Gradient methods in nonconvex optimization

Maxim Balashov[1]

[1]Laboratory 7
Institute of Control Sciences RAS, Moscow

June 18, 2019

Outline

The problem

The problem under consideration is

$$\min_{x \in A} f(x) \qquad (*)$$

$A \subset \mathbb{R}^d$ is closed, $f : \mathbb{R}^d \to \mathbb{R}$ has the Lipschiz continuous gradient.

We are interested in situations when A or/and f are nonconvex.

Panos M. Pardolos, H. Edwin Romeijn (Eds.) Handbook of global optimization. Volume 2 // Nonconvex optimization and its applications. Springer Science + Business Media Dordreht, 2002. 569 P.

Convex algorithms for nonconvex case?

The gradient projection algorithm (GPA) — 1

Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ and $A \subset \mathbb{R}^d$ is a closed subset.
GPA

$$x_0 \in A, \qquad x_{k+1} = P_A(x_k - \alpha_k f'(x_k)).$$

The function f has the Lipschitz continuous gradient and convex, A is a convex closed set.
Goldstein, 1964,
Polyak, Levitin 1966.

The gradient projection algorithm (GPA) — 2

A function $f : \mathbb{R}^d \to \mathbb{R}$ is strongly convex with constant $\varkappa > 0$ iff $f(x) - \frac{\varkappa}{2}\|x\|^2$ is convex.
For a Lipschitz differentiable with Lipschitz constant L_1 strongly convex function and convex set A the GPA process converges with the rate

$$\|x_k - x_*\| \leq \left(1 - \frac{\varkappa}{L_1}\right)^k \|x_0 - x_*\|$$

The Ležanski-Polyak-Lojasiewicz (LPL) and the Error bound (EB) condition

For nonconvex (Frechet) differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ the LPL-condition means that $\exists \mu > 0$ with

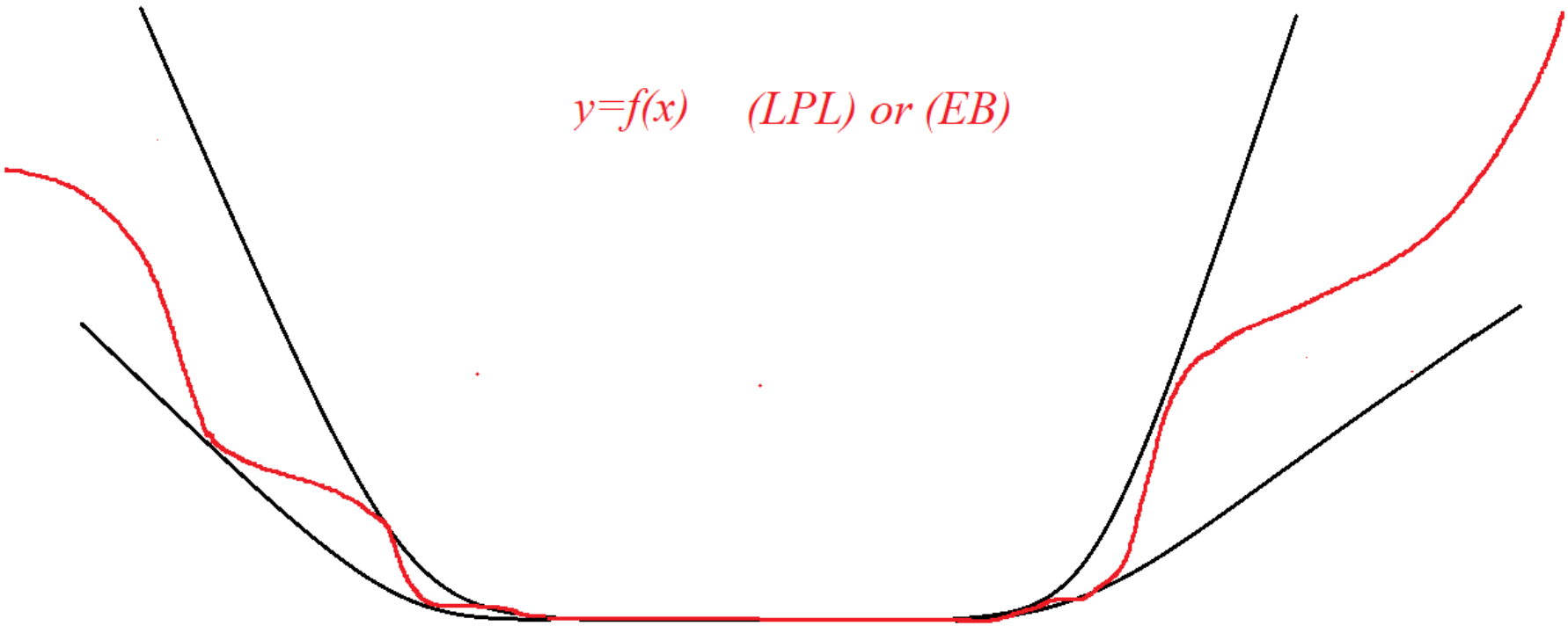
$$\|f'(x)\|^2 \geq \mu(f(x) - f_*), \quad \forall x \in \mathbb{R}^d, \quad f_* = \inf_{\mathbb{R}^d} f.$$

Ležanski 1962; Polyak, Lojasiewicz 1963.
The Error bound (EB) condition means that $\exists \nu > 0$ with

$$\|f'(x)\| \geq \nu \operatorname{dist}(x, \Omega), \quad \forall x \in \mathbb{R}^d.$$

Here $x \in \Omega \Leftrightarrow f(x) = f_*$.
Under (LPL) or (EB) conditions $x_{k+1} = x_k - \frac{1}{L_1} f'(x_k)$ converges with linear rate.

