ISSN 0005-1179, Automation and Remote Control, 2014, Vol. 75, No. 4, pp. 646–656. © Pleiades Publishing, Ltd., 2014.
Original Russian Text © E.A. Nurminskii, D. Tien, 2014, published in Avtomatika i Telemekhanika, 2014, No. 4, pp. 67–80.

## MATHEMATICAL PROGRAMMING PROBLEMS

# Method of Conjugate Subgradients with Constrained Memory

**E. A. Nurminskii**[*,**] **and D. Tien**[***]

[*] *Institute of Automation and Control Processes, Far-Eastern Branch, Russian Academy of Sciences, Vladivostok, Russia*
[**] *Far-Eastern Federal University, Vladivostok, Russia*
[***] *Charles Stuart University, Bathurst, Australia*
*e-mail: nurmi@dvo.ru, dtien@csu.edu.au*

Received November 14, 2013

**Abstract**—A method to solve the convex problems of nondifferentiable optimization relying on the basic philosophy of the method of conjugate gradients and coinciding with it in the case of quadratic functions was presented. Its basic distinction from the earlier counterparts lies in the a priori fixed constraint on the memory size which is independent of the accuracy of the resulting solution. Numerical experiments suggest practically linear rate of convergence of this algorithm.

**DOI:** 10.1134/S0005117914040055

## 1. INTRODUCTION

The present paper considers the convex problem of nondifferentiable optimization

$$\min_{x\in E} f(x) = f_\star, \tag{1}$$

where $E$ is a finite-dimensional Euclidean space with the ordinary scalar product $xy$ and the corresponding norm $\|x\| = \sqrt{xx}$. It is assumed that problem (1) is well-defined in the sense that its solution does exist. For such problems, the paper aims at suggesting a method of the conjugate gradient type with a priori fixed boundary of the memory used, proving its convergence, and presenting some encouraging results of the computer experiments.

The most general methods for solution of problem (1) make use of the so-called subgradient oracles enabling one to calculate at an arbitrary point the value of the objective function $f(x)$ and some subgradient $g$ from the subdifferential set $\partial f(x)$. The simplest of such methods is the subgradient algorithm

$$x^{k+1} = x^k - \lambda_k g^k, \quad g^k \in \partial f(x^k), \quad k = 0, 1, \ldots, \tag{2}$$

actively studied beginning from the pioneering works of Shor [1] and Polyak [2]. It was shown in the most general case that under very weak conditions (2) converges to solution (1) if the step multipliers $\lambda_k$ satisfy the conditions $\sum_k \lambda_k = \infty$, $\lambda_k \to +0$ for "series divergence." However, the numerical experiments and theoretical analysis demonstrated that this rule for selection of the step multipliers usually results in slow convergence. That is why search of new, more efficient algorithms such as, in particular, numerous variants of the quadratic-linear algorithms using the piecewise-linear models of the nonsmooth objective functions and quadratic corrections intentionally increasing the accuracy of these approximations [3–5] was started actually immediately. The concepts of variable metric [6] in practice proved to be very effective. Among the latest ideas in this

field, smoothing combined with the optimal gradient schemes of smooth optimization [7], splitting in the smoothness-nonsmoothness subspaces (the so-called $UV$-algorithms [8]) deserve mentioning.

At the same time, consideration was given beginning from [9, 10] to the analogs of the method of conjugate gradients [11–13]. These studies developed mostly along two lines of research. In the first line descending from the initial works of Ph. Wolfe [9, 10] the algorithms used the accumulated packet of subgradients obtained at some prehistory of the current iteration. This subgradient packet was used in the attempt to construct the direction of descent of the objective function as the solution of the problem

$$\|p^k\|^2 = \min_{\substack{p\in\mathrm{co}\{g^0,g^1,\ldots,g^k\},\\ g^i\in\partial f(x^i),\ i=0,1,\ldots,k}} \|p\|^2, \tag{3}$$

where $x^0, x^1, \ldots, x^k$ is the history of the preceding iterations (tentative and working steps). For simplicity of notation, we consider here the prehistory starting from the zero (initial) iteration.

Depending on various conditions, either the algorithm is restarted which can be regarded as the change of the initial point, or the information is further accumulated using the tentative steps, or a working step like

$$x^{k+1} = x^k - \lambda_k p^k \tag{4}$$

is performed, where $\lambda_k$ is determined by a precise or approximate one-dimensional optimization. The difficulty of this approach lies in the fact that the volume of the accumulated packet of subgradients is defined by the estimated accuracy of the solution obtained in the given cycle, and grows indefinitely with accuracy. Apart from the increased memory overhead and growth in the volume of the processed data, under a sufficiently rapid growth in the volume of calculations this leads to undesirable algorithmic consequences where for many subsequent iterations some bad tentative step may complicate the search of a good direction of descent.

The second line may be regarded as an approximate solution of problem (3) with the use of the Polak–Ribiere formula for construction of the conjugate directions; see, for example, [11, 12] where the constant step multipliers and constant weight coefficients were used. However, the computer experiments with these algorithms demonstrated that, for example, in the case of lack of the acute "minimum" the number of tentative steps grows progressively. Convergence of the algorithm was strongly retarded in terms of the wasted time. Nevertheless, for the smooth extremal problems these methods demonstrated quite good results and currently are actively studied under the name of the shortest residuals (SR) methods [14–16].

