A FAST METHOD FOR SOLVING A CLASS OF TRI-DIAGONAL LINEAR SYSTEMS

Michael A. Malcolm, et al

Stanford University

Prepared for:

Office of Naval Research

November 1972

A FAST METHOD FOR SOLVING A CLASS OF TRI-DIAGONAL LINEAR SYSTEMS

BY

MICHAEL A. MALCOLM

JOHN PALMER

STAN-CS-72-323

NOVEMBER 1972

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Stanford University Computer Science Department Stanford, California 94305	Unclassified
	2b. GROUP

3 REPORT TITLE

A Fast Method For Solving A Class of Tri-Diagonal Linear Systems

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)

technical, November 1972

5 AUTHOR(S) (First name, middle initial, last name)

Michael A. Malcolm and John Palmer

6 REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
November 1972	18	4

8a. CONTRACT OR GRANT NO	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-67-A-0112-0029 b. PROJECT NO.	STAN-CS-72-323
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10 DISTRIBUTION STATEMENT

Distribution Unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY

13 ABSTRACT

The folution of linear systems having real, symmetric, diagonally dominant, tri-diagonal coefficient matrices with constant diagonals is considered. It is proved that the diagonals of the LU decomposition of the coefficient matrix rapidly converge to full floating-point precision. It is also proved that the computed LU decomposition converges when floating-point arithmetic is used and that the limits of the LU diagonals using floating point are roughly within machine precision of the limits using real arithmetic. This fact is exploited to reduce the number of floating-point operations required to solve a linear system from $8n-7$ to $5n+2k-3$, where k is much less than n, the order of the matrix. If the elements of the sub- and superdiagonals are 1, then only $4n+2k-3$ operations are needed. The entire LU decomposition takes k words of storage, and considerable savings in array subscripting are achieved. Upper and lower bounds on k are obtained in terms of the ratio of the coefficient matrix diagonal constants and parameters of the floating-point number system.

Various generalizations of these results are discussed.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

I

Unclassified

Security Classification

A FAST METHOD FOR SOLVING
A CLASS OF TRI-DIAGONAL LINEAR SYSTEMS

Michael A. Malcolm*

John Palmer**

November, 1972

*The work of this author was supported by the Office of Naval Research, Contract N00014-67-A-0112-0029.

**The work of this author was supported by Bell Telephone Laboratories.

II

A FAST METHOD FOR SOLVING
A CLASS OF TRI-DIAGONAL LINEAR SYSTEMS

by

Michael A. Malcolm and John Palmer

ABSTRACT

The solution of linear systems having real, symmetric, diagonally dominant, tridiagonal coefficient matrices with constant diagonals is considered. It is proved that the diagonals of the LU decomposition of the coefficient matrix rapidly converge to full floating-point precision. It is also proved that the computed LU decomposition converges when floating-point arithmetic is used and that the limits of the LU diagonals using floating point are roughly within machine precision of the limits using real arithmetic. This fact is exploited to reduce the number of floating-point operations required to solve a linear system from $8n-7$ to $5n+2k-3$, where k is much less than n, the order of the matrix. If the elements of the sub- and superdiagonals are 1, then only $4n+2k-3$ operations are needed. The entire LU decomposition takes k words of storage, and considerable savings in array subscripting are achieved. Upper and lower bounds on k are obtained in terms of the ratio of the coefficient matrix diagonal constants and parameters of the floating-point number system.

Various generalizations of these results are discussed.

TR

1. Introduction

We will consider the solution of linear algebraic systems having real symmetric, diagonally dominant, tridiagonal coefficient matrices with constant diagonals. This problem occurs frequently in solving certain kinds of partial differential equations, boundary value problems of ordinary differential equations, and cubic spline interpolation problems.