(LPL)/(EB) conditions - a picture



The Frank-Wolfe algorithm (conditional gradient)

The (FW) algorithm: $x_0 \in A$

$$y_k \in \operatorname{Arg}\min_{x \in A}(-f'(x_k), x),$$

$$x_{k+1} \in \operatorname{Arg}\min_{t \in [0,1]} f((1-t)y_k + tx_k), \quad k = 0, 1, \ldots.$$

Sublinear rate of convergence ($O(\frac{1}{k})$) for convex case.

Proximally smooth sets

A closed set $A \subset \mathbb{R}^d$ is called proximally smooth (or prox-regular) with constant $R > 0$ if the distance function $\varrho_A(x) = \inf\limits_{a \in A} \|x - a\|$ is (Frechet) continuously differentiable on the set

$$U_A(R) = \{x \in \mathbb{R}^d \mid 0 < \varrho_A(x) < R\}.$$

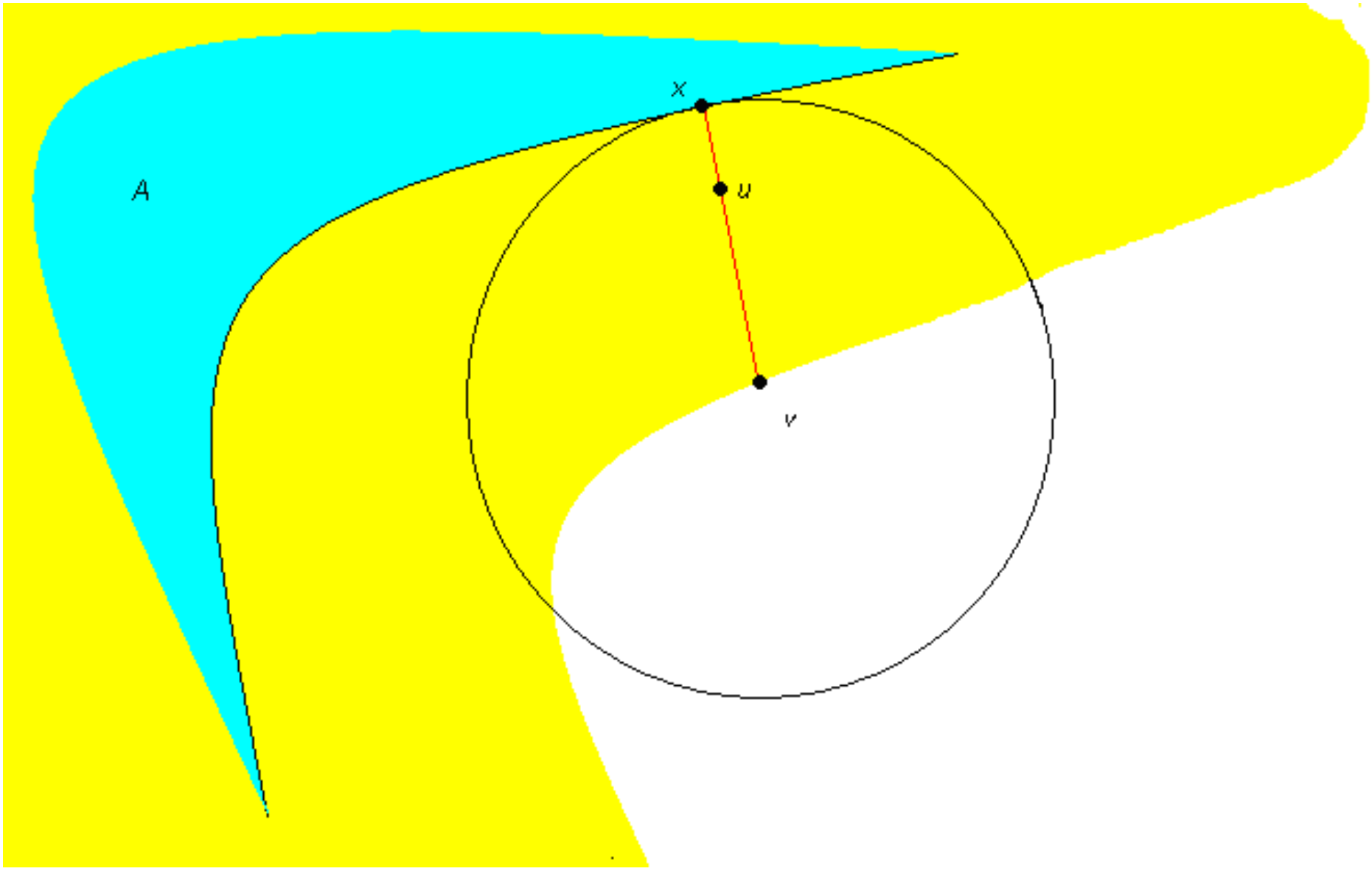
Reshetnyak 1956, Stechkin, Borwein, Clarke, Rockafellar (1990th),...

Proximally smooth sets — the supporting principle

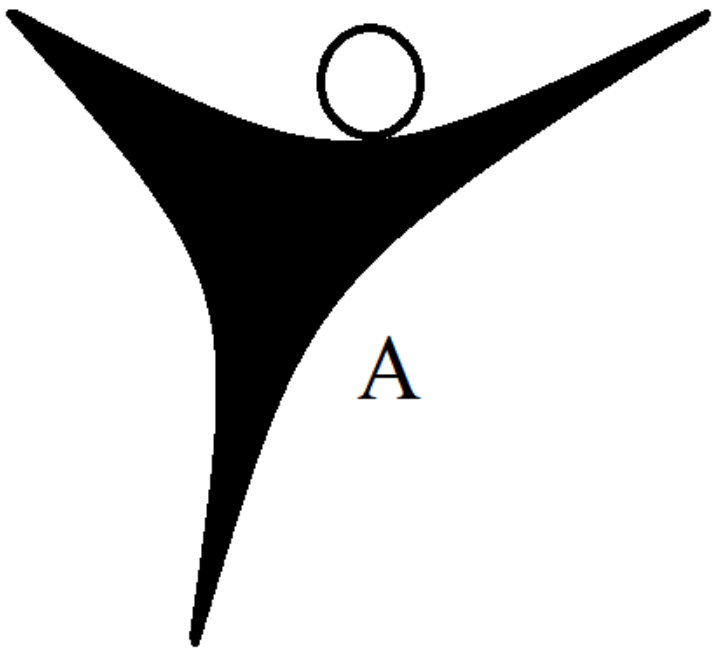
Cone of proximal normals
$N(A, x) = \{p \in \mathbb{R}^d \mid \exists \delta > 0 \; P_A(x + \delta p) = \{x\}\}$.
The set $A \subset \mathbb{R}^d$ is a proximally smooth set with constant R if and only if

