ADAPTIVE NONMONOTONIC METHODS WITH AVERAGING OF SUBGRADIENTS

N.D. Chepurnoj

July 1987
WP-87-62

Working Papers are interim reports on work of the International Institute for Applied Systems Analysis and have received only limited review. Views or opinions expressed herein do not necessarily represent those of the Institute or of its National Member Organizations.

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

FOREWORD

The numerical methods of the nondifferentiable optimization are used for solving decision analysis problems in economic, engineering, environment and agriculture. This paper is devoted to the adaptive nonmonotonic methods with averaging of the subgradients. The unified approach is suggested for construction of new deterministic subgradient methods, their stochastic finite-difference analogs and a posteriori estimates of accuracy of solution.

Alexander B. Kurzhanski
Chairman
System and Decision Sciences Program

CONTENTS

ADAPTIVE NONMONOTONIC METHODS WITH AVERAGING OF SUBGRADIENTS

N.D. Chepurnoj

1. OVERVIEW OF RESULTS IN NONMONOTONIC SUBGRADIENT METHODS

Among the existing numerical methods of solution of nondifferentiable optimization problems, the nonmonotonic subgradient methods hold an important position.

The pioneering work by N.Z. Shor [26] gave impetus to their explosive progress. In 1962, he suggested an iterative process of minimization of convex piecewise-linear function named afterwards the generalized gradient descent (GGD):

$$x^{s+1} = x^s - r_s g^s \quad , \tag{1.1}$$

where $g^s \in \partial f(x^s)$ is a set of subgradients of a function $f(x)$ at a point x^s; $r_s \geq 0$ is a step size.

For the differentiable functions this method agrees very closely with the well-known subgradient method. The fundamental difference between them is that the motion direction $(-g^s)$ in (1.1) is, as a rule, not a descent direction.

At the first attempts to substantiate theoretically the convergence of procedures of the type (1.1) the researchers immediately faced two difficulties. For one thing, the objective function lacked the property of differentiability. For another, method (1.1) was not monotonic. These combined features rendered impractical the use of known gradient procedure convergence theorems.

New theoretical approaches therefore became a must.

One more "misfortune" came on the neck of the others: numerical computations demonstrated that GGD has a low convergence rate.

Initially great hopes were pinned on the step-size selection strategy as a way towards overcoming the crisis.

By the early 1970s difficulties caused by the formal substantiation of convergence of nonmontonic subgradient procedures had been mastered and different approaches to the step-size regulation had been offered [6, 7, 8, 19, 20, 26]. However the computations continued to prove the poor convergence of GGD in practice.

It can be said that the first stage in GGD evolution was over in 1976.

Thereupon the numerical methods of nondifferentiable optimization developed in three directions, i.e., methods with space dilation, monotone, and adaptive nonmonotonic methods were explored.

Let us dwell on each of these approaches.

In an effort to enhance the GGD efficiency, N.Z. Shor elaborated methods where the operation of space dilation in the direction of a subgradient and a difference between two successive subgradients was employed. Literally the next few years were prolific for papers [27, 28, 29] investigating into the space dilation operation in nondifferentiable function minimization problems. A high rate of convergence of the suggested methods was corroborated theoretically.

Computational practice attested convincingly to the advantageousness of application of the algorithms with space dilation, especially the r-algorithm [29], as alternative to GGD, providing dimensions of the space do not exceed 200 to 300. However, if dimensions are ample, first, a considerable amount of computations is spent on the space dilation matrix transformation, second, some extra capacity of computer memory is required.

The monotonic methods became another essential direction.

Even though the first papers on the monotonic methods appeared back in 1968 (V.F. Dem'janov [30]), their progress reached its peak in the early 70's. Two classes of these algorithms should be distinguished here: the ε-steepest descent [5, 30] and the ε-subgradient algorithms [31–34]. We shall not examine them in detail but note, that the monotonic methods offered higher rate of convergence as against GGD. Just as with the methods using the space dilation, vast dimensions of problems to be solved still remained Achilles' heel for the monotonic algorithms.

Thus, the nonmonotonic subgradient methods have come into particular importance in the solution of large-scale nondifferentiable optimization problems.

The nonmonotonic procedures have another important object of application, apart from the large-scale problems, i.e., the problems in which the subgradient cannot be precisely defined at a point. The latter encompass problems of identification, learning, and pattern recognition [1, 21]. The minimized function is there a mathematical expectation whose distribution law is unknown. Errors in subgradient calculation may stem from computation errors and many other real processes.

Ju.M. Ermol'ev and Z.V. Nekrylova [9] were the first to investigate the like procedures. Stochastic programming problems have increasingly drawn the attention to the nonmonotonic subgradient methods.

However, as pointed out earlier, GGD, widely used, resistant to errors in subgradient computations, saving memory capacity, still had a poor rate of convergence. Of great importance therefore was the construction of nonmonotonic methods such that, on the one hand, retain all advantages of GGD and, on the other, possess a high rate of convergence.

It has been this requirement that has let to elaboration of the adaptive nonmonotonic procedures.

An analysis revealed that the Markov nature of GGD is the chief cause of its slow convergence. It is quite obvious that the use of the most intimate knowledge of progress of the computations is indispensable to selection of the direction and regulation of the stepsize.

Several ideas provided the basis for the development of adaptive nonmonotonic methods.

The major concept of all techniques for selecting the direction and regulating the step-size was the use of information about the fulfillment of necessary conditions to have the extremal-value function.

Its implementation are the methods with averaging of the subgradients.

In the most general case by the operation of averaging is meant a procedure of "taking" the convex hull of an arbitrary finite number of vectors.

The operation of averaging in the numerical methods was first applied by Ja.Z. Cypkin [22] and Ju.M. Ermol'ev [11].

The paper by A.M. Gupal and L.G. Bazhenov [3] also dealing with the use of operation of averaging of stochastic estimates of the generalized gradients appeared in 1972.

However all the above papers considered the program regulation of the stepsize, i.e., a sequence $\{r_s\}$ independent of computations was selected such that

$$r_s \geq 0,\ r_s \to 0,\ \sum_{s=0}^{\infty} r_s = \infty,\ \sum_{s=0}^{\infty} r_s^2 < \infty .$$

The next natural stage in the evolution of this concept was the construction of adaptive step-size regulation using the operation of averaging of preceding subgradients.

In 1974, E.A. Nurminskij and L.A. Zhelikovskij [18] suggested a successive program-adaptive regulation of the step-size for the quasigradient method of minimization of weakly convex function.

The crux of this relation consists in the following.

