WORKING PAPER

ADAPTIVE VARIABLE METRIC ALGORITHMS FOR NONSMOOTH OPTIMIZATION PROBLEMS

Stanislav Uryas'ev

July 1988
WP-88-60

IIASA
International Institute
for Applied Systems Analysis

ADAPTIVE VARIABLE METRIC ALGORITHMS FOR NONSMOOTH OPTIMIZATION PROBLEMS

Stanislav Uryas'ev

July 1988
WP-88-60

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

FOREWORD

This paper deals with new variable metric algorithms for nonsmooth optimization problems. The author develops so-called adaptive algorithms. The essence of such algorithms is as follows: there are two simultaneously working gradient algorithms, the first is in the main space and the second with respect to the matrix for modification of the space. The author proves convergence of such algorithms for different cases.

Alexander B. Kurzhanski
Chairman
System and Decision Sciences Program

CONTENTS

ADAPTIVE VARIABLE METRIC ALGORITHMS FOR NONSMOOTH OPTIMIZATION PROBLEMS

Stanislav Uryas'ev

1. INTRODUCTION

Variable metric algorithms are widely used for smooth optimization problems (see for example the review article [1]). As a rule this algorithm can not be generalized for nonsmooth optimization problems. The difficulties are connected with the fact that even if the first and second derivatives exist at some point they do not give the full local description of the function. Because the function is nonsmooth, a point of nonsmoothness can be arbitarily close to the point where derivatives exist. For this reason the quasi-Newton methods can not be automatically generalized for nonsmooth problems.

These difficulties lead to the appearance of new ideas for the construction of variable metric algorithm. In the works of N. Shor (see for example [2]) and his pupils, so-called space-dilatation algorithms were developed. Such an approach gives the opportunity to construct practical and effective algorithms, but the most effective algorithm (r-algorithms) from this family is not sufficiently understood from the theoretical point of view.

This author proposed an alternative "adaptive" approach, that can be applied to optimization and game theoretic problems. This approach stems from the article [3], where a step size control was proposed for the stochastic quasi-gradient algorithm [4]. The first variable metric algorithm based on such approach was proposed in the paper [5]. In the paper [6], a short review of new variable metric algorithms is given for different optimization problems: smooth, nonsmooth, stochastic optimization problem.

2. ESSENCE OF THE APPROACH

Let us consider a convex optimization problem

$$f(x) \to \min_{x \in R^n} , \tag{1}$$

where the function $f(x)$ is convex on the Euclidean space R^n. We use the following recurrent algorithm for the solution of this problem:

$$x^{s+1} = x^s - \rho_s H^s g^s, \quad s = 0, 1, \cdots \tag{2}$$

here s is the iteration number, $\rho_s > 0$ is step size (scalar value); H^s is a $n \times n$ matrix; g^s is a subgradient from the subdifferential $\partial f(x)$ of the function $f(x)$ at the point x^s, i.e. $g^s \in \partial f(x^s)$. We recall that the subdifferential of the function $f(x)$ at the point $y \in R^n$ is given by the formula (see, for example [7])

$$\partial f(y) = \{g \in R^n : f(x) - f(y) \geq <g, x - y> \quad \text{for } \forall x \in R^n\} \ .$$

At the s^{th} iteration the natural criterion defining the best choice of matrix H^s is via the function

$$\varphi_s(H) = f(x^s - \rho_s H g^s) \ .$$

The best matrix is a solution of the problem

$$\varphi_s(H) \rightarrow \min_{H \in R^{n \times n}} \ . \tag{3}$$

It is easy to see that problem (3) is a reformulation of the source problem (1), since if H^* is a solution of (3) then the point $x^s - \rho_s H^* g^s$ is a solution of (1). More than that, the problem (3) is more complex than (1) because the dimension of the problem (3) is in n times higher the dimension of (1). However, at the s^{th} iteration of algorithm (2) we do not need the optimal matrix, it is enough to correct (update) the matrix H^s. If we already have some matrix H_0^s, then the direction of adaptation can be defined by differentiating, in the general sense, the function $\varphi_s(H)$ at the point H_0^s. If the function $F(x)$ is a convex function then the function $\varphi_s(H)$ is also convex. We can use the following formula [8] for the differentiation of the complex function φ_s:

$$\partial \varphi_s(H_0^s) = -\rho_s \{g g^{s^T} : g \in \partial f(x^s - \rho_s H_0^s g^s)\} \ ;$$

here and below the superscript T means transposition. If $g_0^s \in \partial f(x^s - \rho_s H_0^s g^s)$, then $-\rho_s g_0^s g^{s^T} \in \partial \varphi(H_0^s)$. With respect to the matrix H, in the direction $g_0^s g^{s^T}$, one can do a step of the generalized gradient method:

$$H_1^s = H_0^s + \lambda_0^s g_0^s g^{s^T}, \ \lambda_0^s > 0 \ .$$

It is possible either to take $H^s = H_1^s$ or to continue the iterations of the generalized gradient algorithm with respect to H

$$H_{i+1}^s = H_i^s + \lambda_i^s g_i^s g^{s\,T}, \ \lambda_i^s > 0 \ , \tag{4}$$

where $g_i^s \in \partial f(x^s - \rho_s H_i^s g^s)$ and $-\rho_s g_i^s g^{s\,T} \in \partial \varphi_s(H_i^s)$. For some $i(s) \geq 1$ assume $H^s = H_{i(s)}^s$. At the next iteration $H_0^{s+1} = H^s$. The number $i(s)$ can be taken independently of s, for example, $i(s) = 1$ for all s. Generally speaking, the algorithm (2) is not monotone with respect to the objective function $f(x)$. However, one can chose $i(s)$ such that

$$f(x^s - \rho_s H_{i(s)}^s g^s) < f(x^s - \rho_s H_0^s g^s)$$

and on each iteration the objective function decreases.

Note that matrix updating requires additional calculations of objective function subgradients. This can be avoided by taking $g^{s+1} = g_0^s$, $i(s) = 1$ and using the matrix H_1^s at $(s+1)^{\text{th}}$ iteration. Therefore we propose the following formula for matrix updating

$$H^{s++1} = H^s + \lambda_s g^{s+1} g^{s\,T}, \ \lambda^s > 0 \ . \tag{5}$$

In formula (5), additional subgradient calculations are not required.

3. CONVERGENCE FOR SMOOTH FUNCTIONS

At first let us investigate the convergence of algorithm (2) and (5) for the case with a differentiable function $f(x)$. Denote $g^s = \nabla f(x^s)$. For algorithm (2), as direction of motion we choose the normalized subgradient

$$\xi^s = \begin{cases} \dfrac{g^s}{\|g^s\|} \ , & \text{for } \|g^s\| \neq 0 \ ; \\ 0 \ , & \text{for } \|g^s\| = 0 \ . \end{cases}$$

The algorithm can now be written in the following way

$$\left.\begin{aligned} & x^{s+1} = x^s - \rho_s H^s \xi^s, \\ & H^{s+1} = H^s + \rho_s g^{s+1} \xi^{s\,T}, \\ & H^0 = I \end{aligned}\right\} \tag{6}$$

where I is a unit matrix. (Thus, for the parameter λ_s in the formula (5) we choose the value $\rho_s / \|g^s\|$.) Denote by

$$\Theta^s = \sum_{l=0}^{s} \rho_l g^{l+1} \xi^{l\,T}, \ D(K) = \max_{y,\, x \in K} \|x - y\| \ ,$$

$$f^* = \min_{x \in R^n} f(x), \bar{f}_s = \min_{0 \le l \le s} f(x^l) \ ,$$

i.e. $D(K)$ is the diameter of a set K and $\bar{f}_s$ is a record of the lowest value of the function f during the previous s iterations. We also denote by $Tr(Q)$ the trace of matrix Q.

Let us now formulate the theorem about the convergence of algorithm (6).

