Accelerated subgradient method with Polyak's step

Petro Stetsyuk
stetsyukp@gmail.com

V.M.Glushkov Institute of Cybernetics, Kyiv, Ukraine

Workshop Optimization and Project Meeting
Hotel Lindenhof, Rechenberg-Bienenmühle
April 3–6, 2018

Outline

1. Polyak's subgradient method
2. Polyak's method and ravine functions
3. One-rank ellipsoidal operator
4. Accelerated method with Polyak's step
5. Comparison of both methods

Contents

Formulation of the Problem

We consider the following problem

$$\text{to find} \quad x^\star = \operatorname*{argmin}_{x \in R^n} f(x), \quad \text{if } f^\star \text{ is known}, \tag{1}$$

where $f(x)$ is a convex function and $f^\star = f(x^\star) = \min_{x \in R^n} f(x)$.

Main inequality for the subgradient

For convex function $f(x)$ the following inequality is valid

$$(x - x^{\star}, \partial f(x)) \geq f(x) - f^{\star}, \quad \forall x \in R^n, \tag{2}$$

where $\partial f(x)$ is a subgradient (gradient) of the function $f(x)$.

Inequality (2) is used for the calculation of the step in subgradient method, which B.T. Polyak offered in 1969.

Polyak B.T. Minimization of unsmooth functionals // *USSR Comput. Math. Math. Phys.*. 1969. Vol. 9, No. 3, pp. 14-29.

Polyak's subgradient method

Polyak's subgradient method has the iterative form

$$x_{k+1}=x_k-h_k\frac{\partial f(x_k)}{\|\partial f(x_k)\|},\quad h_k=\frac{f(x_k)-f^\star}{\|\partial f(x_k)\|},\quad k=0,1,\ldots, \tag{3}$$

Step h_k is called Polyak's step.

The geometric sense of the method (3) is the following. The convex function $f(x)$ is approximated by a linear function

$$\tilde{f}(x)=f(x_k)+(\partial f(x_k),x-x_k)$$

and the step is selected so that this approximation function becomes equal to $f^\star$ (i. e. $\tilde{f}(x_{k+1})=f^\star$).

Decrease of the distance to the minimum point

Theorem 1 (Polyak,1969)

The sequence $\{x_k\}_{k=0}^{\infty}$, *generated by the method (3), satisfies the inequalities*

$$\|x_{k+1} - x^\star\|^2 \leq \|x_k - x^\star\|^2 - \frac{(f(x_k) - f^\star)^2}{\|\partial f(x_k)\|^2}, \quad k = 0, 1, 2, \ldots$$

Remark. Theorem 1 guarantees that in Polyak's method the distance to the minimum point decreases monotonically.

Why Polyak's step is also called the AMS-step?

For the first time Polyak's step was used in 1954 by Agmon and Motzkin, Schoenberg. They used this step in the relaxation method for finding at least one of the solutions of feasible system of linear inequalities.

AGMON S. The relaxation method for linear inequalities // Canadien Journal of Mathematics. 1954. V. 6. P. 382–392.

MOTZKIN T. AND SCHOENBERG I.J. The relaxation method for linear inequalities // see ibid – P. 393–404.

Eremin (1965) generalized this relaxation method for the systems of convex inequalities.

EREMIN I.I. Generalization of the Motzkin-Agmon relaxational method // Uspekhi Mat. Nauk. 1965. V. 20, № 2. P. 183–187.