Let an iterative sequence be constructed according to the rule

$$x^{s+1} = x^s - r_0 g^s, \; s = 0, 1, 2, \ldots \; ,$$

where $g^s \in \partial f(x^s)$ is a quasi-gradient of the function $f(x)$ at the point x^s, r_0 is a constant step-size.

Assume that there exist $\bar{x} \in E^n$ and numerical parameters $\varepsilon > 0$, $\delta > 0$ such that for any $s = 0, 1, 2, \ldots \; \|x^s - \bar{x}\| \le \delta$. Let us suppose also that a convex combination of subgradients $\{g^i\}_{i=0}^{s_0}$ exists such that $\|e^{s_0}\| \le \varepsilon$,

$$e^{s_0} \in \text{conv} \, \{g^i\}_{i=0}^{s_0} \; .$$

Then the point $\bar{x}$ is sufficiently close to the set $X^* = \text{argmin} \, f(x)$ according to the necessary extremum conditions. In the given case the step-size has to be reduced and the procedure repeated with the new step-size value r_1 starting at the obtained point x^{s_0}. The numerical realization of the described algorithm requires a specific rule for constructing vectors e^{s_i}. In [18] the vector e^{s_i} is constructed by the rule $e^{s_i} = \text{Proj} \; 0/\text{conv} \; \{g^k\}_{k=s_{i-1}}^{s_i}$, that is, all quasi-gradients are included into the convex hull starting from the most recent instant of the step-size change. Numerical computations bore out the expediency of making allowances for such regulation. However a grave disadvantage was inherent in it: the great laboriousness of iteration. Considering that the approach as a whole holds promise, averaging schemes had to be developed for the efficient use when selecting the direction and regulating the step-size.

This paper treats such averaging schemes. They serve as a foundation for new nonmonotonic subgradient methods, for the description of stochastic finite-difference analogs, a posteriori estimates of solution accuracy. Prior to discussing results, let us make some general assumptions. Presume that the minimization problem is being solved on the entire space of the function $f(x)$:

$$\min_{x \in E^n} f(x) \quad (*)$$

where E^n is an n-dimensional Euclidean space. The function $f(x)$ will be everywhere thought of as being the convex eigenfunction $f(x)$, dom $f = E^n$, the sets $\{x : f(x) \leq c\}$ being bounded for any bounded constant C. The set of solutions of the problem (*) will be believed to be the set

$$X^* = \{x^* \in E^n : 0 \in \partial f(x^*)\} \ .$$

2. SUBGRADIENT METHODS WITH PROGRAM-ADAPTIVE STEP-SIZE REGULATION

The concept of adaptive successive step-size regulation has already been set forth. In [23] a way of determining the instant of the step-size variation was suggested. Central to it was the simplest scheme of averaging of the preceding subgradients. This method is easy to implement and effects a saving in computer memory capacity. Compared to the program regulation, the adaptive regulation improves convergence of the subgradient methods.

Description of Algorithm 1

Let x^0 be an arbitrary initial point, $\delta > 0$ be a constant, $\{\varepsilon_k\}$, $\{r_k\}$ be number sequences such that $\varepsilon_k > 0$, $\varepsilon_k \to 0$, $r_k > 0$, $r_k \to 0$. Put $s = 0$, $j = 0$, $k = 0$, $l^0 = g^0 \in \partial f(x^0)$.

Step 1. Construct

$$x^{s+1} = x^s - r_k g^s \ .$$

Step 2. If $f(x^{s+1}) > f(x^0) + \delta$, then select $x^{s+1} \in \{x : f(x) \leq f(x^0)\}$ and go to Step 5.

Step 3. Define

$$e^{s+1} = \frac{s+1}{s-j+2} e^s + \frac{1}{s-j+2} g^{s+1} \ .$$

Step 4. If $\|e^{s+1}\| > \varepsilon_k$, then $s = s+1$ and go to Step 1.

Step 5. Set $k = k+1$, $j = s+1$, $s = s+1$ and go to Step 1.

THEOREM 1.1 *Assume that the problem (*) is solved by algorithm 1. Then all limit points of the sequence* $\{x^s\}$ *belong to* X^*.

PROOF Denote the instants of step-size variations by s_m. Let us prove that the step-size r_k varies an infinite number of times. Suppose it is not so, i.e., the step-size does not vary starting from an instant s_m and is equal r_m. Then the points x^s for $s \leq s_m$ belong to the set

$$\{x : f(x) \leq f(x^0) + \delta\}$$

and are related by

$$x^{s+1} = x^s - r_m g^s \quad .$$

Considering that the step-size does not vary, $\|e^s\| > \varepsilon_m > 0$ for $s \geq s_m$. In passing to the limit by $s \rightarrow \infty$ in the inequality

$$\|x^{s+1} - x^{s_m}\| = r_m \| \sum_{i=s_m}^{s} g^i \| = r_m(s - s_m + 1)\|e^s\| > r_m(s - s_m + 1)\varepsilon_m$$

we obtain a contradiction in the boundedness of the set

$$\{x : f(x) \leq f(x^0) + \delta\} \quad .$$

The further proof of Theorem 1.1 amounts to checking the general conditions of algorithm convergence derived by E.A. Nurminskij [17].

NURMINSKIJ THEOREM *Let the sequence* $\{x^s\}$ *and the set of solutions* X^* *be such that the following conditions are satisfied:*

D1. For any sequence $\{x^{s_k}\}$ such that

$$x^{s_k} \rightarrow x^* \in X^* \quad ,$$

$$\lim_{k \rightarrow \infty} \|x^{s_k+1} - x^{s_k}\| = 0 \quad .$$

D2 There exists the closed bounded set S such that

$$\{x^s\} \subseteq S \quad .$$

D3 For any subsequence $\{x^{n_k}\}$ such that

$$\lim_{k \rightarrow \infty} x^{n_k} = x' \overline{\in} X^*$$

there exists $\varepsilon_0 > 0$ such that for all $0 < \varepsilon \leq \varepsilon_0$ and any k

$$\inf_{m > n_k} m : \{\|x^m - x^{n_k}\| > \varepsilon\} = m_k < \infty \text{ '}.$$

D4 The continuous function $W(x)$ exists such that for an arbitrary subsequence $\{x^{n_k}\}$ such that

$$\lim_{k \to \infty} x^{n_k} = x' \overline{\in} X^*$$

and for the subsequence $\{x^{m_k}\}$ corresponding to it by condition D3 for an arbitrary $0 < \varepsilon \leq \varepsilon_0$

$$\overline{\lim_{k \to \infty}} W(x^{m_k}) < \lim_{k \to \infty} W(x^{n_k}) \ .$$

D5. The function $W(x)$ of condition D4 assumes no more than countable number of values on the set X^*.