THEOREM 1 *Let $f: R^n \to R$ be a convex smooth Lipschitz function*

$$f(x) - f(y) \le L_1 \|x - y\| \quad \text{for } x, y \in R^n \tag{7}$$

with Lipschitz gradient

$$\|\nabla f(x) - \nabla f(y)\| \le L_2 \|x - y\| \quad \text{for } x, y \in R^n \ , \tag{8}$$

and let there exist a compact set K such that

$$\|\nabla f(x)\| \ge \delta_k > 0 \quad \text{for } x \notin K \ . \tag{9}$$

If the step sizes ρ_s, $s = 0, 1, \ldots$, in the algorithm (6) are nonnegative, i.e. and for all s larger than some $\bar{s}$ one has

$$\delta_k \ge \left[\sum_{l=0}^{s} \rho_l^2 \left[L_1^2 2^{-1} + L_2 \left[n + 2(f(x^0) - f^*) + L_1^2 \sum_{t=0}^{l-1} \rho_t^2 \right]^{1/2} \right] + \right.$$

$$\left. + f(x^0) - f^* \right] \left[\sum_{l=0}^{s} \rho_l \right]^{-1} \tag{10}$$

then:

1) $$\bar{f}_s - f^* \le \left[\sum_{l=0}^{s} \rho_l^2 \left[L_1^2 2^{-1} + L_2 \left[n + 2(f(x^0) - f^*) + L_1^2 \sum_{t=0}^{l-1} \rho_t^2 \right]^{1/2} \right] + \right.$$

$$\left. + f(x^0) - f^* - 2^{-1} Tr(\Theta^s \Theta^{s^T}) \right] D(K) \left[\sum_{l=0}^{s} \rho_l \right]^{-1} \quad \text{for } s > \bar{S} \ ;$$

2) for $\rho_l = (s+1)^{-1/2}$, $l = 0, \ldots, s$ and $s > \bar{S}$

$$\bar{f}_s - f^* \le \left[L_1^2 2^{-1} + L_2 (n + 2(f(x^0) - f^*) + L_1^2)^{1/2} + f(x^0) - f^* - \right.$$

$$\left. - 2^{-1} Tr(\Theta^s \Theta^{s^T}) \right] D(K)(s+1)^{-1/2} \le c(s+1)^{-1/2} \ ,$$

where $c = const$;

3) for $\sum_{l=0}^{\infty} \rho_s^2 = \infty$

$$\lim_{s \to \infty} (f(x^s) - f^*) \left[\sum_{l=0}^{s-1} \rho_l^2\right]^{-1/2} \rho_s^{-1} \leq 2L_1 L_2 D(K) \ ,$$

if $\rho_s = s^{-1/2}$ *then*

$$\lim_{s \to \infty} (\bar{f}_s - f^*)(\ln(s))^{-1/2} s^{1/2} \leq \text{const} \ .$$

PROOF First we evaluate the Euclidean norm of the matrix H^s.

LEMMA 1 *The following inequality holds*

$$\|H^{s+1}\| \leq \left[n + 2(f(x^0) - f^*) + L_1^2 \sum_{l=0}^{s} \rho_l^2\right]^{1/2} \ . \tag{11}$$

PROOF Relation (6) implies

$$H^{s+1} H^{s+1^T} = (H^s + \rho_s g^{s+1} \xi^{s^T})(H^{s^T} + \rho_s \xi^s g^{s+1^T}) = H^s H^{s^T} +$$

$$+ \rho_s (g^{s+1} \xi^{s^T} H^{s^T} + H^s \xi^s g^{s+1^T}) + \rho_s^2 \|\xi^s\|^2 g^{s+1} g^{s+1^T} \ .$$

Since the function $f(x)$ satisfies the condition (7) then $\|g^s\| \leq L_1$, $s = 0, 1, \cdots$. Denote by $\Delta x^{s+1} = x^{s+1} - x^s$.

Taking into account the convexity of the function $f(x)$ we can evaluate the trace of $H^s H^{s^T}$

$$Tr(H^{s+1} H^{s+1^T}) = Tr(H^s H^{s^T}) + 2\rho_s < g^{s+1}, H^s \xi^s > + \rho_s^2 \|\xi^s\|^2 \|g^{s+1}\|^2 \leq$$

$$\leq Tr(H^s H^{s^T}) - 2 < g^{s+1}, \Delta x^{s+1} > + \rho_s^2 L_1^2 \leq$$

$$\leq Tr(H^s H^{s^T}) + 2(f(x^s) - f(x^{s+1})) + \rho_s^2 L_1^2 \leq$$

$$\leq Tr(H^0 H^{0^T}) + 2(f(x^0) - f(x^{s+1})) + L_1^2 \sum_{l=0}^{s} \rho_l^2 \leq$$

$$\leq n + 2(f(x^0) - f^*) + L_1^2 \sum_{l=0}^{s} \rho_l^2 \ . \tag{12}$$

Since $\|H^{s+1}\|^2 \leq Tr(H^{s+1} H^{s+1^T})$, then

$$\|H^{s+1}\| \leq \left[n + 2(f(x^0) - f^*) + L_1^2 \sum_{l=0}^{s} \rho_l^2\right]^{1/2} \ .$$

The Lemma is proved. □

The inequality (12) implies

$$Tr(H^{s+1}H^{s+1^T}) = Tr\left[\left(\sum_{l=0}^{s}\rho_l g^{l+1}\xi^{l^T} + I\right)\left(\sum_{l=0}^{s}\rho_l\xi^l g^{l+1^T} + I\right)\right] =$$

$$= Tr(\Theta^s\Theta^{s^T}) + 2\sum_{l=0}^{s}\rho_l <g^{l+1}, \xi^l> +$$

$$+ n \le n + 2(f(x^0) - f^*) + L_1^2\sum_{l=0}^{s}\rho_l^2 \ .$$

Consequently

$$\sum_{l=0}^{s}\rho_l<g^{l+1}, \xi^l> \le 2^{-1}L_1^2\sum_{l=0}^{s}\rho_l^2 - 2^{-1}Tr(\Theta^s\Theta^{s^T}) + f(x^0) - f^* \ . \tag{13}$$

Since the gradient of the function $f(x)$ satisfies Lipschitz condition, then applying Lemma 1 we see that

$$\|g^{l+1} - g^l\| \le L_2\|x^{l+1} - x^l\| = L_2\|\rho_l H^l\xi^l\| \le \rho_l L_2\|H^l\| \le \tag{14}$$

$$\le \rho_l L_2\left[n + 2(f(x^0) - f^*) + L_1^2\sum_{t=0}^{l-1}\rho_t^2\right]^{1/2} \ .$$

Using (13) and (14), it is easy to establish

$$\sum_{l=0}^{s}\rho_l\|g^l\| = \sum_{l=0}^{s}\rho_l<g^l, \xi^l> = \sum_{l=0}^{s}\rho_l<g^l - g^{l+1}, \xi^l> +$$

$$+ \sum_{l=0}^{s}\rho_l<g^{l+1}, \xi^l> \le \sum_{l=0}^{s}\rho_l\|g^l - g^{l+1}\| + \sum_{l=0}^{s}\rho_l<g^{l+1}, \xi^l> \le$$

$$\le L_2\sum_{l=0}^{s}\rho_l^2\left[n + 2(f(x^0) - f^*) + L_1^2\sum_{t=0}^{l-1}\rho_t^2\right]^{1/2} + 2^{-1}L_1^2\sum_{l=0}^{s}\rho_l^2 -$$

$$- 2^{-1}Tr(\Theta^s\Theta^{s^T}) + f(x^0) - f^* \ . \tag{15}$$

Consequently

$$\min_{0\le l\le s}\|g^l\| \le \left[\sum_{l=0}^{s}\rho_l\|g^l\|\right]\left[\sum_{l=0}^{s}\rho_l\right]^{-1} \le \left[\sum_{l=0}^{s}\rho_l^2\left[L_1^2 2^{-1} +\right.\right.$$

$$\left.\left. + L_2\left[n + 2(f(x^0) - f^*) + L_1^2\sum_{t=0}^{l-1}\rho_t^2\right]^{1/2}\right] -\right.$$

$$- 2^{-1} Tr(\Theta^s \Theta^{s^T}) + f(x^0) - f^* \Big] \Big[\sum_{l=0}^{s} \rho_l \Big]^{-1} \quad . \tag{16}$$

Combining inequality (10) with the last inequality we get

$$\min_{1 \le l \le s} \|g^l\| \le \delta_k \quad \text{for} \quad s > \bar{S} \quad .$$

Thus for a number $l(s)$ such that $g^{l(s)} = \min_{1 \le l \le s} \|g^l\|$, the inclusion $x^{l(s)} \in K$ holds because of the definition of δ_k. Let x^* be a minimum point i.e. $f(x^*) = f^*$. Using th convexity of the function $f(x)$ we see that

$$\bar{f}_s - f^* \le f(x^{l(s)}) - f(x^*) \le <g^{l(s)}, x^{l(s)} - x^*> \le$$

$$\le \|g^{l(s)}\| \, \|x^{l(s)} - x^*\| \le \|g^{l(s)}\| D(K) \le D(K) \min_{1 \le l \le s} \|g^l\| \quad .$$

Applying this last relation and (16) we get the statement 1 of the Theorem.