Consider the coefficient matrix

$$A = \begin{bmatrix} a & b & & & & & \\ b & a & b & & & & \\ & b & a & b & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & b & a & b & \\ & & & & b & a & b \\ & & & & & b & a \end{bmatrix},$$

of order n. The usual LU decomposition of A requires $n-1$ divisions, $n-1$ multiplications, and $n-1$ additions. The solution of the equations $LU\underset{\sim}{x} = \underset{\sim}{d}$ requires an additional n divisions, $2n-2$ multiplications, and $2n-2$ additions. With the following observation, the entire LU decomposition of A can be stored in k floating-point words, and the solution of the linear system $A\underset{\sim}{x} = \underset{\sim}{d}$ can be obtained in k divisions, $3n-1$ multiplications, and $2n-2+k$ additions, where k is usually much less than n. Typically, k is on the order of 10. Moreover, k can easily be estimated from the values of a and b and parameters of the floating-point number system used in the solution. If $b = 1$, then n multiplies can be avoided. In addition to a smaller operation count, substantial savings in array indexing are achieved.

2. The Algorithm

Consider the matrix

$$B = \begin{bmatrix} \alpha & 1 & & & & \\ 1 & \alpha & 1 & & & \\ & 1 & \alpha & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & \alpha & 1 \\ & & & & 1 & \alpha \end{bmatrix},$$

where $\alpha = a/b$. Note that $A = bB$. The analysis, as well as the computation is simplified by considering the coefficient matrix to be B and the linear system $bB\underset{\sim}{x} = \underset{\sim}{d}$. B can be factored into the product LU , where

$$L = \begin{bmatrix} 1 & & & & \\ \ell_1 & 1 & & & \\ & \ell_2 & 1 & & \\ & & \ddots & \ddots & \\ & & & \ell_{n-1} & 1 \end{bmatrix}, \quad U = \begin{bmatrix} u_1 & 1 & & & \\ & u_2 & 1 & & \\ & & \ddots & \ddots & \\ & & & & 1 \\ & & & & u_n \end{bmatrix},$$

using the recurrence relations:

$$u_1 = \alpha \ , \quad \ell_{i-1} = 1/u_{i-1} \ , \quad u_i = \alpha - \ell_{i-1} \ , \quad i=2,\ldots,n,$$

or

$$u_i = \alpha - 1/u_{i-1} \ , \quad i=2,\ldots,n \ . \tag{1}$$

Under suitable conditions, to be discussed, the ℓ_i converge and $\ell_k = \ell_{k+1} = \ldots = \ell_n = \ell$ to machine accuracy. In the computer, one simply computes and stores the values of ℓ_i , $i=1,\ldots,k$. The solution vector $\underset{\sim}{x}$ can then be computed as follows:

$$
\begin{aligned}
&y_1 = d_1 \ , \\
&y_i = d_i - \ell_{i-1} y_{i-1} \ , \ i=2,\ldots,k \ , \quad y_i = d_i - \ell y_{i-1} , \ i=k+1,\ldots,n \ , \\
&z_n = \ell y_n \ , \\
&z_i = \ell(y_i - z_{i+1}) \ , \ i=n-1,\ldots,k \ , \quad z_i = \ell_i(y_i - z_{i+1}), \ i=k-1,\ldots,1 \ , \\
&x_i = b^{-1} z_i \ , \ i=1,\ldots,n \ .
\end{aligned}
\tag{2}
$$

3. Convergence of the LU decomposition

We will show that when A is diagonally dominant, the sequences $[u_i]$ and $[\ell_i]$ converge. We will also find an estimate of the rate of convergence which can be used to determine a value for k .

It is sufficient to show that the sequence $[u_i]$ converges, and for this we assume diagonal dominance, or equivalently, $|\alpha| > 2$. The following theorem is a special case of a theorem of Parter (1962) for band matrices.

Theorem 1: If $|\alpha| > 2$, then the sequence $[u_i]$ converges to u where

$$
u = \frac{\alpha + \operatorname{sgn}(\alpha) \sqrt{\alpha^2 - 4}}{2} \ . \tag{3}
$$

Proof: Convergence follows from the fact that the sequence $[\alpha u_i]$ is bounded and monotone:

Lemma 1 (boundness): If $|\alpha| > 2$, then

$$\alpha u_i > 2 , \quad i=1,\ldots . \tag{4}$$

Proof: From (1), $u_1 = \alpha$. Thus $\alpha u_1 = \alpha^2 > 4$. Now assume that (4) holds for some value of $i > 1$. By (1),

$$\alpha u_{i+1} = \alpha^2 - \alpha/u_i > \alpha^2 - \alpha^2/2 > 2 .$$

Lemma 1 follows by induction. ∎

Lemma 2 (monotonicity): If $|\alpha| > 2$, then

$$\alpha u_{i+1} < \alpha u_i , \quad i+1,\ldots .$$

Proof: From (1),

$$u_2 - u_1 = - \frac{1}{\alpha} ,$$

and $$\alpha(u_{i+1} - u_i) = \frac{1}{u_i u_{i-1}} \alpha(u_i - u_{i-1}) , \quad i=2,\ldots .$$

It follows from Lemma 1 that the u_i must all have the same sign. Thus, by induction,

$$\alpha(u_{i+1} - u_i) < 0 .$$

∎

Now, in the limit,

$$u = \alpha - \frac{1}{u} ,$$

or, $$u^2 - \alpha u + 1 = 0 .$$

Equation (3) is the quadratic formula with the sign of the radical chosen to avoid a contradiction with Lemma 1. This completes the proof of Theorem 1.

The following two theorems provide a way to estimate the value of k.

Theorem 2: If $|\alpha| > 2$, then

$$k \leq \left\lceil 1 + \frac{t - 1 - \log_\beta u\alpha}{\log_\beta(\alpha^2 - \frac{\alpha}{u} - 1)} \right\rceil , \tag{5}$$

where β is the floating-point radix, t is the number of digits, and $\lceil \xi \rceil$ denotes the smallest integer not less than ξ.

Proof: We will first prove the following lemma.

Lemma 3: If $|\alpha| > 2$, then

$$\alpha(u_{i+1} - u_i) > -(\alpha^2 - \frac{\alpha}{u} - 1)^{1-i}, \quad i=1,\ldots . \tag{6}$$

Proof: From (1), Lemmas 1 and 2,

$$\alpha(u_{i+1} - u_i) = \frac{1}{u_i u_{i-1}} \alpha(u_i - u_{i-1}) < 0 , \tag{7}$$

and $$\frac{1}{u_i u_{i-1}} = \frac{1}{\alpha u_i - 1} > 0 , \quad i=2,\ldots . \tag{8}$$

Now, $$\alpha u_i = \alpha^2 - \frac{\alpha}{u_{i-1}} , \quad i=2,\ldots . \tag{9}$$

By Lemma 2, and the fact that $u_i u > 0$,

$$\frac{\alpha}{u_i} < \frac{\alpha}{u} , \quad i=1,\ldots .$$

Thus, $\alpha u_i > \alpha^2 - \frac{\alpha}{u}$,

and $$\frac{1}{u_i u_{i-1}} < \frac{1}{\alpha^2 - \frac{\alpha}{u} - 1} , \quad i=2,\ldots .$$

Thus, $$\alpha(u_{i+1} - u_i) > \frac{1}{\alpha^2 - \frac{\alpha}{u} - 1} \alpha(u_i - u_{i-1}), \quad i=2,\ldots .$$

Repeated application of this inequality yields

$$(u_{i+1} - u_i) > (\alpha^2 - \frac{\alpha}{u} - 1)^{1-i}\alpha(u_2 - u_1), \quad i=1,\ldots .$$

Since $\alpha(u_2 - u_1) = -1$, (10)

the Lemma is proved. ∎

Dividing (6) by $\alpha u > 0$ and taking absolute values,

$$\left|\frac{u_{i+1} - u_i}{u}\right| < \frac{1}{\alpha u}(\alpha^2 - \frac{\alpha}{u} - 1)^{1-i} . \tag{11}$$

Requiring the right-side of (11) to be less than β^{1-t} gives a sufficient condition on i for the convergence of $[u_i]$. Taking logarithms yields the sufficient condition

$$i > 1 + \frac{t - 1 - \log_\beta \alpha u}{\log_\beta(\alpha^2 - \frac{\alpha}{u} - 1)} . \tag{12}$$

Thus k need be no larger than the smallest possible value of i given by (12). ∎

Theorem 3: If $|\alpha| > 2$, then

$$k \geq \left\lceil 1 + \frac{t - 1 - \log_\beta \alpha u}{\log_\beta(\alpha^2 - 2)} \right\rceil .$$

Proof: We will first prove the following lemma.