Then all limiting points of the sequence $\{x^s\}$ belong to X^*.

Select the function $f(x)$ as the function $W(x)$. Conditions D1, D5 are satisfied in view of the algorithm structure and the earlier assumptions.

The rest of the conditions will be verified by the following scheme. We will prove that conditions D3, D4 hold the points being the inner points of the set $s = \{x : f(x) \leq f(x^0)\}$. It is therewith obvious that

$$\max_{x \in S} W(x) < \inf W(x)$$

$$x \in \{x : f(x) \geq f(x^0) + \delta\} \ .$$

Then the sequence $\{x^s\}$ falls outside the set S only finite number of times. Consequently, condition D2 is satisfied and this automatically entails the validity of D3 and D4.

So, let the subsequence $\{x^{n_p}\}$ exists such that $x^{n_p} \to x' \overline{\in} X^*$. Assume at this stage of the proof that $x' \in \mathit{int}\, S$. We will prove that there exists $\varepsilon_0 > 0$ such that for all $0 < \varepsilon \leq \varepsilon_0$ at an arbitrary p:

$$\inf_{m > n_p} \{m : \|x^m - x^{n_p}\| > \varepsilon\} = m_p < \infty \ . \tag{2.1}$$

Now suppose condition (2.1) is not satisfied, that is, for any $\varepsilon > 0$ there exists n_p such that $\|x^s - x^{n_p}\| \leq \varepsilon$ for all $s > n_p$.

We have

$$\|x^s - x'\| \le \|x^s - x^{n_p}\| + \|x^{n_p} - x'\| \le 2\varepsilon$$

for sufficiently large n_p and $s > n_p$. By the supposition $0 \overline{\in} \partial f(x')$. By virtue of the closedness, convexity and upper semi-continuity of the many-valued mapping $\partial f(x)$ there exists $\varepsilon > 0$ such that $0 \overline{\in} \operatorname{conv} G_{4\varepsilon}(x')$, where conv $\{\cdot\}$ is a convex hull and $G_{4\varepsilon}(x')$ is a set

$$G_{4\varepsilon}(x') = U\, \partial f(x),\ x \in U_{4\varepsilon}(x') \ .$$

It is easily seen that $\varepsilon > 0$ can be always selected in such a way that $U_{4\varepsilon}(x') \subseteq \operatorname{int} S$, where $U_{4\varepsilon}(x') = \{x : \|x - x'\| \le 4\,\varepsilon\}$. Let $\vartheta = \min \|\tilde{g}\|$, $\tilde{g} \in \operatorname{conv} G_{4\varepsilon}(x')$. Obviously $\vartheta > 0$. As $\varepsilon_k \to 0$, there exists an integer $\tilde{k}(\vartheta)$ such that for $k \ge \tilde{k}(\vartheta)$ we have $\varepsilon_k \le \vartheta/2$. Put $n_p \ge \tilde{k}(\vartheta)$. Then it is readily seen that for $s \ge n_p$ the step-size r_k can vary no more than once within the set $U_{4\varepsilon}(x')$. Examine the sequence $\{x^s\}$ separately on the intervals $n_p \le s < s_p^*$, where

$$s_p^* = \min s_m : s_m \ge n_p{}' .$$

When $n_p \le s < s_p^*$ the points x^s are related as follows

$$x^{s+1} = x^s - r_l g^s \ ,$$

where the index l is reconstructed with respect to s_p^*. Let us consider the scalar products

$$d_s = (x^{n_p} - x^s, g^s) = r_l\Big(\sum_{i=n_p}^{s-1} g^i, g^s\Big) =$$

$$= r_l(s - n^p)\Big(\frac{1}{s - n_p} \sum_{i=n_p}^{s-1} g^i, g^s\Big) = r_l(s - n_p)\,(z^{s-1}, g^s) \ ,$$

where $z^{n_p} = g^{n_p}$,

$$z^s = \left[1 - \frac{1}{s - n_p + 1}\right] z^{s-1} + \frac{1}{s - n_p + 1} g^s \ ,$$

$$s \ge n_p \ .$$

Since $z^s \in \text{conv}\, G_{4\varepsilon}(x')$, $s \ge n_{\rho'}$ it is possible to prove that $(z^{N_1}, g^{N_1+1}) \ge \gamma$, $\gamma = 1/2\vartheta^2$.

Thus,

$$d_{N1+1} = r_l(N_1 - n_\rho + 1)(z^{N_1}, g^{N_1+1}) \ge r_l(N_1 - n_\rho + 1)\gamma \ .$$

We next consider the scalar products

$$d_s = (x^{N_1+1} - x^s, g^s) = r_l(s - N_1 - 1)(z^{s-1}, g^s) \ ,$$

where $s \ge N_1 + 1$.

The index N_2 exists such that $(z^{N_2}, g^{N_2+1}) \ge \gamma$ and $d_{N_2+1} \ge r_l(N_2 - N_1)\gamma$.

Then in a similar way we can prove the existence of indices N_t ($t \ge 3$) such that

$$d_{N_t+1} \ge r_l(N_t - N_{t-1})\gamma \ .$$

It is easy to prove that $N_{t+1} - N_t \le N < \infty$, $t = 1, 2, \ldots$. Let N_t^* be the maximal of indices N_t that does not exceed s_ρ^*. Then

$$f(x^{N_t^*}) \le f(x^{n_\rho}) - \sum_{t=1}^{t} d_{N_t+1} \le f(x^{n_\rho}) - r_l\, \gamma\, (N_t^* - n_\rho) \ .$$

Since $s_\rho^* - N_t^* \le N$, then with $\rho \to \infty$ the last term on the right-hand side of the inequality

$$f(x^{s_\rho^*}) - f(x^{n_\rho}) \le f(x^{N_t^*}) - f(x^{n_\rho}) + |f(x^{N_t^*}) - f(x^{s_\rho^*})|$$

approaches zero. We finally obtain

$$f(x^{s_\rho^*}) - f(x^{n_\rho}) \le -r_l(s_\rho^* - n_\rho)\gamma + \varepsilon'_\rho \ , \tag{2.2}$$

where $\varepsilon'_\rho \to 0$ with $\rho' \to \infty$.