In the present paper, these approaches are in a sense combined for solution of the problem of convex nondifferentiable optimization. The proposed variant of the method of conjugate subgradients, on the one hand, retains the advantages of the Wolfe method lying in the synthetic use of the subgradient packet and, on the other hand, bounds the computational burden for processing this packet. The latter is attained through the a priori constraints on the size of packet upon reaching which the algorithm is restarted.

## 2. METHOD OF CONJUGATE CONSTRAINED-MEMORY SUBGRADIENTS

Let us consider the method of conjugate constrained-memory subgradients as a rule for constructing the sequence $x^k$, $k = 0, 1, \ldots$, converging under certain conditions to the solution of problem (1). This rule makes use of the packet of subgradients accumulated and modified as the algorithm works. In the general form, the packet $G(z, s, t)$ consists of the initializing vector $z$ and the subgradient $g^i \in \partial f(x^i)$, $i = s, s+1, \ldots, t$, calculated at the iterations $s, s+1, \ldots, t$:

$$G(z, s, t) = \{z, g^i \in \partial f(x^i), i = s, s+1, \ldots, t\}.$$

The initializing vector $z$ is used for information transmission at the restarts of the algorithm. To simplify notation, the convex hull of the finite set $G(z,s,t)$ is denoted by $G_{co}(z,s,t)$.

The packet $G(z,s,t)$ consists at most of $N+1$ vector, where $N$ is a fixed input parameter. The control sequence of $\delta_k \to +0$, $k=0,1,\ldots$, which is an equivalent of the accuracy of satisfying the optimality conditions is also defined in the algorithm.

The algorithm constructs sequences of iterations interrupted by the instants of restart during which the packet $G(\cdot,\cdot,\cdot)$ is cleared of the accumulated subgradients and the initializing vector is modified. The algorithm is restarted upon satisfaction of at least one condition: either the number of subgradients in the packet reaches the maximal value $N$ or the norm of the shortest vector in the convex hull of the packet $G_{co}(\cdot,\cdot,\cdot)$ drops below the current estimate of satisfaction of the optimality conditions. With the introduced notation, the method is as follows.

**Initialization.** Set up the initial values of the restart counter $r=0$ and iteration counter $t=0$, and determine the initial instant $t_r=0$ of restart. Define the maximal size $N$ of the subgradient packet, sequence of accuracy estimates $\{\delta_k\}$, and the initial point $x^0$. Calculate $g^0 \in \partial f(x^0)$, set up to $g^0$ the coordinating vector $z^0$, and assume that the initial packet $G_0 = G(z^0,0,0)$ is equal to $\{g^0,g^0\}$.

The current $t+1$st iteration is carried out as follows, provided that $t$ iterations were carried out during which $r$ restarts occurred, the last one at the instant $t_r < t$.

**t + 1st iteration.**

*Step 1.* Solve the problem of finding an element of the minimal Euclidean norm

$$\min_{p\in G_{co}(z^r,t_r,t)} \|p\|^2 = \|p^t\|^2 \tag{5}$$

and go to Step 2 if $\|p^t\| > \delta_r$.

If $\|p^t\| \leqslant \delta_r$, then restart the algorithm with increased requirements on the accuracy of satisfying the optimality conditions:

- increment the restart counter $r = r+1$ and update completely the packet of subgradients:

$$t_r = t, \quad z^r = g^t \in \partial f(x^t), \quad G(z^r,t,t) = \{g^t\}; \tag{6}$$

- repeat Step 1.

*Step 2.* Solve the one-dimensional problem

$$\min_{\lambda} f(x^t - \lambda p^t) = f(x^t - \lambda_t p^t) = f(x^{t+1}) \leqslant f(x^t) \tag{7}$$

and select $g^{t+1} \in \partial f(x^{t+1})$ such that $g^{t+1}p^t = 0$. The optimality condition for this problem guarantees existence of such $g^{t+1}$ even for $\lambda_t = 0$. At determination of the point $x^{t+1}$ by dichotomy as the limit of the embedded intervals $[x^t - \lambda_k^- p^t, x^t - \lambda_k^+ p^t]$, $k=0,1,\ldots$, with $g_-^k p^t = \alpha_k^- < 0$ and $g_+^k p^t = \alpha_k^+ > 0$, where $g_-^k \in \partial f(x^t - \lambda_k^- p^t)$ and $g_+^k \in \partial f(x^t - \lambda_k^+ p^t)$, such vector can be found as the limit of the sequence $\bar{g}^k = \gamma_k g_-^k + (1-\gamma_k) g_+^k$, $k=0,1,\ldots$, where $\gamma_k = \frac{\alpha_k^+}{\alpha_k^+ - \alpha_k^-} \in [0,1]$ and makes $\bar{g}^k p^t$ vanish. By the semicontinuity from above of the subdifferential map $\partial f$, the vector $g^{t+1} = \lim_{k\to\infty} \bar{g}^k \in \partial f(x^{t+1})$ where the limit possibly must be taken in the arbitrary converging subsequence.

*Step 3.* Complement the set $G(z^r,t_r,t)$ by the vector $g^{t+1}$:

$$G(z^r,t_r,t+1) = \{G(z^r,t_r,t), g^{t+1}\},$$

increase the iteration counter to $t \to t+1$ and go to Step 4.