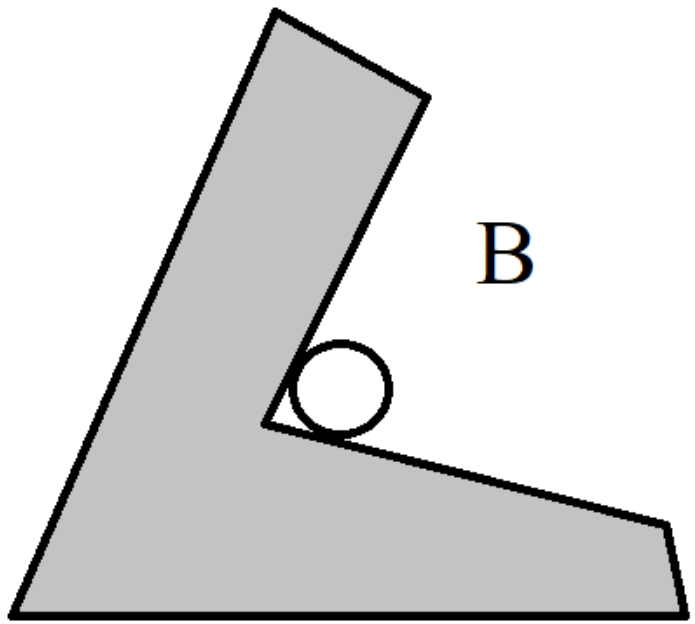
$$A \cap \operatorname{int} B_R(x + Rp) = \emptyset, \quad \forall x \in \partial A, \; \forall p \in N(A, x).$$

Proximally smooth sets — a picture



Proximally smooth sets — a picture





Smooth prox-regular surfaces

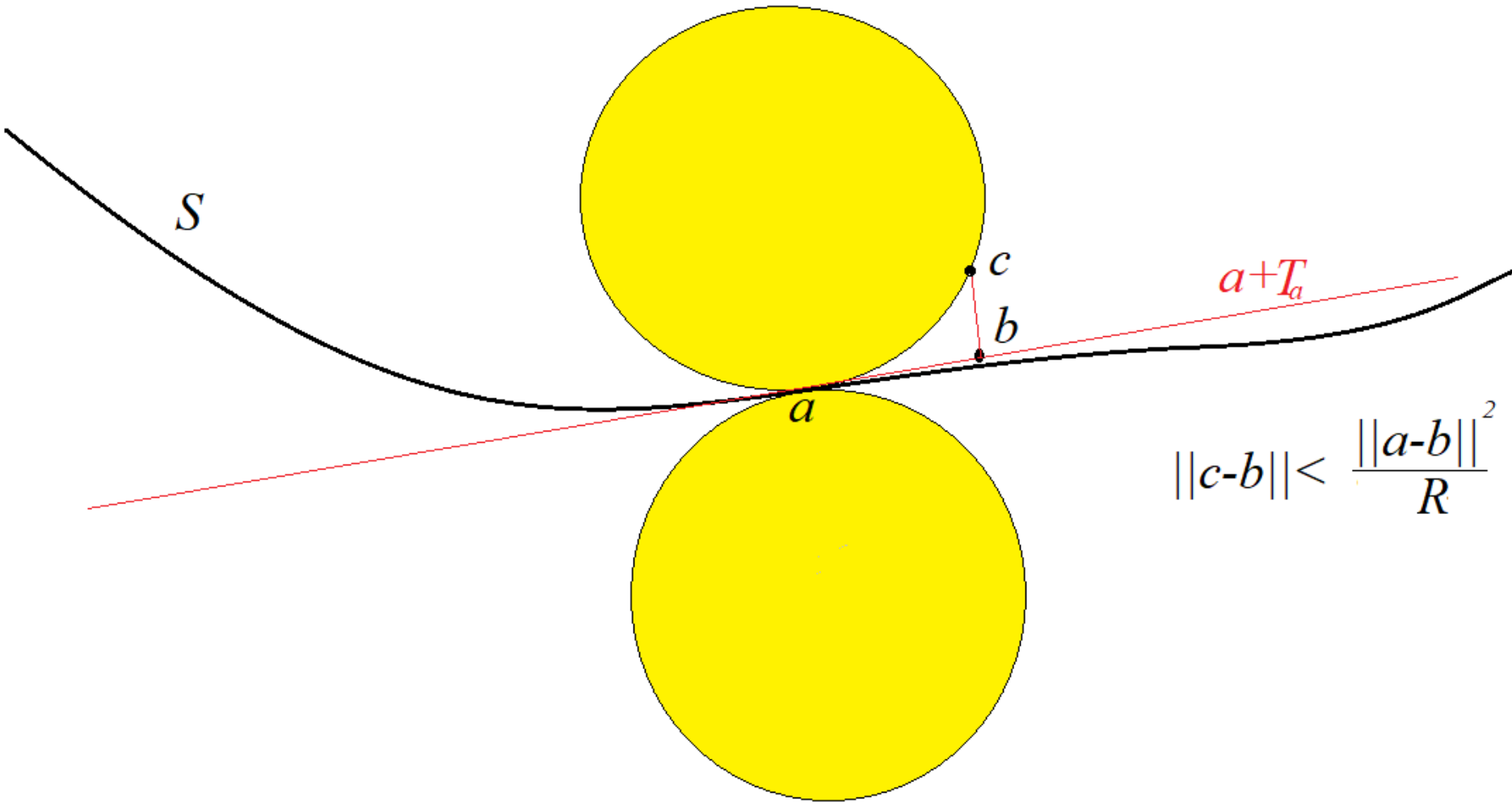
Let $g(x) = (g_1(x), \ldots, g_m(x))$, $x \in \mathbb{R}^d$, $m < d$. The set $S_i = \{x \mid g_i(x) = 0\}$ is proximally smooth with constant R if the continuous unit normal vector to the surface S_i is the Lipschitz continuous with respect to the point:

$$\|n(x_1) - n(x_2)\| \leq R^{-1}\|x_1 - x_2\|,$$

$$\forall x_1, x_2 \in S_i, \ \forall n(x_j) \in N(S_i, x_j), \ j = 1, 2.$$

Under some natural assumptions the surface $\{x \mid g(x) = 0\} = \bigcap\limits_{1 \leq i \leq m} S_i$ is also proximally smooth.

Smooth prox-regular surfaces — a picture

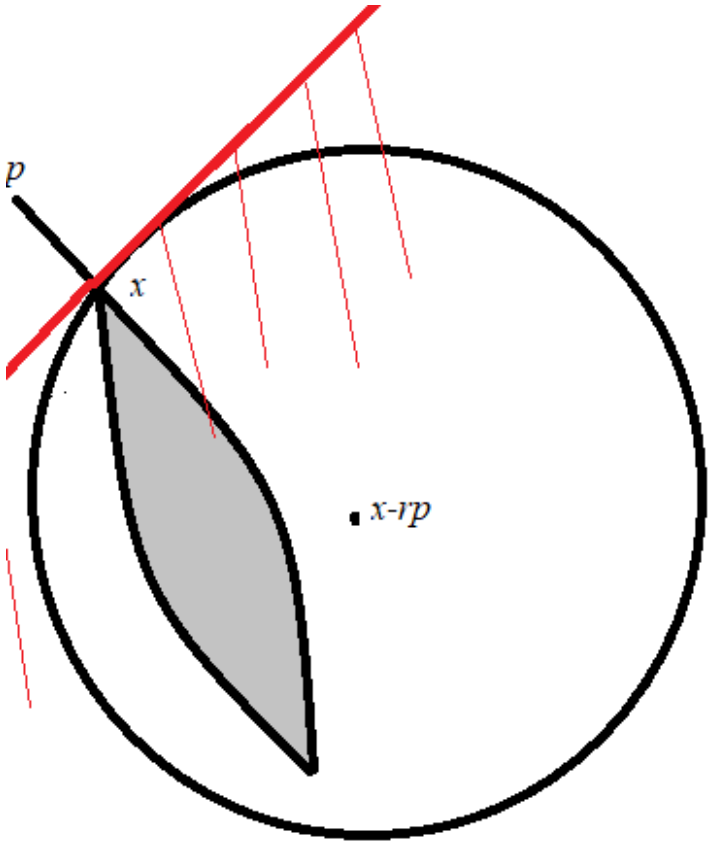


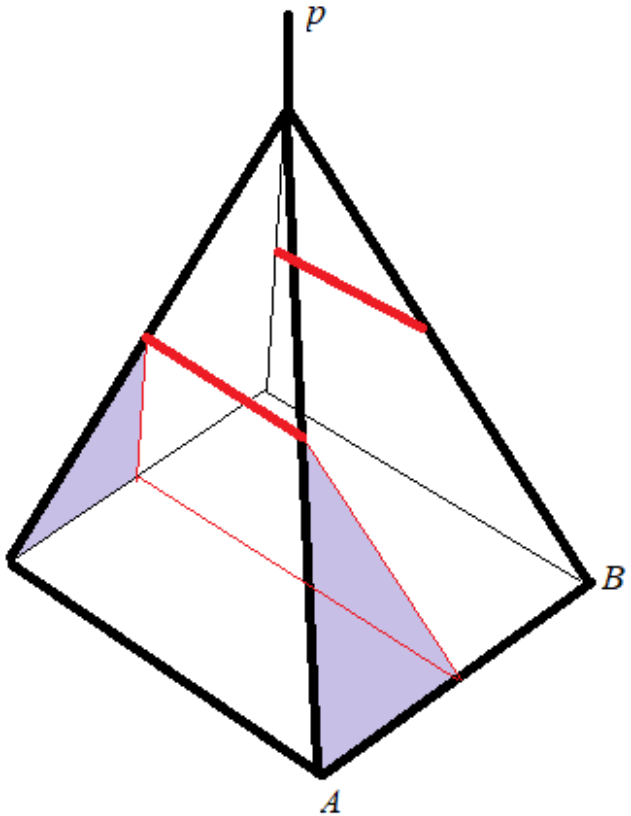
Strongly convex sets and supporting principle

A closed convex set $A \subset \mathbb{R}^d$ is strongly convex of radius $r > 0$ if $A = \bigcap_{x \in X} B_r(x) \neq \emptyset$.

The set $A \subset \mathbb{R}^d$ is strongly convex of radius $r > 0$ if and only if for any $x \in \partial A$ and unit vector $p \in N(A, x)$ $A \subset B_r(x - rp)$.