It is not difficult to notice that the reasoning which underlies the derivation of inequality (2.2) may be also repeated without changes for the interval $s \ge s_\rho^*$ to get

$$f(x^m) - f(x^{n_\rho}) \le -r_{l+1}(m - s_\rho^*)\gamma + \varepsilon''_\rho \ . \tag{2.3}$$

Adding (2.2) to (2.3) we obtain

$$f(x^m) - f(x^{n_p}) \leq - r_l (s_p^* - n_p)\gamma \\ - r_{l+1}(m - s_p^*)\gamma + \varepsilon_p' + \varepsilon_p'' \quad . \tag{2.4}$$

In passing to the limit by $m \longrightarrow \infty$ in inequality (2.4) we are led to a contradiction with respect to the boundedness of continuous function on the closed bounded set $U_{4\varepsilon}(x')$. Consequently, condition (2.1) is proved.

Let

$$m_p = \inf_{m > n_p} m : \|x^m - x^{n_p}\| > \varepsilon \ .$$

By structure $x^{m_p} \overline{\in} U_\varepsilon(x^{n_p})$, but for sufficiently large p

$$x^{m_p} \in U_{4\varepsilon}(x') \ .$$

All the reasoning involved in derivation of inequality (2.4) remains valid for the instant m_p, that is,

$$f(x^{m_p}) - f(x^{n_p}) \leq - r_l (s_p^* - n_p)\gamma - r_{l+1}(m_p - s_p^*)\gamma + \varepsilon_p' + \varepsilon_p'' \quad . \tag{2.5}$$

As

$$\varepsilon < \|x^{m_p} - x^{n_p}\| \leq \|x^{n_p} - x^{s_p^*}\| \\ + \|x^{s_p^*} - x^{m_p}\| \leq r_l(s_p^* - n_p)C + r_{l+1}(m_p - s_p^*)C \ ,$$

we have

$$W(x^{m_p}) \leq W(x^{n_p}) - \frac{\varepsilon\gamma}{2C} \ .$$

In passing to the limit by $p \longrightarrow \infty$ we get

$$\overline{\lim_{p\to\infty}} W(x^{m_p}) < \lim_{p\to\infty} W(x^{n_p}) \ .$$

The further proof of this theorem follows from the Nurminskij theorem.

To fix more precisely the instant when the iteration process gets into the neighborhood of the solution we can employ the following modification of algorithm 1 provided the computer capacity allows.

Let x^0 be an arbitrary initial point, $\delta > 0$ be a constant, $\{\varepsilon_k\}$, $\{r_k\}$ be number sequences such that $\varepsilon_k > 0$, $\varepsilon_k \to 0$, $r_k > 0$, $r_k \to 0$; $k_1, k_2, \ldots, k_m$ be integer positive bounded constants.

Put $s = 0$, $j = 0$, $k = 0$, $e^0 = g^0 \in \partial f(x^0)$.

Step 1 Construct

$$x^{s+1} = x^s - r_k g^s \quad ,$$

$$g^{s+1} \in \partial f(x^{s+1}) \; .$$

Step 2 If $f(x^{s+1}) > f(x^0) + \delta$, then $x^{s+1} \in \{x : f(x) \le f(x^0)\}$ and go to Step 5.

Step 3 Define

$$e_o^{s+1} = \frac{s+1}{s-j+2} e_o^s + \frac{1}{s-j+2} g^{s+1} \quad ,$$

$$e_1^{s+1} = P_1(g^{s-k_1+1}, \ldots, g^{s+1}) \; ,$$

$$e_m^{s+1} = P_m(g^{s-k_m+1}, \ldots, g^{s+1}) \; .$$

Each of the notations $P_t(\cdot, \cdot, \cdot)$ designates an arbitrary convex combination of a finite number of the indicated preceding subgradients.

Find

$$\mu_{s+1} = \min_{0 \le p \le m} \|e_p^{s+1}\| \; .$$

Step 4 If $\mu_{s+1} > \varepsilon_k$, then $s = s + 1$ and go to Step 1.

Step 5 Set $k = k + 1$, $j = s + 1$, $s = s + 1$, $e^s = g^s$ and go to Step 1.

THEOREM 2.1 *Suppose that the problem (*) is solved by the modified algorithm 1. Then all limit points of the sequence $\{x^s\}$ belong to X^*.*

3. METHODS WITH AVERAGING OF SUBGRADIENTS AND PROGRAM-ADAPTIVE SUCCESSIVE STEP-SIZE REGULATION

Successive Step-Size Regulation

As noted in a number of works [2, 3, 12, 16] it is expedient to average subgradients calculated at the previous iterations so that the subgradient methods will be more regular. For instance, when the "ravine"-type functions are minimized, the averaged direction points the way along the bottom of the "ravine".

It will be demonstrated in Section 5 that the operation of averaging enables the improvement of a posteriori estimates of the solution accuracy along with the upgrading of regularity of the described methods.

Methods with averaging of subgradients and consecutive program-adaptive regulation of the step-size are set forth in this section.

Results obtained here stem from [24].

Description of Algorithm 2.

Let x^0 be an arbitrary initial approximation; $\bar{\delta} > 0$ be a constant; $\{\varepsilon_k\}$, $\{r_k\}$ be number sequences such that

$$\varepsilon_k > 0,\ \varepsilon_k \longrightarrow 0,\ r_k > 0,\ r_k \longrightarrow 0\ .$$

Put $s = 0,\ j = 0,\ k = 0,$

$$v^0 = g^0,\ e^0 = g^0,\ g^0 \in \partial f(x^0)\ .$$

Step 1 Construct

$$x^{s+1} = x^s - r_k v^s\ .$$

Step 2 If $f(x^{s+1}) > f(x^0) + \bar{\delta}$, then go to Step 7.

Step 3 Define v^{s+1} according to the schemes a) or b).

Step 4 Construct $e^{s+1} = e^s + (s - j + 2)^{-1}(v^{s+1} - e^s)$.

Step 5 If $\|e^{s+1}\| > \varepsilon_k$, then $s = s + 1$ and go to Step 1.

Step 6 Set $k = k + 1$, $j = s + 1$, $s = s + 1$, $e^s = v^s$ and go to Step 1.

Step 7 Set $x^{s+1} \in \{x : f(x) \leq f(x^0)\}$, $s = s + 1$, $j = s$, $k = k + 1$ and go to Step 1.