Contents

Examples of slow convergence of Polyak's method

1. Ravine piecewise linear function ($t \gg 1$)

$$f_1(x_1, x_2) = |x_1| + t|x_2|, \quad x^\star = (0,0) \quad f^\star = 0.$$

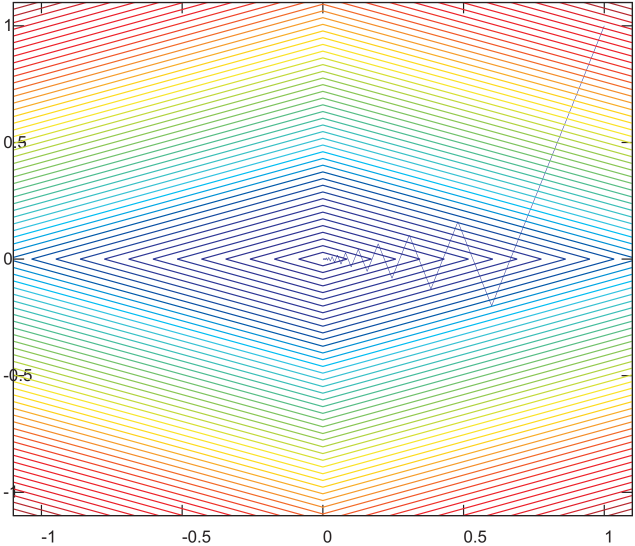
The method (3) converges at a geometric rate with ratio $q(t) = \sqrt{1 - \frac{1}{t^2}}$ (Polyak, 1969). For large t the $q(t)$ is close to one and the method (3) converges very slowly.

2. Essentially ravine piecewise quadratic function

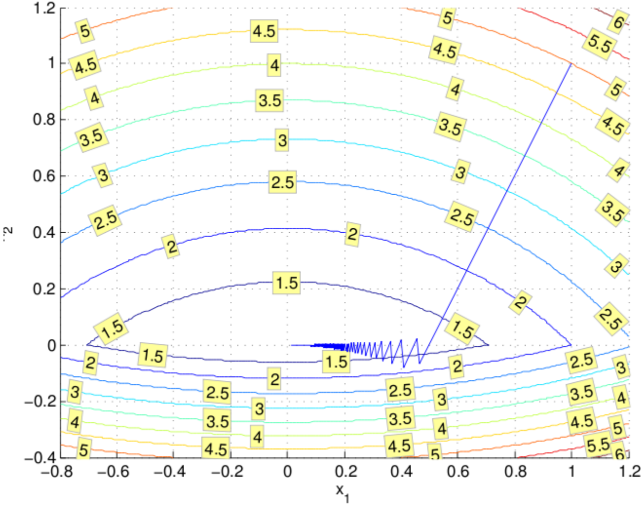
$$f_2(x_1, x_2) = \max\left\{x_1^2 + (2x_2 - 2)^2 - 3, x_1^2 + (x_2 + 1)^2\right\},$$

Degeneration at the minimum point $x^\star = (0,0)$, $f^\star = 1$.

Polyak's method for $f_1(x_1, x_2)$ $(t = 10)$



Polyak's method for $f_2(x_1, x_2)$ (10000 iterations)



How can we handle the ravine?

1. A slow convergence of Polyak's method for ravine convex functions is caused by an obtuse angle between two successive subgradients. The closer the angle to 180 degrees, the slower the method's (3) convergence.

2. It is possible to accelerate the method (3) if we transform the space of variables so that to decrease the obtuse angle between two successive (consecutive) subgradients.

3. An obtuse angle between the vectors ξ and η can be transformed into a straight through „one-rank ellipsoidal operator“. It is accelerates Polyak's subgradient method.

Contents

One-rank ellipsoidal operator

Linear operator from R^n to R^n

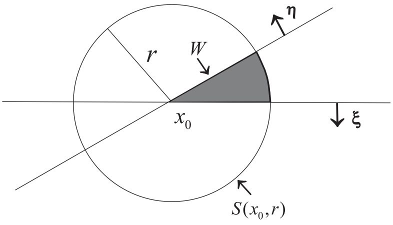
$$T_1(\xi,\eta) = I_n - \left(\frac{1-\sqrt{1-(\xi,\eta)^2}}{1-(\xi,\eta)^2}\eta - \frac{(\xi,\eta)}{1-(\xi,\eta)^2}\xi \right) \eta^T. \qquad (4)$$

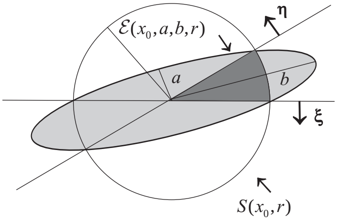
Here $\xi, \eta \in R^n$ are vectors, such that $\|\xi\| = 1, \|\eta\| = 1$ and $(\xi,\eta)^2 \neq 1$, I_n – identity matrix $n \times n$.

Makes a ball from the special 2d-ellipsoid, circumscribed around the set W, which is obtained by the intersection of the ball and two half-spaces, passing through the center of the ball.

