NUREG/CR-1943
LA-8705-MS

TERA

The Incomplete Lower-Upper Conjugate Gradient (ILUCG) Method for Solving Large, Sparse Linear Systems

University of California

LOS ALAMOS SCIENTIFIC LABORATORY
Post Office Box 1663 Los Alamos, New Mexico 87545

This report was not edited by the Technical Information staff.

NUREG/CR-1943
LA-8705-MS

R4

The Incomplete Lower-Upper Conjugate Gradient (ILUCG) Method for Solving Large, Sparse Linear Systems

Charles H. Neil*

Manuscript submitted: January 1981
Date published: February 1981

Prepared for
Division of Reactor Safety Research
Office of Nuclear Regulatory Research
US Nuclear Regulatory Commission
Washington, DC 20555

NRC FIN No. A7016

*Graduate Research Assistant. 3209-30th Street, Lubbock, TX 79410.

UNITED STATES
DEPARTMENT OF ENERGY
CONTRACT W-7405-ENG. 36

CONTENTS

TABLE

THE INCOMPLETE LOWER-UPPER CONJUGATE GRADIENT (ILUCG) METHOD FOR SOLVING LARGE, SPARSE LINEAR SYSTEMS

by

Charles H. Neil

ABSTRACT

The method of Conjugate Gradients is combined with an incomplete factorization of the coefficient matrix to produce an iterative method for approximate solution of systems of linear equations. The method is suited for systems where the coefficient matrix is large, sparse, and nonsymmetric. A comparison is presented in this document of the performance of the method vs the Gauss-Seidel method for test problems arising in the TRAC Code for reactor hydrodynamics.

I. INTRODUCTION

The discretization of partial differential equations in mathematical physics often results in the problem of solving systems of linear equations, the coefficients of which form a large, sparse matrix. The method of Conjugate Gradients (Refs. 1, 2, and 3) was originally proposed as a general iterative technique for solving linear systems but is not competitive with other iterative methods when applied to the full matrix. More recent work (Refs. 4, 5, and 6) has been concerned with applying the Conjugate Gradient method to a system modified by an incomplete decomposition of the coefficient matrix. An accurate approximate solution can thereby be obtained in relatively few iterations.

Much of the work to date deals with symmetric coefficient matrices. In the solution of hydrodynamic equations in the Transient Reactor Analysis Code[7] (TRAC) developed at the Los Alamos National Laboratory, however, the resulting seven-stripe pressure matrix is nonsymmetric. This paper presents a generalized Conjugate Gradient (CG) method applied to an incomplete factorization for such nonsymmetric matrices, and gives results of numerical tests of the method.

II. THE METHOD OF CONJUGATE GRADIENTS

We will be concerned in what follows with solving the system

$$Mx = k \quad . \tag{1}$$

Here, M is an N-dimensional nonsingular matrix, not necessarily symmetric; let us denote the exact solution by h. Suppose we have available a linearly independent set of vectors $\{\pi_0, \pi_1, \ldots\}$. We will be interested in studying an iterative method of the form

$$x_{i+1} = x_i + a_i\pi_i, \; i = 0, 1, \ldots; \; x_0 \text{ arbitrary}, \tag{2}$$

which produces the exact solution after a finite number of steps.

One way to accomplish this is to replace each vector, π_i, by a vector

$$p_i = \sum_{j=0}^{i} c_{ij}\pi_j \quad ,$$

such that

$$\langle p_i, p_j \rangle = 0, \; i \neq j$$

(that is, orthogonalize the set $\{\pi_i\}$), and then select the constants a_i so that the error in the $i+1$ $\underline{st}$ iterate, $\left| h - x_{i+1} \right|^2$, is minimized. This will occur when $h - x_{i+1}$ is orthogonal to the set $\{p_0, p_1, \ldots, p_i\}$ and [illegible] a_i is given by

$$a_i = \frac{\langle h - x_0, p_i \rangle}{\langle p_i, p_i \rangle} \quad . \tag{3}$$

If it ever occurs that the initial error, $e_0 = h - x_0$, lies in the subspace $S_k = \text{span} \left\{ p_0, \ldots, p_k \right\}$, then we have

$$h - x_{k+1} = h - x_0 - \sum_{j=0}^{k} a_j p_j \in S_k \quad .$$

As $h - x_{k+1}$ also was chosen orthogonal to S_k, it must be the zero vector, and therefore the method terminates.

