Approximate methods in nonsmooth optimization

Adil Bagirov

CIAO, University of Ballarat, Victoria, Australia

Nonlinear Programming and Applications, Beijing, April 7-9, 2008

Outline

- Introduction
- Motivation
- Approximation of subgradients
- Computation of descent directions
- Minimization algorithm
- Computational results
- Conclusions

Introduction: Subdifferential

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

The subdifferential:

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

Generalized directional derivatives:

$$f^0(x,g) = \limsup_{y \to x, \; \alpha \to +0} \frac{f(y + \alpha g) - f(y)}{\alpha}$$

Regular functions:

$$f'(x,g) = f^0(x,g)$$

The function f is called quasidifferentiable at a point x if

- it is locally Lipschitz continuous, directionally differentiable at this point;
- there exist convex, compact sets $\underline{\partial} f(x)$ and $\overline{\partial} f(x)$ such that:

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

$\underline{\partial} f(x)$ - a subdifferential, $\overline{\partial} f(x)$ - a superdifferential, the pair $[\underline{\partial} f(x), \overline{\partial} f(x)]$ - a quasidifferential.

Introduction

- For regular functions a calculus exists with equalities which can be used to estimate subgradients.

- f_1 and f_2 are not regular:

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

$$\partial f(x) \subset \partial f_1(x) + \partial f_2(x).$$

Introduction

- Difference of two convex compact sets (Demyanov, 1983):

  A and B are convex sets, p_A, p_B their support functions, T is any full-measure subset.

$$A - B = \text{clco}\{\nabla p_A(x) - \nabla p_B(x) : x \in T\}.$$

- $$\underline{\partial} f(x) - (-\overline{\partial} f(x) \subset \partial f(x).$$

Motivation: Cluster analysis

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

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

We consider the hard unconstrained partition clustering problem, that is the distribution of the points of the set A into a given number k of disjoint subsets $A^j, \ j = 1, \ldots, k$ with respect to predefined criteria such that:

1) $A^j \neq \emptyset, \ j = 1, \ldots, k$;
2) $A^j \bigcap A^l = \emptyset, \ j, l = 1, \ldots, k, \ j \neq l$;
3) $A = \bigcup_{j=1}^{k} A^j$;
4) no constraints on the clusters $A^j, \ j = 1, \ldots, k$.

Motivation: Cluster analysis

$$\text{minimize} \quad f_k(x) \quad \text{subject to } x = (x^1, \ldots, x^k) \in \mathbb{R}^{n \times k}, \tag{1}$$

where

$$f_k(x^1, \ldots, x^k) = \frac{1}{m} \sum_{i=1}^{m} \min_{j=1,\ldots,k} \|x^j - a^i\|^2. \tag{2}$$

(Bagirov, Rubinov, Sukhorukova and Yearwood, TOP, 2003, Bagirov and Yearwood, EJOR, 2006, M. Teboulle, JMLR, 2007)