Strongly convex sets — a picture





The (FW) generalization

Suppose that $B \subset \mathbb{R}^n$ is a strongly convex set of radius r and a function $f : \mathbb{R}^n \to \mathbb{R}$ has the Lipschitz continuous gradient with constant L_1, $m = \inf\limits_{x \in \partial B} \|f'(x)\|$ and $\frac{\mathbf{m}}{\mathbf{L_1}} > \mathbf{r}$. Then the iteration process $x_0 \in \partial B$,

$$x_{k+1} = \arg\max_{x \in B}(-f'(x_k), x), \qquad k = 0, 1, \ldots$$

converges to the unique solution $x_* \in \partial B$ of the problem $\min_B f$ with linear rate:

$$\|x_k - x_*\| \leq \left(\frac{rL_1}{m}\right)^k \|x_0 - x_*\|.$$

The (GPA) generalization

Suppose that $A \subset \mathbb{R}^n$ is a proximally smooth set with constant R, a function $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with constant $\varkappa$ and the Lipschitz continuous with constant L on the level set

$$\{x \in \mathbb{R}^d \mid f(x) \leq \alpha\}, \quad \alpha \in \mathbb{R}.$$

If $\frac{\mathbf{L}}{\varkappa} < \mathbf{R}$, then for any $x_0 \in A$, $f(x_0) \leq \alpha$, the iteration process

$$x_{k+1} = P_A(x_k - tf'(x_k)), \quad t = t(\varkappa, L, R),$$

converges to the unique solution with linear rate.

Necessary conditions of optimality

Suppose that the set $A \subset \mathbb{R}^n$ is proximally smooth with constant $R > 0$, the function f has the Lipschitz continuous gradient f' with constant $L_1 > 0$ and x_0 is a local minimum in the problem $\min_A f$. Then

$$-f'(x_0) \in N(A, x_0).$$

The (GPA) for arbitrary prox-reg set and function

Let $A \subset \mathbb{R}^n$ be a bounded proximally smooth set with constant $R > 0$. Suppose that the function $f : \mathbb{R}^n \to \mathbb{R}$ is the Lipschitz continuous with constant $L > 0$ and its gradient f' is also the Lipschitz continuous with constant $L_1 > 0$. Take $C > 0$ with $\frac{L}{C+L_1} < R$. Then for any $x_0 \in A$ the sequence of the form

$$x_{k+1} = P_A\left(x_k - \frac{1}{C + L_1} f'(x_k)\right)$$

$k = 0, 1, \ldots,$ converges to the set of stationary points Ω: $\lim\limits_{k\to\infty} \varrho_\Omega(x_k) = 0$ and $f(x_{k+1}) + \frac{C}{2}\|x_{k+1} - x_k\|^2 \leq f(x_k)$ for all $k \geq 0$.

The gradient mapping

The gradient mapping ($t > 0$).

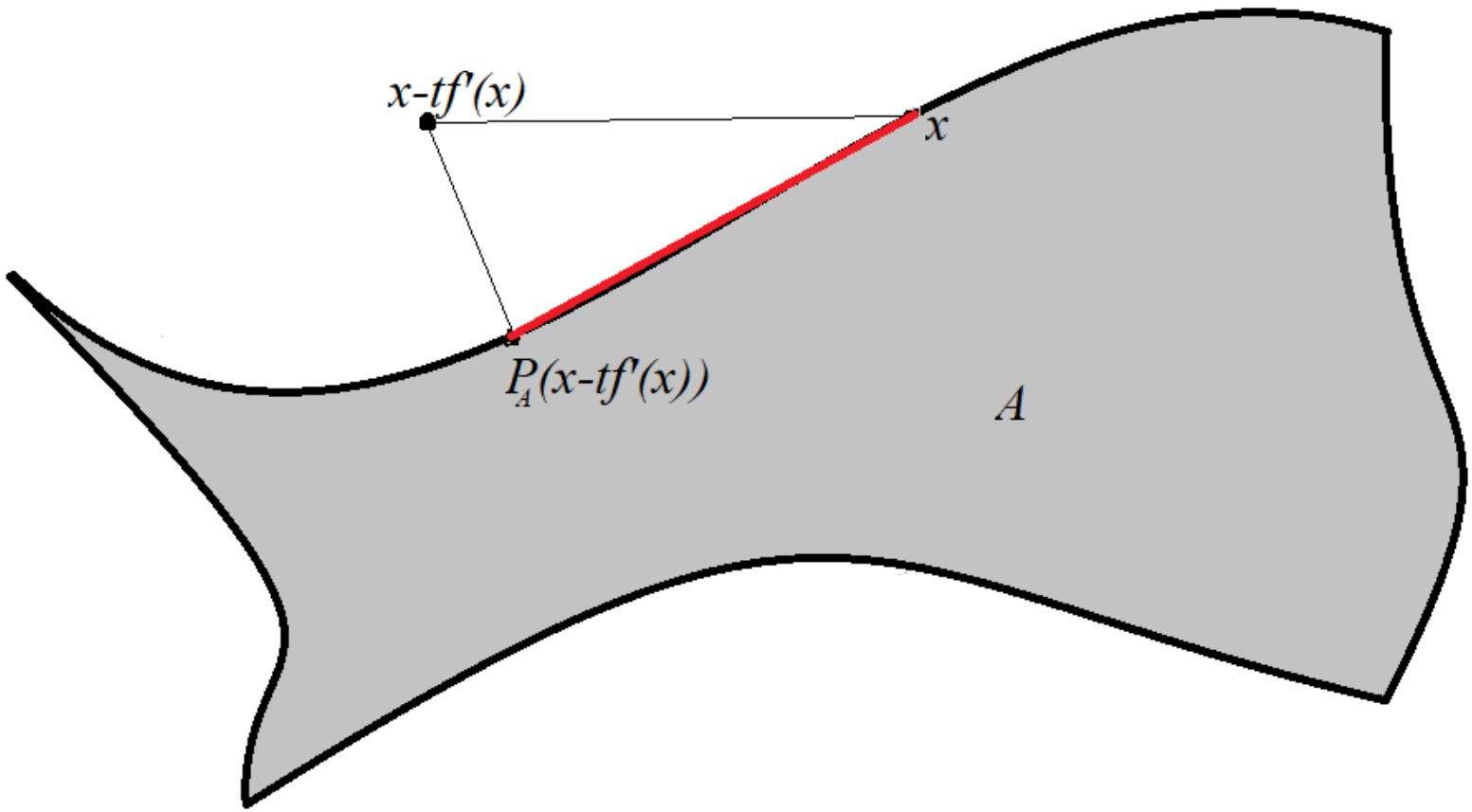
$$x_{k+1} = \operatorname{argmin}_{x \in A} \left(f(x_k) + (f'(x_k), x - x_k) + \frac{1}{2t}\|x - x_k\|^2 \right),$$