Lemma 4: If $|\alpha| > 2$, then

$$\alpha(u_{i+1} - u_i) \leq -(\alpha^2 - 2)^{1-i}, \quad i=1,\ldots . \tag{13}$$

Proof: By Lemma 2 and (1),

$$\alpha u_i \leq \alpha u_1 = \alpha^2 , \quad i=1,\ldots .$$

Since, by Lemma 1, $\alpha u_i > 0$,

$$\frac{\alpha}{u_i} \geq 1 , \quad i=1,\ldots .$$

Substituting into (8) and (9) gives

$$\frac{1}{u_i u_{i-1}} \geq \frac{1}{\alpha^2 - 2} , \quad i=2,\ldots .$$

This inequality and (7) and (10) yield Lemma 4. ∎

Dividing (12) by $\alpha u > 0$ and taking absolute values gives

$$\left| \frac{u_{i+1} - u_i}{u} \right| \geq \frac{1}{\alpha u} (\alpha^2 - 2)^{1-i}, \quad i=1,\ldots . \tag{13}$$

Setting the right-side of (13) greater than β^{1-t} gives a non-convergence condition for i , and thus, a lower bound on k . Taking logarithms yields Theorem 3. ∎

If we denote by $\overline{K}$, the upper bound given in Theorem 2, and by $\underline{K}$, the lower bound given in Theorem 3, we have

$$\underline{K} \leq k \leq \overline{K} .$$

In practice, these bounds are very close. Usually $\underline{K} = k = \overline{K}$. The following table gives values for $\underline{K}$, $\overline{K}$ and k for various values of α for both single and double precision on the IBM 360.

	Short Precision (β=16, t=6)			Long Precision (β=16, t=14)		
α	$\underline{K}$	k	$\overline{K}$	$\underline{K}$	k	$\overline{K}$
2.05	18	27	30	46	77	80
2.1	16	20	22	41	55	57
2.2	14	15	16	35	40	41
2.3	12	13	13	31	33	34
2.4	11	11	11	28	29	29
2.5	10	10	10	25	26	26
3.0	8	8	8	19	19	19
4.0	6	6	6	14	14	14
5.0	5	5	5	12	12	12
6.0	4	4	4	11	11	11
7.0	4	4	4	10	10	10

Upper and Lower Bounds ($\overline{K}$ and $\underline{K}$) and Observed Values for k for the IBM 360

The preceeding theorems characterize the convergence of the sequence $[u_i]$ in the absence of rounding errors. If the computer arithmetic satisfies certain reasonable rules, then the computed sequence $[\tilde{u}_i]$ also converges monotonically to a limit $\tilde{u}$ which is very close to u. We will prove this result for $\alpha > 2$. A similar argument holds for $\alpha < -2$.

Let $\oslash$ denote the operation of floating-point divide, and $\ominus$ denote the operation of floating-point subtraction. For any floating-point numbers a, b, and c, we will assume the following:

(i) $a > 0 \supset 1 \oslash a > 0$

(ii) $a \geq b \geq 1 \supset 1 \geq 1 \oslash b \geq 1 \oslash a$

(iii) $a \geq b \supset c \ominus b \geq c \ominus a$

(iv) $a > 2 \supset a \ominus 1 \geq 1$

(v) $a \ominus 0 = a$

Theorem 4: If $\alpha > 2$, and the computer arithmetic satisfies the above rules, then the computed sequence $[\tilde{u}_i]$ converges monotonically to $\tilde{u}$ and $\tilde{u} = u + 0(\beta^{1-t})$.

Proof: $\tilde{u}_1 = \alpha > 2$ and $\tilde{u}_2 = \alpha \ominus (1 \oslash \alpha)$. Since $\alpha > 2$, (i) yields $1 \oslash \alpha > 0$. From (iii) and (v) we have $\alpha \geq \alpha \ominus (1 \oslash \alpha)$; thus $u_1 \geq u_2$. From (ii), $1 \geq 1 \oslash \alpha$. By (iii) and (iv), $\alpha \ominus 1 \oslash \alpha \geq \alpha \ominus 1 \geq 1$. So $\tilde{u}_1 \geq \tilde{u}_2 \geq 1$.