Statement 2 of the Theorem can be obtained by substituting in the statement 1 the values $\rho_l = (s+1)^{-1/2}$, $l = 0, \ldots, S$.

Let us now prove the statement 3 of the Theorem. It is enough to show that

$$\overline{\lim_{s \to \infty}} \|g^s\| \left[\sum_{l=0}^{s-1} \rho_l^2 \right]^{-1/2} \rho_s^{-1} \le 2 L_1 L_2 \quad , \tag{17}$$

because $\|g^s\| D(K) \ge f(x^s) - f^*$ for $s \ge \bar{S}$. If that inequality (17) does not hold, then there exists a number $\hat{S}$ such that for $s > \hat{S}$

$$\|g^s\| \ge 2 L_1 L_2 \rho_s \left[\sum_{l=0}^{s-1} \rho_l^2 \right]^{1/2} \quad .$$

Substituting the last inequality into the left side of the inequality (15) we get a contradiction for $s \to \infty$, because the left side of inequality tends to infinity faster than the right one. □

If the number of iterations of the algorithm was chosen before the start of the algorithm then statement 2 of the theorem implies that the convergence rate of algorithm (6) is not worse than the generalized gradient algorithm with matrix $H^s = I$, $s = 0, 1, \cdots$. For $H^s = I$, $s = 0, 1, \cdots$ the following estimate (see, for example [9]) is known

$$\bar{f}_s - f^* \le (s+1)^{-1/2} d(x^0, X^*) \quad ,$$

where

$$d(x, X^*) = \min_{y \in X^*} \|x - y\|, \; X^* = \{x^* : f(x^*) = f^*\} \; . \tag{18}$$

Note that in the estimate in statement 2 of the theorem there exists an additional term $Tr(\Theta^s \Theta^{s^T})$ that increases the convergence rate. This term does not let the algorithm "stick" when the objective function is ill-conditioned.

4. CONVERGENCE FOR NONSMOOTH FUNCTIONS

Let us consider algorithm (2). We suppose that at the s^{th} iteration of main algorithm for the updating of the matrix H_0^s, the formula (4) is used $i(s)$ times. At the iteration $s + 1$ we take $H_0^{s+1} = H_{i(s)}^s$.

At the zero iteration $H_0^0 = I$, where I is a unit matrix. Fix some $\epsilon > 0$. We choose $i(s)$ to be the minimal member such that

$$f(x^s - \rho_s H_{i(s)}^s g^s) \leq f(x^s) - \epsilon \; .$$

It is convenient to normalize the test vector g_i^s, therefore denote by

$$\xi_i^s = \begin{cases} 0, \text{ if } g_i^s = 0, \\ g_i^s \|g_i^s\|^{-1}, \text{ otherwise} \; . \end{cases} \tag{19}$$

For each $s = 0, 1, \cdots$ let the sequence $\{\lambda_{si}\}$, $i = 0, 1, \cdots$ of positive values be given. We write the algorithm in more detail.

ALGORITHM I

STEP I Initialization

$$s = 0, \; x^0 = x_{init}, \; g^0 \in \partial f(x^0), \; i = -1, \; H_0^{-1} = I \; .$$

STEP II

1 $H_0^s = H_{i+1}^{s\,-1}$, $i = 0$;

2 $x_i^s = x^s - \rho_s H_i^s g^s$;

3 compute $g_i^s \in \partial f(x_i^s)$, if $g_i^s = 0$ then STOP, otherwise $\xi_i^s = g_i^s \|g_i^s\|^{-1}$.

4 $H_{i+1}^s = H_i^s + \lambda_{si} \xi_i^s g^{s^T}$.

5 if $f(x_i^s) \leq f(x^s) - \epsilon$, then $i(s) = i$ and go to step III;

6 $i = i + 1$, return to point 2.

STEP III $x^{s+1} = x_i^s$, $g^{s+1} = g_i^s$.

STEP IV $s = s + 1$ and return to step II.

We now formulate a theorem about the convergence of algorithm 1.

THEOREM 2 *Let: $f: R^n \to R$ be a convex function; the set of minimum points X^* of the function $f(x)$ be non-empty and bounded; $\{\rho_s\}$ be a sequence of positive numbers; and $\{\lambda_{si}\}$, $i = 0, 1, \cdots$ be a given sequence of positive numbers satisfying*

$$\sum_{i=0}^{\infty} \lambda_{si} = \infty \sum_{i=0}^{\infty} \lambda_{si}^2 < \infty, \lambda_{si} \to 0 \quad \text{for } i \to \infty \ . \tag{20}$$

Then there exists number $\bar{s}$ such that $f(x^{\bar{s}}) \leq \epsilon + f^*$ and $d(x_i^{\bar{s}}, X^*) \to 0$, $f(x_i^{\bar{s}}) \to f^*$ for $i \to \infty$, where f^* equals minimum value of f on R^n.

PROOF Let us prove at first that if $f(x^s) - f^* - \epsilon = \delta > 0$ then there exists number $i(s)$ such that

$$f(x_{i(s)}^s) \leq f(x^s) - \epsilon \ . \tag{21}$$

Let $x^* \in X^*$. Using the formulae of the point 2 and 4 of Step II we get

$$\|x^* - x_{i+1}^s\|^2 = \|x^* - x^s + \rho_s H_{i+1}^s g^s\|^2 =$$

$$= \|x^* - x^s + \rho_s (H_i^s + \lambda_{si} \xi_i^s g^{s^T}) g^s\|^2 = \|x^* - x^s + \rho_s H_i^s g^s +$$

$$+ \rho_s \lambda_{si} \|g^s\|^2 \xi_i^s\|^2 = \|x^* - x_i^s + \rho_s \lambda_{si} \|g^s\|^2 \xi_i^s\|^2 \leq$$

$$\leq \|x^* - x_i^s\|^2 + 2\rho_s \lambda_{si} \|g^s\|^2 <\xi_i^s, x^* - x_i^s> + (\rho_s \lambda_{si} \|g^s\|^2)^2 =$$

$$= \|x^* - x_0^s\|^2 + 2\rho_s \|g^s\|^2 \sum_{l=0}^{i} \lambda_{sl} <\xi_l^s, x^* - x_l^s> + \rho_s^2 \|g^s\|^4 \sum_{l=0}^{i} \lambda_{sl}^2 \ . \tag{22}$$

We prove by contradiction that there exists a number $i(s)$ satisfying the inequality (21). Let for all $i > 0$

$$f(x_i^s) > f^* + \delta \ .$$

Since $<\xi_i^s, x^* - x_l^s> \leq 0$ then it follows from (22) and (20) that

$$\|x^* - x_{i+1}^s\|^2 \leq \|x^* - x_0^s\|^2 + \rho_s^2 \|g^s\|^4 \sum_{l=0}^{\infty} \lambda_{sl}^2 \leq C = \text{const}$$