We are thus motivated to try to select the vectors $\left\{ \pi_i \right\}$ and $\left\{ p_i \right\}$ in such a way that $e_0 \in S_k$ for k is as small as possible. Furthermore, the exact solution, h, is unavailable a priori and so we must choose the vectors without explicit knowledge of e_0. The following allows us to find a way of doing this.

Lemma: Suppose T is a symmetric, positive definite matrix and u is some chosen vector. Then $u \in \text{span} \left\{ Tu, T^2u, \ldots, T^{k-\ell}u \right\}$, where

k = number of distinct eigenvalues of T and

ℓ = number of vanishing components corresponding to distinct eigenvalues in the eigenvector decomposit on of u.

Proof Since T is symmetric and positive definite, there exists a matrix R (of eigenvectors of T) such that $R^{-1}TR = \Lambda$, where $\Lambda = \text{diag} (\lambda_1, \lambda_2, \ldots, \lambda_{k-1}, \lambda_k, \ldots \lambda_k)$, and each λ_j is distinct.

Let d be the vector of coefficients in the eigenvector decomposition of u; thus $u = Rd$. We then have

$$T^j u = R\Lambda^j d, \text{ for } j = 0, 1, \ldots \quad .$$

We wish to be able to write for some constants c_j,

$$u = c_1 Tu + \ldots + c_{k-\ell} T^{k-\ell} u \quad .$$

or

$$Rd = c_1 R\Lambda d + \ldots + c_{k-\ell} R\Lambda^{k-\ell} d \quad .$$

Equivalently, we want

$$d = c_1 \Lambda d + \ldots + c_{k-\ell} \Lambda^{k-\ell} d \quad . \tag{4}$$

Among the first k components of d, there are only $k - \ell$ nonzero entries, say $d_{j_1}, d_{j_2}, \ldots, d_{j_{k-\ell}}$, are nonzero. After dividing both sides of these equations by the components of d, we can re-write the nonzero equations of (4) as

$$Lc = \underline{1}, \text{ where } L = \begin{bmatrix} \lambda_{j_1} & \lambda_{j_1}^2 & \ldots & \lambda_{j_1}^{k-\ell} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \lambda_{j_{k-\ell}} & \lambda_{j_{k-\ell}}^2 & \ldots & \lambda_{j_{k-\ell}}^{k-\ell} \end{bmatrix},$$

$c = (c_1, \ldots, c_{k-\ell})^T$, and $\underline{1}$ is a $k-\ell$ dimensional vector of 1's. As the λ's are all distinct, L is non-singular: c exists and the proof is complete.

Returning now to the problem of choosing the spanning set of vectors $\{\pi_0, \pi_1, \ldots\}$, consider the set

$$\{M^T M e_0, (M^T M)^2 e_0, \ldots\} .$$

We note that $Me_0 = M(h - x_0) = k - Mx_0$, so that once x_0 is selected, all these vectors are available. Furthermore, in view of the lemma, if M^TM has many repeated eigenvalues (or in the fortunate event that x_0 is chosen so that e_0 is defective in many eigendirections of M^TM), then the method (1) terminates in relatively few steps.

In creating the set $\{p_i\}$, orthogonalization of each $(M^TM)^j e_0$ with respect to $(M^TM)^i e_0$, $i = 0, 1, \ldots, j-1$, is too expensive. We produce an equivalent spanning set $\{p_0, p_1, \ldots\}$ as follows.

Denote by r_i the $i^{\underline{th}}$ residual:

$$r_i = k - Mx_i, \quad i = 0, 1, \ldots . \tag{5}$$

Having produced $\{p_0, p_1, \ldots, p_i\}$, $i \geq 0$, obtain p_{i+1} by orthogonalizing $M^TMe_{i+1} = M^Tr_{i+1}$ with respect to p_i (instead of orthogonalizing $(M^TM)^{i+1}e_0 = (M^TM)^iM^Tr_0$ with respect to p_i). The resulting set $\{p_j\}$ has the same span as does $\{(M^TM)^j e_0\}$. The advantage of this procedure is that if b_i is chosen so that $p_{i+1} = M^Tr_{i+1} + b_ip_i$ is orthogonal to p_i, then automatically p_{i+1} is orthogonal to $p_0, p_1, \ldots, p_{i-1}$. For, we have

$$\begin{aligned} \langle p_{i+1}, p_{i-j}\rangle &= \langle M^Tr_{i+1} + b_ip_i, p_{i-j}\rangle \\ &= \langle M^Tr_{i+1}, p_{i-j}\rangle + b_i\langle p_i, p_{i-j}\rangle \\ &= \langle M^TM(h-x_{i+1}), p_{i-j}\rangle \\ &= \langle h - x_{i+1}, M^TMp_{i-j}\rangle \quad , \end{aligned}$$

and $h-x_{i+1}$ was chosen orthogonal to the span $\{M^Tr_0, (M^TM)M^Tr_0, \ldots, (M^TM)^{i-j}M^Tr_0\}$, which contains $(M^TM)\,p_{i-j}$ for $j \geq 1$.