In construction of the direction v^s the following schemes of subgradient averaging are dealt with.

a) The "moving" average. Let $K \geq 1$ be an integer. Then

$$v^s = \begin{cases} \sum\limits_{i=s-K}^{s} \lambda_{i,s} g^i & \text{if } s \geq K, \ \sum\limits_{i=s-K}^{s} \lambda_{i,s} = 1, \\ \sum\limits_{i=0}^{s} \lambda_{i,s} g^i & \text{if } s < K, \ \sum\limits_{i=0}^{s} \lambda_{i,s} = 1 \end{cases}$$

where $g^i \in \partial f(x^i)$, $\lambda_{i,s} \geq 0$.

b) The "weighted" average. Let $M \geq 1$ be an integer. Then $v^s = g^s + \lambda_s(v^{s-1} - g^s)$, where $0 \leq \lambda_s \leq 1$ for $s \not\equiv 0 \pmod{M}$, $0 \leq \lambda_s \leq \delta < 1$ for $s = 0 \pmod{M}$.

THEOREM 3.1 *Assume that the problem (*) is solved by algorithm 2. Then all limit points of the sequence* $\{x^s\}$ *belong to the set* X^*.

4. STOCHASTIC FINITE-DIFFEENCE ANALOGS TO ADAPTIVE NONMONOTONIC METHODS WITH AVERAGING OF SUBGRADIENTS

It should be emphasized that the practical value of the subgradient-type methods essentially depends upon the existence of their finite-difference analogs. Of great importance the finite-difference methods are primarily in situations when subgradient computation programs are unavailable. This generally occurs in the solution of large-scale problems. Construction of the finite-difference methods in the nonsmooth optimization originated two approaches: the nondeterministic and the stochastic ones. Each of them has its own advantages and disadvantages. The stochastic approach is favored here.

One of the advantages of the introduced averaging operation is the fact that the construction of stochastic analogs to subgradient methods presents no special problems.

The offered methods are close to those with smoothing [4] which, in their turn, are closest to the schemes of stochastic quasi-gradient methods [12]. Research into the stochastic quasi-gradient methods with successive step-size regulation is quite a new and underdeveloped field. Ju. M. Ermol'jev spurred first the investigations in this direction. His and Ju. M. Kaniovskij results [13] are undoubtedly of theoretical interest. However implementation of methods described in [14] creates complications as there is no rule to regulate variations in the step-size.

Let us first dwell on functions $f(x, i)$ of the form

$$f(x, i) = \left[\frac{1}{2\alpha_i}\right]^n \int_{-\alpha_i}^{\alpha_i} \cdots \int_{-\alpha_i}^{\alpha_i} f(x + y) dy_1 \cdots dy_n ,$$

where $\alpha_i > 0$.

Properties of the functions $f(x, i)$ have been studied by A.M. Gupal [4] proceeding from the assumption that $f(x)$ satisfies the Lipschitz local condition.

THEOREM 4.1 *If $f(x)$ is a convex eigenfunction, $\mathrm{dom}\, f = E^n$, then $f(x, i)$ is also a convex eigenfunction, $\mathrm{dom}\, f(x, i) = E^n$, for any $\alpha_i > 0$.*

THEOREM 4.2 *A sequence of functions $f(x, i)$ uniformly converges to $f(x)$ with $\alpha_i \to 0$ in any bounded domain X.*

Now we shall go to the description of stochastic finite-difference analogs to algorithms with successive program-adaptive regulation of the step-size and with averaging of the direction.

Description of Algorithm 3 Let x^0 be an arbitrary initial approximation, $\delta > 0$ be a constant, $\{r_i\}$, $\{t_i\}$, $\{\alpha_i\}$, $\{\rho_i\}$ be number sequences.

Put $s = 0$, $i = 0$, $j = 0$.

Step 1 Compute

$$\xi^s = \frac{1}{2\alpha_i} \sum_{k=1}^{n} (f(\tilde{x}_1^s, \ldots, x_k^s + \alpha_i, \ldots, \tilde{x}_n^s)$$

$$- f(\tilde{x}_1^s, \ldots, x_k^s - \alpha_i, \ldots, \tilde{x}_n^s))e^k ,$$

where $\tilde{x}_k^s$, $k = \overline{1, n}$ are independent random values distributed uniformly on intervals $[x_k^s - \alpha_i, x_k^s + \alpha_i]$, $\alpha_i > 0$.

Step 2 Construct e^s in compliance with the schemes a) and b), where the subgradients are replace by their stochastic estimates.

Step 3 Find $x^{s+1} = x^s - r_i e^s$.

Step 4 If $f(x^{s+1}) > f(x^0) + \delta$, then go to Step 9.

Step 5 Define $z^{s+1} = z^s + (s - j + 1)^{-1}(e^s - z^s)$

Step 6 If $s - j < \rho_i$, then $s = s + 1$ and go to Step 1.

Step 7 If $\|z^{s+1}\| > t_i$, then $s = s + 1$ and go to Step 1.

Step 8 Put $i = i + 1$, $j = s + 1$, $s = s + 1$ and go to Step 1.

Step 9 Set $x^{s+1} \in \{x : f(x) \le f(x^0)\}$, $j = s + 1$, $i = i + 1$, $s = s + 1$ and go to Step 1.

THEOREM 4.3 *Let the problem* (*) *be solved by algorithm 3 and the number sequences*

$$\{r_i\},\ \{t_i\},\ \{\lambda_i\},\ \{\delta_i\},\ \{\rho_i\},\ \{\alpha_i\},\ i = 0, 1, 2, \ldots$$

satisfy the following conditions

$$r_i > 0,\ r_i \to 0,\ t_i > 0,\ t_i \to 0,\ \lambda_i > 0,\ \lambda_i \to 0\ ,$$

$$\delta_i > 0,\ \sum_{i=0}^{\infty} \delta_i < \infty,\ \alpha_i > 0,\ \alpha_i \to 0\ ,$$

$$\frac{|\alpha_i - \alpha_{i+1}|}{r_{i+1}} \to 0,\ \frac{|\alpha_i - \alpha_{i+1}|}{r_{i+1}} \to 0\ ,$$

$$\rho_i r_i \to 0,\ \frac{c}{\lambda_i^2 \delta_i} \le \rho_i < \infty\ ,$$

$$0 < c < \infty\ .$$

Then almost for all ω the sequence $f(x^s(\omega))$ converges and all limit points of the sequence $\{x^s(\omega)\}$ belong to the set of solutions X^*. Theorem 4.3 is proved in detail in [25].

5. A POSTERIORI ESTIMATES OF ACCURACY OF SOLUTION TO ADAPTIVE SUBGRADIENT METHODS AND THEIR STOCHASTIC FINITE-DIFFERENCE ANALOGS

In numeric solution of extremum problems of nondifferentiable optimization strong emphasis is placed on the check of obtained solution accuracy. Given the solution accuracy estimates, first, a very efficient rule of algorithm stopping can be formulated, second, the obtained estimates can form the basis for justified conclusions with respect to the strategy of selection of algorithm parameters.