The function $f: R^n \to R$ is convex, consequently on the compact set $\{x \in R^n : \|x^* - x\| \le C\}$ it satisfies Lipschitz condition with a constant L. Therefore

$$\|g_i^s\| \le L \quad \text{for} \quad i \ge 0 \ .$$

Using the convexity of the function $f(x)$ we obtain.

$$2\rho_s\|g^s\|^2 < \xi_i^s,\ x^* - x_i^s > \le 2\rho_s\|g^s\|^2\|g_i^s\|^{-1}(f(x^*) - f(x_i^s)) \le$$

$$\le - 2\rho_s\|g^s\|^2 L^{-1}\delta = - \sigma < 0 \ .$$

Substituting this inequality into the relation (22) we see

$$\|x^* - x_{i+1}^s\|^2 \le \|x^* - x_0^s\|^2 - \sigma\sum_{l=0}^{i}\lambda_{sl} + \rho_s^2\|g^s\|^4\sum_{l=0}^{i}\lambda_{sl}^2 =$$

$$= \|x^* - x_0^s\|^2 + \left[\sum_{l=0}^{i}\lambda_{sl}\right]\left[- \sigma + \rho_s^2\|g^s\|^4\left[\sum_{l=0}^{i}\lambda_{sl}\right]^{-1}\sum_{l=0}^{i}\lambda_{sl}^2\right] \ . \qquad (23)$$

Applying (20) and the Tëplitz lemma we have

$$\left[\sum_{l=0}^{i}\lambda_{sl}\right]^{-1}\sum_{s=0}^{i}\lambda_{sl}^2 \to 0 \quad \text{for} \quad i \to \infty \ .$$

Using (20) we obtain

$$\left[\sum_{l=0}^{i}\lambda_{sl}\right]\left[- \sigma + \rho_s^2\|g^s\|^4\left[\sum_{l=0}^{i}\lambda_{sl}\right]^{-1}\sum_{s=0}^{i}\lambda_{sl}^2\right] \to - \infty \quad \text{for} \quad i \to \infty \ .$$

and this contradicts (23)

It follows from (21) that $f(x^{s+1}) \le f(x^s) - \epsilon$ if $f(x^s) - f^* - \epsilon > 0$. Consequently there exists a number $\bar{s}$ such that $f(x^{\bar{s}}) \le \epsilon + f^*$. Since ϵ was arbitrary, then (21) implies that there exists subsequence i_k for which

$$f(x_{i_k}^{\bar{s}}) \to f^* \quad \text{for} \quad k \to \infty \ . \qquad (24)$$

Let us prove that the convergence of this subsequence leads to the convergence of sequence. Take some $\beta > 0$, then if

$$f(x_i^{\bar{s}}) - f^* \ge \beta \quad \text{then} \quad 2\rho_{\bar{s}}\|g^{\bar{s}}\|^2 < \xi_i^{\bar{s}},\ x^* - x_i^{\bar{s}} > \le - q < 0 \ .$$

It follows from (20) that there exists such number $\hat{I}$ that for all $i > \hat{I}$ inequality $\lambda_i^{\bar{s}} \le q\rho_{\bar{s}}^{-2}\|g^{\bar{s}}\|^{-4}$ holds. Using the inequality (22) we have

$$\|x^* - x_{i+1}^{\bar{s}}\|^2 = \|x^* - x_i^{\bar{s}}\|^2 + \lambda_{\bar{s}i}(2\rho_{\bar{s}}\|g^{\bar{s}}\|^2 < \xi_i^{\bar{s}},\ x^* - x_i^{\bar{s}} > +$$

$$+ \lambda_{\bar{s}i}\rho_{\bar{s}}^2\|g^{\bar{s}}\|^4) \le \|x^* - x_i^{\bar{s}}\|^2 + \lambda_{\bar{s}i}(-q + q) \le \|x^* - x_i^{\bar{s}}\|^2 \qquad (25)$$

for i such that $i > \hat{I}$ and $f(x_i^{\bar{s}}) - f^* \ge \beta$. Denote by

$$U(\mu) = \{x : d(x, X^*) \le \mu\},\ Q_\beta = \{x : f(x) < f^* + \beta\}$$

(see (18)). Let $\mu(\beta)$ be a minimal number such that $Q_\beta \subset U(\mu(\beta))$. Since the function $f: R^n \to R$ is convex and the set X^* is compact then $\mu(\beta) \to 0$ for $\beta \to 0$. Applying (25) we see that if $x_i^{\bar{s}} \notin U(\mu(\beta))$ then

$$\|x^* - x_{i+1}^{\bar{s}}\| < \|x^* - x_i^{\bar{s}}\|$$

for $i > \hat{I}$. Using points 2 and 4 of Step II of the algorithm, we obtain

$$\|x_{i+1}^{\bar{s}} - x_i^{\bar{s}}\| = \|x^{\bar{s}} - \rho_{\bar{s}}H_{i+1}^{\bar{s}}g^{\bar{s}} - x_i^{\bar{s}}\| =$$

$$= \|x_i^{\bar{s}} - \rho_{\bar{s}}\lambda_{\bar{s}i}\xi_i^{\bar{s}}\|g^{\bar{s}}\|^2 - x_i^{\bar{s}}\| \le \lambda_{\bar{s}i}\rho_{\bar{s}}\|g^{\bar{s}}\|^2 \ .$$

The relation (24) implies that beginning with some $\bar{K}$ for $k > \bar{K}$ the inclusion $x_{i_k}^{\bar{s}} \in U_{\mu(\beta)}$ holds. Taking into account the two previous inequalities we get

$$x_i^{\bar{s}} \in U(\mu(\beta) + \rho_{\bar{s}}\|g^{\bar{s}}\|^2 \max_{i_{k+1} > l \ge i_k} \lambda_{\bar{s}l}) \qquad (26)$$

for $i > \max\{\hat{I}, i_{\bar{k}}\}$, $i_k \le i < i_{k+1}$. It follows from the assumption (20) of the theorem that

$$\max_{i_{k+1} > l \ge i_k} \lambda_{\bar{s}l} \to 0 \quad \text{for} \quad k \to \infty \ ,$$

therefore for sufficiently large numbers k, the inequality

$$\rho_{\bar{s}}\|g^{\bar{s}}\|^2 \max_{i_{k+1} > l \ge i_k} \lambda_{\bar{s}l} \le \mu(\beta)$$

holds. Substituting this estimate into (26) we have

$$x_i^{\bar{s}} \in U(2\mu(\beta))$$

for sufficiently large i. Since β can be arbitrarily small and $\mu(\beta) \to 0$ for $\beta \to 0$, then $d(x_i^{\bar{s}}, X^*) \to 0$ for $i \to \infty$. The function $f: R^n \to R$ is convex, consequently it is continuous on R^n. For this reason the convergence $d(x_i^{\bar{s}}, X^*) \to 0$ implies $f(x_i^{\bar{s}}) \to f^*$. The theorem is proved.

□

Algorithm 1 has a substantial deficiency connected with the fact that the step size ρ_s does not change in the internal iterations $i = 1, \ldots, i(s)$. Let us consider an algorithm with a steepest descent control of ρ_s at each iteration $i = 1, \ldots, i(s)$. Such a modification considerably improves the algorithm.

Let $\nu > 0$ be a given number and $\{\lambda_j\}_0^\infty$ be a sequence of positive numbers.

ALGORITHM 2

STEP I Initialization

$$s = 0,\ H_0^{-1} = I,\ j = -1,\ i = -1,\ x^0 = x_{init},\ g^0 \in \partial f(x^0)\ .$$

STEP II

1 $H_0^s = H_{i+1}^{s\,-1},\ i = 0;$

2 $\rho_{si} = \underset{\rho > 0}{\operatorname{argmin}} f(x^s - \rho H_i^s g^s);$

3 $j = j + 1,\ j(s, i) = j;$

4 $x_i^s = x^s - \rho_{si} H_i^s g^s;$

5 compute $g_i^s \in \partial f(x_i^s)$ such that $< g_i^s,\ H_i^s g^s > \leq 0;$

6 $H_{i+1}^s = H_i^s + \lambda_j \xi_i^s g^{s\,T};$

7 if $\|x_i^s - x^s\| \geq \nu$ then $i(s) = i$ and go to Step III;

8 $i = i + 1$, return to the point 2.