Stetsyuk (1997) Orthogonalizing linear operators in convex programming, Cybernetics and System Analysis, **33**, No. 3.

The convex set W and 2d-ellipsoid $\mathcal{E}(x_0, a, b, r)$





the set W is the intersection of a ball $S(x_0, r)$ with half-spaces $P(x_0, \xi)$ and $P(x_0, \eta)$

2d-ellipsoid contains the set W and has the minimum volume

STETSYUK (1996) r-Algorithms and ellipsoids, Cybernetics and System Analysis, **32**, No. 1.

Properties of 2d-ellipsoid $\mathcal{E}(x_0, a, b, r)$

(i) If $(\xi, \eta) < 0$ then 2d-ellipsoid contains the convex set W.

2d-ellipsoid has the following parameters:

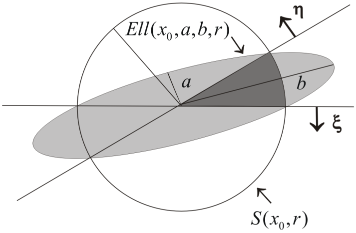
$$a = r\sqrt{1 + (\xi, \eta)} < r; \quad b = r\sqrt{1 - (\xi, \eta)} > r.$$

(ii) The ratio of 2d-ellipsoid volume to ball volume is equal to

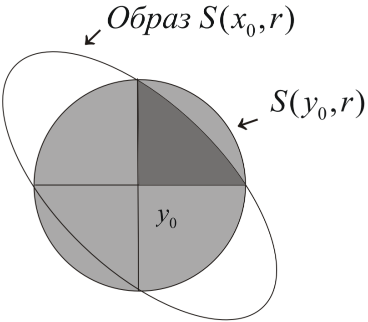
$$q = \frac{\text{vol}(\mathcal{E}(x_0, a, b, r))}{\text{vol}(S(x_0, r))} = \left(\frac{a}{r}\right)\left(\frac{b}{r}\right) = \sqrt{1 - (\xi, \eta)^2} < 1.$$

If the angle between vectors ξ and η becomes closer to 180 degrees then the ratio q becomes smaller.

2d-ellipsoid before and after transformation



The special 2d-ellipsoid



in the transformed space becomes a ball

Similarity of $T_1(\xi, \eta)$ and Shor's r-algorithm

(iii) Images of vectors ξ and η in the transformed space $Y = T_1(\xi, \eta)X \equiv E^n$ are orthogonal

This feature allows to „extend“ cone of feasiable directions of the function decrease for the subgradient process in the transformed space of variables, similar to Shor's r-algorithm.

Contents

Accelerated method with Polyak's step

We have $x_0 \in R^n$ and $n \times n$-matrix $B_0 = I_n$. Then we iterate

$$x_{k+1} = x_k - h_k B_k \frac{B_k^T \partial f(x_k)}{\|B_k^T \partial f(x_k)\|}, \quad h_k = \frac{f(x_k) - f^*}{\|B_k^T \partial f(x_k)\|}, \quad (5)$$

$$B_{k+1} = \begin{cases} B_k + (B_k \eta)\xi_{k+1}^T, & \text{if } \mu_k < 0, \\ B_k, & \text{otherwise}, \end{cases} \quad k = 0, 1, 2, \dots \quad (6)$$

where

$$\mu_k = (\xi_k, \xi_{k+1}), \quad \xi_k = \frac{B_k^T \partial f(x_k)}{\|B_k^T \partial f(x_k)\|}, \quad \xi_{k+1} = \frac{B_k^T \partial f(x_{k+1})}{\|B_k^T \partial f(x_{k+1})\|},$$

$$\eta = \left(\frac{1}{\sqrt{1 - \mu_k^2}} - 1 \right) \xi_{k+1} - \frac{\mu_k}{\sqrt{1 - \mu_k^2}} \xi_k.$$

Decrease of the distance to the minimum point

Theorem 2 (Stetsyuk,1997)

Let $A_k = B_k^{-1}, A_{k+1} = B_{k+1}^{-1}$. *The sequence* $\{x_k\}_{k=0}^{\infty}$, *generated by the method (5)–(6), satisfies the inequalities*

$$\|A_{k+1}(x_{k+1}-x^\star)\|^2 \leq \|A_k(x_k-x^\star)\|^2 - \frac{(f(x_k)-f^\star)^2}{\|B_k^T \partial f(x_k)\|^2}, \qquad k = 0, 1, 2,$$

Remark. Theorem 2 guarantees that in accelerated subgradient method with Polyak's step the distance to the minimum point decreases monotonically in successively transformed spaces of variables.