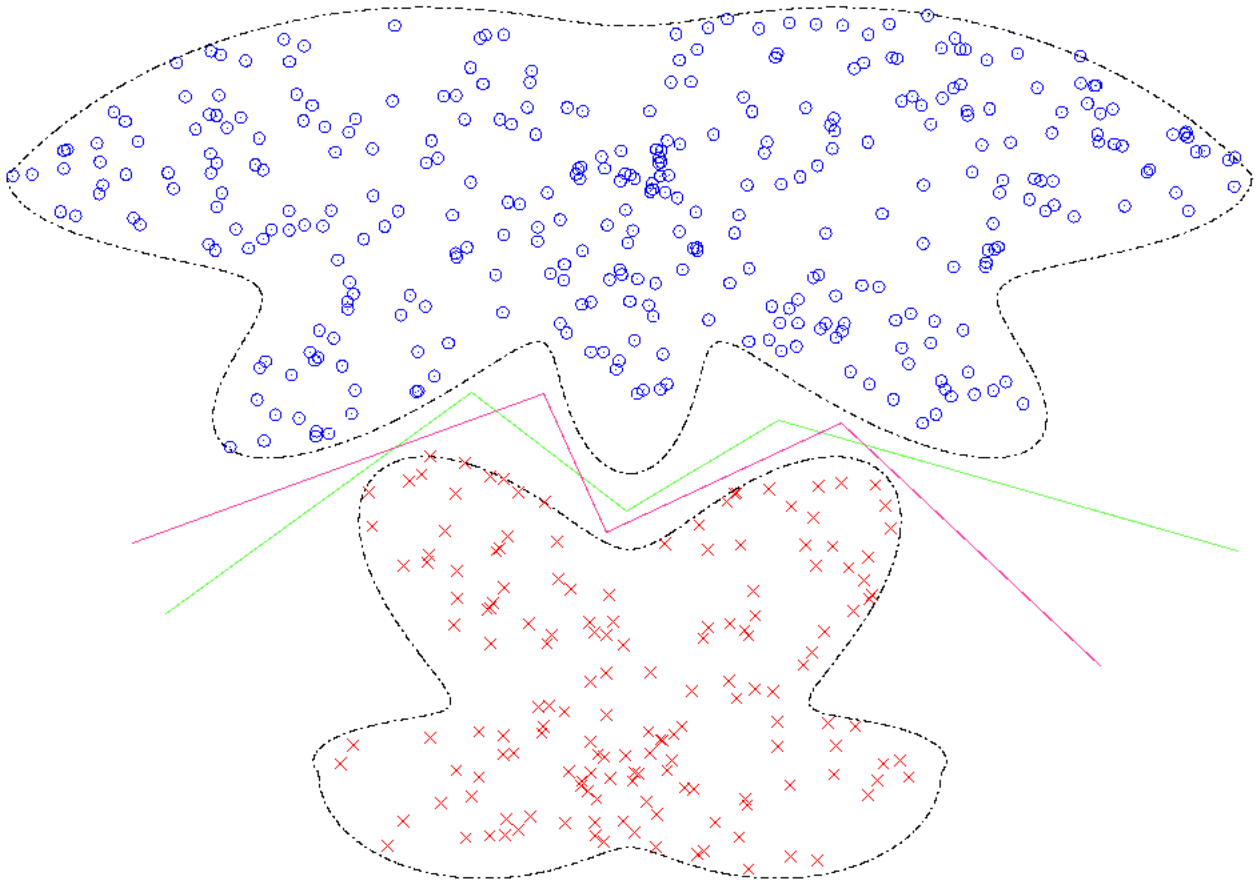
Piecewise linear separability:

Let A and B be given sets containing m and p n-dimensional vectors, respectively:

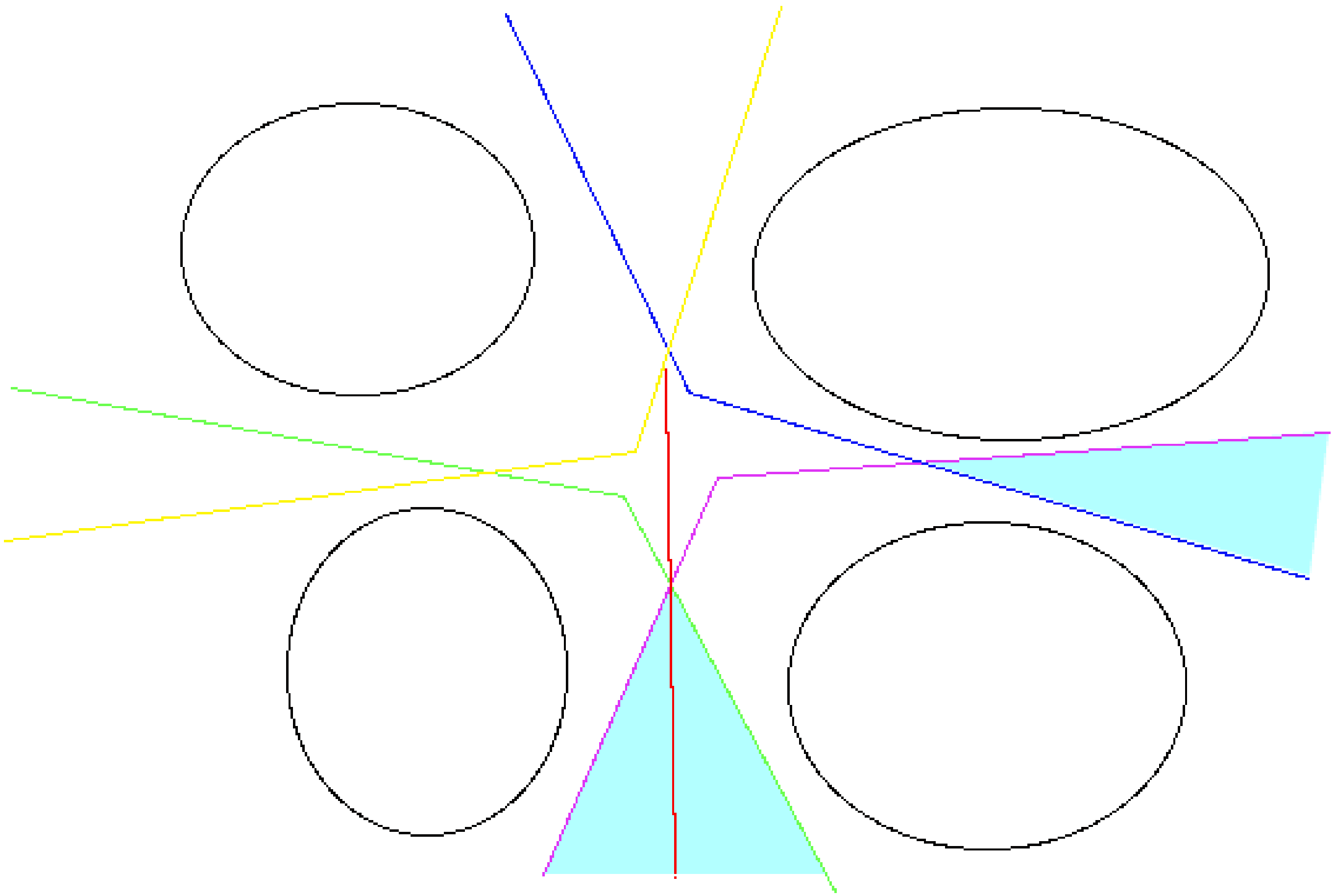
$$A = \{a^1, \ldots, a^m\}, \; a^i \in \mathbb{R}^n, \; i = 1, \ldots, m,$$