STEP III $x^{s+1} = x_i^s,\ g^{s+1} = g_i^s.$

STEP IV $s = s + 1$, return to Step II.

Let us introduce some additional designations: let

$$T(x) = \{y \in R^n : f(y) \leq f(x)\}$$

and let L be a Lipschitz constant of function f on the set $T(x^0)$. Recall that the function f is called strictly convex on a set $T(x^0)$ if

$$\alpha_1 f(x) + \alpha_2 f(y) > f(\alpha_1 x + \alpha_2 y)$$

for all $\alpha_1,\ \alpha_2,\ x,\ y$ such that

$$\alpha_1 + \alpha_2 = 1,\ \alpha_1 > 0,\ \alpha_2 > 0,\ x \in T(x^0),\ y \in T(x^0)\ .$$

We next formulate a theorem about the convergence of algorithm 2.

THEOREM 3 *Let a function $f: R^n \to R$ be strictly convex (possibly nonsmooth) on a set $T(x^0)$, let a number $\nu > 0$ be given, and let the sequence $\{\lambda_j\}$, $j = 0, 1, \cdots$ of positive numbers satisfy the conditions*

$$\sum_{i=0}^{\infty} \lambda_j = \infty, \ \sum_{i=0}^{\infty} \lambda_j^2 > \infty; \ \lambda_j \to 0 \quad \text{for} \quad j \to \infty \ .$$

Then there exists a number $\bar{s}$ such that

$$f(x^{\bar{s}}) - f^* \leq 2\nu L \ .$$

PROOF Let us prove first that the norm of the matrix H_i^s is uniformly bounded for all $s > 0$, $i > 0$.

LEMMA 2 *The inequalities*

$$\|H_i^s\|^2 \leq Tr(H_i^s H_i^{s\,T}) \leq n + L^2 \sum_{l=0}^{\infty} \lambda_l^2 \tag{27}$$

obtain.

PROOF The inequality $\|H_i^s\| \leq Tr(H_i^s H_i^{s\,T})$ follows from the definition of norm and trace of matrix. Point 6 of Step II of the algorithm implies

$$H_{i+1}^s H_{i+1}^{s\,T} = H_i^s H_i^{s\,T} + \lambda_{j(s,i)}(\xi_i^s g^{s\,T} H_i^{s\,T} + H_i^s g^s \xi_i^{s\,T}) +$$

$$+ \lambda_{j(s,i)}^2 \xi_i^s \xi_i^{s\,T} \|g^s\|^2 \ .$$

Using this equality and taking into account that due to the construction of algorithm $<\xi_i^s, H_i^s g^s> \leq 0$, $\|\xi_i^s\| \leq 1$, we have

$$Tr(H_{i+1}^s H_{i+1}^{s\,T}) = Tr(H_i^s H_i^{s\,T}) + 2\lambda_{j(s,i)} <\xi_i^s, H_i^s g^s> + \lambda_{j(s,i)}^2 \|\xi_i^s\|^2 \|g^s\|^2 \leq$$

$$\leq Tr(H_i^s H_i^{s\,T}) + \lambda_{j(s,i)}^2 \|g^s\|^2 \leq Tr(H_0^s H_0^{s\,T}) +$$

$$+ \|g^s\|^2 \sum_{l=j(s,0)}^{j(s,i)} \lambda_l^2 \leq Tr(H_{i(s-1)+1}^{s-1} H_{i(s-1)+1}^{s-1\,T}) +$$

$$+ L^2 \sum_{l=j(s,0)}^{j(s,i)} \lambda_l^2 \leq Tr(H_0^{-1} H_0^{-1\,T}) + L^2 \sum_{l=0}^{j(s,i)} \lambda_l^2 =$$

$$= n + L^2 \sum_{l=0}^{j(s,i)} \lambda_l^2 \leq n + L^2 \sum_{l=0}^{\infty} \lambda_l^2 \ .$$

The lemma is proved. □

LEMMA 3 *There exists a number $\bar{s}$ such that $\|x_i^{\bar{s}} - x^{\bar{s}}\| < \nu$ for all $i \geq 0$.*

PROOF The statement of this lemma follows from the following lemma [10].

LEMMA 4 *Suppose the function $f(x)$ is strictly convex on R^n, the set $T(x^0)$ is bounded, and there is a sequence $\{x^s\}_{s=0}^{\infty}$ such that $\{x^s\}_{s=0}^{\infty} \subset T(x^0)$ and*

$$f(x^{s+1}) = \min_{\alpha \in [0,\,1]} f(x^s + \alpha(x^{s+1} - x^s)) \ . \tag{28}$$

Then $\lim\limits_{s \to \infty} \|x^{s+1} - x^s\| = 0$.

Note that the boundedness of the set $T(x^0)$ follows from the strict convexity of the function $f(x)$ on R^n. If the statement of the lemma 3 does not hold then points 2 and 4 of Step II imply that the sequence $\{x^s\}$ satisfies condition (28) and that beginning with some s^* the inequality

$$\|x^{s+1} - x^s\| < \nu \tag{29}$$

holds. This contradicts point 7 of Step II, since

$$\|x^{s+1} - x^s\| = \|x^s_{i(s)} - x^s\| \geq \nu$$

due to the construction of the algorithm. This contradiction proves Lemma 3. □

To finish the proof we need one more lemma. Recall that the set

$$\partial_\epsilon f(z) = \{g \in R^n : f(x) - f(z) \geq <g,\, x - z> - \epsilon \quad \forall x \in R^n\}$$

is called the ϵ-subdifferential of the convex function $f(x)$ at the point $z \in R^n$ [7]. A vector $g \in \partial_\epsilon f(z)$ is an ϵ-subgradient of the function f at a point z.

LEMMA 5 *Suppose $g \in \partial f(z)$, $z \in T(x^0)$, $y \in T(x^0)$. If $\|z - y\| \leq \nu$ then $g \in \partial_{2\nu L} f(y)$.*

PROOF Using the definition of the subdifferential we have

$$f(x) \geq <g,\, x - z> + f(z) = <g,\, x - y> + f(y) + <g,\, y - z> + f(z) - f(y) \geq$$

$$\geq <g,\, x - y> + f(y) - \|g\|\,\|z - y\| - |f(z) - f(y)| \ . \tag{30}$$

Since the function $f(x)$ satisfies the Lipschitz condition with a constant L then

$$\|g\| \leq L,\ |f(z) - f(y)| \leq L\|z - y\| \ .$$

Consequently

$$\|g\| \|z - y\| + |f(z) - f(y)| \leq 2L\|z - y\| < 2L\nu \ .$$

It follows from (30) and last inequality that

$$f(x) \geq <g, x - y> + f(y) - 2L\nu \ . \qquad \square$$

Let us prove that the number $\bar{s}$ from Lemma 3 satisfies the statement of the theorem. Suppose it does not hold, i.e.

$$f(x^{\bar{s}}) - f^* > 2\nu L \ . \tag{31}$$

Since $\|x_i^{\bar{s}} - x^{\bar{s}}\| < \nu$, then according to Lemma 5

$$g_i^{\bar{s}} \in \partial_{2\nu L} f(x^{\bar{s}}) \ .$$

It follows from (31) that $0 \notin \partial_{2\nu L} f(x^{\bar{s}})$ (see, for example, Lemma 8.1 [10]). The set $\partial_{2\nu L} f(x^{\bar{s}})$ is convex, closed, bounded [10]. Denote by q some vector satisfying

$$q \in \partial_{2\nu L} f(x^{\bar{s}}), \ \|q\| = \min_{g \in \partial_{2\nu L} f(x^{\bar{s}})} \|g\| > 0 \ .$$

