ε-SUBGRADIENT ALGORITHM WITH SPACE TRANSFORMATION ($\alpha(\varepsilon)$-algorithm)

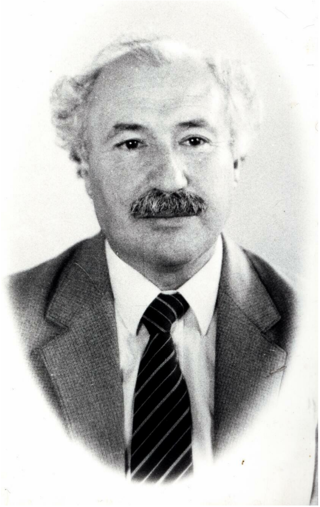
N.G. Zhurbenko

V.M. Glushkov Institute of Cybernetics
Department of non-smooth optimization methods
zhurbnick@gmail.com

5th International Conference "Problems of Cybernetics and Informatics" (PCI 2023)

Baku, August 28-30, 2023

Shor Naum Zuselevich



Academician Naum Zuselevich Shor (1937 - 2006)

OUTLINE

- INTRODUCTION
- $\bar{\varepsilon}$- MINIMIZATION PROBLEM
- ε-SUBUBGRADIENT
- AGGREGATED ε-SUBGRADIENT
- ε-SUBGRADIENT ALGORITHM WITH SPACE TRANSFORMATION ($\alpha(\varepsilon)$-ALGORITHM)
- NUMERICAL EFFICIENCY OF THE $\alpha(\varepsilon)$-ALGORITHM
- CONCLUSION
- REFERENCES

INTRODUCTION

Subgradient (gradient) algorithm of minimization is not invariant to linear transformations (unlike, for example, Newton's method). For functions with strongly elongated level surfaces, the gradient algorithm can have an extremely low convergence rate.
To overcome this problem, Shor N. Z. proposed to use an operation of space transformation (1969). At first this idea had heuristic nature.
But when the ellipsoid method was developed (Nemirovsky-Yudin 1975, Shor 1976), it appeared that it is a special version of subgradient algorithm with transformation of the space.
Hope you will be interested in the interpretation of well-known algorithms of smooth optimization (conjugate directions and quasi-Newton algorithms) from positions of gradient algorithms with the space transformation.

INTRODUCTION

More than 50 years ago, a subgradient minimization algorithm with space dilation in the direction of the difference of two successive gradients was developed (so called r-algorithm) [1, 2] (N. Z. Shor, N. G. Zhurbenko).

The practice of using the r-algorithm shows that it is still one of the most efficient non-smooth optimization algorithms. However, the theoretical study of the efficiency of the algorithm is far from being completed (even for the class of convex functions).

The present work provides a description of the $\bar{\varepsilon}$-algorithm – the results of the author on the development of a subgradient algorithm with comparable efficiency and with its **main characteristics of the r-algorithm:**

the use of one-dimensional minimization procedure;

space dilation with large coefficients (in contrast, for example, to the ellipsoid method)

$\bar{\varepsilon}$- MINIMIZATION PROBLEM

$f(x)$ – convex function in R^n.
$f^* = \min\{f(x)|x \in R^n\}$.
X^* – be the set of minimum points of the function $f(x)$.
For a given number $\bar{\varepsilon} \geq 0$ we introduce
$X^*(\bar{\varepsilon})$ – $\bar{\varepsilon}$-optimal set: $X^*(\bar{\varepsilon}) = \{x \in R^n | f(x) \leq f^* + \bar{\varepsilon}\}$.
Any point $\tilde{x} \in X^*(\bar{\varepsilon})$ and function value $\tilde{f} = f(\tilde{x})$ will be called the solution of the $\bar{\varepsilon}$-optimization problem ($\bar{\varepsilon}$-solution).
Let the ball $D(z, R)$ of initial localization of $\bar{\varepsilon}$-solutions be given, where $z \in R^n -$ is its center, and $R > 0 -$is its radius. We will assume that the ball $D(z, R)$ contains at least one minimum point $x^* \in X^*$.

ε-sububgradient

We will use the definition of ε-subgradient somewhat different from the classical one [3] (C. Lemarechal, K. Mifflin).

Definition. A vector $g \in R^n$ is called $(\varepsilon, \tilde{f})-$subgradient of a function $f(x)$ at the point z, if for $\forall x$ the following inequality holds:

$$f(x) \geq \tilde{f} + (g, x - z) - \varepsilon \tag{1}$$

where $\tilde{f} \geq f^*; \varepsilon \in R^1$.