Using rather simple procedure a posteriori estimates of solution accuracy for the introduced adaptive algorithms are constructed here. The estimates provide a means for strictly evaluating efficiency of the averaging operation use.

Thus, assume that the convex function minimization problem

$$\min_{x \in E^n} f(x) \quad (*)$$

is being solved.

Suppose the set X^* contains only one point x^*.

To solve the problem (*) consider algorithm 1. The spin-off from the proof of theorem 2.1 is the proof that the sequence $\{x^s\}$ falls outside the set $\{x : f(x) \leq f(x^0) + \delta\}$ a finite number of times only. Therefore, $\tilde{s} \geq 0$ exists such that for $s \geq \tilde{s}$

$$x^s \in \{x : f(x) \leq f(x^0) + \delta\} \ .$$

Then the step size will vary only if the condition $\|e^{s+1}\| \leq \varepsilon_k$ is satisfied, where

$$e^{s+1} = e^s + (s - j + 2)^{-1}(g^{s+1} - e^s) \ ,$$

$$g^{s+1} \in \partial f(x^{s+1}) \ .$$

Without loss of generality we will assume that the first instant of the change from the step r_0 to r_1 occurred just because the condition

$$\|e^{s_0}\| \leq \varepsilon_0$$

is satisfied.

From the convexity of the function $f(x)$ it is inferred that

$$f^0 \leq f^* + (g^0, x^0 - x^*) \ , \tag{5.1}$$

$$f^1 \leq f^* + (g^1, x^1 - x^*) \ , \tag{5.2}$$

$$f^{s_0} \leq f^* + (g^{s_0}, x^{s_0} - x^*) \ . \tag{5.3}$$

Summation of inequalities (5.1), (5.2), . . . (5.3) yields

$$\frac{\sum_{i=0}^{s_0} f^i}{s_0 + 1} \leq f^* + \frac{1}{s_0 + 1} \sum_{i=0}^{s_0} (g^i, x^i - x^*) =$$

$$= f^{*} + (e^{s_0}, x^0 - x^{*}) + \frac{1}{s_0 + 1} \sum_{i=1}^{s_0} (g^0, x^i - x^0) \ .$$

Denote the expression $(s_0 + 1)^{-1} \sum_{i=1}^{s_0} (g^i, x^i - x^0)$ by Δ_0.

We have obvious inequalities

$$f(\tilde{x}^{s_0}) = f\left(\frac{\sum_{i=0}^{s_0} x^i}{s_0 + 1} \right) \le \frac{\sum_{i=0}^{s_0} f^i}{s_0 + 1} \le f^{*} + (e^{s_0}, x^0 - x^{*}) + \Delta_0 \ , \tag{5.4}$$

$$f(x_{\min}^{s_0}) \le f^{*} + (e^{s_0}, x^0 - x^{*}) + \Delta_0 \ ,$$

$$x_{\min}^{s_0} \in \{x : f(x) \le \min \{f(x^0), f(x^1), \ldots, f(x^{s_0})\}\} \ ,$$

where with $s_0 \le s \le s_1$ the points x^s are related by $x^{s+1} = x^s - r_1 g^s$. For these values of s it is possible to derive that

$$f(\tilde{x}^{s_1}) \le f\left(\frac{\sum_{i=s_0+1}^{s_1} x^i}{s_1 - s_0} \right) \le f^{*} + (e^{s_1}, x^{s_0+1} - x^{*}) + \Delta_1 \ , \tag{5.5}$$

$$f(x_{\min}^{s_1}) \le f^{*} + (e^{s_1}, x^{s_0+1} - x^{*}) + \Delta_1 \ , \tag{5.6}$$

where $\Delta_1 = (s_1 - s_0)^{-1} \sum_{i=s_0+2}^{s_1} (g^i, x^i - x^{s_0+1})$,

$$x_{\min}^{s_1} \in \{x : f(x) \le \min \{f(x^{s_0+1}), \ldots, f(x^{s_1})\}\} \ .$$

Thus, for $s_k + 1 \le s \le s_{k+1}$ we have

$$f(\tilde{x}^{s_{k+1}}) = f\left(\frac{\sum_{i=s_k+1}^{s_{k+1}} x^i}{s_{k+1} - s_k} \right) \le f^{*} + (e^{s_{k+1}}, x^{s_k+1} - x^{*}) + \Delta_k \ ; \tag{5.7}$$

$$f(x_{\min}^{s_{k+1}}) \le \frac{\sum_{i=s_k+1}^{s_{k+1}} f(x^i)}{s_{k+1} - s_k} \le f^{*} + (e^{s_{k+1}}, x^{s_k+1} - x^{*}) + \Delta_k \ , \tag{5.8}$$

where

$$x_{\min}^{s_{k+1}} \in \{x : f(x) \le \min \{f(x^{s_k+1}) \ , \ldots, f(x^{s_{k+1}})\}\} \ ,$$

$$\Delta_k = \frac{1}{s_{k+1} - s_k} \sum_{i=s_k+2}^{s_{k+1}} (g^i, x^i - x^{s_k+1}) \ .$$

It is easily proved that $\Delta_k \to 0$.

THEOREM 5.1 *Assume that the problem* (*) *is solved by algorithm 1. Then the inequalities*

$$f(x_{\min}^{s_k}) \le f^* + \varepsilon_k \|x^{s_{k-1}+1} - x^*\| + \Delta_k \ ,$$

$$f(\tilde{x}^{s_k}) \le f^* + \varepsilon_k \|x^{s_{k-1}+1} - x^*\| + \Delta_k \ .$$

hold for such instants s_k at which the step-size varies because the condition $\|e^{s_k}\| \le \varepsilon_k$ is satisfied.

REMARK It follows from theorem 5.1 that the same estimate occurs both for the subsequence of "records" $\{x_{\min}^{s_k}\}$ and for Cesaro subsequence $\{\tilde{x}^{s_k}\}$.