By definition of the matrix norm

$$\|H_{i+1}^{\bar{s}}\| = \max_{\|u\| \leq 1} \max_{\|v\| \leq 1} <u, H_{i+1}^{\bar{s}} v> \geq$$

$$\geq \|q\|^{-2} < q, H_{i+1}^{\bar{s}} q> \ . \tag{32}$$

By construction of the algorithm

$$H_{i+1}^{\bar{s}} = H_i^{\bar{s}} + \lambda_{j(s,i)} \xi_i^{\bar{s}} g^{\bar{s}T} = H_0^{\bar{s}} + \sum_{l=0}^{i} \lambda_{j(s,l)} \xi_l^{\bar{s}} g^{\bar{s}T} \ .$$

Inequality (32) implies

$$\|H_{i+1}^{\bar{s}}\| \geq <q, \left[H_0^{\bar{s}} + \sum_{l=0}^{i} \lambda_{j(s,l)} \xi_l^{\bar{s}} g^{sT}\right] q> = <q, H_0^{\bar{s}} q> +$$

$$+ \sum_{l=0}^{i} \lambda_{j(s,l)} <q, \xi_l^{\bar{s}}> <g^{\bar{s}}, q> \ . \tag{33}$$

Since $g_i^{\bar{s}} \in \partial_{2\nu L} f(x^{\bar{s}})$, $i \geq 0$; $g^{\bar{s}} \in \partial_{2\nu L} f(x^{\bar{s}})$ and the set $\partial_{2\nu L} f(x^{\bar{s}})$ is convex then there exists a positive number α such that

$$<q, \xi_l^{\bar{s}}> \geq \alpha, <q, g^{\bar{s}}> \geq \alpha \ .$$

Applying (33) we see

$$\|H_{i+1}^{\bar{s}}\| > <q, H_0^{\bar{s}} q> + \alpha^2 \sum_{l=0}^{i} \lambda_{j(s,l)} \ .$$

By the conditions of the theorem $\sum_{l=0}^{i} \lambda_{j(s,l)} \to \infty$ for $i \to \infty$, consequently

$$\|H_{i+1}^{\bar{s}}\| \to +\infty \quad \text{for} \quad i \to \infty \ .$$

This last statement contradicts Lemma 2 and the proof of the theorem is complete. □

5. ALGORITHM WITH SIMMETRIC MATRICES

The algorithms discussed above have the following deficiency: one must store an $n \times n$ matrix where n is the dimension of the source problem. We next propose an algorithm with simmetrix matrices, to store such matrices requires only $(n^2 + n)/2$ numbers. The function $\varphi_s(H) = f(x^s - \rho_s H g^s)$ characterizes the choice of a matrix H. Denote by G the set of symetric $n \times n$ matrices. The set G is a linear space. For th adaptation of the matrix we can consider the following problem

$$\varphi_s(H) \to \min_{H \in G} \ .$$

Analogously to (4) one can use the gradient algorithm with projection onto the set G

$$H_{i+1}^s = \prod_G (H_i^s + \lambda_i^s g_i^s g^{s\,T}), \ \lambda_i^s > 0 \ ,$$

where $g_i^s \in \partial f(x^s - \rho_s H_i^s g^s)$ and $\prod_G$ is the projection operation onto the set G.

LEMMA 6 *If $H_i^s \in G$, then*

$$\prod_G (H_i^s + \lambda_i^s g_i^s g^{s\,T}) = H_i^s + 2^{-1} \lambda_i^s (g_i^s g^{s\,T} + g^s g_i^{s\,T}) \ .$$

PROOF It is evident that the matrix $H_i^s + 2^{-1}\lambda_i^s(g_i^s g^{s\,T} + g^s g_i^{s\,T})$ is simmetric if $H_i^s \in G$. To prove the Lemma it is enough to show that the matrix

$$[H_i^s + 2^{-1}\lambda_i^s(g_i^s g^{s\,T} + g^s g_i^{s\,T})] - [H_i^s + \lambda_i^s g_i^s g^{s\,T}] =$$

$$= 2^{-1}\lambda_i^s(g^s g_i^{s\,T} - g_i^s g^{s\,T})$$

is ortogonal to any symmetric matrix H. But:

$$<H,\ g^s g_i^{s\,T} - g_i^s g^{s\,T}> = <H,\ g^s g_i^{s\,T}> - <H,\ g_i^s g^{s\,T}> =$$

$$= <H,\ g^s g_i^{s\,T}> - <H^T,\ g^s g_i^{s\,T}> = 0\ . \qquad \square$$

Thus to update the matrix H_0^s one can use the algorithm

$$H_{i+1}^s = H_i^s + \lambda_i^s (g_i^s g^{s\,T} + g^s g_i^{s\,T})\ . \tag{34}$$

It is convenient to normalize the vector g_i^s, therefore we rewrite the formula (34) as

$$H_{i+1}^s = H_i^s + \lambda_i^s (\xi_i^s g^{s\,T} + g^s \xi_i^{s\,T})\ . \tag{35}$$

The symmetric formula for matrix modification can be combined with algorithms 1 and 2. Theorem 2 can be proved for this algorithm without any differences. For this reason we shall not dwell on convergence proofs for algorithms with the matrix modification formula (35).

6. ALGORITHM WITH POSITIVE MATRICES

Note that in the algorithms described above, the matrix H^s can be, generally speaking, non positive. If a function $f(x)$ is convex and $g^s \in \partial f(x^s)$ then the minimum point of the problem (1) belong to the subspace $A_s = \{x \in R^n : <x - x^s,\ g^s> \leq 0\}$. It is possible that the point $x^{s+1} = x^s - \rho_s H g^s$ does not belong to the subspace A_s if H^s is not positive. To guarantee positiveness of the matrix H^s let us consider the case when matrix H^s can be represented as follows

$$H^s = B^s B^{s\,T}\ ,$$

where B^s is a $n \times n$ matrix. In this case the iteration of the algorithm is given by the formula

$$x^{s+1} = x^s - \rho_s B^s B^{s\,T} g^s\ ,$$

where $g^s \in \partial f(x^s)$. The function $\varphi_s(B) = f(x^s - \rho_s B B^T g^s)$ defines the choice of a matrix B. If the function $f(x)$ is convex then it can be proved that the function $\varphi_s(B)$ is weakly convex. Next we will study the family of weakly convex functions which were investigated in the paper [11], (however other analogous families of functions can be used; see, for example, paper [12]).

Let X be a convex subset of R^n (possibly $X = R^n$). A continuous function f on X is called weakly convex on the set X if for all $x \in X$ the set $\partial f(x)$ consisting of the vectors g such that

$$f(y) - f(x) \geq <g, y - x> + \varsigma(x, y) \quad \text{for all} \quad y \in X$$

is not empty, where $\varsigma(x, y)$ is uniformly small with respect to $\|x - y\|$ on each compact subset $K \subset X$, i.e. for each $\epsilon > 0$ there exists $\delta > 0$ such that

$$\|\varsigma(x, y)\|/\|x - y\| < \epsilon$$

for $x, y \in K$, $\|x - y\| < \delta$.