*Step 4.* Upon reaching the limits on the memory size, this step restarts the algorithm without modifying the restart counter and the current estimate of solution accuracy. If $t - t_r \geqslant N$, then change the coordinating vector $z^r = p^t$, redefine the instant of the last restart $t_r = t$, initialize the packet of subgradients $G(z^r, t_r, t) = \{z^r, g^t\}$, $g^t \in \partial f(x^t)$, and go to Step 1.

## 3. CONVERGENCE OF THE ALGORITHM

Convergence of the algorithm is studied using the convergence conditions discussed in detail in [17]. From the point of view of these conditions, the algorithm to solve the optimization problem is a kind of rule for constructing the sequence of approximate solutions $\{x^k\}$ which should converge to some desired set $X_\star$ defined usually by the corresponding optimality conditions.

The weak form of convergence (convergence in subsequence) is ensured if the following conditions are met:

**A1.** Sequence $\{x^k\}$ is restricted.

**A2.** There exists a continuous function $W(x) : E \to \mathbb{R}$ such that if $\{x^k\}$ has the limit point $x' \notin X_\star$, then this sequence has another limit point $x''$ such that $W(x'') < W(x')$.

If these conditions are met, then the sequence $\{x^k\}$ has the limit point $x^\star \in X_\star$.

The strong form of convergence of $\{x^k\}$ to the set $X_\star$ exists under somewhat stronger conditions:

**B1.** The sequence $\{x^k\}$ is limited.

**B2.** For an arbitrary subsequence $\{x^{k_t}\} \to x' \notin X_\star$ and $t \to \infty$, there exists $\epsilon > 0$ such that for any $t$ there is an instant of leaving the neighborhood of $x'$:

$$m_t = \inf\left\{m : \|x^{k_t} - x^m\| > \epsilon\right\} < \infty. \tag{8}$$

**B3.** There exists a continuous function $W(x) : E \to \mathbb{R}$ such that

$$\limsup_{t\to\infty} W(x^{m_t}) < \lim_{t\to\infty} W(x^{k_t}) = W(x') \tag{9}$$

for all subsequences $\{x^{k_t}\}, \{x^{m_t}\}$ satisfying condition **B2**.

**B4.** The set $W_\star = \{W(x^\star), x^\star \in X_\star\}$ such that $\mathbb{R} \setminus W_\star$ is dense everywhere.

**B5.** If $\{x^{k_t}\} \to x^\star \in X_\star$, then $\|x^{k_t+1} - x^{k_t}\| \to 0$ for $t \to \infty$.

If these conditions are met, then all limit points of $\{x^k\}$ belong to $X_\star$ [17].

The following theorem can be proved using the above conditions for convergence.

**Theorem 1.** *Let $f$ be finite and strongly convex, and the Lebesgue set $\{x : f(x) \leqslant f(x^0)\}$ be bounded. Then, $\{x^t\}$ converges to the single solution of problem* (1).

**Proof.** To apply conditions **B1**–**B5**, we first of all define the set $X_\star$ as the point of $x^\star$ satisfying the necessary, and in the case at hand also sufficient, optimality conditions $0 \in \partial f(x^\star)$.

In virtue of monotonicity of the method, all elements of the sequence $\{x^t\}$ belong to the bounded set $\{x : f(x) \leqslant f(x^0)\}$, so that condition **B1** is satisfied trivially. Now we assume that condition **B2** is not met, the entire sequence $\{x^t\}$ converges to some point $x' \notin X_\star$ where correspondingly $0 \notin \partial f(x')$. By virtue of the upper semicontinuity of $\partial f$ and finiteness of $f$, there exists a sufficiently small $\epsilon > 0$ such that for some $0 < \gamma < \Gamma < \infty$ there exist estimates $\gamma \leqslant \|g\| \leqslant \Gamma$ for any $g \in \partial f(x)$, $\|x - x'\| \leqslant 4\epsilon$. We omit the dependence of $\gamma$ and $\Gamma$ on $x'$ and $\epsilon$ because in the reasoning below they are fixed.

We show that under the above assumption there arises an infinite sequence $k = 0, 1, \ldots$ of iterations such that $\|z^k\| \leqslant \delta_k \to 0$. Indeed, if this is not the case, then there exists $\bar{k}$ such that $\|p^t\| \geqslant \delta_{\bar{k}}$ for all $t > t_{\bar{k}}$. We can assume without loss of generality that $t$ is so great that $\|x^t - x'\| \leqslant \epsilon$.

The assumption of boundedness of $\bar{k}$ as a matter of fact means that, as soon as the maximal size of the vector packet $G(\cdot)$ is reached, all subsequent restarts occur according only to Step 4.