Thus, $\langle p_{i+1}, p_{i-j} \rangle = 0$, $j \geq 1$.

Therefore, to carry out the iterations we need only save the vectors x_i, r_i, and p_i.

Thus we write the Conjugate Gradient Algorithm as follows. Choose x_0 arbitrarily; set $r_0 = k - Mx_0$, and $p_0 = M^T r_0$.

For $i = 0, 1, \ldots,$ set

$$a_i = \frac{\langle r_i, r_i \rangle}{\langle p_i, p_i \rangle}$$

$$x_{i+1} = x_i + a_i p_i$$

$$r_{i+1} = r_i - a_i M p_i$$

$$b_i = \frac{\langle r_{i+1}, r_{i+1} \rangle}{\langle r_i, r_i \rangle}$$

$$p_{i+1} = M^T r_{i+1} + b_i p_i .$$

It should be noted that these formulae for a_i and r_{i+1} are equivalent to Eqs. (3) and (5), respectively (see Refs. 2 and 3).

As previously observed, the CG method terminates in relatively few steps if the matrix M has large clustering of eigenvalues. Furthermore, at each step of the iteration the vector x_i minimizes $\left| h - x \right|^2$ within a certain subspace. With our choice of vectors $\{p_i\}$, each one being a linear combination of $\{(M^T M)^j e_0\}_{j=0}^{i-1}$, much of the change in x_i occurs in eigendirections corresponding to the largest eigenvalues of M. Accordingly, a large reduction in

error can be anticipated within the first few iterations. It is therefore possible that x_i may be an accurate approximation to h for $i \ll N$, and it is this situation we will try to produce.

III. The INCOMPLETE L-U FACTORIZATION CONJUGATE GRADIENT METHOD

Let us now turn to the problem of solving the system

$$Ax = b, \tag{6}$$

where A is a large, sparse, nonsymmetric matrix. Suppose we can carry out the usual L-U decomposition of A (that is, where L is unit lower triangular and U is upper triangular), with the exception that elements in L or U corresponding to zero elements in A are considered zero and are neither computed nor stored. The resulting incomplete L-U decomposition of A will require no more storage than does A (and will not be as costly as the complete L-U decomposition wherein L and U are generally no longer sparse).

We have $LU = A+E$, where the matrix E contains errors because of the incomplete decomposition. Hence,

$$L^{-1} AU^{-1} = I - L^{-1}EU^{-1} \quad .$$

If the matrix E is in some sense small, then the matrix $L^{-1}AU^{-1}$ is an approximate identity -- in particular, $L^{-1}AU^{-1}$ should tend to have many eigenvalues close to unity, and by previous remarks, the conjugate gradient method, applied to a system with $L^{-1}AU^{-1}$ as the coefficient matrix, should converge rapidly.

We therefore re-write Eq. (6) as

$$L^{-1}AU^{-1}Ux = L^{-1}b \quad .$$

Letting $M = L^{-1}AU^{-1}$, and writing the algorithm in terms of x (instead of Ux), we can write the ILUCG algorithm as follows.

Choose x_0 arbitrarily; set $r_0 = b - Ax_0$ and

$$p_0 = (U^TU)^{-1}A^T(LL^T)^{-1}r_0 \quad .$$

For i = 0,1, ..., until satisfied, set

$$a_i = \frac{\langle r_i, (LL^T)^{-1} r_i \rangle}{\langle p_i, U^TUp_i \rangle}$$

$$x_{i+1} = x_i + a_i p_i$$

$$r_{i+1} = r_i - a_i A p_i$$

$$b_i = \frac{\langle r_{i+1}, (LL^T)^{-1} r_{i+1} \rangle}{\langle r_i, (LL^T)^{-1} r_i \rangle}$$

$$p_{i+1} = (U^TU)^{-1} A^T(LL^T)^{-1} r_{i+1} + b_i p_i \quad .$$

It should be noted that the inversion and vector multiplication are carried out quickly because of the triangular and sparse nature of L and U.