LEMMA 7 *Let the function $f: R^n \to R$ be convex on R^n, and the set $\mathcal{B} \in R^{n \times n}$ be convex. Then the function $\varphi(B) = f(x - pBB^T\xi)$ is weakly convex on $\mathcal{B}$ and*

$$\partial\varphi(B) = \{ - \rho(\eta\xi^T + \xi\eta^T)B : \eta \in \partial f(x - \rho BB^T\xi)\} \ .$$

PROOF Let K be a compact subset of $\mathcal{B} \subset R^{n \times n}$ and $B \in K$, $\Delta B \in \mathcal{B}$, $B + \Delta B \in K$, and $\eta \in \partial f(x - \rho BB^T\xi)$. Denote by

$$X_{\mathcal{B}} = \{y \in R^n : y = x - \rho BB^T\xi, \ B \in \mathcal{B}\} \ .$$

The function $f: R^n \to R$ is Lipschitz with some constant $L_{\mathcal{B}}$ on $X_{\mathcal{B}}$, because the function f is convex on R^n and the set $X_{\mathcal{B}}$ is compact. Using the Lipschitz and convexity properties of $f(x)$ we have

$$\begin{aligned}
\varphi(B + \Delta B) &= f(x - \rho(B + \Delta B)(B^T + \Delta B^T)\xi) = \\
&= f(x - \rho(BB^T + B\Delta B^T + \Delta BB^T + \Delta B^T\Delta B^T)\xi) \geq \\
&\geq f(x - \rho(BB^T + B\Delta B^T + \Delta BB^T)\xi) - L_{\mathcal{B}}\|\rho\Delta B\Delta B^T\xi\| \geq \\
&\geq f(x - \rho BB^T\xi) + <-\rho(B\Delta B^T + \Delta BB^T)\xi, \eta> - L_{\mathcal{B}}\|\rho\Delta B\Delta B^T\xi\| = \\
&= \varphi(B) - \rho<(\xi\eta^T + \eta\xi^T)B, \Delta B> - L_{\mathcal{B}}\|\rho\Delta B\Delta B^T\xi\| \ .
\end{aligned}$$

Since the value $\|\Delta B\Delta B^T g\|$ is uniformly small with respect to $\|\Delta B\|$, the lemma is proved. □

Lemma 7 gives a formula for the subdifferential of the function $\varphi_s(B)$. For the adaptation of matrix B^s, the following gradient method can be used

$$B^s_{i+1} = B^s_i + \gamma_{si}(\xi^s_i g^{s\,T} + g^s \xi^{s\,T}_i)B^s_i, \ i = 0, 1, \cdots \ , \tag{36}$$

where ξ_i^s denotes the normalized vector $g_i^s \in \partial f(x_i^s)$ (see (19)). Analogously to algorithm 2, we write an algorithm with the matrix modification formula (36).

ALGORITHM 3

STEP I Initialization

$$s = 0,\ B_0^{-1} = I,\ i = -1,\ j = -1,\ x^0 = x_{init},\ g^0 \in \partial f(x^0)\ .$$

STEP II

1 $B_0^s = \begin{cases} I, \text{ if } \|B_{i+1}^{s\,-1} g^s\| = 0\ , \\ B_{i+1}^{s\,-1}, \text{ otherwise}\ ; \end{cases}$

2 $i = 0;$

3 $\rho_{si} = \underset{\rho > 0}{\operatorname{argmin}} f(x^s - \rho B_i^s B_i^{s\,T} g^s);$

4 $j = j + 1,\ j(s, i) = j;$

5 $x_i^s = x^s - \rho_{si} B_i^s B_i^{s\,T} g^s;$

6 compute $g_i^s \in \partial f(x_i^s)$ such that $< g_i^s,\ B_i^s B_i^{s\,T} g^s > \leq 0;$

7 $B_{i+1}^s = B_i^s + \lambda_j (\xi_i^s g^{s\,T} + g^s \xi_i^{s\,T}) B_i^s;$

8 if $\|x_i^s - x^s\| \geq \nu$, then $i(s) = i$ and go to step III.

9 $i = i + 1$ and return to the point 3 of Step II.

STEP III $x^{s+1} = x_i^s,\ g^{s+1} = g_i^s.$

STEP IV $s = s + 1$ and return to Step II.

We formulate a theorem about convergece of algorithm 3 for smooth objective functions.

THEOREM 4 *Let the function $f: R^n \to R$ be strictly convex and smooth, L_1 be a Lipschitz constant of the function f on the set $T(x^0)$, and L_2 be a Lipschitz constant for gradient $\nabla f(x)$ on the set*

$$T_\nu(x^0) \overset{\text{def}}{=} \{x: \min_{y \in T(x^0)} \|x - y\| \leq \nu\}\ .$$

Let there be given a value $\nu > 0$ and a sequence of positive numbers $\{\lambda_j\}_0^\infty$ satisfying

$$\sum_{j=0}^{\infty} \lambda_j = \infty,\ \sum_{j=0}^{\infty} \lambda_j^2 < \infty,\ \lambda_j > 0 \quad \text{for } j \geq 0\ .$$

Then for algorithm 8 there exists number $\bar{s}$ such that

$$\|g^{\bar{s}}\| \leq 2\nu L_2 \ .$$

PROOF To begin with we evaluate the norm of the matrix $\|B^s\|$.

LEMMA 8 *The inequality*

$$\|B_i^s\|^2 \leq Tr(B_i^s B_i^{s\,T}) \leq n\prod_{l=0}^{\infty}(1+4L_1^2\lambda_l^2) < \infty \tag{37}$$

holds for all integer $s > 0$, $i \geq 0$.

PROOF It follows from point 7 of Step II that

$$\begin{aligned} B_{i+1}^s B_{i+1}^{s\,T} = B_i^s B_i^{s\,T} + \lambda_{j(s,i)}[(\xi_i^s g^{s\,T} + g^s \xi_i^{s\,T}) B_i^s B_i^{s\,T} + \\ + B_i^s B_i^{s\,T}(\xi_i^s g^{s\,T} + g^s \xi_i^{s\,T})] + \\ + \lambda_{j(s,i)}^2 (\xi_i^s g^{s\,T} + g^s \xi_i^{s\,T}) B_i^s B_i^{s\,T} (\xi_i^s g^{s\,T} + g^s \xi_i^{s\,T}) \ . \end{aligned} \tag{38}$$

We denote by $k(s)$ the maximal number of iterations such that $k(s) \leq s$ and $B_0^{k(s)} = I$. Using (38) and taking into account that due to the construction of the algorithm $<\xi_i^s,\ B_i^s B_i^{s\,T} g^s> = 0$ and $\|\xi_i^s\| \leq 1$ we obtain

$$\begin{aligned} Tr(B_{i+1}^s B_{i+1}^{s\,T}) = Tr(B_i^s B_i^{s\,T}) + \lambda_{j(s,i)} 4 <\xi_i^s,\ B_i^s B_i^{s\,T} g^s> + \\ + \lambda_{j(s,i)}^2 [<\xi_i^s,\ g^s><\xi_i^s,\ B_i^s B_i^{s\,T} g^s> + \|\xi_i^s\|^2 <g^s,\ B_i^s B_i^{s\,T} g^s> + \\ + \|g^s\|^2 <\xi_i^s,\ B_i^s B_i^{s\,T} \xi_i^s> + <g^s,\ \xi_i^s><\xi_i^s,\ B_i^s B_i^{s\,T} g^s>] \leq \\ \leq Tr(B_i^s B_i^{s\,T}) + \lambda_{j(s,i)}^2 [\|\xi_i^s\|^2 \|g^s\|^2 \|B_i^s B_i^{s\,T}\| + \|\xi_i^s\|^2 \|g^s\|^2 \|B_i^s B_i^{s\,T}\| + \\ + \|g^s\|^2 \|\xi_i^s\|^2 \|B_i^s B_i^{s\,T}\| + \|g^s\|^2 \|\xi_i^s\|^2 \|B_i^s B_i^{s\,T}\|] \leq Tr(B_i^s B_i^{s\,T}) + \\ + 4\lambda_{j(s,i)}^2 \|g^s\|^2 \|B_i^s B_i^{s\,T}\| \leq Tr(B_i^s B_i^{s\,T}) + 4\lambda_{j(s,i)}^2 L_1^2 Tr(B_i^s B_i^{s\,T}) = \\ = Tr(B_i^s B_i^{s\,T})(1 + 4L_1^2 \lambda_{j(s,i)}^2) = Tr(B_0^s B_0^{s\,T}) \prod_{l=0}^{i}(1 + 4L_1^2 \lambda_{j(s,l)}^2) = \\ = Tr(B_{i(s-1)+1}^{s-1} B_{i(s-1)+1}^{s-1\,T}) \prod_{l=0}^{i}(1 + 4L_1^2 \lambda_{j(s,l)}^2) \leq \end{aligned}$$

$$\leq Tr(B_0^{k(s)}B_0^{k(s)T}) \prod_{t=j(k(s),0)}^{j(s,i)} (1 + 4L_1^2\lambda_t^2) \leq n \prod_{t=0}^{\infty} (1 + 4L_1^2\lambda_t^2) \ .$$