Then, for $t_{\bar{k}} < t \leqslant t_{\bar{k}} + N$

$$\|z^t\|^2 = \min_{z \in G_{\mathrm{co}}(z^{\bar{k}},\ t_{\bar{k}},t)} \|z\|^2 \leqslant \min_{z \in \mathrm{co}\{z^{\bar{k}}, g^{t_{\bar{k}}+1}\}} \|z\|^2 .$$

The concluding minimum can be easily estimated from above as follows:

$$\begin{aligned} &\min_{\lambda\in[0,1]} \left\|\lambda z^{\bar{k}} + (1-\lambda) g^{t_{\bar{k}}+1}\right\|^2 \\ &= \min_{\lambda\in[0,1]} \left\{\lambda^2 \|z^{\bar{k}}\|^2 + (1-\lambda)^2 \|g^{t_{\bar{k}}+1}\|^2\right\} = \lambda_\star \|z^{\bar{k}}\|^2, \end{aligned} \tag{10}$$

where

$$\lambda_\star = \frac{\|g^{t_{\bar{k}}+1}\|^2}{\|z^{\bar{k}}\|^2 + \|g^{t_{\bar{k}}+1}\|^2}$$

solves (10).

Taking into consideration that $\|z^{\bar{k}}\|$ is globally bounded by some constant $C \geqslant \|g\|$, $g \in \partial f(z)$, $f(z) \leqslant f(x^0)$ and $\|g^{t_{\bar{k}}+1}\| \geqslant \gamma$, for $\lambda_\star$, it is easy to establish the estimate

$$\lambda_\star \leqslant 1/(1+\gamma^2/C^2) = \theta < 1,$$

which means that $\|z^k\|^2$, $k = \bar{k}, \bar{k}+1, \ldots$, decreases at least at the rate of geometric progression with the denominator $\theta$, and, consequently, tends to 0, which contradicts the initial assumption.

This contradiction implies that

(a) either $\{x^t\} \to x^\star$ (and the theorem is proved);

(b) or $\{x^t\} \to x' \notin X^\star$, but at that $\|z^k\| \to 0$ for $k \to \infty$;

(c) or for any limit point $x' \notin X_\star$, the sequence $\{x^t\}$ leaves any its sufficiently small neighborhood an infinite number of times.

It follows from (c) that condition **B2** is satisfied at least in this case. To make sure once and for all that there is no doubt that condition **B2** is satisfied, it remains to demonstrate that case (b) is excluded.

For that we assume that $\epsilon > 0$ is so small that the set $\tilde{G} = \mathrm{co}\{\partial f(x), \|x - x'\| \leqslant 4\epsilon\}$ is strictly separable from zero, that is, there exist the vector $p$, $\|p\| = 1$, and $\delta > 0$ such that $pg \geqslant \delta$ for all $g \in \tilde{G}$. Since for sufficiently great $t$ the points $\|x^t - x'\| \leqslant \epsilon$, from the instant of some restart with full update (6) also $z^r \in \tilde{G}$, that is, $pz^r \geqslant \delta > 0$, which rules out $z^r \to 0$.

Now we demonstrate that condition **B3** is also satisfied, but first we ensure satisfaction of conditions **B4, B5** and define for that the convergence indicator $W(x) = \|x - x^\star\|^2$ which is traditional for the convex problems, where $x^\star$ is a single element of the set $X_\star$ in virtue of strong convexity. We notice that this mechanically entails satisfaction of **B4**.

It is easy to demonstrate that for $t \to \infty$ the sequence $\|x^{t+1} - x^t\| \to 0$, which is even stronger than **B5**. Indeed, the elements of the sequence $\{x^t\}$ are related by $x^{t+1} = x^t - \alpha_t p^t$, where the step multiplier $\alpha_t$ is selected so that there exists the subgradient $\bar{g}^{t+1} \in \partial f(x^{t+1})$ such that

$$\bar{g}^{t+1} p^t = 0 = g^{t+1}(x^{t+1} - x^t). \tag{11}$$

Since $f(x^{t+1}) \leqslant f(x^t)$, in virtue of boundedness $f(x^t) \to \bar{f}$ for $t \to \infty$.

Assuming that $x^t \to x'$, $x^{t+1} \to x''$, $\bar{g}^{t+1} \to \bar{g}$, we notice that $f(x') = f(x'') = \bar{f}$ and $\bar{g} \in \partial f(x'')$ in virtue of upper semicontinuity of $df(x)$. By passing in (11) to the limit and using the strong convexity with some constant of strong convexity $\sigma > 0$

$$f(x^t) - f(x^{t+1}) \geqslant \bar{g}^{t+1}(x^t - x^{t+1}) + \sigma\|x^t - x^{t+1}\|^2 = \sigma\|x^t - x^{t+1}\|^2 \geqslant 0$$

we obtain

$$0 = f(x') - f(x'') = \sigma\|x' - x''\|^2 \geqslant 0$$

or $\|x^{t+1} - x^t\| \to 0$ for $t \to \infty$, which proves **B5.**

To pass to condition **B3**, we denote by $\{m_k\}$ and $\{n_k\}$, $k = 0, 1, \ldots$, the index sequences such that $x^{n_k} \to x' \neq x^\star$ there exists $\epsilon > 0$ such that $\|x^{m_k} - x^{n_k}\| > \epsilon$ and $\|x^t - x^{n_k}\| \leqslant \epsilon$ for all $n_k \leqslant t < m_k$. To satisfy the above estimates, this $\epsilon$ may be regarded arbitrarily small.

The sequence $\{x^{n_k}\}$ by construction converges to $x'$, and $\{x^{m_k}\}$ is the sequence of the first exits from the $\epsilon$-neighborhoods of the corresponding points $x^{n_k}$.