Let the problem (*) be solved by algorithm 2 where the operation of averaging of proceeding subgradients is used. Denote instants of changes in the step-size by s_i, $i = 0, 1, 2, \ldots$. Suppose the first instant of the change from r_0 to r_1 takes place because the inequality $\|e^{s_0}\| \le \varepsilon_0$ holds. Examine the scheme of averaging by "moving" average. We have

$$f^0 \le f^* + (g^0, x^0 - x^*) \ ,$$

$$f^s \le f^* + (g^s, x^s - x^*) \ ,$$

$$\sum_{i=0}^{s} \lambda_{i,s} f^i \le f^* + (v^s, x^0 - x^*) + \sum_{i=0}^{s} \lambda_{i,s} (g^i, x^i - x^0) \ .$$

Designate the expression $\sum_{i=0}^{s} \lambda_{i,s} f^i$ by $\tilde{f}^s$.

Then

$$\tilde{f}^0 \le f^* + (v^0, x^0 - x^*) \ ,$$

$$\tilde{f}^s \le f^* + (v^s, x^0 - x^*) + \sum_{i=0}^{s} \lambda_{i,s} (g^i, x^i - x^0) \ .$$

Whence for $s \le K$ we have

$$\sum_{i=0}^{s} \tilde{f}^s \le f^* + (e^s, x^0 - x^*) + \frac{\sum_{j=1}^{s} \sum_{i=0}^{j} \lambda_{i,j} (g^i, x^i - x^0)}{s+1} \ .$$

For $s > K$ we shall have

$$\tilde{f}^s \leq f^* + (v^s, x^0 - x^*) + \sum_{i=s-K}^{s} \lambda_{i,s}(g^i, x^i - x^0) \ .$$

Thus,

$$\sum_{i=0}^{K} \tilde{f}^i + \sum_{i=K+1}^{s} \tilde{f}^i \leq f^* + (e^s, x^0 - x^*)$$

$$+ \frac{1}{s+1}\left[\sum_{j=1}^{K}\sum_{i=0}^{j} \lambda_{i,j}(g^i, x^i - x^0) + \sum_{j=K+1}^{s}\sum_{i=s-K}^{j} (g^i, x^i - x^0)\right] .$$

From the formula

$$\tilde{f}^s \leq f^* + (v^s, x^0 - x^*) + \sum_{i=0}^{s} \lambda_{i,s}(g^i, x^i - x^0)$$

the following recommendations can be offered with respect to the selection of parameters $\lambda_{i,s}$:

(1) $\min\limits_{\lambda_{i,s} \geq 0} \{\| \sum_{i=0}^{s} \lambda_{i,s} g^i \| + \sum_{i=1}^{s} \lambda_{i,s}(g^i, x^i - x^0), \sum_{i=0}^{s} \lambda_{i,s} = 1\}$

(2) $\min\limits_{\lambda_i, s \geq 0} \sum_{i=0}^{s} \lambda i, s\,(g^i, x^i - x^0) \ , \ \sum_{i=0}^{s} \lambda i, s = 1$

The subgradient averaging thereby allows improving a posteriori estimates of the solution accuracy. This may substantiate formally that it is of advantage to introduce and study the operation of subgradient averaging.

For an arbitrary instant of step-size variation $s_i > K$ we can easily obtain the estimate

$$\frac{\sum\limits_{l=s_{i-1}}^{s_i} \tilde{f}^l}{s_i - s_{i-1} + 1} \leq f^* + (e^{s_i}, x^{s_{i-1}} - x^*)$$

$$+ (s_i - s_{i-1} + 1)^{-1} \sum_{l=s_{i-1}}^{s_i} \sum_{j=l-K}^{l} \lambda_{l,j}(g^l, x^l - x^{s_{i-1}}) \ . \qquad (5.9)$$

THEOREM 5.1 *Let the problem (*) be solved by algorithm 2 with the use of averaging scheme a). Then for the instants s_i, for which $\|e^{s_i}\| \leq \varepsilon_i$, inequality (5.9) holds. The scheme of averaging by "weighted" average b) is treated in a* *similar way.*

The a posteriori estimates of the solution accuracy attained for the adaptive subgradient methods can be extended to their stochastic finite-difference analogs with the minimum of alterations. The way of getting them is illustrated with algorithm 3. We will use notations introduced in Section 4. When proving theorem 4.3 it is possible to demonstrate that the step-size r_i varies an infinite number of times. As algorithm 3 converges with a probability of unity, then for almost all ω it is possible to indicate $\tilde{s}(\omega)$ such that with $s \geq \tilde{s}$

$$x^s \in \{x : f(x) \leq f(x^0) + \delta\} \ .$$

Therefore, with $s \geq \tilde{s}(\omega)$ the step-size r_i varies because the condition

$$\|z^{s_i}\| \leq t_i$$

holds, where $s_i \geq \rho_i + j$, $z^{s_i} = z^{s_i - 1} + (s_i - j)^{-1}(\xi^{s_i} - z^{s_i - 1})$ sequences $\{t_i\}$ and $\{\rho_i\}$ comply with properties formulated in theorem 4.3, j is reconstructed by s_i.

Consider the event

$$A^i = \left\{ \max_{k \geq \rho_i} \frac{1}{k} \| \sum_{l=1}^{k} (\xi^{s_{i-1}+l} - \nabla f(x^{s_{i-1}+l}, i-1)) \| > \lambda i \right\} ,$$

where s_{i-1} is the instant of step-size change that precedes s_i. There exists the constant $0 < c < \infty$ such that with the probability greater than $1 - C\delta_i$ it is possible to state that

$$\frac{1}{s_i} \| \sum_{l=1}^{s_i} (\xi^{s_{i-1}+l} - \nabla f(x^{s_{i-1}+l}, i-1)) \| \leq \lambda_i \ .$$

Then for the instant s_i the inequality

$$\frac{1}{s_i} \| \sum_{l=1}^{s_i} \nabla f(x^{s_{i-1}+l}, i-1) \| \leq \| \sum_{l=1}^{s_i} \nabla f(x^{s_{i-1}+l}, i-1)$$

$$- z^{s_i} \| + \| z^{s_i} \| \leq \lambda_i + t_i$$

holds with the same probability.

Theorem 5.3 is readily formulated and proved. Assume that the problem (*) is solved by algorithm 3. Then for almost all ω it is possible to isolate a subsequence of points $\{x^{s_i}(\omega)\}$ for which with the probability greater than $1 - C\delta_i$ the inequalities hold

$$\frac{\sum\limits_{l=s_{i-1}}^{s_i} f(x^l, i-1)}{s_i - s_{i-1} + 1} \le f_{i-1}^* + (\lambda_i + t_i)\|x^{s_{i-1}} - x_{i-1}^*\|$$

$$+ c\|x_{\max}^{s_i} - x^{s_{i-1}}\| ,$$

where $f_{i-1}^* = \min\limits_{x \in E^n} f(x, i-1),$

$$x_{i-1}^* \in \text{Argmin}\, f(x, i-1) \ .$$

REFERENCES

1 Ajzerman, M.A., E.M. Braverman and L.I. Rozonoer: Potential Functions Method in Machine Learning Theory. M.: Nauka, 1970, p. 384.

