Error Propagation in the Solution of Tridiagonal Liner Equations

HIDEKO NAGASAKA*

The degree of ill-conditioning of a matrix A is measured by its condition number $\|A\|\cdot\|A^{-1}\|$, where $\|A\|$ is the norm of A. In particular, if A is symmetric, the norm can be taken as $\|A\|=|\lambda_1|$ and $\|A^{-1}\|=|\lambda_n|^{-1}$, where λ_1 denotes the numerically largest eigenvalue and λ_n the smallest one. If the condition number takes a large value, it leads to the inaccuracy of the numerical solution of linear equations $AX=b$, and it is usually thought that the inaccuracy of the numerical solution of linear equations is due mainly to such a sort of ill-conditioning.

We want to show, in the following remarkable examples, that there is an another cause which gives the inaccuracy of the solution of linear equations.

1. *Problems*

We study in this paper linear equations which are of the following tridiagonal form

$$\begin{cases} bx_1+cx_2 & =k_1\,, \\ ax_1+bx_2+cx_3 & =k\;\;, \\ \qquad ax_2+bx_3+cx_4 & =k\;\;, \\ \qquad\qquad \cdots\cdots & \\ \qquad\qquad\qquad ax_{n-1}+bx_n & =k_e\,, \end{cases} \tag{1}$$

where $k_1=b+c$, $k=a+b+c$, $k_e=a+b$. We know previously that the exact values of the unknowns are

$$x_1=x_2=\ldots.=x_n=1\,,$$

which can be taken as the standard of comparison against the numerical solution obtained by several methods of solution.

Now let us try to solve the above equations by the Gauss elimination method. First the process of forward-elimination is done eliminating the unknows in the order $x_1, x_2, \ldots, x_n$. At the last stage of this process we obtain the value of the unknown x_n. Next starting from this value of x_n, the process of back-substitution gives the values of the unknowns in the order $x_{n-1}, x_{n-2}, \ldots, x_1$.

2. *Examples*

When the above method is applied to (1), many types of numerical solutions

This paper first appeared in Japanese in Joho Shori (the Journal of the Information Processing Society of Japan), Vol. 5, No. 4 (1964), pp. 195~202.

* College of Science and Engineering, Nihon University, Tokyo.

appear corresponding to the combinations of values of a, b, c. Three of these types are given in Table 1. (Numerical value such as $-3.169+22$ denotes $-3.169 \times 10^{+22}$)

Table 1.

Case	I	II	III
(a, b, c)	(1, 6, 8)	(8, 6, 1)	(12, 25, 12)
a/α, c/α	1/4, 2	2, 1/4	3/4, 3/4
n	100	48	100
X_1	7.8532415 −01	1.0000000 +00	1.0000000 +00
X_2	1.1610069 +00	9.9999988 −01	1.0000000 +00
X_3	9.0607937 −01	1.0000002 +00	1.0000000 +00
X_4	1.0503146 +00	9.9999970 −01	1.0000000 +00
|	|	|	|
|	|	|	|
|	|	|	|
X_{46}	|	−1.2022295 +06	|
X_{47}	|	2.0610703 +06	|
X_{48}	|	−2.7480914 +06	|
|	|		|
|	|		|
X_{97}	9.9999996 −01		1.0000000 +00
X_{98}	1.0000000 +00		1.0000000 +00
X_{99}	9.9999996 −01		1.0000000 +00
X_{100}	1.0000000 +00		1.0000000 +00

Table 2.

Case	I	II	III
(a, b, c)	(1, 6, 8)	(8, 6, 1)	(12, 25. 12)
a/α, b/α	1/4, 2	2, 1/4	3/4, 3/4
n	100	48	100
X_1	−3.1691204 +22	1.0000000 +00	9.9999996 −01
X_2	2.3768403 +22	9.9999988 −01	1.0000001 +00
X_3	−1.3864901 +22	1.0000002 +00	9.9999995 −01
X_4	7.4276254 +21	9.9999970 −01	1.0000001 +00
|	|	|	|
|	|	|	|
|	|	|	|
X_{46}	|	6.2499995 +14	|
X_{47}	|	−2.4999999 +15	|
X_{48}	|	1.0000000 +16	|
|	|		|
|	|		|
X_{97}	9.9999916 −01		9.9999994 −01
X_{98}	1.0000004 +00		1.0000001 +00
X_{99}	9.9999976 −01		9.9999987 −01
X_{100}	1.0000001 +00		1.0000002 +00

The main character of these three types is as follows. In case I, during the forward-elimination, the rounding errors introduced in an intermediate stage do not propagate to the last stage, and we obtain $x_n=1$ which is the exact value of the last unknown. While at the process of back-substitution, the errors grow with the advancement of the process, and we obtain the erroneous results for the values of earlier unknowns. In case II, the tendency is quite opposite to that of the case I. Errors grow during the forward-elimination, while they vanish in the process of back-substitution. Finally case III gives an example in which errors do not appear in both processes.