Let $q_k < n_k$ be the maximal index not exceeding $n_k$ such that $G(\cdot, \cdot, \cdot)$ was updated and $p_k$ is the minimal index exceeding $n_k$ when $G(\cdot, \cdot, \cdot)$ was updated for the next time. Independently of the form of restart, $p_k - n_k \leqslant p_k - q_k \leqslant N$ and, consequently,

$$\|x^{p_k} - x^{n_k}\| \leqslant \sum_{t=n_k}^{p_k-1} \|x^{t+1} - x^t\| \leqslant N \sup_{t \geqslant n_k} \|x^{t+1} - x^t\| \to 0$$

for $k \to \infty$.

Therefore, $x^{p_k} \to x'$ and, consequently, $\lim_{k\to\infty} W(x^{p_k}) = \lim_{k\to\infty} W(x^{n_k}) = W(x')$. The proof of

$$\lim_{k\to\infty} W(x^{p_k}) > \limsup_{k\to\infty} W(x^{m_k}) \tag{12}$$

is equivalent to the proof of **B3.**

To prove (12), we notice that for all $t$ such that $p_k \leqslant t < m_k$ we have $p^t \in \mathrm{co}\{\partial f(x), \|x - x'\| \leqslant 4\epsilon\} = \tilde{G}$ for sufficiently small $\epsilon > 0$.

By virtue of the Carathéodori theorem, an arbitrary $p \in \tilde{G}$ is representable as $p = \sum_{i=1}^{n+1} \lambda_i g^i$, where $g^i \in \partial f(y^i)$, $\|y^i - x^\star\| \leqslant 4\epsilon$, $\lambda_i \geqslant 0$, $\sum_{i=1}^{n+1} \lambda_i = 1$. In virtue of convexity, $0 < \gamma < f(y^i) - f(x^\star) \leqslant g^i(y^i - x^\star)$ which by virtue of its continuity can be rearranged in

$$0 < \gamma/2 \leqslant g^i(x' - x^\star). \tag{13}$$

By multiplying (13) by $\lambda_i$ and summing up, we obtain $p(x' - x^\star) \geqslant \gamma/2$.

Now we consider $t$ such that $p_k \leqslant t < m_k$ and

$$\begin{aligned} W(x^{t+1}) - W(x^t) &= \|x^t - \alpha_t p^t - x^\star\|^2 - \|x^t - x^\star\|^2 \\ &= -\alpha_t p^t(x^t - x^\star) + \alpha_t^2\|p^t\|^2. \end{aligned}$$

Again, $p^t(x^t - x^\star) \geqslant \gamma/4$ in virtue of continuity, and, consequently,

$$W(x^{t+1}) - W(x^t) \leqslant -\alpha_t\gamma/2 + \alpha_t^2\|p^t\|^2. \tag{14}$$

Since $\alpha_t\|p^t\| \to 0$ but $\|p^t\| \geqslant \kappa > 0$ ($\tilde{G}$ can be strictly separated from 0 in virtue of convexity), $\alpha_t \to 0$ which implies that the last term in (14) can be disregarded. Therefore, for $p_k \leqslant t < m_k$

$$W(x^{t+1}) - W(x^t) \leqslant -\alpha_t\gamma/4,$$

which after summation provides

$$W(x^{m_k}) - W(x^{p_k}) \leqslant -\gamma \sum_{t=p_k}^{m_k-1} \alpha_t/4 < 0.$$

Since

$$\epsilon/2 < \|x^{m_k} - x^{p_k}\| \leqslant \sum_{t=p_k}^{m_k-1} \alpha_t \|p^t\| \leqslant C \sum_{t=p_k}^{m_k-1} \alpha_t,$$

$\sum_{t=p_k}^{m_k-1} \alpha_t \geqslant \epsilon/2C$ and

$$W(x^{m_k}) - W(x^{p_k}) \leqslant -\gamma\epsilon/8C < 0.$$

The passage to the limit for $k \to \infty$ proves **B3** and, consequently convergence of the algorithm. Interestingly, the strong convexity of the objective function in problem (1) played a technical role in the proof and may be removed by a slight complication of both the algorithm and the proof. The assumption of finite size $N$ of the packet and the full update mechanism (6) plays an essential role in the proof enabling one to "forget" the possibly unlucky prehistory of search. At the same time, the complete loss of information at full update (6) plays, possibly, the part of the bad guy forcing one to waste a certain number of calls to the subgradient oracle in order to accumulate a sufficiently representative packet $G$. Here, search of an acceptable compromise seems to be of interest for future research.

## 4. RELATION WITH THE METHOD OF CONJUGATE GRADIENTS

It is known from the theory of quasi-Newton algorithms that some variants of algorithms of the Broyden type represent corrections of the quasi-Newton matrices having the least matrix norm, the Frobenius norm, in particular. Interestingly, the principle of constructing the direction of search as a minimal Euclidean-norm element that was used in the proposed algorithm in the classical case of strongly convex quadratic objective functions also gives the traditional algorithm of conjugate gradients [19].