IV. NUMERICAL RESULTS

This algorithm was applied to several test cases designed to simulate

numerical problems in the TRAC code for reactor hydrodynamics. In this application, the matrix A was a nonsymmetric 440 x 440 real matrix with unit main diagonal and with negative off-diagonal entries occurring in a pattern following a seven-point discretization in cylindrical coordinates of the Laplace operator.

The matrix was loaded uniformly across rows such that $\sum_i = 1 - \rho_i$, where $\sum_i$ denotes the sum of the absolute values of the off-diagonal entries in the i^{th} row, and ρ_i is a pseudo-random number in the interval $[0, \delta]$, with δ an input parameter. By allowing $\delta \to 0^+$, the degree of diagonal dominance of the matrix A is reduced (and its condition is worsened).

Table I contains results of a comparison of the performance of the ILUCG method and the Gauss-Seidel method on the same linear system for values of δ ranging from 0.500 to 0.001. As shown in the table, below a certain level of diagonal dominance, the Gauss-Seidel method failed to converge to the desired residual norm within 1000 iterations, whereas in all cases the ILUCG method converged. Because of the relative expense of the ILUCG method per iteration (it is roughly 8.2 times as costly per iteration as Gauss-Seidel for matrices of this type) it is important to observe the column labeled "Ratio," which is the ratio of the number of iterations of Gauss-Seidel (GS) to ILUCG; only when this figure is greater than 8.2 will ILUCG be the preferred method.

Figure 1 illustrates the norm of the residual of the two methods, plotted vs number of multiplications, for the case $\delta = 0.060$. As this figure shows, the error decreases faster for ILUCG than for GS, although it is initially greater for ILUCG than for GS. The plot is typical of all cases, with the divergence of the curves more pronounced for smaller values of δ.

Finally, Fig. 2 is a plot of the 440 eigenvalues of the original matrix A and of the matrix used in the ILUCG algorithm, $L^{-1} AU^{-1}$, for $\delta = 0.060$. The eigenvalues of A are uniformly spread from 1.97 to 0.029; but those of $L^{-1} AU^{-1}$ are, for the most part, quite close to 1.0, indicating a reason for the effectiveness of the ILUCG method.

V. CONCLUSION

The reported results indicate that the ILUCG method warrants consideration in situations requiring high accuracy or when the coefficient matrix is poorly

conditioned (or both). However, the cost per iteration of ILUCG in its present form makes it not a preferred method in less severe instances.

TABLE I

ILUCG METHOD vs GAUSS-SEIDEL METHOD FOR VALUES OF δ

	Number of iterations		
δ	GS	ILUCG	Ratio (Break-even = 8.2)
0.500	49	16	3.1
0.400	61	19	3.2
0.300	81	22	3.7
0.200	122	26	4.7
0.100	241	34	7.1
0.080	301	36	8.4
0.060	401	39	10.3
0.040	601	42	14.3
0.020	a	47	-
0.010	a	49	-
0.008	a	50	-
0.004	a	52	-
0.001	a	55	-

[a]Failed to converge within 1000 iterations.