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

Motivation: supervised data classification



Motivation: supervised data classification



Motivation: supervised data classification

Max-min separability:

An averaged error function is defined as

$$F(x, y) = (1/m)\sum_{k=1}^{m} \max\left[0, \max_{i\in I}\min_{j\in J_i}\left\{\langle x^{ij}, a^k\rangle - y_{ij} + 1\right\}\right]$$

$$+(1/p)\sum_{t=1}^{p} \max\left[0, \min_{i\in I}\max_{j\in J_i}\left\{-\langle x^{ij}, b^t\rangle + y_{ij} + 1\right\}\right].$$

(**Bagirov, Optimization Methods and Software, 2005)**

Motivation: The estimation of a regression function

In *regression analysis* an $\mathbb{R}^p \times \mathbb{R}^1$-valued random vector (U, V) with $EV^2 < \infty$ is considered and the dependency of V on the value of U is of interest. More precisely, the goal is to find a function $\varphi : \mathbb{R}^p \to \mathbb{R}^1$ such that $\varphi(U)$ is a "good approximation" of V.

Main aim of the analysis is minimization of the mean squared prediction error or L_2 *risk*

$$\mathbf{E}\{|\varphi(U) - V|^2\}.$$

In this case the optimal function is the so-called *regression function* $m : \mathbb{R}^p \to \mathbb{R}^1$, $m(u) = \mathbf{E}\{V|U = u\}$.

Motivation: The estimation of a regression function

In applications, usually the distribution of (U, V) (and hence also the regression function) is unknown. But often it is possible to observe a sample of the underlying distribution. This leads to the *regression estimation problem*. Here (U, V), (U_1, V_1), (U_2, V_2), ... are independent and identically distributed random vectors. The set of data

$$\mathcal{D}_l = \{(U_1, V_1), \dots, (U_l, V_l)\}$$

is given, and the goal is to construct an estimate

$$m_l(\cdot) = m_l(\cdot, \mathcal{D}_l) : \mathbb{R}^p \to \mathbb{R}^1$$

of the regression function such that the L_2 error

$$\int |m_l(u) - m(u)|^2 \mu(du)$$

is small.

Motivation: The estimation of a regression function

For least squares estimates the given data is used to estimate the L_2 risk by the so-called empirical L_2 risk

$$\frac{1}{l}\sum_{i=1}^{l}|\varphi(U_i) - V_i|^2,$$

and the regression estimate is defined by minimizing this function.

