrete gradients and the origin is computed in Step 2. This problem can be solved using the algorithm from [22] (for more recent approaches to this problem, see also [23, 24]). If this distance is less than the tolerance $\delta > 0$ then we accept the point x as an approximate stationary point (Step 2), otherwise we compute another search direction in Step 3. In Step 4 we check whether this direction is a descent direction. If it is we stop and the descent direction has been computed, otherwise we compute another discrete gradient with respect to this direction in Step 5 and update the set $\overline{D}_k(x)$. At each iteration k we improve the approximation of the subdifferential of the function f.

In the next proposition we prove that Algorithm 1 terminates after a finite number of iterations.

Proposition 8 *Let f be a locally Lipschitz function defined on $\mathbb{R}^n$. Then for $\delta \in (0, \bar{C})$ either the condition (19) or the condition (20) satisfy after m computations of the discrete gradients where*

$$m \leq 2(\log_2(\delta/\bar{C})/\log_2 r + 1), \quad r = 1 - [(1-c)(2\bar{C})^{-1}\delta]^2,$$

$\bar{C} = C(n)L$ and $C(n)$ is a constant from Proposition 4.

Proof: First, we will show that if both conditions for the termination of the algorithm do not hold, then a new discrete gradient $v^{k+1} \notin \overline{D}_k(x)$. Indeed, in this case $\|w^k\| > \delta$ and

$$f(x + \lambda g^{k+1}) - f(x) > -c\lambda\|w^k\|.$$

It follows from (10) that

$$\begin{aligned} f(x + \lambda g^{k+1}) - f(x) &= \lambda\langle \Gamma^i(x, g^{k+1}, e, z, \lambda, \alpha), g^{k+1}\rangle \\ &= \lambda\langle v^{k+1}, g^{k+1}\rangle \\ &> -c\lambda\|w^k\|. \end{aligned}$$

Then we have

$$\langle v^{k+1}, w^k\rangle < c\|w^k\|^2. \tag{21}$$

On the other hand, since $w^k = \operatorname{argmin}\{\|w\|^2 : w \in \overline{D}_k(x)\}$, necessary condition for a minimum implies that for any $w \in \overline{D}_k(x)$

$$\langle w^k, w - w^k\rangle \geq 0$$

or

$$\langle w^k, w\rangle \geq \|w^k\|^2.$$

The latter and (21) imply that $v^{k+1} \notin \overline{D}_k(x)$.

Now we will show that the described algorithm is a terminating. For this purpose, it is sufficient to get upper estimation for the number of the computed discrete gradients m, after which:

$$\|w^m\| \leq \delta. \tag{22}$$

It is clear that for all $t \in [0, 1]$

$$\|w^{k+1}\|^2 \leq \|tv^{k+1} + (1 - t)w^k\|^2$$

or

$$\|w^{k+1}\|^2 \leq \|w^k\|^2 + 2t\langle w^k, v^{k+1} - w^k\rangle + t^2\|v^{k+1} - w^k\|^2.$$

It follows from Proposition 4 that

$$\|v^{k+1} - w^k\| \leq 2\bar{C}.$$

Hence taking into account the inequality (21), we have

$$\|w^{k+1}\|^2 < \|w^k\|^2 - 2t(1-c)\|w^k\|^2 + 4t^2\bar{C}^2.$$

For $t = (1-c)(2\bar{C})^{-2}\|w^k\|^2 \in (0,1)$ we get

$$\|w^{k+1}\|^2 < \left\{1 - [(1-c)(2\bar{C})^{-1}\|w^k\|]^2\right\}\|w^k\|^2. \tag{23}$$