To emphasize the above tendency, we made the experiments shown in Table 2. In these experiments, in place of the unknown x_n, we put the values which are different from the results of the last stage of the forward-elimination. Table 2 gives the values of the unknowns obtained by the back-substitution starting from these hypothetical values of x_n. It exaggeratedly shows error-growing tendency of the back-substitution for the case I, and also shows its error-diminishing tendency for the case II. Meanwhile for the case III, errors remain to be of the same order as at the starting point.

3. *Condition Number and Pivots*

Let us examine the reason for the above results. First we notice that, in these examples, the coefficient matrix of (1) is far from singular and its condition number is not large. This is seen in the following way. The coefficient matrix (which is symmetric) of (1) has the eigenvalues*

$$\lambda_i=b-2\sqrt{ac}\cos\frac{i\pi}{n+1} \qquad (i=1, 2, \ldots, n),$$

and its condition number is

$$P=\frac{\max\limits_i |\lambda_i|}{\min\limits_i |\lambda_i|}=\frac{|b|+2\sqrt{ac}\cos\dfrac{\pi}{n+1}}{\min\limits_i\left|b-2\sqrt{ac}\cos\dfrac{i\pi}{n+1}\right|}.$$

If $b^2-4ac>0$ (three cases in Table 1 and 2 satisfy this condition), the numerator and the denominator of the above expression do not vanish, and we obtain

$$P<\frac{|b|+2\sqrt{ac}}{|b|-2\sqrt{ac}}$$

and from this inequality we can see that in this case the equations (1) do not belong to the ill-conditioned type in the sense explained at the beginning of this paper.

For the sake of the further examination, we give the scheme of transition from the $(m-1)$-th stage to the m-th of the forward-elimination. This is as follows:

* Proof of this relation is due to Y. Itagaki (The National Aerospace Laboratory, Tokyo).

$$\begin{array}{cccc|c}
b_{m-1} & c & & & k_{m-1} \\
a & b & c & & k \\
 & a & b & c & k \\
 & & \cdots\cdots & & \vdots \\
\hline
 & b_m & c & & k_m \\
 & a & b & c & k \\
 & & \cdots\cdots & & \vdots
\end{array}$$

where

$$b_m = b - \frac{ac}{b_{m-1}}, \tag{2}$$

$$k_m = k - \frac{ak_{m-1}}{b_{m-1}}. \tag{3}$$

Thus the value of x_m to be obtained in the back-substitution is

$$x_m = \frac{k_m}{b_m} - \frac{c}{b_m} x_{m+1}. \tag{4}$$

If we regard the relation (2) as the linear transformation from b_{m-1} to b_m, its fixed points α, β are the roots of the equation of the second order

$$z^2 - bz + ac = 0 \tag{5}$$

and the relation (2) can be expressed in the form

$$\frac{b_m - \alpha}{b_m - \beta} = \frac{\beta}{\alpha} \frac{b_{m-1} - \alpha}{b_{m-1} - \beta}. \tag{6}$$

If $b^2 - 4ac > 0$, (5) has two distinct real roots. If we put $|\alpha| > |\beta| \geqq 0$, we see from (6) that b_m converges to the positive value α, when $m \to \infty$. This fact shows that the pivot b_m in each stage of the forward-elimination tends to a steady positive value. Hence, in this process, the errors due to the cancellation of figures in pivots do not occur.

4. *Error Propagation in the Forward-elimination*

Now let ε_i be the rounding error introduced on the right-hand side k_i in the i-th stage of elimination. At the last stage of this process, this error grows to the magnitude

$$(-1)^{n-i} \frac{a}{b_i} \frac{a}{b_{i+1}} \cdots \frac{a}{b_{n-1}} \varepsilon_i \tag{7}$$

which can be shown by the relation (3).