First a class $\mathcal{F}_l$ of functions $\varphi : \mathbb{R}^p \to \mathbb{R}^1$ is chosen and then the estimate is defined by minimizing the empirical L_2 risk over $\mathcal{F}_l$, i.e.,

$$m_l(\cdot) = \arg\min_{\varphi\in\mathcal{F}_l}\frac{1}{l}\sum_{i=1}^{l}|\varphi(U_i) - V_i|^2.$$

Motivation: The estimation of a regression function

$$\mathcal{F}_l = \Bigg\{ \varphi : \mathbb{R}^p \to \mathbb{R}^1 \ : \ \varphi(u) = \max_{k=1,\ldots,K_l} \min_{j=1,\ldots,L_{k,l}} \left(\langle x^{k,j}, u \rangle + y_{k,j} \right) (u \in \mathbb{R}^p),$$

$$\text{for some } x^{k,j} \in \mathbb{R}^p, y_{k,j} \in \mathbb{R}^1 \Bigg\}$$

$$\text{minimize} \quad F(x,y) = \frac{1}{l} \sum_{i=1}^{l} \left(\max_{k=1,\ldots,K_l} \min_{j=1,\ldots,L_{k,l}} \left(\langle x^{k,j}, U_i \rangle + y_{k,j} \right) - V_i \right)^2 .$$

(Bagirov, Clausen and Kohler, COAP, 2008).

Motivation: Wireless local area networks planning

Wireless local area networks' (WLAN) access points are common in large public areas. Network planning is essential in cellular networks to warrant substantial investment savings.

Consider outdoor compact scenario characterized by valleys and hills. All points $x = (x_1, x_2)$ belong a well defined compact set $X \subset \mathbb{R}^2$ and the surface $\varphi(\cdot) : X \to \mathbb{R}$ is known continuous function. Distance $\delta(x, y)$ between two points $x, y \in X$ is defined by

$$\delta^2(x, y) = \|x - y\|^2 + (\varphi(x) - \varphi(y))^2.$$

Motivation: Wireless local area networks planning

$$S = \{s_1, \dots, s_p\} \subset X.$$

Given a set $S \subset X$ and a point $x \in X$ we say that x is visible from the set S if there exists $s \in S$ such that

$$\varphi(\lambda x + (1-\lambda)s)) \leq \lambda\varphi(x) + (1-\lambda)\varphi(s)), \quad \forall\lambda \in [0,1].$$

The Path Loss $g(S,x)$ is given by

$$g(S,x) = 10 \min_{s \in S} \log_{10}\left[\frac{4\pi}{\lambda}(\delta^2(s,x)+\theta)\right].$$

Motivation: Wireless local area networks planning

$$Y = \{x^1, \ldots, x^q\} \subset X$$

is a discretized set. $V(S)$ is the set of all points $x \in Y$ visible from the set S.

$$\text{minimize} \sum_{x \in Y \bigcap V(S)} [g(S,x) + \mu_M \max(0, g(S,x) - g_M)]$$

subject to

$$S = (s^1, \ldots, s^p) \in \mathbb{R}^{2p}.$$

(F.J. Gonzalez-Castano, et al., COAP, 2007)

Motivation: More examples

- Localization of sensor networks (Bagirov, Lai and Palaniswami, 2008)
- Minimization of eigenvalue products (Burke, Lewis and Overton, SIAMOPT, 2005)
- Computation of a separating set (Grzybowski, Pallaschke and Urbanski, OMS, 2005)

Approximation of subgradients

Approximating subdifferentials by random sampling of gradients (Burke, Lewis and Overton).

Approximation of subgradients

- f is quasidifferentiable.

- its subdifferential $\underline{\partial} f(x)$ and superdifferential $\overline{\partial} f(x)$ at any $x \in \mathbb{R}^n$ are polytopes:

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

and

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

such that

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

Approximation of subgradients

$$G = \{e \in \mathbb{R}^n : \quad |e_i| = 1, \ i = 1, \ldots, n\}$$

Take $e \in G$. Consider sets:

$$\underline{R}_0 = A, \quad \overline{R}_0 = B,$$

$$\underline{R}_j = \left\{v \in \underline{R}_{j-1} : v_j e_j = \max\{w_j e_j : \ w \in \underline{R}_{j-1}\}\right\},$$

$$\overline{R}_j = \left\{v \in \overline{R}_{j-1} : v_j e_j = \min\{w_j e_j : \ w \in \overline{R}_{j-1}\}\right\}.$$