Indeed, let $\{g^1, \ldots, g^k\}$ be a collection of the gradients of the strongly convex quadratic objective function obtained at the $k$ precious iterations regarded as mutually orthogonal and obtained as the result of one-dimensional minimization along the corresponding conjugate directions $p^1, \ldots, p^k$. The corresponding problem $P_k$ of finding the search direction is given by

$$\min_{\sum_{i=1}^k \lambda_i=1,\ \lambda_i\geqslant 0,\ i=1,\ldots,k} \frac{1}{2}\left\| \sum_{i=1}^k \lambda_i g^i \right\|^2 \quad = \quad \min_{\sum_{i=1}^k \lambda_i=1,\ \lambda_i\geqslant 0,\ i=1,\ldots,k} \frac{1}{2} \sum_{i=1}^k \lambda_i^2 \|g^i\|^2 \quad . \tag{15}$$

If $g^i \geqslant 0$, $i = 1, \ldots, n$, are mutually orthogonal, then the solutions of problem $P_k$ and $P_{k+1}$ represent the conjugate vectors. To demonstrate this fact, we consider the solution of $P_{k+1}$ defined by the optimality conditions in (15):

$$\lambda_i \|g^i\|^2 + \theta = 0, \quad i = 1, 2, \ldots, k+1; \quad \theta \sum_{i=1}^{k+1} \|g^i\|^{-2} = -1. \tag{16}$$

We disregard the constraints on nonnegativeness of $\lambda_i$ because, as will be shown below, they are satisfied mechanically.

It follows from (16) that

$$\lambda_j = \|g^j\|^{-2} \left( \sum_{i=1}^{k+1} \|g^i\|^{-2} \right)^{-1}, \quad j = 1, \ldots, k+1 \geqslant 0$$

and, consequently, the nonnegativeness conditions are satisfied mechanically.

We denote

$$\sigma_{k+1} = \sum_{i=1}^{k+1} \|g^i\|^{-2} = \sigma_k + \|g^{k+1}\|^{-2}.$$

Then,

$$p^{k+1} = \sum_{i=1}^{k+1} \lambda_i g^i = \sum_{i=1}^{k} \lambda_i g^i + \lambda_{k+1} g^{k+1}$$

$$= \sum_{i=1}^{k} \|g^i\|^{-2}(\sigma_k + \|g^{k+1}\|^{-2})^{-1} g^i + \|g^{k+1}\|^{-2}(\sigma_k + \|g^{k+1}\|^{-2})^{-1} g^{k+1}$$

$$= (\sigma_k + \|g^{k+1}\|^{-2})^{-1} \left( g^{k+1} + \|g^{k+1}\|^{-2} \sum_{i=1}^{k} \|g^i\|^{-2} g^i \right)$$

$$= \theta_{k+1} \left( g^{k+1} + \|g^k\|^{-2}\|g^{k+1}\|^2 \left( \|g^k\|^2 \sum_{i=1}^{k} \|g^i\|^{-2} g^i \right) \right)$$

$$= \theta_k (g^{k+1} + \|g^k\|^{-2}\|g^{k+1}\|^2 z^k) = \theta_k (g^{k+1} + \mu_{k+1} p^k),$$

where $\mu_{k+1} = \|g^{k+1}\|^2/\|g^k\|^2$ and, consequently, $p^{k+1}$ was determined using the classical Polak-Ribiere formula to within the scaling multiplier $\theta_k$. Whence it follows that $p^{k+1}$ is conjugate to $p^1, \ldots, p^k$. Since the scaling multiplier is of no importance in the subsequent one-dimensional minimization along $p^{k+1}$, the gradient $g^{k+2}$ is orthogonal to all preceding gradients and the resulting sequence of iterations coincides with the method of conjugate gradients.

## 5. NUMERICAL EXPERIMENT

To demonstrate in practice the computational characteristics of the propose method, we consider the results of numerical experiments with the well-known piecewise-quadratic test function `maxqfg` given by

$$f(x) = \max_{1 \leqslant k \leqslant 5} \phi_k(x),$$

where $\phi_k(x) = xA_kx - b^kx$ and $A^{(k)}$, $k = 1, \ldots, 5$, are the symmetrical positive definite $10 \times 10$ matrices

$$A_{ij}^{(k)} = \begin{cases} \exp(\min(i,j)/\max(i,j))\cos(ij)\sin(k), & i \neq j \\ i|\sin(k)|/10 + \sum\limits_{l=1,\ldots,10, l\neq i} |A_{il}^{(k)}|, & i = j, \end{cases} \qquad i,j = 1, \ldots, 10,$$

$$b_i^k = \exp(i/k)\sin(ik), \quad i = 1, \ldots, 10, \quad k = 1, \ldots, 5.$$

This problem clearly demonstrate the difficulties of solving the piecewise-quadratic problems of convex nonsmooth optimization because it combines both the problems of nonsmooth nature (discontinuity of gradients) and smooth optimization (ravinity of the Lebesgue sets). The attempts to solve this problem by simple subgradient algorithms encounter serious difficulties because the acute minimum condition is not satisfied here and at the optimal point the subdifferential set has an empty interior, only three of the five functions $\phi_k(x)$ being active at the best. This means that in the subspace orthogonal to the linear hull of the subdifferential the function behaves quadratically and there exists degeneracy in terms of the relation of the "inscribed and circumscribed spheres" for the subdifferential sets.