As $b_1 = b = \alpha + \beta$, we obtain from (6)

$$\frac{b_i - \alpha}{b_i - \beta} = \left(\frac{\beta}{\alpha}\right)^{i-1} \frac{b_1 - \alpha}{b_1 - \beta} = \left(\frac{\beta}{\alpha}\right)^i,$$

and, furthermore,

$$b_i = \frac{\alpha \left\{ 1 - \left(\frac{\beta}{\alpha}\right)^{i+1} \right\}}{1 - \left(\frac{\beta}{\alpha}\right)^i}. \tag{8}$$

Hence (7) becomes

$$\left(-\frac{a}{\alpha}\right)^{n-i}\frac{1-\left(\frac{\beta}{\alpha}\right)^{i}}{1-\left(\frac{\beta}{\alpha}\right)^{n}}\varepsilon_i\,, \tag{9}$$

which represents the fact that last k_n has an error whose component due to ε_i is given by the expression (9).

By (9), we see that, when n is large, the errors introduced in the way of the forward-elimination effect a large influence on the last unknown x_n if $\left|\frac{a}{\alpha}\right|>1$. This is the case II shown in Table 1. On the contrary, if $\left|\frac{a}{\alpha}\right|\leqq 1$ the error which once appeared diminish immediately or do not grow in the subsequent stages. These are the cases I and III in Table 1.

5. *Error Propagation in the Back-substitution*

Next let us examine the back-substitution. During this process, if an error δ_j is introduced in the unknown x_j, its effect on the unknown $x_i (i<j)$ is given by

$$(-1)^{j-i}\frac{c}{b_i}\frac{c}{b_{i+1}}\cdots\frac{c}{b_{j-1}}\delta_j=\left(-\frac{c}{\alpha}\right)^{j-i}\frac{1-\left(\frac{\beta}{\alpha}\right)^{i}}{1-\left(\frac{\beta}{\alpha}\right)^{j}}\delta_j \tag{10}$$

by virtue of the relations (4) and (8).

1) In the case $\left|\frac{a}{\alpha}\right|\leqq 1$, the forward-elimination gives the correct values of k_m, so that from (4) we see that the main errors are due to the above effect. Hence if $\left|\frac{c}{\alpha}\right|\leqq 1$, we obtain generally correct values of x_i,while if $\left|\frac{c}{a}\right|>1$, (10) shows that we obtain the erroneous values of the earlier unknows. These circumstances are shown by the cases I and III in Table 1 and 2.

2) In the case $\left|\frac{a}{\alpha}\right|>1$, the forward-elimination gives considerable errors on the values of k_i, so that we must pay attention to them when we consider the errors of the unknowns x_i in view of the formula (4) of back-substitution. If k_j admits an error ε_j during the forward-elimination, it gives the error

$$(-1)^{m-j}\frac{a}{b_j}\frac{a}{b_{j+1}}\cdots\frac{a}{b_{m-1}}\varepsilon_j \tag{11}$$

measured by the unit of the right-hand side k_m in the m-th stage of elimination. On the other hand, we obtain from (4)

$$x_i=\frac{k_i}{b_i}-\frac{c}{b_i}\frac{k_{i+1}}{b_{i+1}}+\frac{c}{b_i}\frac{c}{b_{i+1}}\frac{k_{i+2}}{b_{i+2}}-\cdots\cdot+(-1)^{n-i}\frac{c}{b_i}\frac{c}{b_{i+1}}\cdots\frac{c}{b_{n-1}}\frac{k_n}{b_n}\,, \tag{12}$$

In view of (11), we see that the right hand side of (12) contains the error due to ε_j given by

$$E_{ij}=\varepsilon_j\sum_{m=i}^{n}\frac{c_m a_m}{b_m}$$

where

$$c_m=(-1)^{m-i}\frac{c}{b_i}\frac{c}{b_{i+1}}\cdots\frac{c}{b_{m-1}}$$

$$a_m=(-1)^{m-j}\frac{a}{b_j}\frac{a}{b_{j+1}}\cdots\frac{a}{b_{m-1}}\,.$$

In view of the relation (8), we obtain after some calculations

$$E_{ij}=\frac{\left\{1-\left(\frac{\beta}{\alpha}\right)^{j}\right\}\left\{\left(\frac{\beta}{\alpha}\right)^{n+1-i}-1\right\}\left(-\frac{a}{\alpha}\right)^{i-j}}{\alpha\left(\frac{\beta}{\alpha}-1\right)\left\{1-\left(\frac{\beta}{\alpha}\right)^{n+1}\right\}}\varepsilon_j\,, \tag{13}$$

When i is small compared with n, we obtain approximately

$$E_{ij}\doteqdot\left(-\frac{a}{\alpha}\right)^{i-j}\frac{1-\left(\frac{\beta}{\alpha}\right)^{j}}{\beta-\alpha}\varepsilon_j \tag{14}$$

This represents the effect of the error ε_j on the unknown x_i. As $\left|\frac{a}{\alpha}\right|>1, |E_{ij}|$ may become large if $i>j$. But if the back-substitution advances, the number i decreases, so that $i-j$ becomes to be negative finally, and in view of the factor $\left(-\frac{a}{\alpha}\right)^{i-j}$, $|E_{ij}|$ diminishes with the advancement of the process. For this reason, if we perform the back-substitution, we generally obtain the correct values for the earlier unknowns. These circumstances are shown by the case II in Table 1 and 2.