2 Glushkova, O.V. and A.M. Gupal: About Nonmonotonic Methods of Nonsmooth Function Minimization with Averaging of Subgradients. Kibernetika, 1980, No. 6, pp. 128–129.

3 Gupal, A.M. and L.G. Bazhenov: Stochastic Analog to Conjugate Gradient Method. Kibernetika, 1972, No. 1, pp. 125–126.

4 Gupal, A.M.: Stochastic Methods of Solution of Nonsmooth Extremum Problems. Kiev: Naukova dumka, 1979, p. 152.

5 Dem'janov, V.F. and V.N. Malozemov: Introduction to Minimax. M.: Nauka, 1972, p. 368.

6 Eremin, I.I.: The Relaxation Method of Solving Systems of Inequalities with Convex Functions on the Left Side. Dokl. AN SSSR, 1965, Vol. 160, No. 5, pp. 994–996.

7 Ermol'ev, Ju.M.: Methods of Solution of Nonlinear Extremum Problems. Kibernetika, 1966, No. 4, pp. 1–17.

8 Ermol'ev, Ju.M. and N.Z. Shor: On Minimization of Nondifferentiable Functions. Kibernetika, 1967, No. 1, pp. 101–102.

9 Ermol'ev, Ju.M. and Z.V. Nekrylova: Some Methods of Stochastic Optimization. Kibernetika, 1966, No. 6, pp. 96–98.

10 Ermol'ev, Ju.M.: On the Method of Generalized Stochastic Gradients and Stochastic Quasi-Fejer Sequences. Kibernetika, 1969, No. 2, pp. 73–83.

11 Ermol'ev, Ju.M.: On One General Problem of Stochastic Programming. Kibernetika, 1971, No. 3, pp. 47–50.

12 Ermol'ev, Ju.M.: Stochastic Programming Methods. M.: Nauka, 1976, p. 240.

13 Ermol'ev, Ju.M. and Ju.M. Kaniovskij: Asymptotic Properties of Some Stochastic Programming Methods with Constant Step-Size. Zhurn. Vych. Mat. i Mat. Fiziki, 1979, Vol. 19, No. 2, pp. 356–366.

14 Kaniovskij, Ju.M., P.S. Knopov and Z.V. Nekrylova: Limit Theorems for Stochastic Programming. Kiev: Naukova dumka, 1980, p. 156.

15 Loev, M.: Probability Theory. M.: Izd-vo inostr. lit., 1967, p. 720.

16 Norkin, V.N.: Method of Nondifferentiable Function Minimization with Averaging of Generalized Gradients. Kibernetika, 1980, No. 6, pp. 86–89, 102.

17 Nurminskij, E.A.: Convergence Conditions for Nonlinear Programming Algorithms. Kibernetika, 1973, No. 1, pp. 122–125.

18 Nurminskij, E.A. and A.A. Zhelikovskij: Investigation of One Regulation of Step in Quasi-Gradient Method for Minimizing Weakly Convex Functions. Kibernetika, 1974, No. 6, pp. 101–105.

19 Poljak, B.T.: Generalized Method of Solving Extremum Problems. Dokl. AN SSSR, 1967, Vol. 174, No. 1, pp. 33–36.

20 Poljak, B.T.: Minimization of Nonsmooth Functionals. Zhurn. vychisl. mat. i mat. fiziki, 1969, Vol. 9, No. 3, pp. 509–521.

21 Tsypkin, Ja.Z.: Adaptation and Learning in Automatic Systems. M.: Nauka, 1968.

22 Tsypkin, Ja.Z.: Generalized Learning Algorithms. Avtomatika i telemekhanika, 1970, No. 1, pp. 97–103.

23 Chepurnoj, N.D.: One Successive Step-Size Regulation for Quasi-Gradient Method of Weakly Convex Function Minimization. Collection: Issledovanie Operacij i ASU. Kiev: Vyshcha shkola, 1981, No. 19, pp. 13–15.

24 Chepurnoj, N.D.: Averaged Quasi-Gradient Method with Successive Step-Size Regulation to Minimize Weakly Convex Functions. Kibernetika, 1981, No. 6, pp. 131–132.

25 Chepurnoj, N.D.: One Successive Step-Size Regulation in Stochastic Method of Nonsmooth Function Minimization. Kibernetika, 1982, No. 4, pp. 127–129.

26 Shor, N.Z.: Application of Gradient Descent Method for Solution of Network Transportation Problem. In: Materialy nauchnogo seminara po prikladnym voprosam kibernetiki i issledovanija operacij. Nauchnyj sovet po kibernetike IK AN USSR, Kiev, 1962, vypusk 1, pp. 9–17.

27 Shor, N.Z.: Investigation of Space Dilation Operation in Convex Function Minimization Problems. Kibernetika, 1970, No. 1, pp. 6–12.

28 Shor, N.Z. and N.G. Zhurbenko: Minimization Method Using Space Dilation, in the Direction of Difference of Two Successive Gradient. Kibernetika, 1971, No. 3, pp. 51–59.

29 Shor, N.Z.: Nondifferentiable Function Minimization Methods and Their Applications. Kiev, Nauk. dumka, 1979, p. 200.

30 Demjanov, V.F.: Algorithms for Some Minimax Problems. Journal of Computer and Systems Science, 1968, 2, No. 4, pp. 342–380.

31 Lemarechal, C.: An Algorithm for Minimizing Convex Functions. In: Information Processing'74 /ed. Rosenfeld/, 1974, North-Holland, Amsterdam, pp. 552–556.

32 Lemarechal, C.: Nondifferentiable Optimization: Subgradient and Epsilon Subgradient Methods. Lecture Notes in Economics and Mathematical Systems /ed. Oettli W./, 1975, 117, Springer, Berlin, pp. 191–199.

33 Bertsekas, D.P. and S.K. Mitter: A Descent Numerical Method for Optimization Problems with Nondifferentiable Cost Functions. SIAM Journal on Control, 1973, 11, No. 4, pp. 637–652.

34 Wolfe, P.: A Method of Conjugate Subgradients for Minimizing Nondifferentiable Functions. In: Nondifferentiable Optimization /eds. Balinski M.L., Wolfe P./, Mathematical Programming Study 3, 1975, North-Holland, Amsterdam. pp. 145–173.