Now assume $\tilde{u}_{k-1} \geq \tilde{u}_k \geq 1$. By (ii), $1 \geq 1 \oslash \tilde{u}_k \geq \tilde{u}_{k-1}$. By (iii) and (iv), $\alpha \ominus (1 / \tilde{u}_{k-1}) \geq \alpha \ominus (1 / \tilde{u}_k) \geq \alpha \ominus 1 \geq 1$. So, $\tilde{u}_k \geq \tilde{u}_{k+1} \geq 1$. By induction, the sequence $[\tilde{u}_i]$ is bounded and monotone. Therefore, since there are a finite number of floating-point representations between α and 1, the sequence converges to a limit $\tilde{u} \geq 1$. In the

limit, we have

$$\tilde{u} = \alpha \ominus (1 \oslash \tilde{u}) \ .$$

Following the techniques of Wilkinson (1965), we have

$$\tilde{u} = (\alpha - \tilde{u}^{-1}\ (1+\epsilon))(1+\eta)$$

for some values of ϵ and η satisfying

$$|\epsilon| \leq \beta^{1-t} \quad \text{and} \quad |\eta| \leq \beta^{1-t} \ .$$

So,

$$\tilde{u} = \alpha - \tilde{u}^{-1} + \delta \ ,$$

where $\delta = \alpha\eta - \tilde{u}^{-1}(\epsilon + \eta + \epsilon\eta)$.

Therefore,

$$\tilde{u} = \tfrac{1}{2}[(\alpha + \delta) + \sqrt{(\alpha + \delta)^2 - 4}\ ].$$

From Theorem 1 we see that

$$\tilde{u} - u = O(\delta) = O(\beta^{1-t}) \ .$$

Since $u > 1$, Theorem 4 provides a bound on the relative error in $\tilde{u}$.

We would like to remark that the algorithm (2) is nothing more than Gaussian elimination which is known to be very stable for positive definite systems. The condition number of the matrix B is easily calculated to be

$$\text{cond}(B) = \frac{|\alpha| + 2\cos\frac{\pi}{n+1}}{|\alpha| - 2\cos\frac{\pi}{n+1}} \leq \frac{|\alpha| + 2}{|\alpha| - 2}$$

Using the error bound given in Forsythe and Moler (1967): If $B\underset{\sim}{y} = \underset{\sim}{d}$ and $(B+E)\underset{\sim}{z} = \underset{\sim}{d}$, then

$$\frac{\|\underset{\sim}{y} - \underset{\sim}{z}\|}{\|\underset{\sim}{z}\|} \leq \text{cond}(B)\ \frac{\|E\|}{\|B\|} \ ,$$

where $\|\cdot\|$ denotes the spectral norm. If E is due to roundoff error in representing α, then $\|E\| \leq \epsilon = |\alpha|\beta^{1-t}$, and

$$\frac{\|\underset{\sim}{y} - \underset{\sim}{z}\|}{\|\underset{\sim}{z}\|} \leq \frac{\epsilon}{|\alpha| - 2\cos\frac{\pi}{n+1}} .$$

4. Generalizations

An important extension of Theorem 1 is that the LU decomposition will converge even if some of the upper left elements of the matrix are changed. If a tri-diagonal matrix contains a Toeplitz sub-matrix, then that portion of the LU decomposition converges. Problems of this sort occur, for example, with cubic spline interpolation with prescribed derivatives at the ends. This is a result of the following.

Theorem 5: If $\alpha > 2$ and $u_1 = \gamma$ where γ has any value except 0, $1/\alpha$, or u_-, then the sequence $u_i = \alpha - 1/u_{i-1}$, $i=2,\ldots,$ converges to u_+ where

$$u_+ = \frac{\alpha + \sqrt{\alpha^2 - 4}}{2} ,$$

and

$$u_- = \frac{\alpha - \sqrt{\alpha^2 - 4}}{2} .$$

(A similar result holds for $\alpha < -2$.)