Let $\delta \in (0, \bar{C})$. It follows from (23) and the condition $\|w^k\| > \delta, k = 1, \ldots, m-1$ that

$$\|w^{k+1}\|^2 < \left\{1 - [(1-c)(2\bar{C})^{-1}\delta]^2\right\}\|w^k\|^2.$$

We denote by $r = 1 - [(1-c)(2\bar{C})^{-1}\delta]^2$. It is clear that $r \in (0,1)$. Then we have

$$\|w^m\|^2 < r\|w^{m-1}\|^2 < \ldots < r^{m-1}\|w^1\|^2 < r^{m-1}\bar{C}^2.$$

If $r^{m-1}\bar{C}^2 \leq \delta^2$, then the inequality (22) is also true and therefore,

$$m \leq 2(\log_2(\delta/\bar{C})/\log_2 r + 1).$$

$\triangle$

Remark 7 Proposition 4 and the equality (10) are true for any $\lambda > 0$ and for any locally Lipschitz continuous functions. This means that Algorithm 1 can compute descent directions for any $\lambda > 0$ and for any locally Lipschitz continuous functions in a finite number of iterations. Sufficiently small values of λ give an approximation to the subdifferential and in this case Algorithm 1 computes local descent directions. For such directions $f(x + \alpha g) < f(x)$ for all $\alpha \in (0, \lambda]$. Larger values of λ do not give an approximation to the subdifferential and in this case for descent directions computed by Algorithm 1 $f(x + \lambda g) < f(x)$, however for sufficiently small $\alpha > 0$ one can expect that $f(x + \alpha g) \geq f(x)$. Such directions exist at local minimizers which are not global minimizers. We call them global descent directions.

7 The discrete gradient method

In this section we describe the discrete gradient method. Let sequences $\delta_k > 0,\ z_k \in P,\ \lambda_k > 0,\ \delta_k \to +0,\ z_k \to +0,\ \lambda_k \to +0,\ k \to +\infty$, sufficiently small number $\alpha > 0$ and numbers $c_1 \in (0,1), c_2 \in (0, c_1]$ be given.

Algorithm 2 Discrete gradient method

Step 1. Choose any starting point $x^0 \in \mathbb{R}^n$ and set $k = 0$.

Step 2. Set $s = 0$ and $x_s^k = x^k$.

Step 3. Apply Algorithm 1 for the computation of the descent direction at $x = x_s^k, \delta = \delta_k, z = z_k, \lambda = \lambda_k, c = c_1$. This algorithm terminates after a finite number of iterations $l > 0$. As a result we get the set $\overline{D}_l(x_s^k)$ and an element v_s^k such that

$$\|v_s^k\|^2 = \min\{\|v\|^2 :\ v \in \overline{D}_l(x_s^k)\}.$$

Furthermore either $\|v_s^k\| \le \delta_k$ or for the search direction $g_s^k = -\|v_s^k\|^{-1} v_s^k$

$$f(x_s^k + \lambda_k g_s^k) - f(x_s^k) \le -c_1 \lambda_k \|v_s^k\|. \tag{24}$$

Step 4. If

$$\|v_s^k\| \le \delta_k \tag{25}$$

then set $x^{k+1} = x_s^k, k = k + 1$ and go to Step 2. Otherwise go to Step 5.

Step 5. Construct the following iteration $x_{s+1}^k = x_s^k + \sigma_s g_s^k$, where σ_s is defined as follows

$$\sigma_s = \text{argmax}\left\{\sigma \ge 0 :\ f(x_s^k + \sigma g_s^k) - f(x_s^k) \le -c_2 \sigma \|v_s^k\|\right\}.$$

Step 6. Set $s = s + 1$ and go to Step 3.

For the point $x^0 \in \mathbb{R}^n$ we consider the set $M(x^0) = \{x \in \mathbb{R}^n : f(x) \le f(x^0)\}$.

Theorem 1 *Assume that the function f is semismooth quasidifferentiable, its subdifferential and superdifferential are polytopes at any $x \in \mathbb{R}^n$, Assumption 1 is fulfilled and the set $M(x^0)$ is bounded for starting points $x^0 \in \mathbb{R}^n$. Then every accumulation point of $\{x^k\}$ belongs to the set $X^0 = \{x \in \mathbb{R}^n : 0 \in \partial f(x)\}$.*