The inequality

$$\prod_{t=0}^{\infty} (1 + 4L_1^2\lambda_t^2) < \text{const}$$

follows from the convergence of the series $\sum_0^{\infty}\lambda_t^2$ in the conditions of the theorem. □

LEMMA 9 *There exists a number $\bar{s}$ such that $\|x_i^{\bar{s}} - x^{\bar{s}}\| < \nu$ for all $i \geq 0$.*

PROOF We prove the lemma by contradiction. Suppose the statement of lemma does not hold. By the construction of the algorithm the sequence $\{x^s\}$ satisfies the assumptions of Lemma 4. Consequently

$$\|x^{s+1} - x^s\| \to 0 \quad \text{for} \quad s \to \infty \ .$$

Applying point 8 of Step II of the algorithm we see that

$$\|x^{s+1} - x^s\| \geq \nu \ ,$$

and obtain a contradiction. □

Now let us prove, by contradiction, the statement of the theorem. We wish to show that for the number $\bar{s}$ from Lemma 9 the statement of the theorem holds. Suppose that it is not the case, i.e.

$$\|g^{\bar{s}}\| \geq 2\nu L_2 \ . \tag{39}$$

Since $\|x_i^{\bar{s}} - x^{\bar{s}}\| < \nu$, then $\|g_i^{\bar{s}} - g^{\bar{s}}\| \leq L_2\nu$ because the gradient of the function $f(x)$ satisfies a Lipschitz condition. Write the following inequalities

$$<\xi_i^{\bar{s}}, g^{\bar{s}}> = <\frac{g_i^{\bar{s}}}{\|g_i^{\bar{s}}\|}, g^{\bar{s}}> = <\frac{g_i^{\bar{s}} - g^{\bar{s}} + g^{\bar{s}}}{\|g_i^{\bar{s}}\|}, g^{\bar{s}}> \geq \frac{\|g^{\bar{s}}\|^2}{\|g_i^{\bar{s}}\|} -$$

$$- \frac{\|g_i^{\bar{s}} - g^{\bar{s}}\| \, \|g^{\bar{s}}\|}{\|g_i^{\bar{s}}\|} = (\|g^{\bar{s}}\| - \|g_i^{\bar{s}} - g^{\bar{s}}\|)\frac{\|g^{\bar{s}}\|}{\|g_i^{\bar{s}}\|} \geq$$

$$\geq (\|g^{\bar{s}}\| - L_2\nu)\|g^{\bar{s}}\| \, \|g_i^{\bar{s}}\|^{-1} \geq L_2\nu\|g^{\bar{s}}\|L_1^{-1} \geq 2L_2^2\nu^2L_1^{-1} \ . \tag{40}$$

We evaluate from below the value $\|B_{i+1}^{\bar{s}}B_{i+1}^{\bar{s}}\|$. Using relations $<\xi_i^{\bar{s}}, B_i^{\bar{s}}B_i^{\bar{s}T}g^{\bar{s}}> = 0$ and

(38) and (40) we get

$$\|g^{\bar{s}}\|^2 \|B_{i+1}^{\bar{s}} B_{i+1}^{\bar{s}T}\| \geq \langle g^{\bar{s}}, B_{i+1}^{\bar{s}} B_{i+1}^{\bar{s}T} g^{\bar{s}} \rangle = \|B_{i+1}^{\bar{s}T} g^{\bar{s}}\|^2 =$$

$$= \langle g^{\bar{s}}, B_i^{\bar{s}} B_i^{\bar{s}T} g^{\bar{s}} \rangle + \lambda_{j(\bar{s},i)} \langle g^{\bar{s}} [(\xi_i^{\bar{s}} g^{\bar{s}T} + g^{\bar{s}} \xi_i^{\bar{s}T}) B_i^{\bar{s}} B_i^{\bar{s}T} +$$

$$+ B_i^{\bar{s}} B_i^{\bar{s}T} (\xi_i^{\bar{s}} g^{\bar{s}T} + g^{\bar{s}} \xi_i^{\bar{s}T})] g^{\bar{s}} \rangle +$$

$$+ \lambda_{j(\bar{s},i)}^2 \langle g^{\bar{s}}, (\xi_i^{\bar{s}} g^{\bar{s}T} + g^{\bar{s}} \xi_i^{\bar{s}T}) B_i^{\bar{s}} B_i^{\bar{s}T} (\xi_i^{\bar{s}} g^{\bar{s}T} + g^{\bar{s}} \xi_i^{\bar{s}T}) g^{\bar{s}} \rangle \geq$$

$$\geq \|B_i^{\bar{s}T} g^{\bar{s}}\|^2 + 2\lambda_{j(\bar{s},i)} \langle g^{\bar{s}}, \xi_i^{\bar{s}} \rangle \langle g^{\bar{s}}, B_i^{\bar{s}} B_i^{\bar{s}T} g^{\bar{s}} \rangle =$$

$$\geq (1 + 2\lambda_{j(\bar{s},i)} \langle g^{\bar{s}}, \xi_i^{\bar{s}} \rangle) \|B_i^{\bar{s}T} g^{\bar{s}} \geq (1 + 4\lambda_{j(\bar{s},i)} L_2^2 \nu^2 L_1^{-1}) \|B_i^{\bar{s}T} g^{\bar{s}}\|^2 \geq$$

$$\geq \|B_0^{\bar{s}T} g^{\bar{s}}\|^2 \prod_{l=0}^{i} (1 + 4\lambda_{j(\bar{s},l)} L_2^2 \nu^2 L_1^{-1}) \ .$$

Since

$$\sum_0^\infty \lambda_j = \infty, \ \lambda_i > 0, \ i = 0, 1, \cdots$$

then

$$\prod_{l=0}^{i} (1 + 4\lambda_{j(\bar{s},l)} L_2^2 \nu^2 L_1^{-1}) \to +\infty \quad \text{for } i \to \infty \ .$$

Consequently $\|B_{i+1}^{\bar{s}} B_{i+1}^{\bar{s}T}\| \to +\infty$, and this contradicts Lemma 8. □

REFERENCES

[1] Dennis, J.N. and J.J. Moré: Quasi-Newton methods, motivation and theory. SIAM Review, 1977, 19, 46–89.

[2] Shor, N.Z.: Minimization Methods for Non-Differentiable Functions. Springer-Verlag, 1985.

[3] Uryas'ev, S.P.: A Step Size Rule for Direct Methods of Stochastic Programming. Kibernetika (Kiev), 1980, No. 6, 96–98.

[4] Ermoliev, Ju.M.: Stochastic Quasi-Gradient Methods and their Applications to Systems Optimization. Stochastics, 1983, No. 4.

[5] Uryas'ev, S.P.: Stochastic quasigradient algorithm with adaptively controlled parameters. Austria, Laxenburg, IIASA, 1986, WP-86-32, 27pp.

[6] Uryas'ev, S.P.: Adaptive variable metric algorithms for different optimization problems. Proceedings of IV simposium "Soluton methods for nonlinear equations and optimization problems", USSR, Viliandi, 1987.

[7] Rockafellar, R.T.: Convex Analysis, Princeton Mathematics, 1970, Vol. 28, Princeton Univ. Press.

[8] Pshenychnyi, B.N.: Necessary conditions for an extremum. Dekker, New York, 1971.

[9] Nesterov, Yu.E.: Minimization methods for nonsmooth convex and quasiconvex functions. Economika i mat. metodi, USSR, 1984, XX, 519–531.

[10] Dem'janov, V.F. and L.V. Vasil'ev: Nondifferentiable optimization. New York, Springer, 1985.

[11] Nurminskij, E.: Numerical Methods for Solving Deterministic and Stochastic Minimax Problems. Naukova Dumka, Kiev, 1979.

[12] Hoffman, A.: Weak convex functions, multifunctions and optimization. 27. IWK d.TH Ilmenau, 1982, Heft 5, 33–36.