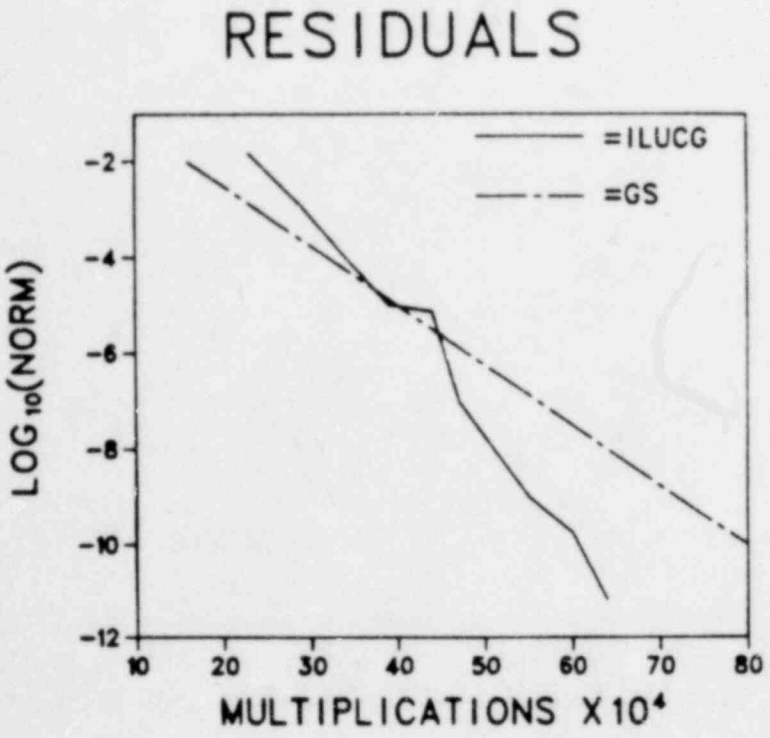

Fig. 1. Residual error.

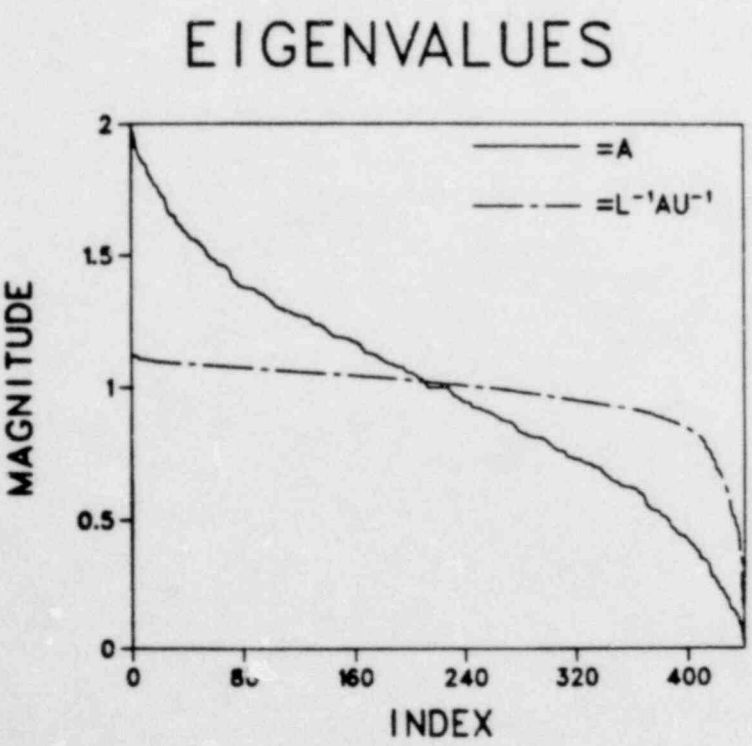

Fig. 2. Matrix eigenvalues.

REFERENCES

1. M. R. Hestenes and E. Steifel, "Methods of Conjugate Gradients for Solving Linear Systems," J. of Research of the National Bureau of Standards 59, 409-436(December, 1952).

2. M. R. Hestenes, "The Conjugate-Gradient Method for Solving Linear Systems," Proc. Sympos. Appl. Math., Numerical Analysis, (McGraw-Hill, New York, 1956) Vol. VI, pp. 83-102.

3. J. K. Reid, "On the Method of Conjugate Gradients for the Solution of Large Sparse Systems of Linear Equations," Proc. Conf. on Large Sparse Systems of Linear Equations (Academic Press, New York, 1971).

4. J. A. Meijerink and H. A. van der Vorst, "An Iterative Solution Method for Linear Systems of Which the Coefficient Matrix is a Symmetric M-Matrix," Math. Comp. 31, 148-162 (January, 1977).

5. P. Concus, G. Golub, and D. O'Leary, "A Generalized Conjugate Gradient Method for the Numerical Solution of Elliptic Partial Differential Equations," Lawerence Berkeley Laboratory report LBL-4604 (1975).

6. D. S. Kershaw, "The Incomplete Cholesky-Conjugate Gradient Method for the Iterative Solution of Systems of Linear Equations," J. Comp. Phys 26, 43-65 (1978).

7. "TRAC-P1A, An Advanced Best Estimate Computer Program for PWR LOCA Analysis," Los Alamos Scientific Laboratory report LA-7777-MS (May 1979).

DISTRIBUTION

	Copies
Nuclear Regulatory Commission, R4, Bethesda, Maryland	388
Technical Information Center, Oak Ridge, Tennessee	2
Los Alamos National Laboratory, Los Alamos, New Mexico	50
	440