Proof: Since the function f is locally Lipschitz continuous and the set $M(x^0)$ is bounded

$$f_* = \inf \{f(x) : x \in \mathbb{R}^n\} > -\infty. \tag{26}$$

First we will show that the loop between Steps 3 and 5 stops after a finite number of steps. In other words for any $k > 0$ there exists $s \geq 0$ such that $\|v_s^k\| \leq \delta_k$. Indeed, since $c_2 \in (0, c_1]$ it follows from (24) that $\sigma_s \geq \lambda_k$. Then we can write

$$\begin{aligned} f(x_{s+1}^k) - f(x_s^k) &\leq -c_2 \sigma_s \|v_s^k\| \\ &\leq -c_2 \lambda_k \|v_s^k\|. \end{aligned}$$

If $\|v_s^k\| > \delta_k$ for all $s \geq 0$ then we have

$$f(x_{s+1}^k) - f(x_s^k) \leq -c_2 \lambda_k \delta_k$$

or

$$f(x_{s+1}^k) \leq f(x_0^k) - (s+1) c_2 \lambda_k \delta_k. \tag{27}$$

Since $\lambda_k > 0$ and $\delta_k > 0$ are fixed for any $k > 0$ it follows from (27) that $f(x_s^k) \to -\infty$ as $s \to +\infty$. This contradicts (26), that is the loop between Steps 3 and 5 stops after a finite number of steps and as a result we get a point x^{k+1} where

$$\min_{v \in \overline{D}_l(x^{k+1})} \|v\| \leq \delta_k.$$

Since $\overline{D}_l(x^{k+1}) \subset D_0(x^{k+1}, \lambda_k)$,

$$\min_{v \in D_0(x^{k+1}, \lambda_k)} \|v\| \leq \delta_k.$$

Replacing $k+1$ by k we get

$$\min_{v \in D_0(x^k, \lambda_{k-1})} \|v\| \leq \delta_{k-1}. \tag{28}$$

Since $\{f(x^k)\}$ is a decreasing sequence $x^k \in M(x^0)$ for all $k > 0$. Then the sequence $\{x^k\}$ is bounded and therefore it has at least one accumulation point. Assume x^* is any accumulation point of the sequence $\{x^k\}$ and $x^{k_i} \to x^*$ as $i \to +\infty$. Then we have from (28)

$$\min_{v \in D_0(x^{k_i}, \lambda_{k_i - 1})} \|v\| \leq \delta_{k_i - 1}. \tag{29}$$

According to Assumption 1 at the point x^* for any $\varepsilon > 0$ there exist $\beta > 0$ and $\lambda_0 > 0$ such that

$$D_0(y, \lambda) \subset \partial f(x^* + \bar{S}_\varepsilon) + S_\varepsilon \tag{30}$$

for all $y \in S_\beta(x^*)$ and $\lambda \in (0, \lambda_0)$. Since the sequence $\{x^{k_i}\}$ converges to x^* for $\beta > 0$ there exists $i_0 > 0$ such that $x^{k_i} \in S_\beta(x^*)$ for all $i \geq i_0$. On the other hand since $\delta_k, \lambda_k \to 0$ as $k \to +\infty$ there exists $k_0 > 0$ such that $\delta_k < \varepsilon$ and $\lambda_k < \lambda_0$ for all $k > k_0$. Then there exists $i_1 \geq i_0$ such that $k_i \geq k_0 + 1$ for all $i \geq i_1$. Thus it follows from (29) and (30) that

$$\min_{v \in \partial f(x^* + \bar{S}_\varepsilon)} \|v\| \leq 2\varepsilon$$

Since $\varepsilon > 0$ is arbitrary and the mapping $\partial f(x)$ is upper semicontinuous $0 \in \partial f(x^*)$. $\triangle$

Remark 8 Since Algorithm 1 can compute descent directions for any values of $\lambda > 0$ we take $\lambda_0 \in (0, 1)$, some $\beta \in (0, 1)$ and update $\lambda_k,\ k \geq 1$ as follows:

$$\lambda_k = \beta^k \lambda_0, \quad k \geq 1.$$

Thus in the discrete gradient method we use approximations to subgradients only at the final stage of the method which guarantees convergence. In most of iterations we do not use explicit approximations of subgradients. Therefore it is a derivative-free method.