$$g(x_k) = \frac{1}{t}(x_k - x_{k+1}).$$

More generally for any $x \in A$

$$g(x) = \frac{x - P_A(x - tf'(x))}{t}.$$

The gradient mapping — a picture



The (GPA) with the (EB) condition

Suppose that in the problem (*) the gradient mapping satisfies the (EB) condition: $\exists \mu > 0$

$$\varrho_\Omega(x) \leq \mu \|g(x)\|, \quad \forall x \in A.$$

Then the (GPA) for the problem (*) converges with linear rate.

The (EB) condition for the gradient mapping — theoretical essence

If the set-valued mapping

$$F(x,z) = (tf'(x) + N(A,z), x - z)^T$$

is (uniformly) metrically regular on the set $\operatorname{dom} F = \{(x,z) \mid x \in A, z = P_A(x - tf'(x))\}$ **if and only if** the (EB) condition ($\exists \mu > 0$ $\varrho_\Omega(x) \le \mu \|g(x)\|, \ \forall x \in A$) takes place.

Here by (uniform) metric regularity we mean that $\exists \nu > 0$ with

$$\varrho(0, F^{-1}(x,z)) \le \nu \varrho((x,z), F(0)) \quad \forall (x,z) \in \operatorname{dom} F.$$

The (LPL)-condition on the surface

Let $\mathcal{L}_f(\alpha) = \{x \mid f(x) \leq \alpha\}$. A is a smooth surface without edge of m dimensions.
The (LPL) condition on A means that $\exists \mu > 0$ such that

$$\|P_x f'(x)\|^2 \geq \mu(f(x) - f_*), \qquad f_* = \inf_S f$$

for all $x \in A \cap \mathcal{L}_f(\alpha)$. Here P_x is the metric projection on the tangent plane T_x.

The (GPA) for a smooth and proximally smooth surface

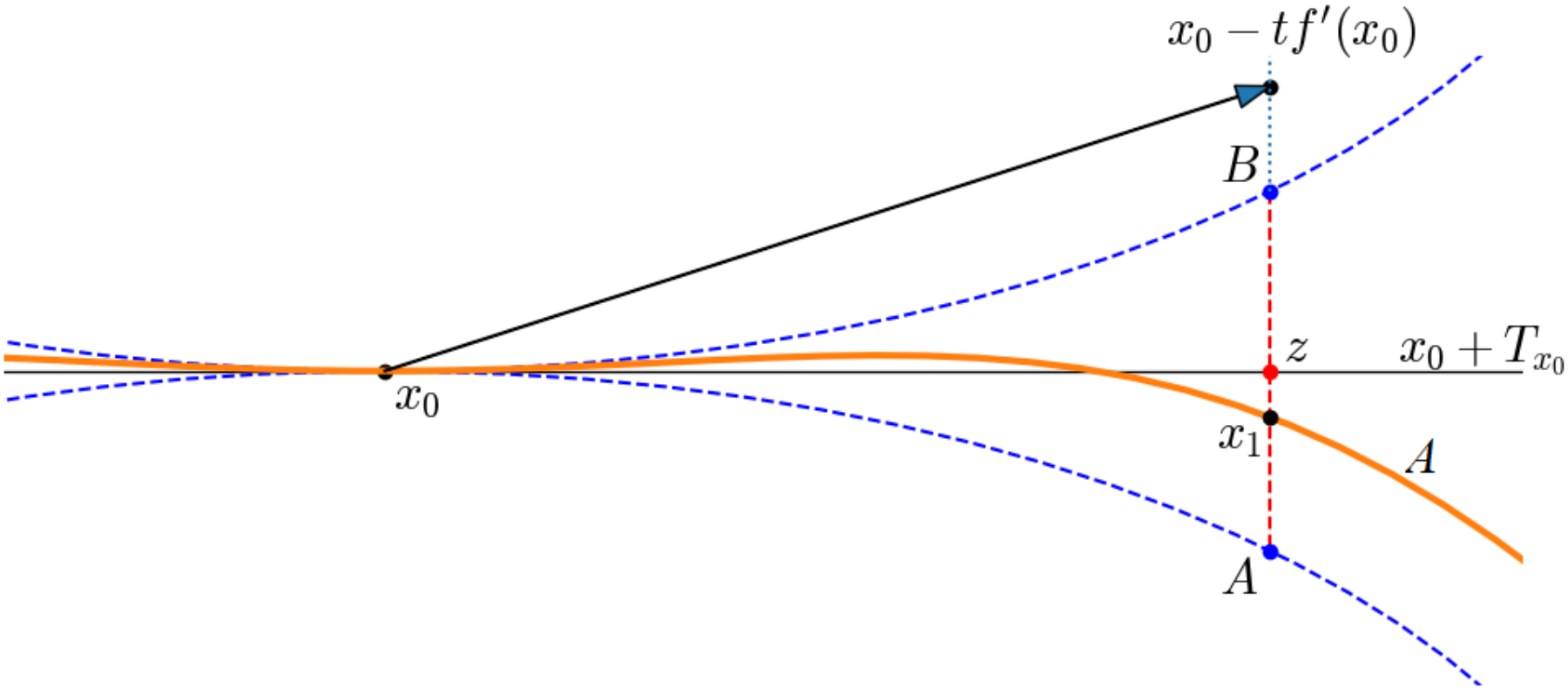
Put $A = \{x \in \mathbb{R}^d \mid g(x) = 0\}$. Case $g : \mathbb{R}^d \to \mathbb{R}$.
Then the gradient projection algorithm in the form
$x_0 \in A \cap \mathcal{L}_f(\alpha)$, $z_k = x_k - tP_{T_{x_k}} f'(x_k)$, $p_k = \frac{g'(x_k)}{\|g'(x_k)\|}$,

$$x_{k+1} = \left[z_k - p_k \left(R - \sqrt{R^2 - \|x_k - z_k\|^2} \right) , \right.$$

$$\left. z_k + p_k \left(R - \sqrt{R^2 - \|x_k - z_k\|^2} \right) \right] \cap A$$

$k = 0, 1, \ldots$ converges with linear rate with respect to the function and with respect to the point.