Proposition 1 *The sets $\underline{R}_n$ and $\overline{R}_n$ are singletons.*

Approximation of subgradients

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

$$\begin{array}{rcl}
e^1 & = & (\alpha e_1, 0, \ldots, 0), \\
e^2 & = & (\alpha e_1, \alpha^2 e_2, 0, \ldots, 0), \\
\ldots & = & \ldots\ldots\ldots \\
e^n & = & (\alpha e_1, \alpha^2 e_2, \ldots, \alpha^n e_n).
\end{array}$$

Approximation of subgradients

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

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

Proposition 2 *There exists* $\alpha_0 > 0$ *such that*

$$\underline{R}(x, e^j(\alpha)) \subset \underline{R}_j, \quad \overline{R}(x, e^j(\alpha)) \subset \overline{R}_j, \quad j = 1, \ldots, n, \quad \forall \alpha \in (0, \alpha_0).$$

Corollary 1 *There exits* $\alpha_0 > 0$ *such that*

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

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

Approximation of subgradients

Take $e \in G$ and define the following points

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

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

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

$$v_j = (\lambda \alpha^j e_j)^{-1} \left[f(x^j) - f(x^{j-1}) \right], \quad j = 1, \ldots, n. \tag{3}$$

Introduce the following set:

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

Proposition 3 *There exists* $\alpha_0 > 0$ *such that*

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

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

Approximation of subgradients

Let $x \in \mathbb{R}^n$ be a given point. The above described scheme allows us to easily check whether the function is strictly differentiable at this point.

- Take any $e \in G$, a sufficiently small $\alpha \in (0,1]$ and compute a subgradient $v^1 \in \partial f(x)$.

- Then we take a vector $e^2 \in G$ such that $e^2 = -e$ and compute a subgradient $v^2 \in \partial f(x)$.

Proposition 4 *If $v^1 = v^2$ then the function f is strictly differentiable at x, otherwise f is nondifferentiable at x.*

Approximation of subgradients

Let $x^0 \in \mathbb{R}^n$ be a given point.

- There exist $\alpha_0 \in (0, 1]$ and $\lambda_0 > 0$ such that the funstion f is strictly differentiable at points $x^n(e) = x^0 + \lambda e^n(\alpha)$ for all $e \in G$, $\alpha \in (0, \alpha_0]$ and $\lambda \in (0, \lambda_0]$.

Approximation of subgradients

$$S_1 = \{g \in \mathbb{R}^n : \|g\| = 1\},$$

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

We take any $g \in S_1$ and define $|g_i| = \max\{|g_k|, \quad k = 1, \ldots, n\}$. We define a sequence of $n+1$ points as follows:

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

Approximation of subgradients

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:*

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

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

Approximation of subgradients

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

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

$$v = \lim_{k \longrightarrow +\infty} \Gamma^i(x, g, e, z_k, \lambda_k, \alpha)\}. \qquad (4)$$

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*

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

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

Approximation of subgradients

Consider two polytopes A and B. Given a vector $e \in G$ we can construct sets:

$$\underline{R}_j(e, A), \quad \underline{R}_j(e, B) \quad j = 1, \ldots, n.$$

The sets $\underline{R}_n(e, A)$, $\underline{R}_n(e, B)$ are singletons. For $g \in S_1$ define the sets:

$$Q_A(g) = \text{Argmax}\,\{\langle v, g\rangle : \ v \in A\}\,, \quad Q_B(g) = \text{Argmax}\,\{\langle v, g\rangle : \ v \in B\}\,.$$

The difference between two polytopes A and B:

$$A \hat{-} B = \{v \in \mathbb{R}^n : \ \exists (g \in S_1, \ e \in G) : \ v = w^1 - w^2,$$

$$w^1 \in \underline{R}_n(e, Q_A(g)), \ w^2 \in \underline{R}_n(e, Q_B(g))\}$$

$$\partial f(x) = \underline{\partial} f(x) \hat{-} (-\overline{\partial} f(x)).$$

Computation of descent directions

Let $e \in G$, $\lambda > 0$, the number $c \in (0,1)$ and a tolerance $\delta > 0$ be given.

Algorithm 1 Computation of the descent direction.

Step 1. Choose $g^1 \in S_1$ and compute $v^1 = \Gamma^i(x, g^1, e, z, \lambda, \alpha)$. Set $\overline{D}_1(x) = \{v^1\}$ and $k = 1$.

Step 2. Compute $\|w^k\|^2 = \min\{\|w\|^2 : w \in \overline{D}_k(x)\}$. If $\|w^k\| \le \delta$, stop.

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

Step 4. If $f(x + \lambda g^{k+1}) - f(x) \le -c\lambda\|w^k\|$, stop.