Remark 9 One can see some similarity between the discrete gradient and bundle-type methods. More specifically, the method presented in this paper can be considered as a derivative-free version of the bundle method introduced in [25]. The main difference is that in the proposed method discrete gradients are used instead of subgradients at some points from neighborhood of a current point. The discrete gradients need not to be approximations to subgradients at all iterations.

Remark 10 It follows from (24) and $c_2 \leq c_1$ that always $\sigma_s \geq \lambda_k$ and therefore $\lambda_k > 0$ is a lower bound for σ_s. This leads to the following rule for the computation of σ_s. We define a sequence:

$$\theta_m = m\lambda_k, \quad m \geq 1$$

and σ_s is defined as the largest θ_m satisfying the inequality in Step 5.

8 Numerical experiments

The efficiency of the proposed algorithm was verified by applying it to some unconstrained nonsmooth optimization problems. In numerical experiments we use 20 unconstrained test problems from [26]: Problems 2.1-7 (P1-P7), 2.9-12 (P9-P12), 2.14-16 (P14-P16), 2.18-21 (P18-P21), 2.23-24 (P23, P24).

Objective functions in these problems are discrete maximum functions and they are regular. Objective functions in Problems 2.1, 2.5, 2.23 are convex and they are nonconvex in all other problems. This means that most of problems are problems of global optimization and the same algorithm may find different solutions starting from different initial points and/or different algorithms may find different solutions starting from the same initial point. The brief description of these problems is given in Table 1 where the following notation is used:

- n - number of variables;
- n_m - number of functions under maximum;
- f_{opt} - optimum value (as reported in [26]).

For the comparison we use DNLP model of CONOPT solver from The General Algebraic Modeling System (GAMS) and CONDOR solver. DNLP is nonsmooth optimization solver and it is based on smoothing techniques. More details on DNLP can be found in [27]. CONDOR is a derivative free solver based on quadratic interpolation and trust region approach (see, [28] for more details).

Numerical experiments were carried out on PC Pentium 4 with CPU 1.6 MHz. We used 20 random initial points for each problem and initial points are the same for all three algorithms.

The results of numerical experiments are presented in Table 2. We use the following notation:

- f_{best} and f_{av} - the best and average objective function values over 20 runs, respectively;

Table 1: The brief description of test problems

Prob.	n	n_m	f_{opt}	Prob.	n	n_m	f_{opt}
P1	2	3	1.95222	P12	4	21	0.00202
P2	2	3	0	P14	5	21	0.00012
P3	2	2	0	P15	5	30	0.02234
P4	3	6	3.59972	P16	6	51	0.03490
P5	4	4	-44	P18	9	41	0.00618
P6	4	4	-44	P19	7	5	680.63006
P7	3	21	0.00420	P20	10	9	24.30621
P9	4	11	0.00808	P21	20	18	133.72828
P10	4	20	115.70644	P23	11	10	261.08258
P11	4	21	0.00264	P24	20	31	0.00000

- nfc - the average number of the objective function evaluations (for the discrete gradient method (DGM) and CONDOR);
- $iter$ - the average number of iterations (for DNLP);
- DN stands for DNLP and CR for CONDOR;
- F means that an algorithm failed for all initial points.

One can draw the following conclusions from Table 2:

1. The discrete gradient method finds the best solutions for all problems whereas the CONDOR solver could find the best solutions only for Problems 1.1-3 and the DNLP solver only for Problems 1.1, 1.4.
2. Average results over 20 runs by the discrete gradient method are better than those by the DNLP and CONDOR solvers, except Problems 2.2, 2.3, 2.6, 2.7 and 2.24 where the CONDOR solver produces better results.
3. For three convex problems 2.1, 2.5 and 2.23 the discrete gradient method always finds the same solutions, that is the best and average results by this

method are the same with respect to some tolerance. However this is not the case for the DNLP and CONDOR solvers.

4. For many test problems results by the DNLP solver are significantly worse than those by the CONDOR solver and the discrete gradient method. In these problems the values of objective functions or their gradients is too large and the DNLP solver fails to solve such problems. Results for Problems 2.6 and 2.23 demonstrate it. However the CONDOR solver and the discrete gradients method are quite efficient to solve such problems.