The (GPA) for a smooth and proximally smooth surface — a picture



The (FW) method for the case of gradient domination

Let $B \subset \mathbb{R}^d$ be a strongly convex set of radius r, $m > 1$, $\alpha \in \mathbb{R}$. Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is a function with the Lipschitz continuous gradient with constant L_1 and for any point $x \in \partial B \cap \mathcal{L}_f(\alpha)$ $\|f'(x)\| \geq mrL_1$. Then for any choice of $x_0 \in \partial B \cap \mathcal{L}_f(\alpha)$ the iterations

$$x_{k+1} = \arg\max_{x \in \partial B}(-f'(x_k), x), \quad k = 0, 1, \ldots$$

converge to the global strict minimum x_* with linear rate:

$$\|x_{k+2} - x_{k+1}\| \leq \frac{1}{m}\|x_{k+1} - x_k\|$$

for all k.

The Stiefel manifold

Let $n, k \in \mathbb{N}$, $k \leq n$.
$S = S_{n,k} = \{X \in \mathbb{R}^{n \times k} \mid X^T X = I_k\}$;
$X = (X_1 \dots X_k)$, $X_i \in \mathbb{R}^n$.
$S \ni X \Leftrightarrow x = (X_1^T \dots X_k^T) \in \mathbb{R}^{nk}$.
S can be treated as the solution of the system

$$g_{ij}(x) = (X_i, X_j) - \delta_{ij} = 0, \quad 1 \leq i \leq j \leq k$$

i.e. $\frac{1}{2}k(k+1)$ equations.
If $(i,j) \neq (i',j')$ and $x \in S$ then $g'_{ij}(x) \perp g'_{i'j'}(x)$.

Proximal smoothness of the Stiefel manifold. Alternating Projections

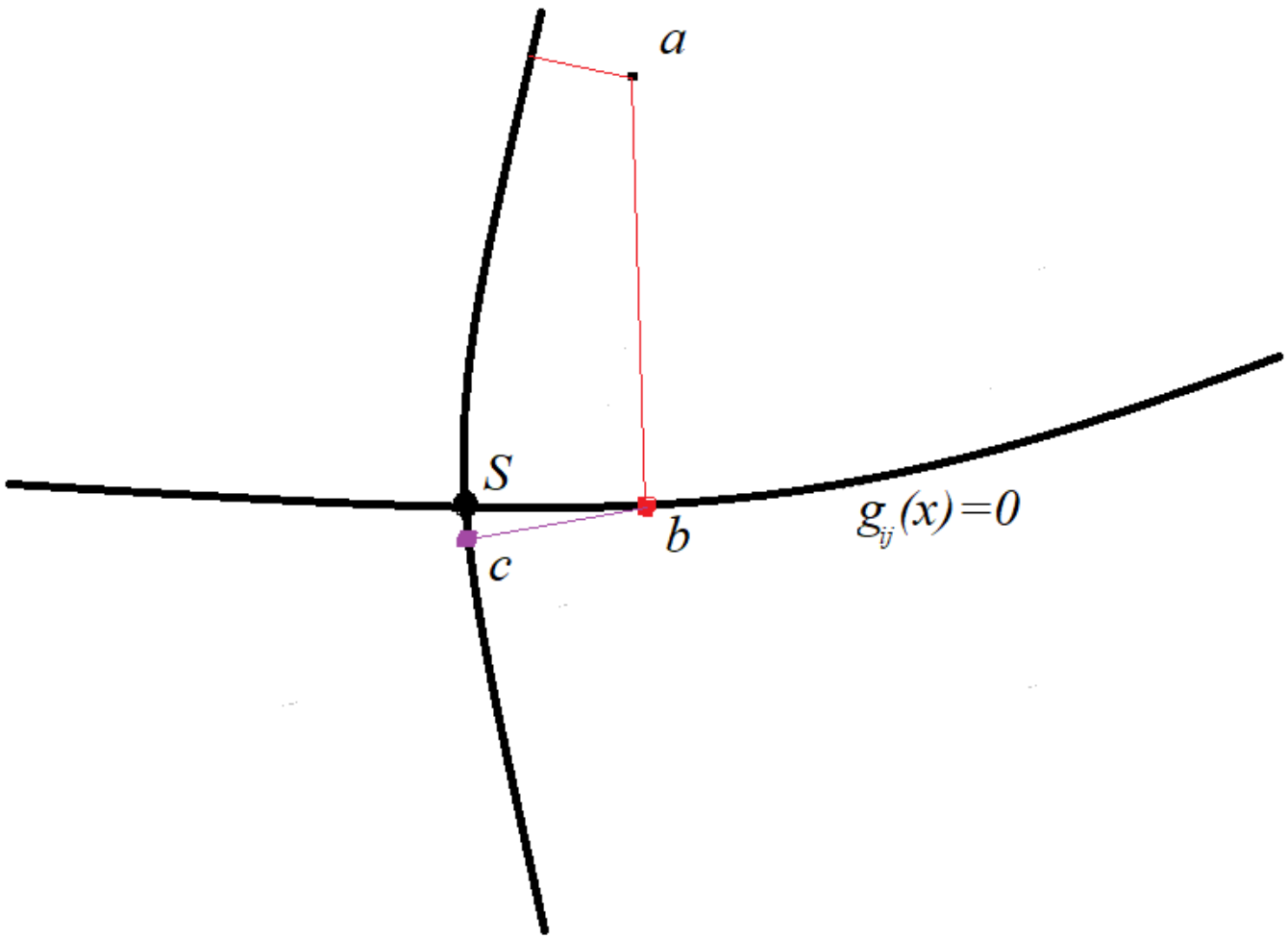
$S = S_{n,k}$ is proximally smooth with constant

