V. I. BURDINA

1958

SovietRxiv

Abstract

Full Text

V. I. BURDINA

ON A METHOD FOR SOLVING SYSTEMS OF LINEAR ALGEBRAIC EQUATIONS

(Presented by Academician S. L. Sobolev, 13 I 1958)

In the present note a method is set forth for the successive reduction of accumulated rounding errors as applied to one of the modifications of the method of conjugate directions ([1])—the method of orthogonalization of columns, equivalent to the Gaussian elimination method. An estimate is also given for the error of the solution through intermediate computational data.

1. Let

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= k_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= k_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= k_n \end{aligned} \tag{1}$$

be the system of equations to be solved. It is assumed that the coefficients of the system are real, and that a solution exists and is unique.

By

$$a_1,\ a_2, \ldots, a_n \tag{2}$$

we denote the column vectors of the matrix $a = \|a_{ij}\|$, and by k the column of free terms.

We shall carry out the orthogonalization of the system of vectors (2) by the formulas

$${}^*b_i = a_i + {}^*f_1^{(i)}a_1 + \cdots + {}^*f_{i-1}^{(i)}a_{i-1}, \tag{3}$$

$${}^*f_i = e_i - {}^*\gamma_1^{(i)}{}^*f_1 - \cdots - {}^*\gamma_{i-1}^{(i)}{}^*f_{i-1}, \tag{4}$$

$${}^*\gamma_s^{(i)} = \frac{(a_i, {}^*b_s)}{|{}^*b_s|^2}, \qquad s = 1, 2, \ldots, i-1, \tag{5}$$

$$i = 1, 2, \dots, n,$$

where $e_1, \dots, e_n$ is the system of unit vectors $(1, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, 0, \dots, 0, 1)$; the coefficients ${}^*f_s^{(i)}$ from (3) are the components of the vector ${}^*f_i = \{{}^*f_1^{(i)}, {}^*f_2^{(i)}, \dots, {}^*f_{i-1}^{(i)}, 1, 0, \dots, 0\}$; (a, c) is the scalar product of the vectors a and c.

As a result of applying formulas (3)–(5), the system of vectors (2) passes into the system of orthogonal vectors

$$ {}^*b_1, \dots, {}^*b_n. \tag{6}$$

If to the vectors of the system (2) one adjoins the column of free terms, putting $k = a_{n+1}$, then after its orthogonalization a zero vector is obtained

$$0 = k + {}^*f_1^{(n+1)} a_1 + \dots + {}^*f_n^{(n+1)} a_n,$$

where

$$ {}^*f_{n+1} = -\alpha_1 {}^*f_1 - \dots - \alpha_n {}^*f_n. \tag{7}$$

where

$$\alpha_i = \frac{(k, {}^*b_i)}{({}^*b_i, {}^*b_i)}, \qquad i = 1, \dots, n. \tag{8}$$

The components of the vector ${}^*f_{n+1}$ give the solution of system (1), taken with the opposite sign:

$$x_1 = -{}^*f_1^{(n+1)}, \dots, x_n = -{}^*f_n^{(n+1)}. \tag{9}$$

2. In reality, because of unavoidable rounding errors, computations by formulas (3)–(5), (7)–(9) will lead not to the exact solution, but only to an approximate solution of system (1). Instead of the vectors (6), one obtains certain vectors $b_1, \dots, b_n$, close to orthogonal; this may be judged from the quantities

$$\Delta_{i,j} = (b_i, b_j), \qquad i, j = 1, \dots, n; \quad i \neq j. \tag{10}$$

If system (1) is ill-conditioned, then the deviation from zero of the quantity $\Delta_{i,j}$ will be significant, and the solution will be obtained in a strongly distorted form. These same quantities $\Delta_{i,j}$ can be used to eliminate the growing error.

Suppose that, as a result of orthogonalization by formulas (3)–(5), from the vector a_i the vector $b_i = {}^{\mathrm{I}}b_i$ has been obtained. Taking this vector instead of the initial one and repeating the process of "orthogonalization," we again obtain the vector ${}^{\mathrm{II}}b_i$; "orthogonalizing" this vector, we obtain the vector ${}^{\mathrm{III}}b_i$, and so on.