5. As it was mentioned above the most of the test problems are global optimization problems. Results presented clearly demonstrate that the derivative-free methods are better than Newton-like methods to solve global optimization problems. Overall the discrete gradient method outperforms other two methods.

6. One can see from Table 2 that the number of function calls by the CONDOR solver is significantly less than those by the discrete gradient method. However, there is no any significant difference in the CPU time used by different algorithms.

Since the most of test problems are nonconvex we suggest the following scheme to compare the performance of algorithms for each run. Let $\bar{f}$ be the best value obtained by all algorithm starting from the same initial point. Let f^1 be the value of the objective function at the final point obtained by an algorithm. If

$$f^1 - \bar{f} \leq \varepsilon(|\bar{f}| + 1)$$

then we say that this algorithm finds the best result with respect to the tolerance $\varepsilon > 0$. Tables 3 and 4 present pairwise comparison and the comparison of three algorithms, respectively. The numbers in these tables show how many times an algorithm could find the best solution with respect to the tolerance $\varepsilon = 10^{-4}$. Results presented in Table 3 demonstrate that the CONDOR solver outperforms the DNLP solver in 90 % of runs, the discrete gradient method outperforms the DNLP solver

in more than 95 % of runs and finally the discrete gradient method outperforms the CONDOR solver in almost 80 % of runs.

Results presented in Table 4 show that the discrete gradient method outperforms other two solvers in all problems except Problems 2.3 and 2.7 where the CONDOR solver outperforms others.

Overall results presented in this paper demonstrate that the discrete gradient method is more efficient than the DNLP and CONDOR solvers for solving nonsmooth nonconvex optimization problems.

9 Conclusions

In this paper we have proposed a derivative free algorithm, the discrete gradient method for solving unconstrained nonsmooth optimization problems. This algorithm can be applied to a broad class of nonsmooth optimization problems including problems with non-regular objective functions.

We have tested the new algorithm on some nonsmooth optimization problems. In the most of these problems the objective functions are regular nonconvex functions. For comparison we used one nonsmooth optimization algorithm: DNLP solver from GAMS which is based on the smoothing of the objective function and one derivative free CONDOR solver which is based on the quadratic approximation of the objective function. Preliminary results of our numerical experiments show that the discrete gradient method outperforms other two algorithms for the most of test problems considered in this paper. Our results also demonstrate that derivative-free methods are better than Newton-like methods for solving global optimization problems. We can conclude that the discrete gradient method is a good alternative to existing derivative-free nonsmooth optimization algorithms.

Acknowledgements

The authors are grateful to the anonymous referee for comments improving the quality of the paper. The research by the first author was supported by the Australian Research Council.

References

[1] HIRIART-URRUTY, J.B. and LEMARCHAL, C., *Convex Analysis and Minimization Algorithms*, Springer Verlag, Heidelberg, Vol. 1 and 2, 1993.

[2] KIWIEL, K.C., *Methods of Descent for Nondifferentiable Optimization,* Lecture Notes in Mathematics, Springer-Verlag, Berlin, Vol. 1133, 1985.

[3] KIWIEL, K.C., *Proximal control in bundle methods for convex nondifferentiable minimization*, Mathematical Programming, Vol. 29, pp. 105-122, 1990.

[4] LEMARCHAL, C., *An extension of Davidon methods to nondifferentiable problems*, Nondifferentiable Optimization, Balinski, M.L. and Wolfe, P. (eds.), Mathematical Programming Study, Vol. 3, pp. 95-109, North-Holland, Amsterdam, 1975.

[5] LEMARCHAL, C., and ZOWE, J., *A condensed introduction to bundle methods in nonsmooth optimization*, In: Algorithms for Continuous Optimization, Spedicato, E. (ed.,), pp. 357 - 482, Kluwer Academic Publishers, 1994.

[6] MIFFLIN, R., *An algorithm for constrained optimization with semismooth functions*, Math. Oper. Res., Vol. 2, pp. 191-207, 1977.