Proof: The nonlinear difference equation, $u_i = \alpha - \frac{1}{u_{i-1}}$, can be solved explicitly by using the substitution $u_i = \frac{w_i}{w_{i-1}}$ to produce a linear second-order difference equation. For $\alpha > 0$ and $u_1 = \gamma$, the solution is:

$$u_i = u_+ \left[\frac{1 + \xi \left(\frac{u_-}{u_+} \right)^{i+1}}{1 + \xi \left(\frac{u_-}{u_+} \right)^{i}} \right],$$

where $\xi = \frac{\sqrt{\alpha^2 - 4} - \gamma + u_-}{\gamma - u_-}$. Since $\alpha > 2$, the positive quantity (u_-/u_+) is less than unity. Convergence follows immediately. ∎

The results we have given for scalars can also be generalized to matrices.

Theorem 6: If a matrix can be partitioned as

$$\mathcal{A} = \begin{bmatrix} A & B & & & \\ B & A & B & & \\ & B & A & B & \\ & & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot \end{bmatrix}$$

where both A and B are symmetric and positive definite, and if the eigenvalues of $B^{-1}A$ are greater than 2 in modulus, then the block Gaussian elimination of $\mathcal{A}$ converges.

Proof: Block elimination is equivalent to constructing the sequence of matrices $U_1 = A$, $U_{i+1} = A - BU_i^{-1}B$, $i=1,2,\ldots$. But $A = P\Delta P^T$ and $B = PP^T$ where Δ is the diagonal matrix of eigenvalues of $B^{-1}A$. Define $\Lambda_1 = \Delta$ and $\Lambda_{i+1} = \Delta - \Lambda_i^{-1}$. Then $U_1 = P\Delta_1 P^T$ and if $U_i = P\Delta_i P^T$ then

$$U_{i+1} = P\Delta P^T - PP^T(P\Delta_i P^T)^{-1}PP^T$$

$$= P[\Delta - \Delta_i^{-1}]P^T = P\Delta_{i+1}P^T .$$

The convergence of Δ_i (as well as the rate of convergence) under the

conditions stated follows from the results for scalars given in Theorems 1-3.

An example of a matrix that satisfies the required conditions for convergence is the matrix that arises from the five-point finite difference approximation to Laplace's operator in a rectangle:

$$a = \begin{bmatrix} A & -I & & & \\ -I & A & -I & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots \end{bmatrix}$$

where

$$A = \begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 4 \end{bmatrix}$$

However, this method does not appear to be competitive with existing methods for this particular matrix.

5. Conclusions

Many of the observations which lead to the simplification in computing the LU decomposition for tri-diagonal Toeplitz matrices generalize to Toeplitz band matrices. Bauer (1955) states that the Cholesky decomposition of band symmetric matrices converges in the sense that each diagonal of the triangular matrix converges. We know of no rate-of-convergence results for the band case.

An alternate proof of Theorem 1 can be easily constructed by considering the analytical solution to the difference equation (1). Bounds

on k similar to those given in Theorems 2 and 3, but not quite as close, can be obtained similarly.

Acknowledgement

We would like to thank Professors Gene H. Golub and Gerald Taylor for criticizing the manuscript and for several stimulating discussions. In addition we thank Professor Golub for bringing the work of Parter and Bauer to our attention and for suggesting the proof technique used in Theorem 4.

REFERENCES

Bauer, Friedrich L. (1955), "Ein direktes Iterationsverfahren zur Hurwitz-Zerlegung eines Polynoms," Archiv der Elektrischen Uebertragung, Wiesbaden, Vol. 9, 285-290. Translated from the German by M. Morf.

Forsythe, G. E. and Moler, C. B. (1967), Computer Solution of Linear Algebraic Equations, Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Parter, Seymour V. (1962), "An Observation on the Numerical Solution of Difference Equations and a Theorem of Szegö," Num. Math. 4, 293-295.

Wilkinson, J. H. (1963), Rounding Errors in Algebraic Processes. Englewood Cliffs, N.J.: Prentice-Hall, Inc.