The value $\tilde{f}$ on iteration k of the algorithm is usually equal to the obtained record value of the function $f(x)$. If the value f^* is known a priori, then $\tilde{f} = f^*$. The usual classical definition of ε-subgradient coincides with the definition of $(\varepsilon, \tilde{f})$-subgradient (1) at $\tilde{f} = f(z)$, $\varepsilon \geq 0$.

AGGREGATED ε-SUBGRADIENT

Notation. $\tilde{X}(\bar{\varepsilon}, \tilde{f}) = \{x \in R^n | f(x) \leq \tilde{f} - \bar{\varepsilon}\}$.
$G(\tilde{f}, z)$ – the set of $(\varepsilon, \tilde{f})$-subgradients of the function $f(x)$ in the point z.
Let $g_i \in G(\tilde{f}, z), i = 1, \ldots, m; \tilde{\varepsilon}_i = \bar{\varepsilon} - \varepsilon_i$.
Then $\tilde{X}(\bar{\varepsilon}, \tilde{f})$ is contained in the intersection of subspaces: $(x - z, g_i) \leq -\tilde{\varepsilon}_i$, (it is not excluded that this intersection is empty).
The "shift" vector y for a set ε-subgradients $G(\tilde{f}, z)$ from a point z defines the solution to the following problem:

$$\min ||y||^2 \tag{2}$$

$$(y, g_i) \leq -\tilde{\varepsilon}_i; i = 1, \ldots, m. \tag{3}$$

If the system (3) is inconsistent, then the vector$y$ is indefined.

AGGREGATED ε-SUBGRADIENT

Statement 1. Let the system of constraints (3) is inconsistent. Then $\tilde{f}$ is a solution to the $\bar{\varepsilon}$-optimizations problem and the convex hull of the set of the ε-subgradients $G(\tilde{f}, z)$ contains 0.

Statement 2. Let's suppose:

a). $\max\{\varepsilon_i; i = 1, \ldots, m\} \leq \bar{\varepsilon}$;

b). system (3) is not empty;

c). λ are the optimal values of dual variables of the problem (2)-(3);

d). $g = \sum \lambda_i g_i / \sum \lambda_i$;

e). $\varepsilon = \sum \lambda_i \varepsilon_i / \sum \lambda_i$.

Then: $g \in G(\tilde{f}, z)$, $y = -hg/||g||$, $h = (\bar{\varepsilon} - \varepsilon)/||g||$.

If system (3) is empty, then we set $g = 0$.

Vector g, defined in the statement 2, will be called **aggregated** ε-**subgradient** of a set of subgradients $G(\tilde{f}, z)$.

AGGREGATED ε-SUBGRADIENT

The geometric sense of the aggregated ε-subgradient: aggregated ε-subgradient belongs to the convex hull of the set of subgradients $G(\tilde{f}, z)$ and provides maximum clipping plane shift. Using this notion of aggregated ε-subgradient allows to get modifications of ε-subgradient minimization methods according to the usual scheme of their construction.

Comment.

1. If all $\varepsilon_i = \varepsilon$ – constant, then $g = Nr\{g_i\} \Rightarrow$ the usual classical aggregate subgradient.
2. Let $\tilde{f} = f^*$, then $g \in \partial f(z)$ is (ε, f^*)-subgradient with $\varepsilon = -(f(z) - f^*)$. For $\bar{\varepsilon} = 0$ the value of "shift" $h = (\bar{\varepsilon} - \varepsilon)/||g|| = (f(z) - f^*)/||g||$ exactly corresponds to the step multiplier of Polyak's method [4] (B. T. Polyak).

$\alpha(\varepsilon)$-algorithm

The $\alpha(\varepsilon)$-algorithm is based on the use of ε-subgradients and space dilation operators [2] (N. Z. Shor).
At each iteration of the algorithm, the transformation consists of applying the space dilation operators along the orthogonal directions.
The transformation parameters are determined by the construction of localization ellipsoids of ε-solution.
Localization ellipsoids are constructed on the basis of information obtained as a result of applying the one-dimensional minimization procedure.
At each iteration the localization volume of the solution is reduced by at least a factor $q < 1$ (algorithm parameter).

$\alpha(\varepsilon)$-algorithm

PROCEDURE FOR GENERATING ε-SUBGRADIENTS.

$G(\tilde{f}, z)$ – the set of $(\varepsilon, \tilde{f})$-subgradients of the function $f(x)$ in z.