On accelerated convergence of method (5)–(6)

If at k-th step the transformation of the space is realized, then

$$\det(B_{k+1}) = \det(B_k)\sqrt{1-\mu_k^2} = \det(B_k)\sqrt{1-\cos^2\varphi_k}.$$

where φ_k is an obtuse angle between two successive subgradients.

For ravine functions determinant of B_k decreases and, consequently, the volume of the ellipsoid localizing the point x^* also decreases.

This provides an accelerated convergence of the method (5)–(6) for ravine convex (smooth or nonsmooth) functions in comparison with the method (3).

Example 1: ravine function $f_1(x_1, x_2)$

For function

$$f_1(x_1, x_2) = |x_1| + t|x_2|, \ \forall t > 1, \ \forall x_0 = (x_0^{(1)}, x_0^{(2)})$$

the method (5)–(6) finds the minimum of $x^* = (0, 0)$ in no more than three iterations:

1) in one iteration, if $|x_0^{(2)}| = t|x_0^{(1)}|$. Then it turns into method(3).
2) in two iterations, if $|x_0^{(2)}| < t|x_0^{(1)}|$. One transformation.
3) in three iterations, if$|x_0^{(2)}| > t|x_0^{(1)}|$. One transformation.

Remark. If $|x_0^{(2)}| \neq t|x_0^{(1)}|$, then the method (3) converges at a geometric rate with ratio $q(t) = \sqrt{1 - 1/t^2}$ and requires a significant number of iterations for large values of t.

Example 2: ravine function $f_2(x_1, x_2)$

Essentially ravine piecewise quadratic function

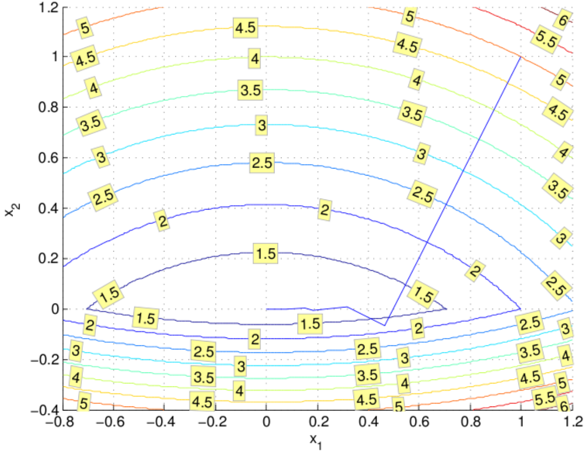
$$f_2(x_1, x_2) = \max\left\{x_1^2 + (2x_2 - 2)^2 - 3, x_1^2 + (x_2 + 1)^2\right\},$$

degenerated at the minimum point $x^\star = (0, 0)$, $f^\star = 1$.

If $x_0 = (1, 1)$, then the method(5)–(6) finds:

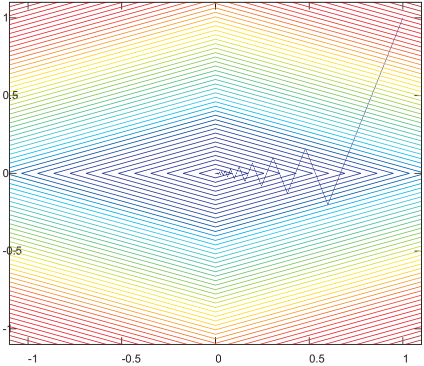
- in 16 iterations – the point x_{16}, where $f_2(x_{16}) \le 1 + 10^{-6}$;
- in 31 iterations – the point x_{31}, where $f_2(x_{31}) \le 1 + 10^{-10}$.