Theorem 1. *If for any p and q, $p < i$, $q < i$, $p \neq q$, the inequality*

$$\frac{\Delta_{p,q}}{\Delta_{p,p}} < \frac{1}{n}, \tag{11}$$

where n is the order of system (1), is satisfied, then the sequence of vectors

$$ {}^{\mathrm{I}}b_i,\ {}^{\mathrm{II}}b_i,\ {}^{\mathrm{III}}b_i, \dots, \tag{12}$$

generated by the vector a_i under repeated application of formulas (3)–(5), will converge to the vector ${}^{*}b_i$ of system (6).

We write the vector ${}^{*}b_i$ of system (6) in the form

$$ {}^{*}b_i = a_i - \tilde{\gamma}_1^{(i)} b_1 - \tilde{\gamma}_2^{(i)} b_2 - \dots - \tilde{\gamma}_{i-1}^{(i)} b_{i-1}. \tag{13}$$

After multiplying both sides of (13) scalarly by b_s, $s = 1, \dots, i-1$, taking into account that $({}^{*}b_i, b_s) = 0$ for $i \neq s$, and solving the s-th equation with respect to $\tilde{\gamma}_s^{(i)}$, we arrive at the following system of equations for determining $\tilde{\gamma}_1^{(i)}, \dots, \tilde{\gamma}_{i-1}^{(i)}$:

$$\tilde{\gamma}_s^{(i)} = \frac{(a_i, b_s)}{(b_s, b_s)} - \tilde{\gamma}_1^{(i)} \frac{\Delta_{1,s}}{\Delta_{s,s}} - \dots - \tilde{\gamma}_{s-1}^{(i)} \frac{\Delta_{s-1,s}}{\Delta_{s,s}} - \tilde{\gamma}_{s+1}^{(i)} \frac{\Delta_{s+1,s}}{\Delta_{s,s}} - \dots$$

$$\dots - \tilde{\gamma}_{i-1}^{(i)} \frac{\Delta_{i-1,s}}{\Delta_{s,s}}, \tag{14}$$

where $\Delta_{i,j}$ is taken from (10).

The vector ${}^{\kappa}b_i$ of the sequence (12) can be represented in the form

$$ {}^{\kappa}b_i = a_i - {}^{\kappa}\gamma_1^{(i)} b_1 - \dots - {}^{\kappa}\gamma_{i-1}^{(i)} b_{i-1}, \qquad \kappa = \mathrm{I}, \mathrm{II}, \mathrm{III}, \dots. \tag{15}$$

It is easy to see that the numbers ${}^{\kappa}\gamma_1^{(i)}, \dots, {}^{\kappa}\gamma_{i-1}^{(i)}$ are an approximate solution of the system of equations (14), which is obtained by solving this system by the method of iterations ([2]) at the κ-th step, if as the zero-order

as the approximation the trivial system of values. Condition (12) guarantees convergence of the iteration process. From comparison of equalities (13) and (15) it then follows that ${}^{x}b_i \to {}^{*}b_i$, and the theorem is proved.

Suppose that the computation of the vector ${}^*b_i$ has been carried to such a degree of accuracy that inequalities (11) prove to be satisfied also for $p = i,\ q = 1, \ldots, i-1$. Then, on the basis of Theorem 1, one can proceed to the construction of the next vector ${}^*b_{i+1}$ of system (6). Starting with $i = 1$, in this way one can obtain a sufficiently good approximation to the whole system (6). Extending the indicated operations to the vector $a_{n+1} = k$ will lead to a series of more and more accurate approximations to the solution (9) of system (1).