Step 5. Compute $v^{k+1} = \Gamma^i(x, g^{k+1}, e, z, \lambda, \alpha)$, 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.

Minimization algorithm

Let sequence $\delta_k, \ \lambda_k > 0, \delta_k, \ \lambda_k \to 0, \ k \to \infty$ 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. Apply Algorithm 1 for the computation of the descent direction at $x = x^k, \delta = \delta_k, \lambda = \lambda_k$. This algorithm terminates after a finite number of iterations. As a result it either finds the descent direction or that the point x^k is δ_k-stationary point.

Step 3. If it finds the descent direction do line search and update the point x^k, otherwise update λ_k and δ_k and go to Step 2.

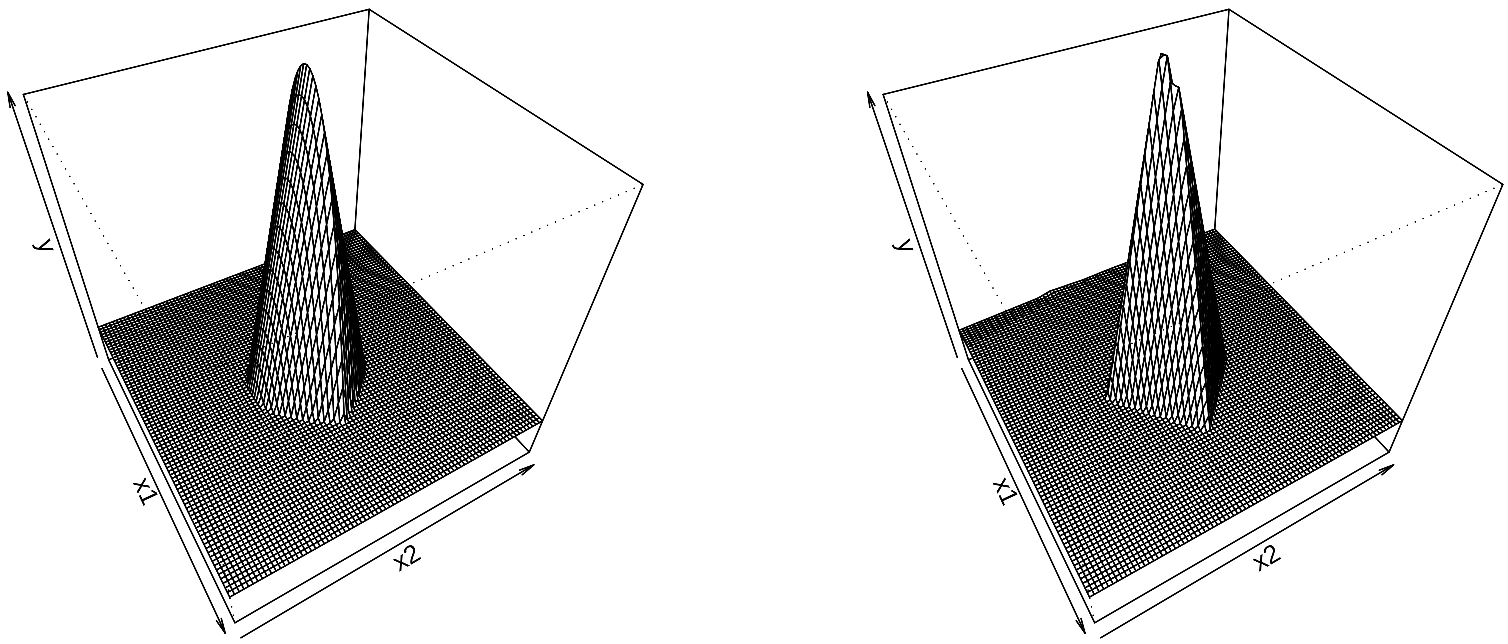
If the function f is semismooth quasidifferentiable, its subdifferential and superdifferential are polytopes then the algorithm converges to Clarke stationary points.

Computational results

The method was applied to solve the following problems:

1. Cluster analysis problems (Bagirov and Yearwood, EJOR, 2006, 2008);
2. Supervised data classification problems (Bagirov, OMS, 2005, Bagirov, Ugon and Webb, 2008) ;
3. Estimation of regression functions (Bagirov, Clausen and Kohler, COAP, 2008).
4. Localization of sensor networks (Bagirov, Lai and Palaniswami, 2008).

The estimation of a regression function



THANK YOU!