The method (5)–(6) for $f_2(x_1, x_2)$ (31 iterations)

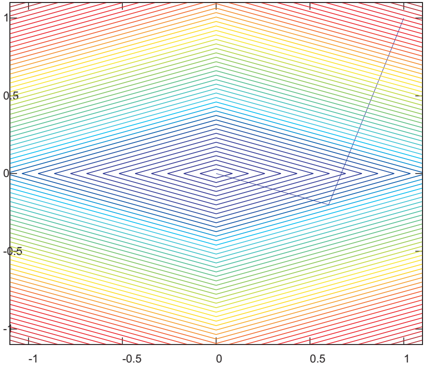


Contents

Example 1. $f_1(x_1, x_2)$: piecewise linear function

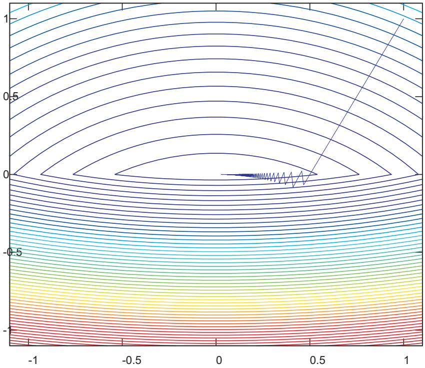


1. Trajectory of Polyak's method

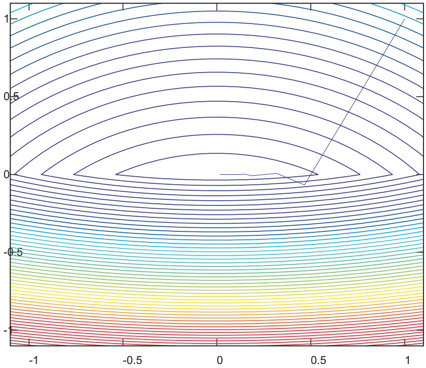


2. ...accelerated Polyak's method

Example 2. $f_2(x_1, x_2)$: piecewise quadratic function



1. Trajectory of Polyak's method



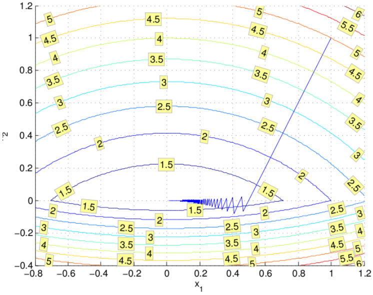
2. ...accelerated Polyak's method

Example 2. More details

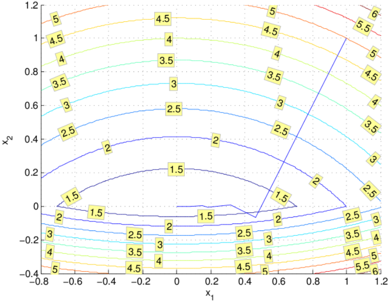
Essentially ravine piecewise quadratic function

$$f_2(x_1, x_2) = \max\left\{x_1^2 + (2x_2 - 2)^2 - 3, x_1^2 + (x_2 + 1)^2\right\}$$

degenerated at the minimum point $x^\star = (0, 0)$, $f^\star = 1$.



Method (3) (10000 iter.)



Method (5)–(6) (31 iterations)

Piecewise linear function, different ravine degrees

eps	$t = 3$		$t = 9$		$t = 27$	
	itn1	itn2	itn1	itn2	itn1	itn2
1.e-01	31	15	220	37	1645	64
1.e-02	72	24	458	44	3257	73
1.e-03	113	29	695	49	4871	78
1.e-04	155	38	933	54	6481	80
1.e-05	196	43	1170	59	8083	84
1.e-06	237	50	1407	62	9633	91
1.e-07	279	54	1642	66	10000	93
1.e-08	320	59	1874	74	10000	100
1.e-09	362	62	2101	78	10000	108
1.e-10	403	65	2322	85	10000	113

$$f(x_1, \ldots, x_{10}) = \sum_{i=1}^{10} t^{(i-1)/9} |x_i - 1|, \;\; x_0 = (0, \ldots, 0)^T$$

Conclusion

On the basis of one-rank ellipsoidal operator,

accelerated version of subgradient methods can be constructed for other techniques of step adjustment.

A remarkable feature of these methods is

automatic choice of space transformation parameters.

Thanks

Volkswagen Foundation
grant No 90 306

Questions?

THANK YOU
FOR YOUR ATTENTION!