$$R \geq \frac{2}{\sqrt{k^2 + 3k}}.$$

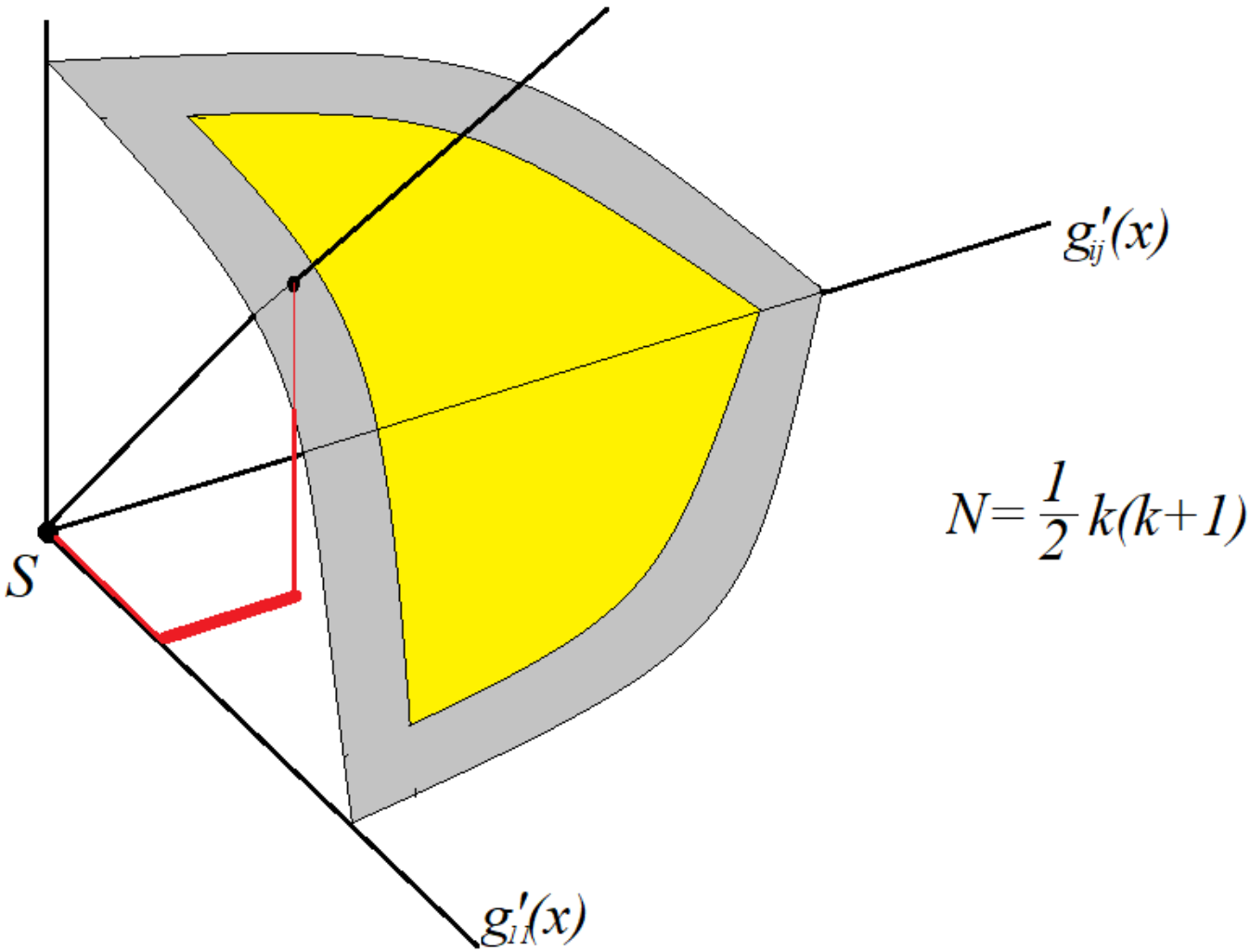
The alternating projections method. Let $N = \frac{1}{2}k(k+1)$, $\varrho_0 = \varrho_S(a) < \frac{1}{2\sqrt{N}}$. Then

$$\varrho_1^2 \leq \varrho_0^2 \left[\frac{1}{1 - \varrho_0 \left(\frac{1}{\sqrt{N}} - \varrho_0 \right)} \left(1 - \left(\frac{1}{\sqrt{N}} - \varrho_0 \right)^2 \right) \right].$$

The alternating projections method — a picture



The alternating projections method — a picture



$$N = \frac{1}{2}k(k+1)$$

Some references

- B. T. Polyk, CMMPh etc 1962-1969.
- J.-Ph. Vial, Strong and weak convexity of sets and functions, Math. Oper. Res., 8:2 (1983), 231-259.
- F. H. Clarke, R. J. Stern, P. R. Wolenski, Proximal smoothness and lower-C2 propoerty, J. Convex Anal., 2:12 (1995), 117144.
- R. A. Poliquin, R. T. Rockafellar, L. Thibault, Local differentiability of distance functions, Trans. Amer. Math. Soc., 352:11 (2000), 5231-5249
- S. Rolewicz, C. Olech, H. Frankowska, A. Pliś 1975-1983 .
- E. Polovinkin, M. Balashov, Elements ..., M. Fizmatlit, 2007.

Acknowledgements

The author is grateful to B.T. Polyak for introduction to the subject area.

Special thanks to A. Tremba for discussions.

Thank you!