6. *Other Cases*

Table 3.

Case	IV-1	IV-2	V-1	V-2		Pivot of Case IV, V	VI
(a, b, c)	(3, 4, 5)	(3, 4, 5)	(5, 4, 3)	(5, 4, 3)			(4, 3, 4)
a/c	3/5	3/5	5/3	5/3			1
n	108	109	108	109			109
X_1	7.3886100 +03	−5.1674815 +08	1.0000000 +00	9.9952240 −01	b_1	4.0000000 +00	1.0000023 +00
X_2	−5.9090880 +03	4.1339852 +08	1.0000000 +00	1.0006368 +00	b_2	2.5000000 −01	9.9999828 −01
X_3	2.9650440 +02	−2.0669923 +07	1.0000000 +00	9.9994694 −01	b_3	−5.6000000 +00	9.9999903 −01
X_4	3.3106492 +03	−2.3150316 +08	9.9999983 −01	9.9900920 −01	b_4	4.2678572 +00	1.0000023 +00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_{53}	9.8698509 −01	9.1137673 +02	1.0124134 +00	2.9025149 +02	b_{53}	3.7500861 +00	9.9999960 −01
X_{54}	1.0097614 +00	−6.8179827 +02	1.0010882 +00	−3.6055605 +02	b_{54}	9.1900000 −05	9.9999961 −01
X_{55}	9.9999982 −01	1.0125908 +00	9.7786007 −01	9.8893576 −01	b_{55}	−1.6321690 +05	1.0000007 +00
X_{56}	9.9342091 −01	4.1067064 +02	1.0295193 +00	6.0361107 +02	b_{56}	4.0000919 +00	9.9999991 −01
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_{105}	1.0000000 +00	9.9840291 −01	−7.5055566 +03	−1.7439803 +08	b_{105}	3.5148074 +00	9.9999952 −01
X_{106}	9.9999999 −01	1.0011227 +00	−6.6923138 +02	2.0431570 +08	b_{106}	−2.6765930 −01	1.0000006 +00
X_{107}	9.9999999 −01	1.0000601 +00	1.3405570 +04	1.8242464 +07	b_{107}	6.0041395 +01	9.9999981 −01
X_{108}	1.0000000 +00	9.9927816 −01	−1.6754713 +04	−3.6484947 +08	b_{108}	3.7501724 +00	9.9999956 −01
X_{109}		1.0005414 +00		4.5606188 +08	b_{109}	1.8370000 −04	1.0000005 +00

In the preceding, we have considered the case $b^2-4ac>0$. If this condition fails, the numerically smallest eigenvalue of the coefficient matrix can take a sufficiently small absolute value, and the condition number may become very large. Yet the propagation of errors can be studied in view of the similar relations as those used in the above consideration. To avoid complexity, we have omitted the detailed proofs. The main results are shown in the general scheme given in Table 4.

Table 3 gives some numerical examples for the case $b^2-4ac<0$.

7. *Conclusion*

Finally, we shall give in Table 4 the main results over all the cases in a table form. It represents the general behavior of the solution of (1) obtained by the Gauss elimination method when n is large.

Table 4.

		$\|x_n\|$	$\|x_1\|$	referred Example
$b^2-4ac\geqq 0$	$\left\|\frac{a}{\alpha}\right\|<1,\ \left\|\frac{c}{\alpha}\right\|\geqq 1$	correct value	increasing tendency	Case I
	$\left\|\frac{a}{\alpha}\right\|\leqq 1,\ \left\|\frac{c}{\alpha}\right\|\leqq 1$	correct value	correct value	Case III
	$\left\|\frac{a}{\alpha}\right\|>1,\ \left\|\frac{c}{\alpha}\right\|<1$	increasing	correct value	Case II
$b^2-4ac<0$	$\frac{a}{c}<1$	correct value	increasing	Case IV
	$\frac{a}{c}>1$	increasing	correct value	Case V
	$\frac{a}{c}=1$	generally correct value		Case VI