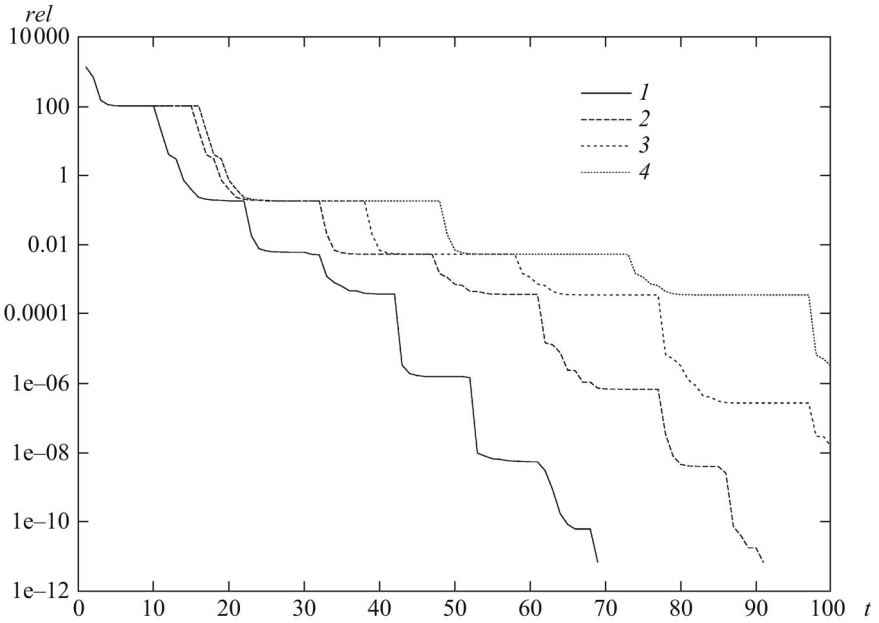


**Fig. 1.** Convergence of the conjugate subgradient method for the piecewise-quadratic function `maxquad`. Shown is the relative accuracy of determining the minimum of $rel = (f(x^k) - f_\star)/|f_\star|$ vs. the number $t$ of iterations of the algorithm at increasing the size of the maximal packet of subgradients `nb`: (*1*) `nb` = 10, (*2*) `nb` = 15, (*3*) `nb` = 20, (*4*) `nb` = 40.

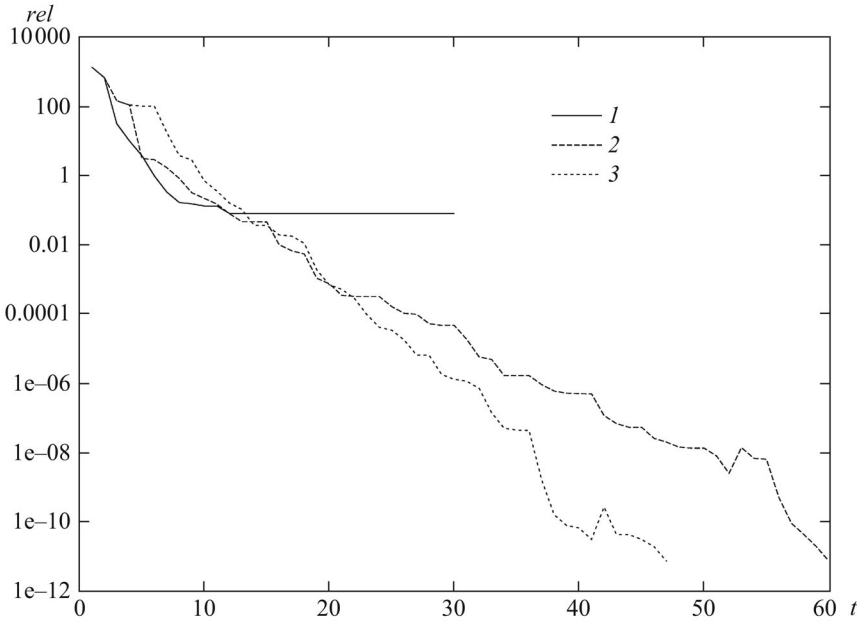


**Fig. 2.** Convergence of the conjugate subgradient method for the piecewise-quadratic function `maxquad`. Shown is the relative accuracy of determining the minimum of $rel = (f(x^k) - f_\star)/|f_\star|$ vs. the number $t$ of iterations of the algorithm at reducing the size of the maximal packet of subgradients `nb`: (*1*) `nb` = 2, (*2*) `nb` = 6, (*3*) `nb` = 10.

Figures 1 and 2 demonstrate convergence in the objective function of the method of conjugate subgradient for different maximal dimensions of the subgradient packets. To avoid complications of graphics, the trajectories of algorithm's operation is divided into two classes: from the best

choice of the maximal packet of subgradients toward its reduction (Fig. 2) and toward increase (Fig. 1). As can be seen from the presented graphs, in all cases the method's convergence remains approximately linear and practically unattainable for the subgradient algorithms. Despite all efforts to select the step multipliers in (2), the ordinary subgradient method was capable to provide the accuracy of $10^{-2}$ only after 5000 iterations.

It deserves noting that the method of conjugate subgradients somewhat improved also the previous record for this function. The optimal value in this problem is $-0.8414083345821985$ to within sixteen significant digits and is reached at the point shown in the table.