1. Compute the vector $p = Nr\{g_i/||g_i|| : g_i \in G(\tilde{f}, z)\}$.
(Bisector of a set of vectors).

2. Based on the one-dimensional descent along $-p$, we determine a new subgradient g_{new}, such that: $(g_{nuw}, p) \leq 0$; $\varepsilon_{nuw} \leq \bar{\varepsilon}$.

3. Include the vector g_{new} in the set $G(\tilde{f}, z)$ and remove inactive subgradients from $G(\tilde{f}, z)$.

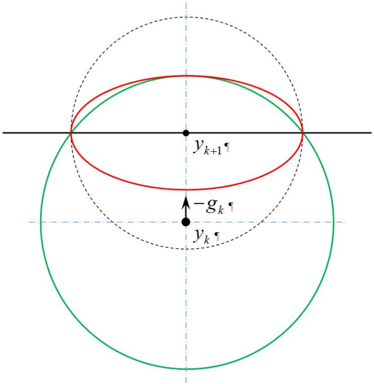
4. If the decrease in the volume of solution localization decreases by a factor of q, then STOP.

5. GO TO 1.

For $k \leq n^2 \Rightarrow q \leq \frac{\sqrt{2}}{2} \approx 0.707$.

(For the method of centers of gravity $q \leq 1 - 1/e \approx 0.632$).

$\alpha(\varepsilon)$-algorithm



Case 1.

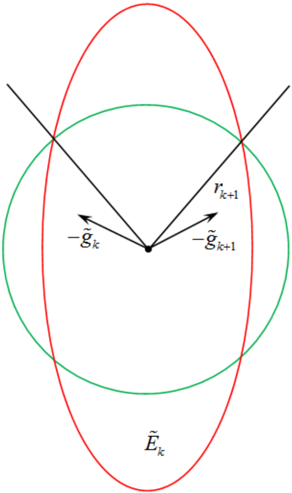
The record value of the function has been significantly improved.

The localization – the segment of the ball.

The space transformation is determined by the dilation operator along the direction of the aggregate subgradient.

Moving to a new point.

$\alpha(\varepsilon)$-algorithm



Case 2.

There is no improvement in the record value of the function.

("ravine" or proximity to the optimal point).

The localization – the sector of the ball.

The space transformation is determined by the dilation operators along $n-1$ orthogonal directions.

The most significant is the dilation operator in the direction of the difference of two normalized subgradients.

The point does not change.

$\alpha(\varepsilon)$-algorithm

The following statement defines a theoretical estimate of the efficiency of the $\alpha(\varepsilon)$-algorithm.

Statement 3. For the number k of iterations after which the $\alpha(\varepsilon)$-algorithm provides the 2ε-optimization solution, the following estimate holds:

$k \leq n \ln(1/\gamma)/\ln(1/q)$,

where γ is the relative accuracy of the solution of the problem.

$\gamma = \varepsilon/(RC)$.

R – радиус шара начальной локализыции решения $D(x_0, R)$.

C – верхняя граница норм субградиентов в $D(x_0, R)$.

$\alpha(\varepsilon)$-algorithm

Remarks.
$q = 1/e \Rightarrow k \leq n \ln(1/\gamma)$ (the method of centers of gravity!)
$q = e^{-1/n} \Rightarrow k \leq n^2 \ln(1/\gamma)$ (the ellipsoid method!)
This result is not as good as it might seem.
It is necessary to take into account the complexity of the iteration to ensure a significant reduction in the size of the ε-solution localizer! The theoretical estimate for the number of one-dimensional descents is depressing (n^2).
However, we note that the behavior of the algorithm, in contrast to the ellipsoid method, can significantly depend on the specific characteristics of the function being minimized.
And we hope for the "simplex algorithm effect" for linear programming problems.

NUMERICAL EFFICIENCY OF THE $\alpha(\varepsilon)$-ALGORITHM

Test problems (for demonstration purposes).

$$f(x) = \sum_{i=1}^{n} \rho_n^{i-1} x_i^2$$

where $\rho_n = 10^{6/(n-1)}$.

The degree of elongation of the level lines ("ravines") of functions does not depend on the dimension and is equal to 10^6.

Starting point $x_i = 1.0, i = 1, \ldots n$. Functional solution accuracy parameter is $\bar{\varepsilon} = 10^{-6}$.