[7] ZOWE, J., *Nondifferentiable optimization: A motivation and a short introduction into the subgradient and the bundle concept*, In: NATO SAI Series, Vol. 15, Computational Mathematical Programming, Schittkowski, K., (ed.), pp. 323-356, Springer-Verlag, New York, 1985.

[8] POLAK, E. and ROYSET, J.O., *Algorithms for finite ans semi-infinite min-max-min problems using adaptive smoothing techniques*, Journal of Optimization Theory and Applications, Vol. 119, pp. 421-457, 2003.

[9] BURKE, J.V., LEWEIS, A.S. and OVERTON, M.L., *Approximating subdifferentials by random sampling of gradients*, Mathematics of Operations Research, Vol. 27, pp. 567-584, 2002.

[10] AUDET, C. and DENNIS, J.E., Jr., *Analysis of generalized pattern searches*, SIAM Journal on Optimization, Vol. 13, pp. 889-903, 2003.

[11] TORZCON, V., *On the convergence of pattern search algorithms*, SIAM Journal on Optimization, Vol. 7, pp. 1-25, 1997.

[12] CLARKE, F.H., *Optimization and Nonsmooth Analysis,* New York: John Wiley, 1983.

[13] MIFFLIN, R., *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control and Optimization, Vol. 15, pp. 959-972, 1977.

[14] DEMYANOV, V.F. and RUBINOV, A.M., *Constructive Nonsmooth Analysis*, Peter Lang, Frankfurt am Main, 1995.

[15] BAGIROV, A.M., RUNINOV, A.M. and YEARWOOD, J., *A global optimisation approach to classification*, Optimization and Engineering, Vol. 3, pp. 129-155, 2002.

[16] BAGIROV, A.M., RUBINOV, A.M., SOUKHOROUKOVA, A.V., and YEARWOOD, J., *Supervised and unsupervised data classification via nonsmooth and global optimisation*, TOP: Spanish Operations Research Journal, Vol. 11, pp. 1-93, 2003.

[17] BAGIROV, A.M. and UGON, J., *An algorithm for minimizing clustering functions*, Optimization, Vol. 54, pp. 351-368, 2005.

[18] BAGIROV, A.M. and YEARWOOD, J., *A new nonsmooth optimisation algorithm for minimum sum-of-squares clustering problems*, European Journal of Operational Research, Vol. 170, pp. 578-596, 2006.

[19] BAGIROV, A.M. and GASANOV, A.A., *A method of approximating a quasidifferential*, Journal of Computational Mathematics and Mathematical Physics, Vol. 35, pp. 403-409, 1995.

[20] BAGIROV, A.M., *Minimization methods for one class of nonsmooth functions and calculation of semi-equilibrium prices*, In: A. Eberhard et al. (eds.),

Progress in Optimization: Contribution from Australasia, pp. 147-175, Kluwer Academic Publishers, 1999.

[21] BAGIROV, A.M. , *Continuous subdifferential approximations and their applications*, Journal of Mathematical Sciences, Vol. 115, pp. 2567-2609, 2003.

[22] WOLFE, P.H., *Finding the nearest point in a polytope*, Mathematical Programming, Vol. 11, pp. 128-149, 1976.

[23] FRANGONI, A., *Solving semidefinite quadratic problems within nonsmooth optimization algorithms*, Comput. Oper. Res. Vol. 23, pp. 1099-1118, 1996.

[24] KIWIEL, K.C., *A dual method for certainpositive semidefinite quadratic programming problems*, SIAM J. Sci. Statist. Comput. Vol. 10, pp. 175-186, 1989.

[25] WOLFE, P.H., *A method of conjugate subgradients of minimizing nondifferentiable convex functions*, Mathematical Programming Study, Vol. 3, pp. 145-173, 1975.

[26] LUKSAN, L. and VLECK, J., *Test Problems for Nonsmooth Unconstrained and Linearly Constrained Optimization*, Technical Report No. 78, Institute of Computer Science, Academy of Sciences of the Czech Republic, 2000.