Point of minimum for `maxquad`

| | | | |
|---|---|---|---|
| 1 | 0.1262565919226512 | 6 | 0.2783995015309495 |
| 2 | 0.0343783011310847 | 7 | –0.0742186640960634 |
| 3 | 0.0068571878440697 | 8 | –0.1385240462792682 |
| 4 | –0.0263606695458208 | 9 | –0.0840312187567561 |
| 5 | –0.0672949264854349 | 10 | –0.0385803073994817 |

It deserves to pay attention to the horizontal parts of the graphs of Figs. 1 and 2 which correspond to the zero solutions of the problem of one-dimensional minimization (7). During such iterations information is in fact collected for seeking the direction of decrease of the objective function. The fact that the number of tentative steps actually remains invariable over the entire accuracy range and does not increase with approaching the extremum and the corresponding degradation of the computation conditionality is the remarkable distinction of the algorithm.

## 6. CONCLUSIONS

For the problems of convex nondifferentiable optimization, it was possible to carry out a complete theoretical substantiation of the analog of the method of conjugate gradients with the a priori limited memory. The proposed algorithm is free of two basic disadvantages of the previously suggested its counterparts such as the unlimited requirements on the memory and/or substantial growth in the number of tentative iterations to seek the direction to improve the objective function at approaching the extremum. The computer experiment with the test problem of nondifferentiable optimization presenting essential difficulties to the subgradient algorithms proved to be quite encouraging, which allows one to anticipate the practical use of the results obtained.

## ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project no. 13-07-12010. The present authors are deeply indebted to the anonymous reviewers whose remarks and comments contributed much to better presentation of the results.

## REFERENCES

1. Shor, N.Z., Using the Gradient Descent Method to Solve the Transport Problem, in *Mater. nauchn. seminara po teoret. i prikl. vopros. kibernetiki i issledov. operatsii* (Proc. Workshop on Theor. Appl. Issues of Cybernetics and Operation Research), Kiev: Nauchn. Sovet po Kibern. Akad. Nauk USSR, 1962, vol. 1, pp. 9–17.

2. Polyak, B.T., A General Method to Solve Extremal Problems, *Dokl. Akad. Nauk SSSR*, 1967, vol. 174, no. 1, pp. 33–36.

3. Lemarechal, C., An Extension of Davidon Methods to Non-differentiable Problems, *Math. Program. Study*, 1975, vol. 3, pp. 95–109.

4. Hiriart-Urruty, J.-B. and Lemarechal, C., *Convex Analysis and Minimization Algorithms II. Advanced Theory and Bundle Methods*, New York: Springer, 1993.

5. Lemarechal, C., Nemirovskii, A., and Nesterov, Ju., New Variants of Bundle Methods, *Math. Program.*, 1995, vol. 69, nos. 1–3, pp. 111–147.

6. Shor, N.Z., Kiwiel, K.C., Ruszcayǹski, A., *Minimization Methods for Non-differentiable Functions*, New York: Springer, 1985.

7. Nesterov, Yu.E., Smooth Minimization of Non-smooth Functions, *Math. Program.*, 2005, vol. 103, no. 3, pp. 127–152.

8. Mifflin, R. and Sagastizabal, C., A VU-algorithm for Convex Minimization, *Math. Program.*, 2005, vol. 104, nos. 2–3, pp. 583–608.

9. Wolfe, Ph., Note on a Method of Conjugate Subgradients for Minimizing Nondifferentiable Functions, *Math. Program.*, 1974, vol. 7, no. 1, pp. 380–383.

10. Wolfe, Ph., A Method of Conjugate Subgradients for Minimizing Nondifferentiable Functions, in *Nondiffer. Optim. Math. Program. Studies*, Berlin: Springer, 1975, vol. 3, pp. 145–173.

11. Konnov, I.V., Method of Conjugate Subgradient Type for Minimization of Functionals, *Issled. Prikl. Mat.*, 1984, no. 12, pp. 59–62.

12. Konnov, I.V., *Combined Relaxation Methods for Variational Inequalities*, Berlin: Springer, 2001.

13. Pytlak, R., On the Convergence of Conjugate Gradient Algorithms, *IMA J. Numer. Anal.*, 1994, no. 14, pp. 443–460.

14. Pytlak, R. and Tarnawski, T., On the Method of Shortest Residuals, *J. Optim. Theory Appl.*, 2007, vol. 133, pp. 99–110.

15. Dai, Y.H. and Yuan, Y., Global Convergence of the Method of Shortest Residuals, *Numer. Math.*, 1999, no. 82, pp. 581–598.

16. Dai, Y.H., Convergence of Conjugate Gradient Methods with Constant Stepsizes, *Optim. Method Software*, 2011, vol. 26, no. 6, pp. 895–909.

17. Nurminskii, E.A., *Chislennye metody vypukloi optimizatsii* (Numerical Methods of Convex Optimization), Moscow: Nauka, 1991.

18. Nurminski, E.A., Envelope Step-size Control for Iterative Algorithms Based on Fejer Processes with Attractants, *Optim. Method Software*, 2010, vol. 25, no. 1, pp. 97–108.

19. Hesten, M.R. and Stiefel, E., Methods of Conjugate Gradients for Solving Linear Systems, *J. Res. Natl. Bureau Standards*, 1952, vol. 49, no. 6, pp. 409–436.

*This paper was recommended for publication by A.I. Kibzun, a member of the Editorial Board*