3. In the direct construction of the solution of system (1), as is seen from formulas (7)–(9), it is not the vectors (6) themselves that participate, but only their scalar products. Therefore, in the case when system (1) has been reduced to the normal form

$$\begin{gathered} A_{11}x_1 + \ldots + A_{1n} = k_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ A_{n1}x_1 + \ldots + A_{nn}x_n = k_n, \end{gathered} \tag{16}$$

where $A_{ij} = A_{ji} = (a_i, a_j),\ k_i = (a_i, k),\ i, j = 1, \ldots, n$, it is more expedient to dispense with the explicit computation of the vectors of system (6), using, for the determination of $\gamma_s^{(i)}$ instead of (5), the formula

$$\gamma_s^{(i)} = \frac{(Ae_i, f_s)}{(Ae_s, f_s)}, \tag{17}$$

where $A = \|A_{ij}\|$ is the coefficient matrix of system (16).

To the process of orthogonalization of the vectors (2) there corresponds the A-orthogonalization of the unit vectors $e_1, \ldots, e_n$. Sequence (12) will correspond to a sequence of vectors ${}^I f_i, {}^{II} f_i, \ldots$, converging to the vector ${}^* f_i$. Each subsequent vector of this sequence is obtained in the left-hand side of (4) (the asterisk sign is omitted), if in its right-hand side, and also in (17), e_i is replaced by the preceding vector of this sequence, and the quantities ${}^*\gamma_s^{(i)}$ are replaced by those which are thereby obtained in the left-hand side of (17), i.e. by iteration according to formulas (5), (17). The quantities $\gamma_s^{(i)}$ could also be computed starting from system (14), taking into account that $(a_i, b_s) = (Ae_i, f_s)$, $\Delta_{ij} = (Af_i, f_j)$. However, this way is accompanied by a greater loss of accuracy. As before, one must monitor the fulfillment of inequalities (11).

Let us note that what has been said in this paragraph also extends to systems (1) for which it is known only that their coefficient matrix is symmetric and nonnegative.

Let us also note that by the method described in paragraphs 1–3 one can solve any systems of equations (1), not necessarily with a unique solution. In the case of an indeterminate system one can obtain one of the possible solutions, and

in the case when the system is inconsistent—a solution in the sense of the least quadratic deviation [3].

4. Let $x_1, \dots, x_n$ be the solution of system (1), which is unique, $\bar{x}_1, \dots, \bar{x}_n$ the approximate solution obtained by the method described above; let $\bar{k}_1, \dots, \bar{k}_n$ be the residuals

$$\bar{k}_i = k_i - \sum_{s=1}^{n} a_{is} x_s.$$

Put

$$\varepsilon = \max_i |\bar{k}_i|. \tag{18}$$

Theorem 2. If, in the process of obtaining the solution of system (1), the inequality

$$\frac{\Delta_{i,j}}{\Delta_{i,i}} < \frac{1}{2n}, \qquad i \neq j;\ i, j = 1, \dots, n, \tag{19}$$

is maintained throughout, then the estimate

$$|\overline{x_i} - x_i| < \frac{\sqrt{n}\, F \varepsilon}{\min_p \sqrt{\Delta_{p,p}}},$$

will be valid, where

$$F = \max_i \left(|f_1^{(i)}| + \dots + |f_{i-1}^{(i)}| + 1 \right)$$

(with the best of the attained approximations to the vector ${}^* f_i$ being taken as the vector f_i); ε is the modulus of the residual (18).

5. As an example of the application of the method described, a system of 10 equations from [4] was solved; the Schreider method [5] (orthogonalization of rows) is unsuitable for solving it. The computations were carried out with 4–6 decimal digits; the solution was obtained to an accuracy of 0.001.

All-Union Correspondence Electrotechnical
Institute of Communications

Received
4 VII 1957

CITED LITERATURE

1. M. R. Hestenes, E. Stiefel, J. Res. Nat. Bur. Stand., **49**, 409 (1952).

2. V. N. Faddeeva, *Computational Methods of Linear Algebra*, Moscow–Leningrad, 1950, pp. 120–128.

3. A. N. Krylov, *Lectures on Approximate Calculations*, 1954, p. 381.

4. A. I. Vzorova, *Computational Mathematics and Computational Technology*, issue 1, 92 (1953).

5. Yu. A. Schreider, DAN, **76**, No. 5 (1951).

Note: Figure translations are in progress. See original paper for figures.