[27] GAMS: *The Solver Manuals*, GAMS Development Corporation, Washington, D.C., 2004.

[28] BERGEN, F. V., *CONDOR: a constrained, non-linear, derivative-free parallel optimizer for continuous, high computing load, noisy objective functions*, PhD thesis, Université Libre de Bruxelles, Belgium, 2004.

Table 2: Results of numerical experiments: best and average values

Pr.	f_{best}			f_{av}			iter	nfc	
	DN	CR	DGM	DN	CR	DGM	DN	CR	DGM
P1	1.9523	1.9522	1.9522	18.6631	1.9563	1.9522	44	88	314
P2	0.0014	0.00000	0.0000	12.6534	0.0021	0.9075	145	97	5018
P3	0.0000	0.0000	0.0000	66.6476	0.0000	0.2200	161	848	8943
P4	3.5998	3.6126	3.5997	9.7501	3.80643	3.5997	44	233	1079
P5	-43.9236	-43.9970	-44	-41.1455	-43.8039	-44	48	352	2862
P6	F	-43.9966	-44	F	-43.8198	-42.8657	F	433	10120
P7	0.0475	0.0042	0.0042	0.0586	0.0054	0.0416	6	215	1316
P9	0.0152	0.02581	0.0081	0.2142	0.0443	0.0179	47	279	5441
P10	116.4907	115.7644	115.7064	256.2033	116.7891	115.7064	51	283	2152
P11	0.0354	0.0103	0.0029	5.3347	0.1679	0.0032	46	343	2677
P12	0.0858	0.0398	0.0125	0.3957	0.1269	0.0628	48	314	2373
P14	2.1643	0.0350	0.0011	2.6490	0.2886	0.1826	34	937	3575
P15	0.7218	0.1405	0.0223	42.9887	0.3756	0.2787	40	592	4656
P16	0.3957	0.0548	0.0349	1.1809	0.4786	0.2872	25	796	7410
P18	0.3085	0.0814	0.0356	0.7024	0.2460	0.1798	43	1289	9694
P19	716.6131	686.0436	680.6301	803.8032	689.7816	680.6301	50	725	2654
P20	35.4217	24.9150	24.3062	60.0370	27.0316	24.3062	121	1892	12926
P21	118.5468	97.8171	93.9073	286.5167	102.7173	94.4516	120	9301	43633
P23	F	3.7053	3.7035	F	3.7107	3.7035	F	3054	3886
P24	0.8953	0.4278	0.3987	15.2643	0.6523	0.8337	58	7418	17928

Table 3: Pairwise comparison of algorithms

Prob.	First pair		Second pair		Third pair	
	DNLP	CONDOR	DNLP	DGM	CONDOR	DGM
P1	5	16	1	20	7	20
P2	1	19	3	17	8	17
P3	3	20	5	16	20	4
P4	10	10	2	20	0	20
P5	2	18	0	20	1	20
P6	0	20	0	20	3	18
P7	0	20	6	20	15	5
P9	4	16	1	19	0	20
P10	0	20	0	20	0	20
P11	5	15	0	20	0	20
P12	6	14	1	19	3	17
P14	0	20	0	20	2	18
P15	2	18	0	20	6	14
P16	1	19	0	20	4	16
P18	1	19	1	20	5	15
P19	0	20	0	20	0	20
P20	0	20	0	20	0	20
P21	0	20	0	20	1	19
P23	0	20	0	20	0	20
P24	0	20	1	19	8	12
Total	40	363	21	390	83	335

Table 4: Comparison of algorithms

Prob.	DNLP	CONDOR	DGM	Prob.	DNLP	CONDOR	DGM
P1	1	7	20	P12	1	3	16
P2	1	7	17	P14	0	2	18
P3	3	20	4	P15	0	6	14
P4	2	0	20	P16	0	4	16
P5	0	1	20	P18	0	5	15
P6	0	3	18	P19	0	0	20
P7	0	15	5	P20	0	0	20
P9	1	0	19	P21	0	1	19
P10	0	0	20	P23	0	0	20
P11	0	0	20	P24	0	8	12
Total	9	82	333				