Table column designations: **nVarbl** is the number of variables; **qVolum** is the value of localization volume reduction per iteration (algorithm parameter); **nIter** is the number of iterations; **nLStep** is the average number of applications of the one-dimensional minimization algorithm per iteration; **Alpha** is the average value of the space expansion coefficient; **r-agorithm** is the number of iterations of r-algorithm.

NUMERICAL EFFICIENCY OF THE $\alpha(\varepsilon)$-ALGORITHM

TABLE. Minimization $f(x)$

nVarbl	qVolum	niter	nLStep	Alpha	r-agorithm
10	0.99	82	1.793	1.817	72
10	0.7	78	2.333	2.850	72
20	0.99	120	1.925	1.738	103
20	0.7	93	2.817	2.837	103
40	0.99	170	1.876	1.678	157
40	0.7	120	3.425	2.701	157
50	0.99	186	1.855	1.709	190
50	0.7	131	3.328	2.764	190
100	0.99	253	1.763	1.670	577
100	0.7	158	3.785	2.724	577

NUMERICAL EFFICIENCY OF THE $\alpha(\varepsilon)$-ALGORITHM

The numerical experiments presented in the Table demonstrate a rather high efficiency of the algorithm, comparable to the efficiency of the r-algorithm (according to the number of iterations).
The results reveal the following interesting features. The average number of applications of the one-dimensional minimization algorithm per iteration turns out to be small compared to its guaranteed estimate (n^2).
Even if the guaranteed localization volume reduction parameter ε-solution at each iteration is set small (0.99), however, the average value of the space expansion coefficient turns out to be significantly greater than unity.
The results also provide some justification for choosing the value of the space dilation coefficient in the r-algorithm in practice: $\alpha \approx 2$.

CONCLUSION

An estimate of the complexity of the algorithm for solving the optimization problem is obtained. The numerical efficiency of the algorithm is comparable to the efficiency of the r-algorithm. However, the complexity of one iteration is significantly greater than the complexity of the r-algorithm.

It is expedient to use the $\alpha(\varepsilon)$-algorithm for problems in which guaranteed accuracy of the solution is required, for example, to obtain dual estimates for non-convex quadratic extremal problems [9] (N. Z. Shor, S. I. Stetsenko).

The development of more efficient variants of the algorithm consists in reducing the numerical complexity of the iteration of the algorithm.

Currently, a software version of the algorithm is being developed without using the solution of the quadratic programming problem (2)-(3). In this version, the aggregate subgradient is determined by averaging the sum of normalized subgradients.

REFERENCES

1. N. Z. Shor and N. G. Zhurbenko, "A minimization method using space dilation in the direction of the difference of two successive gradients," Cybernetics, 1971, №3, pp.51–59 (in Russian).
2. N. Z. Shor, "Methods for minimizing non-differentiable functions and their application," Kyiv, Nauk.dumka, 1979, 200 p. (in Russian).
3. C. Lemarechal, and K. Mifflin, "Nonsmooth Optimization," Oxford, Pergamon Press, 1978, 180 p.
4. B. T. Polyak, "Minimization of non-smooth functionals," Journal of comput. mathematics and math. physics, 1969, V.9, №3, pp. 507–521 (in Russian).
5. N. G. Zhurbenko, "On ε-subgradient minimization algorithm," Theory of optimal solutions. Kyiv, V. M. Glushkov Institute of Cybernetics of the NAS of Ukraine, 2002, pp. 111–118 (in Russian).

REFERENCES

6. N. G. Zhurbenko, “Subgradient minimization algorithm with space transformation $\alpha(\varepsilon)$-algorithm,” in: “Stochastic programming and its applications,” P. S. Knopov, V. B. Zorkaltsev et al. Irkutsk: L. A. Melentiev Institute of Power Systems of the SB RAS, 2012, pp.36–51 (in Russian).
7. N. Z. Shor, “Cut-off method with space extension in convex programming problems,” Cybernetics, 1977, 13(1), pp.94–96 (in Russian).
8. D. Yudin and A. Nemirovskii, “Informational complexity and efficient methods for the solution of convex extremal problems,” Economics and mathematical methods, 1976, №2, pp.357-369 (in Russian).
9. N. Z. Shor and S. I. Stetsenko, “Quadratic extremal problems and non-differentiable optimization,” Kyiv, Nauk.dumka, 1989, 208 p. (in Russian).

ACKNOWLEDGEMENTS

The work was supported by a grant from the Volkswagen Foundation (grant № 97775).

Thank you for attention!

zhurbnick@gmail.com